research
reports

No. 474 March 2006
2006 / 03 / 23

Jens Ledet Jensen

Maximum likelihood classifiers in microarray studies

department of
theoretical
statistics

university of
aarhus

ISSN 0902-9893

Maximum likelihood classifiers in microarray studies

Jens Ledet Jensen
Department of Mathematical Sciences, University of Aarhus
Ny Munkegade, DK-8000 Aarhus C, Denmark

Contents

1 Introduction **1**

2 Maximum likelihood classifier **3**
2.1 $n = m$ 5
2.2 $n \neq m$ 6

3 Thresholding **9**
3.1 Data dependent threshold 11
3.2 Shrunken centroids 17
3.3 Partial Least Squares 19

4 Estimating the threshold: crossvalidation **22**

5 Estimating POCC: crossvalidation **24**

6 Bayes modeling **25**

7 Three groups **30**

8 Summary **31**

A Variance of ξ_N when $n \neq m$ **32**

B Appendix: Bayes classifier **32**

1 Introduction

This paper is motivated by the use of microarray investigations in the biological and medical sciences. The microarray technology is less than ten years old and therefore still developing. Using a microarray one measures simultaneously the *expression levels* of a large number of genes in a biological sample. Simplifying, the microarray chip is a small glass plate divided into a large number of spots (1000–50000), and in

each spot copies of a short string of letters are fixed. After extraction of the DNA material of a biological sample this is put on the chip and molecules that match the text in a spot will bind to these. The reading consists of an image of light intensities, where the light intensity of a spot reflects the amount of DNA material that binds to the spot. Thus each measurement is a high dimensional vector giving the intensities of the different spots.

An example of the use of microarray experiments is in cancer research (see e.g. Alon et al. (1999), Golub et al. (1999), and Andersen et al. (2003)). The aim is to improve the treatment given to cancer patients, in particular to make the treatment more person specific by identifying suitable subgroups. Colon cancer is divided into Dukes A, B, C, and D. Dukes C and D are serious cases that all receive chemotherapy, whereas Dukes B usually do not receive chemotherapy after the removal of the cancer cells. However, some of the Dukes B cases develop a more serious cancer, and it is therefore of considerable interest to identify these cases beforehand.

In the just mentioned example the aim is to be able to classify a new case as belonging to one of two groups. Such classification problems are common in microarray investigations. Compared to traditional classification problems the distinct feature here is the large number of variables. Another distinctive feature is that of the many genes being interrogated only a very small fraction is giving information on the difference between the two groups.

A classical textbook as Mardia et al. (1979) treats exclusively the case where the number of observations N is larger than the number of variables p measured on each unit. In the microarray setting typical values of N are of the order 100 or less, whereas p is in the range 1000–50000. Mardia et al. (1979) mainly treats what is known as the *maximum likelihood classifier* which is also the one of main interest to us in this paper. The more recent book Hastie et al. (2001) is directed also toward cases with p larger than n and consider many other methods than the maximum likelihood classifier. Dudoit and Fridlyand (2003) give an excellent survey of different classification methods in relation to microarray applications, and perform a comparison of the different methods using microarray datasets.

Our interest in this paper is to investigate the influence of a very large number of variables p on the performance of a classifier. In particular we look at the possibility of removing variables in order to improve the performance. In most of the paper we consider two groups with observations $x_1, \ldots, x_n$ from group 1 and observations $y_1, \ldots, y_m$ from group 2. On each unit p variables are measured so that $x_i = (x_{i1}, \ldots, x_{ip})$ and $y_i = (y_{i1}, \ldots, y_{ip})$. A variable j is called differentiable expressed if the mean of the variable in group 1 is different from the mean in group 2. When constructing a classifier the data are often divided into two sets, the *training data* and the *test data*. Only the training data are used to estimate the parameters of the classifier, whereas the test data are used to evaluate the performance of the classifier.

In the figures the abbreviation POCC is used for the probability of correct classification.

2 Maximum likelihood classifier

To begin with we consider the case of two homogeneous groups. An observation $z = (z_1, \ldots, z_p)$ from group 1 has density $f_1(z)$ and an observation from group 2 has density $f_2(z)$. If f_1 and f_2 are known the maximum likelihood classifier assigns an observation z to the group

$$\arg\max_I \{\pi_I f_I(z)\},$$

where π_1 and π_2 are prior probabilities for the two groups, $\pi_1 + \pi_2 = 1$. In medical applications the prior probabilities reflect the knowledge one has on the composition of the population. For convenience in the investigations below we take $\pi_1 = \pi_2 = \frac{1}{2}$. Typically, the densities f_1 and f_2 are not known and must be estimated from training data. When p is much larger than the sample size it is difficult to estimate a completely general density. In this paper we make the simplifying assumption that the coordinates of x are independent, with the jth coordinate having density f_{Ij}. In many applications this assumption is not at all realistic. However, it still serves our purpose of investigating the influence of having the number of variables very large. Also, the classifiers that we consider are still of relevance even though the independence assumption is not valid. In the case of independent variables the classification rule becomes

$$\arg\max_I \left\{ \prod_{j=1}^{p} f_{Ij}(z_j) \right\}.$$

In practice f_{Ij} contains parameters that need to be replaced by estimates based on the training data. We consider exclusively the case with

$$z_j \sim \begin{cases} N(\mu_j, \sigma_j^2) & \text{group 1,} \\ N(\mu_j + \delta_j \sigma_j, \sigma_j^2) & \text{group 2.} \end{cases} \tag{1}$$

Note that we here assume that the variance σ_j^2 is the same in the two groups, and that we have scaled the difference in the two means by the standard deviation σ_j. Under the model (1) the maximum likelihood classification of a new observation z is based on the distances $\sum_j (z_j - \mu_j)^2/\sigma_j^2$ and $\sum_j (z_j - \mu_j - \delta_j\sigma_j)^2/\sigma_j^2$. Setting

$$\tilde{D} = \tilde{D}(z) = \frac{1}{2}\left(\sum_{j=1}^{p} \frac{(z_j - \mu_j)^2}{\sigma_j^2} - \sum_{j=1}^{p} \frac{(z_j - \mu_j - \delta_j\sigma_j)^2}{\sigma_j^2} \right), \tag{2}$$

the observation z is classified as belonging to group 1 if $\tilde{D} < 0$ and to group 2 if $\tilde{D} > 0$.

Let $x_1, \ldots, x_n$ be the observations from group 1 and $y_1, \ldots, y_m$ the observations from group 2 in the training data. We estimate μ_j by $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$ and $\mu_j + \delta_j\sigma_j$ by $\bar{y}_j = \sum_{i=1}^{m} y_{ij}/m$. The variance σ_j^2 is estimated by $s_j^2 = \{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 + \sum_{i=1}^{m}(y_{ij} - \bar{y}_j)^2\}/(n + m - 2)$. This gives the classifier statistic

$$D_0 = D_0(z) = \frac{1}{2}\sum_{j=1}^{p} \frac{(z_j - \bar{x}_j)^2 - (z_j - \bar{y}_j)^2}{s_j^2} = \sum_{j=1}^{p} z_j \frac{\bar{y}_j - \bar{x}_j}{s_j^2} - \frac{1}{2}\sum_{j=1}^{p} \frac{\bar{y}_j^2 - \bar{x}_j^2}{s_j^2}. \tag{3}$$

For a given training set, that is, the x_is and y_is, the probability of correct classification (POCC) is the probability that $D_0(z)$, as a function of z, has the correct sign. To highlight that probabilities and other quantities are calculated with respect to the distribution of a new observation z, we use a lower subscript N, N for new. Thus, if the new observation z is from group 1 we have $\text{POCC} = P_N(D_0 < 0)$.

Lemma 1. *Consider a new observation z from group 1. Define $\xi_N = E_N(D_0) = \sum_{j=1}^p \{\mu_j(\bar{y}_j - \bar{x}_j) - \frac{1}{2}(\bar{y}_j^2 - \bar{x}_j^2)\}/s_j^2$ and $\tau_N^2 = V_N(D_0) = \sum_{j=1}^p \sigma_j^2(\bar{y}_j - \bar{x}_j)^2/s_j^4$. Then the probability of correct classification of z is*

$$POCC = P_N(D_0 < 0) = \Phi\Big(-\frac{\xi_N}{\tau_N}\Big), \tag{4}$$

where Φ is the standard normal distribution function.

Proof. The normality of z_j trivially implies that the distribution of D_0 for fixed training data is $D_0 \sim N(\xi_N, \tau_N^2)$. □

To study the statistical properties of $-\xi_N/\tau_N$, as a function of the training data, we note the following simple results.

Lemma 2. *Let $f = n + m - 2$ and let $\delta_\bullet^k = \sum_j \delta_j^k$. Then*

$$E(\xi_N) = \frac{p}{n}c_{11} - c_{12}\delta_\bullet^2, \quad c_{12} = \frac{f}{2(f-2)}, \quad c_{11} = \big(1 - \frac{n}{m}\big)c_{12},$$

$$E(\tau_N) = \frac{p}{n}c_{21} + c_{22}\delta_\bullet^2, \quad c_{22} = \frac{f^2}{(f-2)(f-4)}, \quad c_{21} = \big(1 + \frac{n}{m}\big)c_{22},$$

and

$$V(\xi_N) = \frac{p}{n^2}c_{31} + \frac{1}{n}c_{32}\delta_\bullet^2 + c_{33}\delta_\bullet^4, \quad c_{31} = \frac{f^2}{2(f-2)(f-4)}\Big\{1 + \frac{n^2}{m^2} + \frac{(1-\frac{n}{m})^2}{f-2}\Big\},$$
$$c_{32} = \frac{f^2}{(f-2)(f-4)}\Big\{\frac{n}{m} - \frac{1-\frac{n}{m}}{f-2}\Big\}, \quad c_{33} = \frac{f^2}{2(f-2)^2(f-4)}.$$

Proof. The results follow from the independence of $\bar{x}$, $\bar{y}$, and s^2, and from the moments of $1/a$, where a has a χ^2 distribution with f degrees of freedom. □

Using a central limit theorem for $\xi_N/\sqrt{p}$ and a law of large numbers for τ_N^2/p in the limit $p \to \infty$, one sees from Lemma 2 that

$$-\frac{\xi_N}{\tau_N} \xrightarrow{\sim} N\left(\frac{-\sqrt{\frac{p}{n}}c_{11} + \sqrt{\frac{n}{p}}c_{12}\delta_\bullet^2}{\sqrt{c_{21} + c_{22}\frac{n}{p}\delta_\bullet^2}}, \frac{1}{n}\frac{c_{31} + c_{32}\frac{n}{p}\delta_\bullet^2 + c_{33}\frac{n^2}{p}\delta_\bullet^4}{c_{21} + c_{22}\frac{n}{p}\delta_\bullet^2}\right), \tag{5}$$

as $p \to \infty$. Let ν and ω^2 be the mean and variance in this approximating normal distribution. Then we use

$$\Phi(\nu) \quad \text{and} \quad \Phi(\nu \pm \omega) \tag{6}$$

as an approximation for the median probability of correct classification and for the upper and lower 15% quantiles of the probability of correct classification, respectively.

2.1 $n = m$

The two formulae (4) and (5) together give a simple way of seeing the size and variation in the probability of correct classification (POCC). The median value of POCC (6) is obtained as the standard normal distribution function evaluated at the mean value in (5). The major controlling variable in this expression is $\sqrt{n/p}\sum_j \delta_j^2$. The POCC median value is illustrated in Figure 1. The vertical bars give the standard normal distribution function evaluated at the mean plus and minus one standard deviation, corresponding to the upper and lower 15% quantiles.

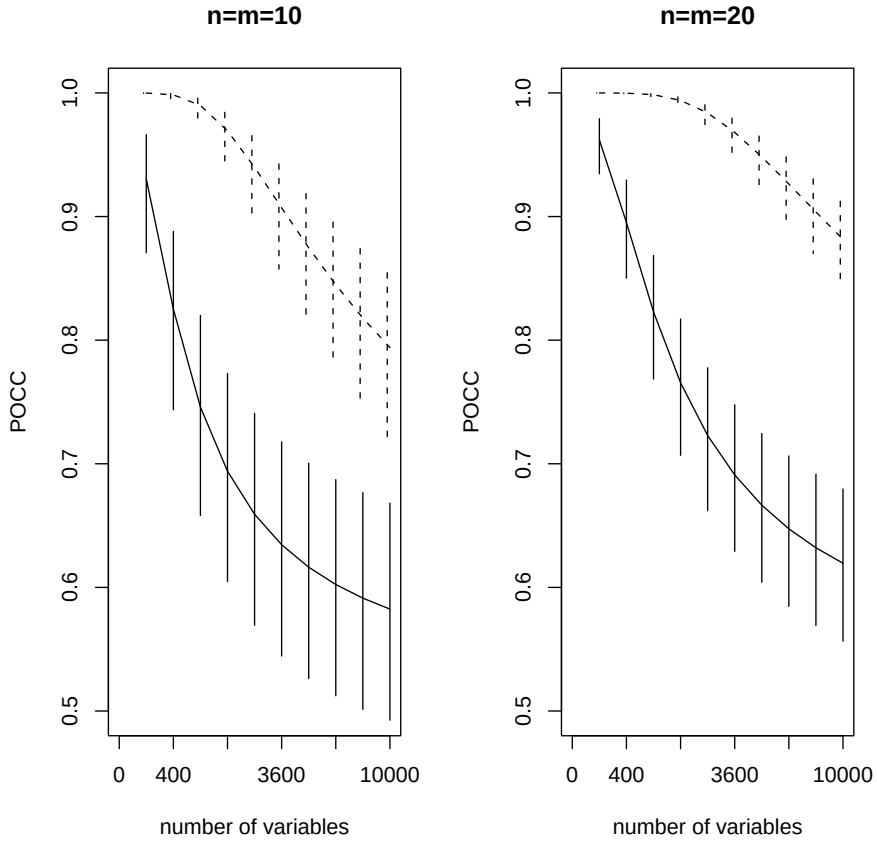

Figure 1: Probability of correct classification (POCC) for a new observation from group 1 as a function of the number of variables p. The curve gives the median value and the vertical bars give upper and lower 15% quantiles of POCC, see (6). Full drawn line: number of expressed variables is 20; dashed line: number of expressed variables is 80. In both cases the average values of δ^2 and δ^4, entering (5), for the expressed variables are 1 and 3, respectively.

The simple (and well-known) message from Figure 1 is that when the number of variables is much larger than the number of expressed variables the classifier has a poor performance. This is also clear from the expression $\sqrt{n/p}\sum_j \delta_j^2$, entering the mean value, which tends to zero as $1/\sqrt{p}$ as p increases and $\sum_j \delta_j^2$ is kept fixed.

Above, the probability of correct classification for a new sample from group 1 was considered. Of interest is also the joint probability of correct classification for

a sample from group 1 and a sample from group 2. It turns out that there is a strong negative correlation between the two probabilities, so that if the classifier works well in one group the performance is not as good in the other group. Defining $\tilde{\xi}_N = E_N(D_0)$ and $\tilde{\tau}_N = V_N(D_0)$, when the new observation z is from group 2, expression (3) shows, trivially, that $\tilde{\tau}_N = \tau_N$ and that $\tilde{\xi}_N = \xi_N + \sum_j \delta_j \sigma_j (\bar{y}_j - \bar{x}_j)/s_j^2$. The probability of correct classification is in this case $P_N(D_0 > 0) = \Phi(\tilde{\xi}_N/\tau_N)$. Supplementing Lemma 2, we have for $n = m$ the following result.

Lemma 3. *Let $n = m$. Then*

$$E(\tilde{\xi}_N) = -E(\xi_N) = c_{12}\delta_\bullet^2,$$

$$V(\tilde{\xi}_N) = V(\xi_N) = \frac{p}{n^2}c_{31} + \frac{1}{n}c_{32}\delta_\bullet^2 + c_{33}\delta_\bullet^4,$$

and

$$Cov(-\xi_N, \tilde{\xi}_N) = -\frac{p}{n^2}c_{31} + c_{33}\delta_\bullet^4.$$

In Figure 2 are simulated values of $(-\xi_N/\tau_N, \tilde{\xi}_N/\tau_N)$ illustrating the strong negative correlation between the two.

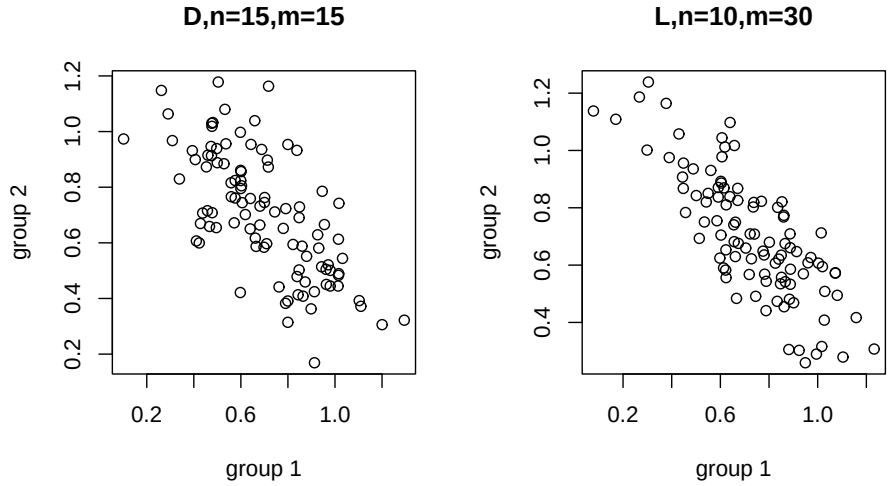

Figure 2: Plot of $\tilde{\xi}_N/\tau_N$ against $-\xi_N/\tau_N$ where the probabilities of correct classification are $\Phi(-\xi_N/\tau_N)$ for group 1 and $\Phi(\tilde{\xi}_N/\tau_N)$ for group 2. The left plot is based on the classification statistic from (3) for the case of 15 observations in each group, and the right plot is based on the classification statistic from (8) for the case of 10 observations in group 1 and 30 observations in group 2. In both plots there are $p = 1000$ variables and $k = 20$ differentiable expressed variables, 10 with the value $\delta = 0.5$, 7 with the value $\delta = 1.0$, 2 with the value $\delta = 1.5$, and 1 with the value $\delta = 2.0$.

2.2 $n \neq m$

Because of the term $c_{11}\sqrt{p/n} = (m-n)f\sqrt{p/n}/[m(f-2)]$ in the mean value in (5), the classifier based on D_0 cannot be used when $n \neq m$. If $m < n$ and p is large the probability of correct classification for an observation from group 1 is close to zero.

The problem is that when $\delta_j = 0$ and $m \neq n$ we have $E[(z_j - \bar{x}_j)^2 - (z_j - \bar{y}_j)^2] = \sigma_j^2(1/n - 1/m) \neq 0$. This very simple observation is not always taken into account (see e.g. Mardia et al. (1979) and Tibshirani et al. (2003)). There are various ways to remedy the problem. Looking at (3) we write

$$L_0 = L_0(z) = \sum_j z_j \frac{\bar{y}_j - \bar{x}_j}{s_j^2} \quad \text{and} \quad \kappa_0 = \frac{1}{2}\sum_j \frac{\bar{y}_j^2 - \bar{x}_j^2}{s_j^2}. \tag{7}$$

The classification based on the sign of $D_0 = L_0 - \kappa_0$ corresponds to the rule that z belongs to group 1 if $L_0 < \kappa_0$ and belongs to group 2 if $L_0 > \kappa_0$. The solution to the above bias problem when $n \neq m$ is to replace κ_0 by a better value. We can formulate this in the way that L_0 is our classifier statistic, with small values indicating that the observation belongs to group 1 and large values indicating group 2, and the problem is to find a good boundary between the two.

One possibility is to evaluate the classifier statistic L_0 on the training data and use these values to separate the two groups. Thus, we calculate $a_i = L_0(x_i)$, $i = 1, \ldots, n$, and $b_i = L_0(y_i)$, $i = 1, \ldots, m$, and use $\frac{1}{2}(\bar{a} + \bar{b})$ as the boundary point. However, this approach simply gives back κ_0: $\frac{1}{2}(\bar{a}+\bar{b}) = \frac{1}{2}\sum_j(\bar{x}_j+\bar{y}_j)(\bar{y}_j+\bar{x}_j)/s_j^2 = \kappa_0$. To obtain a more useful boundary point consider a *leave one out* method, where a classifier is constructed based on the reduced data set with one data point left out, and this classification statistic is evaluated at the point left out. Thus, if x_i is left out let $\bar{x}(i)$ be the average of the remaining $n-1$ observations, and let $s^2(x_i)$ be the within group variance for the two groups with x_i taken out. Define the classification statistic $L_0^{x_i}(z) = \sum_j z_j(\bar{y}_j - \bar{x}_j(i))/s_j^2(x_i)$, and let $a_i = L_0^{x_i}(x_i)$ be the value when $L_0^{x_i}$ is evaluated at the point x_i left out. Similarly, define $b_i = L_0^{y_i}(y_i)$, where $L_0^{y_i}$ is the classification statistic when y_i has been removed form the training set. We then use $\kappa = \frac{1}{2}(\bar{a} + \bar{b})$ as the boundary point between the two groups. Written explicitly this gives

$$\kappa = \frac{1}{2}\sum_j \left(\frac{1}{n}\sum_{i=1}^n x_{ij}\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)} + \frac{1}{m}\sum_{i=1}^m y_{ij}\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)} \right).$$

However, using this value is still not quite satisfactory. Going back to the case $n = m$ the mean of $D_0 = L_0(z) - \kappa_0$, with respect to both z and the training set, is $-f\delta_\bullet^2/[2(f-2)]$ when the new observation z belongs to group 1. This expression is independent of the mean μ, which is clear from the first form of D_0 in (3). Contrary to this the mean of $L_0 - \kappa$ is $-(f-1)\delta_\bullet^2/[2(f-3)] - 2\sum_j \mu_j\delta_j/[(f-2)(f-3)]$, which depends on μ. To overcome the dependency on μ we write also the classification statistic L as an average of the statistics obtained when leaving out one observation. Thus, we use

$$\tilde{L}_0 = \tilde{L}_0(z) = \frac{1}{2}\sum_j z_j \left(\frac{1}{n}\sum_{i=1}^n \frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)} + \frac{1}{m}\sum_{i=1}^m \frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)} \right),$$

so that

$$L(z) = \tilde{L}_0(z) - \kappa = \frac{1}{2}\sum_j \left(\frac{1}{n}\sum_{i=1}^n (z_j - x_{ij})\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)} + \frac{1}{m}\sum_{i=1}^m (z_j - y_{ij})\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)} \right). \tag{8}$$

This statistic clearly has a mean independent of μ, the mean being $E(L) = -(f-1)\delta_\bullet^2/[2(f-3)]$ for an observation z from group 1 and minus this value for an observation z from group 2.

Using notation as for the classifier D_0 we define

$$\xi_N = L(\mu), \quad \text{and} \quad \tau_N^2 = \frac{1}{4}\sum_j \sigma_j^2 \Big(\frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)} + \frac{1}{m}\sum_{i=1}^{m} \frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)} \Big)^2. \quad (9)$$

Then, as in Lemma 1, the probability of correct classification for an observation from group 1 is $\Phi(-\frac{\xi_N}{\tau_N})$. To establish a result as in Lemma 2 some notation is needed. Let $u_1, \ldots, u_n, v_1, \ldots, v_m$ be independent variables with a standard normal distribution. We use the same notation as in the construction of L above so that for example $\bar{u}(i)$ is the average of the u variables with u_i left out, and $s^2(u_i)$ is the corresponding variance estimate based on the remaining u_js and on all the v_js. Define also

$$tu(i) = \frac{\bar{v} - \bar{u}(i)}{s^2(u_i)}, \quad tv(i) = \frac{\bar{v}(i) - \bar{u}}{s^2(v_i)}.$$

Lemma 4. *The means $E\xi_N$ and $E\tau_N^2$ are given as in Lemma 2 with the following coefficients*

$$c_{11} = 0, \quad c_{12} = \frac{f-1}{2(f-3)},$$

$$\begin{aligned} c_{21} &= \frac{n-1}{4}E[tu(1)tu(2)] + \frac{n(m-1)}{4m}E[tv(1)tv(2)] + \frac{n}{2}E[tu(1)tv(1)] \\ &\quad + \frac{f(f+1)(f-1)^2}{4m(m-1)(n-1)(f-3)(f-5)}, \\ c_{22} &= \frac{n-1}{4n}E[(s^2(u_1)s^2(u_2))^{-1}] + \frac{m-1}{4m}E[(s^2(v_1)s^2(v_2))^{-1}] \\ &\quad + \frac{1}{2}E[(s^2(u_1)s^2(v_1))^{-1}] + \frac{(f+2)(f-1)^2}{4nm(f-3)(f-5)}. \end{aligned}$$

The variance $V(\xi_N)$ is given in Appendix A.

Proof. We illustrate the calculations leading to $E\tau_N^2$. Taking one of the terms of the sum (9) for τ_N^2 we need to evaluate

$$E\Big(\frac{1}{n}\sum_{i=1}^{n} \frac{\bar{v} + \delta - \bar{u}(i)}{s^2(u_i)} + \frac{1}{m}\sum_{i=1}^{m} \frac{\bar{v}(i) + \delta - \bar{u}}{s^2(v_i)} \Big)^2.$$

The typical terms here are

$$E\Big(\frac{(\bar{v} + \delta - \bar{u}(i))^2}{s^4(u_i)} \Big) = \Big(\frac{1}{n-1} + \frac{1}{m} + \delta^2 \Big) \frac{(f-1)^2}{(f-3)(f-5)},$$

$$E\Big(\frac{(\bar{v} + \delta - \bar{u}(1))(\bar{v} + \delta - \bar{u}(2))}{s^2(u_1)s^2(u_2)} \Big) = E\Big(\frac{(\bar{v} - u(1))(\bar{v} - \bar{u}(2))}{s^2(u_1)s^2(u_2)} \Big) + \delta^2 E([s^2(u_1)s^2(u_2)]^{-1}),$$

and

$$E\Big(\frac{(\bar{v}+\delta-\bar{u}(1))(\bar{v}(1)+\delta-\bar{u})}{s^2(u_1)s^2(v_1)}\Big)$$
$$= E\Big(\frac{(\bar{v}-u(1))(\bar{v}(1)-\bar{u})}{s^2(u_1)s^2(v_1)}\Big) + \delta^2 E([s^2(u_1)s^2(v_1)]^{-1}),$$

where in the second and third symmetry is used to remove terms with δ. Counting the number of terms of a particular form the result in the lemma for $E\tau_N^2$ is obtained. The expression for $V(\xi_N)$ is calculated in a similar manner. □

The mean values entering the expressions for the coefficients in Lemma 4 do not seem to have a closed form expression, and we calculate these by simulations. For the case where $n = m$ the approach based on D_0 from (3) and the approach based on L from (8) give almost the same probability of correct classification. Actually, the asymptotic mean in (5) is slightly larger when using L instead of D_0: if we consider $n = m = 10$ we have $c_{21}/c_{12}^2 = 9.1429$ and $c_{22}/c_{12}^2 = 4.5714$ when considering D_0, and the corresponding numbers are 8.1 and 4.0 when using L. As for D_0, the use of L for classification gives a strong negative correlation between the probabilities of correct classification for a new observation from group 1 and group 2. An example can be seen in Figure 2. In Table 1 are a few instances of the median probability of correct classification of an observation from group 1. As in (6) the median is taken as $\Phi(-E\xi_n/\sqrt{E\tau_N^2})$.

			POCC, $k = 20$		POCC, $k = 80$	
	c_{21}/c_{12}^2	c_{22}/c_{12}^2	$p = 1000$	$p = 10000$	$p = 1000$	$p = 10000$
$n = m = 10$	8.1	4.0	0.75	0.59	0.99	0.81
$n = 15, m = 5$	15.9	4.0	0.72	0.58	0.98	0.78

Table 1: Median of the probability of correct classification (POCC) for the case with k differentiable expressed variables. The mean of δ^2 for the expressed variables is 1. The classification is based on L from (8).

3 Thresholding

The differentiable expressed variables are those with $\delta_j \neq 0$. When the number of differentiable expressed variables is small as compared to the total number of variables p, the use of the classification statistic D_0 from (3) or L from (8) fails. The intuitive reason is that most variables contribute noise only, and this eventually drowns the signal of interest. An obvious way to try to remedy this problem is to use a subset of the variables only in the classification statistic. For the simple model that is considered in this paper, with independence between the variables, the selection of variables is naturally based on the observed difference between the two groups for each variable. This difference is expressed through the t statistic

$$t_j = \frac{\bar{y}_j - \bar{x}_j}{\sqrt{s_j^2(\frac{1}{n}+\frac{1}{m})}}.$$

In the microarray setting, where variables are genes, there can be dependence between the variables due to the existence of biological pathways. In such a case one can consider including a whole pathway and not only those genes within a pathway that show a large difference between the two groups.

Let $w(t_j)$ be a weight function, where $w = 1$ means that the variable is included and $w = 0$ means that the variable is not included in the classification. Using the classification statistic D_0 from (3) the new classification based on a subset of the variables only is

$$D = D(z) = \frac{1}{2}\sum_{j=1}^{p} \frac{(z_j - \bar{x}_j)^2 - (z_j - \bar{y}_j)^2}{s_j^2} w(t_j) \tag{10}$$

When w can take on the two values zero and one only, the selection of the variables is called *hard thresholding*. For a parameter Δ we can in this case write

$$w(t) = \begin{cases} 1 & |t| \geq \Delta, \\ 0 & |t| < \Delta. \end{cases} \tag{11}$$

When instead w is a continuous function of t one uses the term *soft thresholding*. A typical choice of w is

$$w(t) = \frac{|t| - \Delta}{\theta + |t|} 1(|t| > \Delta), \tag{12}$$

where θ is a parameter. In the soft thresholding case we keep the idea that $w = 0$ below some cutoff Δ. This is because the aim is both to make a good classifier and to have this classifier based on a small list of variables. The selected variables can then be investigated in new experiments.

To begin with let us consider the classification based on (10) when $n = m$ and when the weight function w is fixed, that is, the parameters of the weight function are fixed as opposed to being determined by the data. Then the performance analysis of the classifier in Section 2 can be repeated. For a new observation z from group 1 let $\xi_N = E_N(D_0)$ and $\tau_N^2 = V_N(D_0)$, and let $\tilde{\xi}_N$ be the mean when the observation is from group 2. The probability of correct classification is $\Phi(-\xi_N/\tau_N)$ and $\Phi(\tilde{\xi}_N/\tau_N)$, respectively. Using a central limit theorem for $(\xi_N, \tilde{\xi}_N)$, and the law of large numbers for τ_N^2, we find that $-\xi_N/\tau_N$ and $\tilde{\xi}_N/\tau_N$ have the same asymptotic distribution. To state this let $u \sim N(0, \frac{2}{n})$, $r^2 \sim \chi^2(f)/f$, and define

$$m_1(\delta) = E\Big[\frac{u+\delta}{r^2} w\Big(\frac{u+\delta}{r\sqrt{2/n}}\Big)\Big] \quad \text{and} \quad m_2(\delta) = E\Big\{\Big[\frac{u+\delta}{r^2} w\Big(\frac{u+\delta}{r\sqrt{2/n}}\Big)\Big]^2\Big\}.$$

Then

$$-\frac{\xi_N}{\tau_N} \approx N\left(\frac{\sum_j \frac{\delta_j}{2} m_1(\delta_j)}{\sqrt{\sum_j m_2(\delta_j)}}, \frac{\sum_j [\frac{1}{4}(\frac{2}{n} + \delta_j^2) m_2(\delta_j) - \frac{1}{4}\delta_j^2 m_1(\delta_j)^2]}{\sum_j m_2(\delta_j)}\right),$$

Furthermore, the covariance is

$$\mathrm{Cov}\Big(-\frac{\xi_N}{\tau_N}, \frac{\tilde{\xi}_N}{\tau_N}\Big) \approx -\frac{\sum_j [\frac{1}{4}(\frac{2}{n} - \delta_j^2) m_2(\delta_j) + \frac{1}{4}\delta_j^2 m_1(\delta_j)^2]}{\sum_j m_2(\delta_j)}.$$

It does not seem possible to calculate m_1 and m_2 analytically, and we therefore find these by simulations. In Table 2 is a small illustration of the effect of using a fixed threshold. There are k expressed variables, all with the same value $\delta = 1$ of the scaled difference between the two groups. A small improvement when using a fixed threshold can be seen, and a soft threshold is slightly better than a hard threshold. Furthermore, one sees that without thresholding there is a strong negative correlation between $-\xi_N/\tau_N$ and $\tilde{\xi}_N/\tau_N$, corresponding to the probabilities of correct classification in group 1 and group 2, and that this correlation is reduced considerably after thresholding. The negative correlation has already been illustrated in Figure 2.

	$p = 1000$			$p = 10000$		
	no	hard	soft	no	hard	soft
mean, $k = 20$	0.63	0.74	0.81	0.21	0.27	0.31
Φ(mean)	0.74	0.77	0.79	0.58	0.61	0.62
sd	0.24	0.29	0.31	0.23	0.24	0.25
corr	−0.74	−0.22	−0.05	−0.97	−0.76	−0.59
Δ		2.2	1.6		2.6	2.2
mean, $k = 80$	2.24	2.38	2.46	0.82	1.01	1.14
Φ(mean)	0.99	0.99	0.99	0.79	0.84	0.87
sd	0.27	0.31	0.33	0.23	0.26	0.28
corr	−0.36	−0.01	0.08	−0.88	−0.44	−0.25
Δ		1.7	1.0		2.4	1.9

Table 2: Asymptotic mean and standard deviation of $-\xi_N/\tau_N$ where the probability of correct classification is $\Phi(-\xi_N/\tau_N)$. There are k differentiable expressed variables, all with $\delta = 1$. The row with Δ gives the threshold function: $1(|t_j| > \Delta)$ (hard thresholding) and $1(|t_j| > \Delta)(|t_j| - \Delta)/|t_j|$ (soft thresholding). The thresholding parameter Δ was chosen so as to optimize the probability of correct classification.

As in the case of no thresholding the statistic D cannot be used when $n \neq m$. Turning to the statistic (8) instead the thresholding idea gives

$$L(z) = \frac{1}{2}\sum_j \Big(\frac{1}{n}\sum_{i=1}^{n}(z_j - x_{ij})\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)}w(t_j(x_i)) + \frac{1}{m}\sum_{i=1}^{m}(z_j - y_{ij})\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)}w(t_j(y_i))\Big), \tag{13}$$

where $t(x_i)$ is the t-statistic for the two groups with x_i left out and, similarly, $t(y_i)$ is the t-statistic with y_i left out. Note that the distributional properties of this classification statistic are independent of the unknown means μ_j, and that the mean of one of the terms in the sum is zero when $\delta_j = 0$.

3.1 Data dependent threshold

In this subsection we illustrate by some examples the optimal improvement one can obtain using a thresholding idea. The optimal choice of the threshold is calculated using the (unknown) values of δ_j. In practice the threshold is estimated from the data giving a suboptimal classifier. We simulate data from the model and calculate

for the statistic D in (10)

$$\eta_1(\Delta) = -\frac{\xi_N}{\tau_N} \qquad \text{and} \qquad \eta_2(\Delta) = \frac{\tilde{\xi}_N}{\tau_N}, \tag{14}$$

where

$$\begin{aligned}
\xi_N &= \sum_{j=1}^{p} \big[\mu_j(\bar{y}_j - \bar{x}_j) - \frac{1}{2}(\bar{y}_j^2 - \bar{x}_j^2)\big] \frac{w(t_j)}{s_j^2}, \\
\tilde{\xi}_N &= \sum_{j=1}^{p} \big[(\mu_j + \delta_j\sigma_j)(\bar{y}_j - \bar{x}_j) - \frac{1}{2}(\bar{y}_j^2 - \bar{x}_j^2)\big] \frac{w(t_j)}{s_j^2}, \\
\tau_N^2 &= \sum_{j=1}^{p} \frac{\sigma_j^2}{s_j^4}(\bar{y}_j - \bar{x}_j)^2 w(t_j)^2,
\end{aligned} \tag{15}$$

as a function of the threshold Δ. The probability of correct classification is $\Phi(\eta_1(\Delta))$ for a new observation from group 1 and $\Phi(\eta_2(\Delta))$ for a new observation from group 2. The corresponding formulae when using the statistic L are obtained from (13). When looking for the best value of Δ we need to decide whether to maximize $\eta_1(\Delta)$, $\eta_2(\Delta)$, or a combination of the two. In some applications it it much more important to classify samples from one of the two groups correctly than samples from the other group. Thus in a setting of cancer patients one group can be patients needing a particular treatment for improving survival, whereas treatment is not needed in the second group. In such a situation we aim at maximizing $\eta_1(\Delta)$, say, subject to an upper bound for $\eta_2(\Delta)$. An important aspect in this situation is, however, that the strong negative correlation between $\eta_1(\Delta)$ and $\eta_2(\Delta)$ diminishes when the threshold Δ is increased. When the two groups are equally important it becomes of interest to maximize $\min\{\eta_1(\Delta), \eta_2(\Delta)\}$ or, alternatively, to maximize the average $(\eta_1(\Delta) + \eta_2(\Delta))/2$.

Figures 3 – 5 show examples of the probability of correct classification as a function of the threshold. In Figure 3 are examples of the probability of correct classification for both groups using the statistics D. As can be seen in this figure the choice of threshold depends on which group is considered.

In Figure 4 hard and soft thresholding are compared and the use of the statistics D and L are compared. Typically the soft thresholding allows for a broader interval of Δ values for which the probability of correct classification is large. However, the price for this is that more variables are used in the classifier. Also, Figure 4 shows that the use of the statistics D and L when $n = m$ often give comparable results.

In Figure 5 the case of unequal sample sizes $n \neq m$ is illustrated. In the first subplot D is included to show the bias problem mentioned in subsection 2.2. Also included is an obvious modification of D, that solves the bias problem in the case of no thresholding, namely

$$\tilde{D} = \frac{1}{2}\sum_{j=1}^{p} \big[(z_j - \bar{x}_j)^2/(1 + 1/n) - (z_j - \bar{y}_j)^2/(1 + 1/m)\big] \frac{w(t_j)}{s_j^2}. \tag{16}$$

As can be seen from this subplot the bias when using D is very serious and this is not solved by the use of $\tilde{D}$ in the case of thresholding. Generally, there is a large

variation in the optimal improvement that can be achieved using the thresholding idea.

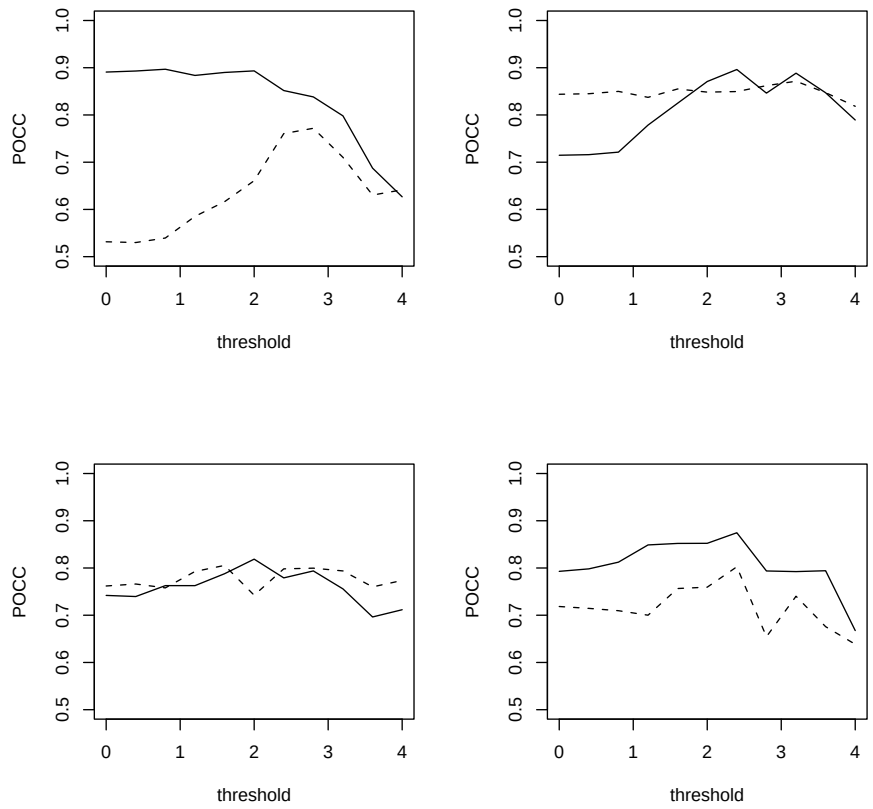

Figure 3: Probability of correct classification (POCC) based on D in (10) as a function of the threshold Δ. The simulations are with $n = m = 15$, $p = 1000$, and with 20 variables having a nonzero differential expression of size $\delta = 1$. Full drawn line: POCC for group 1; dashed line: POCC for group 2.

Tables 3 and 4 give median values, based on 100 simulations, for the probability of correct classification using the optimal threshold. A rough rule is that the distribution of the improvement $\max_\Delta \eta(\Delta) - \eta(0)$ is independent of the probability of correct classification without thresholding, $\Phi(\eta(0))$. When comparing Table 3 with Table 2 we see that the use of an optimal data dependent threshold gives a somewhat larger improvement in the probability of correct classification than the use of a fixed threshold. Another difference to Table 2 is that there is very little difference between the use of a hard and a soft threshold.

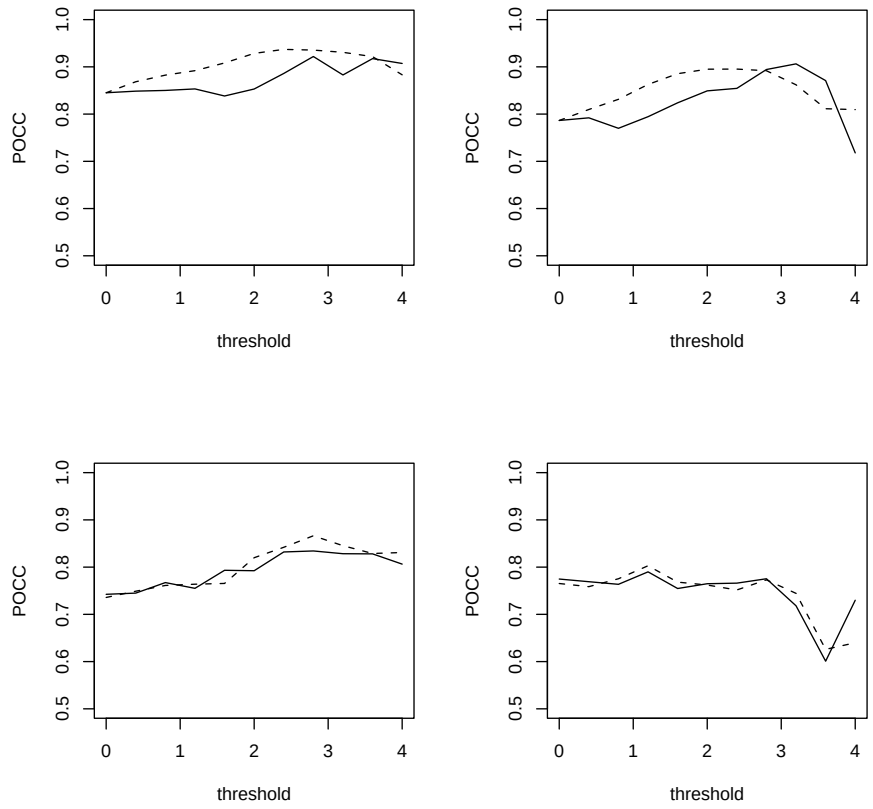

Figure 4: Probability of correct classification (POCC) for group 1 as a function of the threshold Δ. The simulations are with $n = m = 15$, $p = 1000$, and with 20 variables having a nonzero differential expression of size $\delta = 1$. In the two upper subplots hard (full drawn line) and soft (dashed line) thresholding are compared for the statistic D. The soft thresholding is (12) with $\theta = 0$. In the two lower subplots hard thresholding for D (full drawn line) and L (dashed line) are compared.

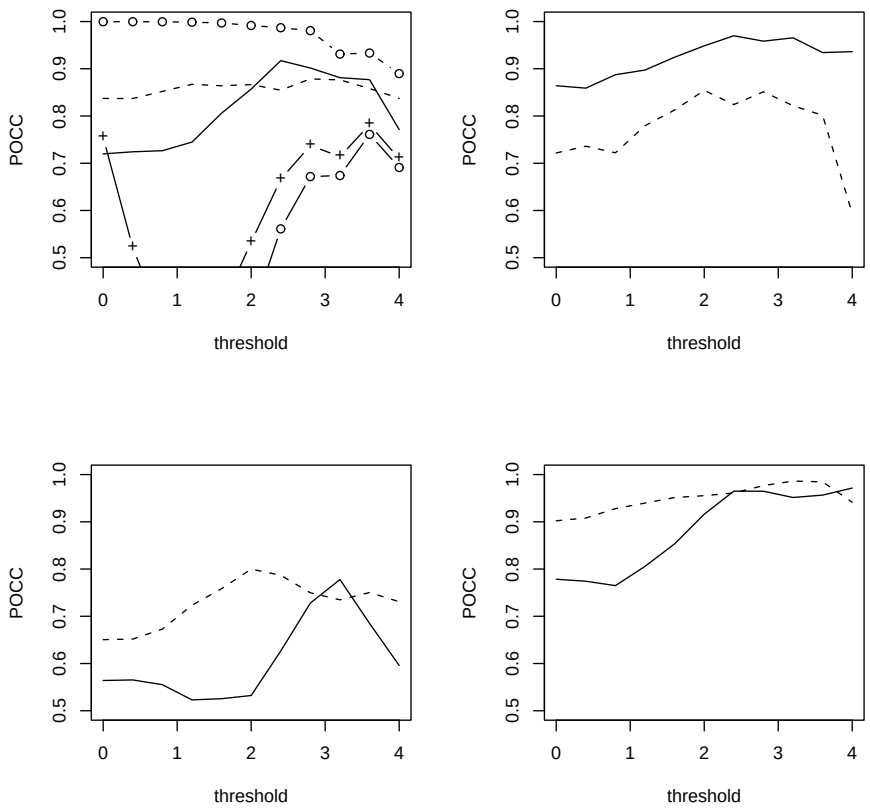

Figure 5: Probability of correct classification (POCC) using the statistic L in the case of unequal sample sizes. In all four subplots $n = 10$ and $m = 30$, and the full drawn lines are for group 1 and the dashed lines are for group 2. Except for the last subplot there are 20 variables having a nonzero differential expression of size $\delta = 1$. In the last subplot there are 80 variables instead of 20. The two upper subplots have $p = 1000$ and the two lower ones have $p = 10000$. The upper left subplot compares hard thresholding for D (lines and points) and L (lines only). Included also is the statistic $\tilde{D}$ in (16) (dashed line with +).

k	δ	p	D/L	no	h	s	av	ha	sa
20	1	1000	D	74	82	81	73	79	79
			L	73	84	82	73	79	80
20	*	1000	D	71	84	86	71	83	84
			L	71	87	87	71	83	84
20	1	10000	D	58	68	67	58	62	62
			L	58	69	68	58	62	62
20	*	10000	D	57	71	72	57	67	69
			L	57	72	74	57	68	70
80	1	10000	D	80	88	88	79	85	87
			L	79	89	89	79	86	87
80	*	10000	D	77	92	94	77	91	94
			L	77	94	94	77	92	93

Table 3: Median probability of correct classification in percent using optimal threshold for the two classification statistics D and L, and for the case of equal sample sizes $n = m = 10$. Numbers are based on 100 simulated values. Each pair of rows give the results for D and L. The column no is for the case of no thresholding, whereas h and s signify hard and soft thresholding, respectively. The last three columns give the values for the average probability for the two groups, av is for the case of no thresholding, ha is hard thresholding, and sa is soft thresholding. Hard thresholding is as in (11) and soft thresholding is given in (12) with $\theta = 0$. The number of differentiable expressed variables is k. When the δ entry is 1 all the nonzero δs are equal to 1, and when the entry is an asterisk the δ values $(0.5, 1.0, 1.5, 2.0)$ are used in the proportions $(0.5, 0.35, 0.10, 0.05)$.

k	δ	p	no	$h1$	$s1$	$h2$	$s2$	av	ha	sa
20	1	1000	78	92	91	91	89	78	88	89
20	*	1000	76	94	94	92	91	76	92	92
80	1	10000	85	98	98	98	98	84	97	98
80	*	10000	82	100	100	99	99	82	99	99

Table 4: Median probability of correct classification in percent using optimal threshold for the classification statistic L, and for the case of unequal sample sizes $n = 10$ and $m = 30$. Numbers are based on 100 simulated values. The column no is for the case of no thresholding, whereas $h1$ and $h2$ signify hard thresholding for the two groups, $s1$ and $s2$ soft thresholding, and ha and sa are hard and soft thresholding for the average of the two groups. Hard thresholding is as in (11) and soft thresholding is given in (12) with $\theta = 0$. The number of differentiable expressed variables is k. When the δ entry is 1 all the nonzero δs are equal to 1, and when the entry is an asterisk the δ values $(0.5, 1.0, 1.5, 2.0)$ are used in the proportions $(0.5, 0.35, 0.10, 0.05)$.

3.2 Shrunken centroids

Tibshirani et al. (2003) consider classification based on a statistic similar to D_0 in (3), but with the group averages $\bar{x}$ and $\bar{y}$ replaced by shrunken averages. This means that the group averages are shifted towards the overall average. Let the overall average for gene j be $a_j = (n\bar{x}_j + m\bar{y}_j)/(n+m)$ and let $w_x = \sqrt{1/n - 1/(n+m)}$ and $w_y = \sqrt{1/m - 1/(n+m)}$. As before $t_j = (\bar{y}_j - \bar{x}_j)/\sqrt{s_j^2(1/n + 1/m)}$ is the t-statistic for difference between the two groups. We can write the shrunken averages as

$$\tilde{x}_j = \begin{cases} a_j & \text{if } |t_j| < \Delta, \\ \bar{x}_j + \text{sign}(\bar{y}_j - \bar{x}_j)\Delta w_x s_j & \text{if } |t_j| \geq \Delta, \end{cases}$$

and

$$\tilde{y}_j = \begin{cases} a_j & \text{if } |t_j| < \Delta, \\ \bar{y}_j - \text{sign}(\bar{y}_j - \bar{x}_j)\Delta w_y s_j & \text{if } |t_j| \geq \Delta, \end{cases}$$

and the classification is based on

$$D_T(z) = \frac{1}{2}\sum_{j=1}^{p} \frac{(z_j - \tilde{x}_j)^2 - (z_j - \tilde{y}_j)^2}{s_j^2} = \sum_{j=1}^{p} z_j \frac{\tilde{y}_j - \tilde{x}_j}{s_j^2} - \frac{1}{2}\sum_{j=1}^{p} \frac{\tilde{y}_j^2 - \tilde{x}_j^2}{s_j^2}. \tag{17}$$

Presumably the motivation for this kind of approach goes back to the James-Stein estimator in statistics. The latter involves shrinkage and has the property that the mean square error (taking the sum over all the variables) is reduced.

Proposition 5. *The classification statistic $D_T(z)$ can be written as*

$$D_T(z) = D(z) + \frac{1}{2}\sum_{j=1}^{p} |t_j| w(t_j)\Delta\left(\frac{1}{m} - \frac{1}{n}\right), \tag{18}$$

where $D(z)$ is the threshold statistic from (10) with the weight function $w(t)$ given in (12) with $\theta = 0$.

Proof. From the definition of $\tilde{x}_j$ and $\tilde{y}_j$ it follows that

$$\begin{aligned} \tilde{y}_j - \tilde{x}_j &= \{\bar{y}_j - \bar{x}_j - \text{sign}(\bar{y}_j - \bar{x}_j)s_j\Delta(m_x + m_y)\}1(|t_j| > \Delta) \\ &= (\bar{y}_j - \bar{x}_j)\Big\{1 - \frac{s_j}{|\bar{y}_j - \bar{x}_j|}\Delta(m_x + m_y)\Big\}1(|t_j| > \Delta) \\ &= (\bar{y}_j - \bar{x}_j)\Big\{1 - \frac{\Delta}{|t_j|}\Big\}1(|t_j| > \Delta), \end{aligned}$$

and when $|t_j| > \Delta$ the average is

$$\tilde{y}_j + \tilde{x}_j = \bar{y}_j + \bar{x}_j - \text{sign}(\bar{y}_j - \bar{x}_j)s_j\Delta(m_x - m_y).$$

This gives

$$\sum_{j=1}^{p} z_j \frac{\tilde{y}_j - \tilde{x}_j}{s_j^2} = \sum_{j=1}^{p} z_j \frac{\bar{y}_j - \bar{x}_j}{s_j^2} w(t_j),$$

with $w(t_j) = (1 - \Delta/|t_j|)1(|t_j| > \Delta)$, and

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{p}\frac{\tilde{y}_j^2 - \tilde{x}_j^2}{s_j^2} &= \frac{1}{2}\sum_{j=1}^{p}\frac{\bar{y}_j - \bar{x}_j}{s_j^2}w(t_j)\{\bar{y}_j + \bar{x}_j - \text{sign}(\bar{y}_j - \bar{x}_j)s_j\Delta(m_x - m_y)\} \\ &= \frac{1}{2}\sum_{j=1}^{p}\frac{\bar{y}_j^2 - \bar{x}_j^2}{s_j^2}w(t_j) - \frac{1}{2}\sum_{j=1}^{p}|t_j|w(t_j)\Delta\left(\frac{1}{m} - \frac{1}{n}\right).\end{aligned}$$

Combining these two expressions the result of the proposition is obtained. □

Proposition 5 shows that when there is the same number of observations in the two groups, $n = m$, the statistic D_T is the same as D from (10), and when $n \neq m$ the two differ by an additive term not dependent on z. Also, when $n \neq m$ the statistic D_T suffers from the same problem as D, that is, when $\delta_j = 0$ the mean of $\{(z_j - \tilde{x}_j) - (z_j - \tilde{y}_j)\}/s_j^2$ is not zero. However, the correction term in (18) goes a long way in alleviating this problem for larger values of the threshold Δ, so that D_T has better properties than D when $n \neq m$. Tibshirani et al. (2003) do not mention the bias problem for $n \neq m$. They do, however, introduce a method called "adaptive choice of threshold" that can help in removing this problem. The classification statistic is still D_T from (17), but now the definition of $\tilde{y}$ is changed to

$$\tilde{y}_j = \begin{cases} a_j & \text{if } |t_j| < \theta\Delta, \\ \bar{y}_j - (1 - \frac{1}{\theta})\frac{n}{n+m}(\bar{y}_j - \bar{x}_j) - \text{sign}(\bar{y}_j - \bar{x}_j)\Delta w_y s_j & \text{if } |t_j| \geq \theta\Delta, \end{cases}$$

where θ is a positive parameter. The definition of $\tilde{x}$ is as before. When $\theta = 1$ we recover the situation from before. To derive a result similar to Proposition 5 define

$$w_1(t) = \begin{cases} \{\frac{n/\theta+m}{n+m} - \frac{\Delta}{|t|}\} & |t| \geq \max\{1,\theta\}\Delta, \\ \frac{n}{n+m}(1/\theta - \frac{\Delta}{|t|}) & \min\{1,\theta\}\Delta < |t| < \max\{1,\theta\}\Delta, \\ 0 & |t| \leq \min\{1,\theta\}\Delta, \end{cases}$$

and

$$h_1(t) = \begin{cases} 1 - \frac{1}{\theta} & |t| \geq \max\{1,\theta\}\Delta, \\ 1 & |t| < \max\{1,\theta\}\Delta, \end{cases} \quad \text{and} \quad h_2(t) = \begin{cases} \frac{1}{m} - \frac{1}{n} & |t| \geq \max\{1,\theta\}\Delta, \\ -\frac{1}{\sqrt{n}} & |t| < \max\{1,\theta\}\Delta. \end{cases}$$

A simple calculation reveals that

$$\tilde{y}_j - \tilde{x}_j = (\bar{y}_j - \bar{x}_j)w_1(t_j),$$

and

$$\tilde{y}_j + \tilde{x}_j = \bar{y}_j + \bar{x}_j - \frac{n}{n+m}h_1(t_j)(\bar{y}_j - \bar{x}_j) - \text{sign}(\bar{y}_j - \bar{x}_j)s_j\Delta h_2(t_j).$$

Inserting these in the classification statistic we find

$$\begin{aligned}D_T(z) = &\sum_{j=1}^{p} z_j\frac{\bar{y}_j - \bar{x}_j}{s_j^2}w_1(t_j) - \frac{1}{2}\sum_{j=1}^{p}\frac{\bar{y}_j^2 - \bar{x}_j^2}{s_j^2}w_1(t_j) \\ &+ \frac{1}{2m}\sum_{j=1}^{p} t_j^2 w_1(t_j)h_1(t_j)\Delta h_2(t_j) + \frac{1}{2}\sum_{j=1}^{p}|t_j|w_1(t_j)\Delta h_2(t_j).\end{aligned} \tag{19}$$

Comparing this with (18) the weight function w has been changed and an extra term has appeared. For a fixed threshold Δ one can presumably choose the parameter θ so as to remove the bias problem for the terms with $\delta_j = 0$. For the case $n = 10$ and $m = 20$ we find the following relation between Δ and θ needed to remove the bias,

θ	0.70	0.80	0.85	0.90
Δ	1.1	2.0	2.8	4.1

Another way of looking at D_T is that the basic classification statistic $\sum_j z_j w_1(t_j)(\bar{y}_j - \bar{x}_j)/s_j^2$ is a thresholded version of L_0 from (7), and the three other terms in D_T in (19) give a cutoff value κ_T for this statistic. To choose a value of θ in practice a crossvalidation step is necessary, and this is basically also what happens in the statistic L from (13).

Tibshirani et al. (2003) also suggest to replace the ordinary t–statistic with a modified version where one adds a constant term to the standard deviation s in the denominator. It seems that this gives a slight reduction in the false discovery rate, and this will be advantageous when selecting genes to include in the classifier. We will, however, not investigate this in detail here.

3.3 Partial Least Squares

A number of papers have appeared recently advocating the use of partial least squares in classifications with many variables and few observations (Nguyen and Roche (2002), Pérez-Enciso and Tenenhaus (2003), Ding and Gentleman (2004), Boulesteix (2004)). The partial least squares idea originates in chemometrics where, as an example, the variables x_j, $j = 1, \ldots, p$, correspond to signals at different wavelengths and the response r is the chemical composition of some compound. Thus, in this setting, the response r is continuous and a natural model is to have (r, x) multivariate normal. This does not quite capture the situation here of two distinct groups, where $r = (1, \ldots, 1, 2, \ldots, 2)$ is the vector of group labels. Nevertheless, in applications the method has been quite successful.

The partial least squares approach (Stone and Brooks (1990)) with K components is based on K linear combinations of the data vector x,

$$w_1'x, w_2'x, \ldots, w_K'x,$$

where $w_1, \ldots, w_K$ are the weight vectors. The first weight vector w_1 is chosen as the unit vector maximizing

$$\mathrm{Cov}(w_1'x, r) = \sum_{j=1}^{p} w_{1j}\mathrm{Cov}(x_j, r). \tag{20}$$

The theoretical solution to this is of course the vector $w_1 = c_1\mathrm{Cov}(x, r)$, where c_1 is a normalizing constant. For a data set $(x_1, r_1), \ldots, (x_M, r_M)$ we replace the theoretical covariance with the estimate $\sum_{i=1}^{M} x_i(r_i - \bar{r})/(M-1)$. In the situation of this paper, with $M = n + m$ and with $r_i = 1$ for an observation from group 1 and $r_i = 2$ for an observation from group 2, the first weight vector w_1 becomes

$$w_{1j} = c_1 \frac{nm}{(M-1)M}(\bar{y}_j - \bar{x}_j) = \tilde{c}_1(\bar{y}_j - \bar{x}_j).$$

If we use only one component, $K = 1$ in the partial least squares approach, the classification of a new observation z is based on the statistic

$$u_1 = \sum_{j=1}^{p} z_j(\bar{y}_j - \bar{x}_j). \tag{21}$$

Thus, the difference to the statistic L_0 in (7) is that in (21) there is no standardization by the within group variance estimate s_j^2. Calculating the mean of u_1 with respect to the distribution of z, $E_N u_1$, when z belongs to group1 and when z belongs to group 2, the difference is $\sum_{j=1}^{p} \delta_j \sigma_j (\bar{y}_j - \bar{x}_j)$. The same difference of means for L_0 in (7) is $\sum_{j=1}^{p} \delta_j \sigma_j (\bar{y}_j - \bar{x}_j)/s_j^2$.

The second partial least squares component, given by the weight vector w_2, is also obtained by maximizing the correlation as in (20), but now w_2 has to be orthogonal to w_1 in the sense

$$w_2' \Sigma w_1 = 0,$$

where $\Sigma = \mathrm{Var}(x)$. The solution to this is

$$w_2 = c_2 \Big\{ w_1 - \Big(\frac{w_1' \Sigma w_1}{w_1' \Sigma^2 w_1} \Big) \Sigma w_1 \Big\}, \tag{22}$$

which follows from

$$\mathrm{Cov}(w_2' x, r) = \mathrm{Cov}\Big(w_2' \Big\{ x - \Big(\frac{x' \Sigma w_1}{w_1' \Sigma^2 w_1} \Big) \Sigma w_1 \Big\}, r \Big),$$

whenever $w_2' \mathrm{Var}(x) w_1 = 0$. In practice Σ is replaced by the sample variance $S = X'X$, where X is the $M \times p$ data matrix with the average subtracted for each variable. It is clear from (22) that the space spanned by $\{w_1, w_2\}$ is the same as that spanned by $\{w_1, Sw_1\}$, and the second classification statistic supplementing (21) can be taken as

$$u_2 = \sum_{j=1}^{p} z_j \Big\{ \sum_{r=1}^{p} S_{jr} (\bar{y}_r - \bar{x}_r) \Big\}.$$

For K components the space spanned by $\{w_1, \ldots, w_K\}$ is the same as that spanned by $\{w_1, Sw_1, \ldots, S^{K-1} w_1\}$. Having selected the number of partial least squares components K the classification involves a second stage where a classifier is being build from $u_1, \ldots, u_K$. The methods used most are either logistic regression or a maximum likelihood classifier (see the references at the start of this subsection). This step corresponds in our setting to choosing a value for the cutoff point κ_0 in (7). The bias problem mentioned in subsection 2.2 reappears here, and it seems that this is not addressed in papers like Nguyen and Roche (2002) and Boulesteix (2004).

We now give an interpretation of the partial least squares method within the classification setting. In the setup of (1) it is assumed that the variables in x are independent. The variance matrix is then $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. In the case where the variables are scaled to have standard deviation one (a step that is often done in partial least squares applications), the theoretical partial least squares components $\Sigma^k w_1$ are therefore equal to the first component w_1 and, furthermore, the maximum

likelihood classifier is based on $z'w_1$. In this case the optimal number of partial least squares components is 1. In the more general model with an arbitrary variance matrix Σ the maximum likelihood classifier is based on the statistic $z'\Sigma^{-1}w_1$. In practice with $p \gg M$ we cannot invert the sample variance S. However, we can view $I, S, S^2, \ldots$ as terms in an expansion of a generalized inverse. Formally, if we write $S^{-1} = [I + (S - I)]^{-1} = I + (S - I) + (S - I)^2 + \cdots$ we see that $S^{-1}w_1$ gives the terms $w_1, Sw_1, S^2w_1, \ldots$ Wold et al. (1984) explain the relation between partial least squares and generalized inverse. When looked at this way the partial least squares approach with more than one component may potentially give a small improvement over the classifier based on the first component only.

In the classification setting the usual variance estimate $S = X'X$ does not seem to be the most appropriate. This does not take into account the division of the data into two groups. Thus, a better estimate, using the notation of this paper, is

$$\frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' + \sum_{i=1}^{m}(y_i - \bar{y})(y_i - \bar{y})'\right\}.$$

In Boulesteix (2004) the partial least squares method is seen as a way of reducing the dimension by itself. However, using one component only, which resembles the method studied in this paper, we have seen a strong need to reduce the number of variables. Most other authors (Nguyen and Roche (2002), Pérez-Enciso and Tenenhaus (2003), Ding and Gentleman (2004)) actually reduce the number of variables before applying the partial least squares idea, including only variables with a large value of a t-statistic.

Nguyen and Roche (2002) standardize variables to have variance 1, select genes based on the two-sample t-test with unequal variances, and use three components in the partial least squares approach. In the second step of the classifier both logistic regression and discriminant analysis are investigated. Five data sets (ovarian, leukemia, lymphoma, colon) are being considered. The classifier generally works well in the range of 50–1000 genes included.

Pérez-Enciso and Tenenhaus (2003) use a data set on breast cancer. As compared to Nguyen and Roche (2002) an alternative (called VIP) to the t-statistic is used to select genes.

Boulesteix (2004) uses the same approach as Nguyen and Roche (2002), and makes a comparison with a number of other methods for five data sets with two groups and four data sets with more than two groups. There is an indication in this paper that the partial least squares approach can be of help in the situation where the basic assumption of two groups no longer holds. Thus in the situation where one of the two groups really contains two subgroups the use of one component in the partial least squares method sometimes gives poor results, whereas the use of two components can improve the performance.

In the above mentioned papers the original partial least squares approach was used together with a second step based on the selected partial least squares components. In Ding and Gentleman (2004) the two steps are integrated using ideas from estimation in generalized linear models and taking into account that the response variable is not continuous. A prior gene filtering is performed based on the usual t–statistic.

4 Estimating the threshold: crossvalidation

In the previous section we considered what is theoretically achievable using the optimal threshold to improve the classifier. In practice the optimal value of the threshold Δ is not available to us, and the latter needs to be estimated from the data.

A common approach to estimation of Δ is to use crossvalidation. In the *leave one out* (or $(n+m)$-fold) crossvalidation one observation is taken out and used as test set, while the classifier is constructed from the remaining $n+m-1$ observations. To describe this in detail consider the classifier based on D from (10). When the observation x_i is taken out let $D^{x_i}(x_i, \Delta)$ be the classification statistic based on the training set $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, y_1, \ldots, y_m)$ and evaluated at the "new observation" x_i, and with a similar notation when one of the y observations is left out. We then look at the number of correctly classified samples

$$\left(\sum_{i=1}^{n} 1(D^{x_i}(x_i, \Delta) < 0), \sum_{i=1}^{m} 1(D^{y_i}(y_i, \Delta) > 0)\right)$$

as a function of Δ, and choose a suitable value of Δ based on these numbers. When the two groups are equally important one possibility is to use

$$\frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} 1(D^{x_i}(x_i, \Delta) < 0) + \frac{1}{m}\sum_{i=1}^{m} 1(D^{y_i}(y_i, \Delta) > 0))\right\} \tag{23}$$

To avoid the discrete nature of the estimates in (23), we can use $D^{x_i}(x_i, \Delta)$ and $D^{y_i}(y_i, \Delta)$ to estimate ξ_N, $\tilde{\xi}_N$, and τ_N, and use the latter to estimate the probability of correct classification. Thus we take

$$\hat{\xi}_N = \frac{1}{n}\sum_{i=1}^{n} D^{x_i}(x_i, \Delta), \qquad \hat{\tilde{\xi}}_N = \frac{1}{m}\sum_{i=1}^{n} D^{y_i}(y_i, \Delta),$$

$$\hat{\tau}_N^2 = \frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(D^{x_i}(x_i, \Delta) - \hat{\xi}_N)^2 + \sum_{i=1}^{n}(D^{y_i}(y_i, \Delta) - \hat{\tilde{\xi}}_N)^2\right\},$$

and use

$$\tfrac{1}{2}\left\{\Phi(-\hat{\xi}_N/\hat{\tau}_N) + \Phi(-\hat{\tilde{\xi}}_N/\hat{\tau}_N)\right\} \tag{24}$$

as our estimate of the probability of correct classification. Here we give the same weight to the two groups. In Figure 6 are some examples of the use of (23) and (24). Generally these measures cannot be trusted in terms of giving a useful value for the probability of correct classification, but they can still be helpful for choosing a value of the threshold parameter Δ.

In Table 5 are average properties based on 1000 simulations. The table shows that an appreciable part of the optimal improvement is achieved also in the case where the threshold parameter is being estimated. The use of (23) or (24) give almost the same results. The hard and soft threshold give different results. The hard threshold, with an estimated threshold parameter, gives almost always an improvement as compared to the case of no threshold. This is contrary to the

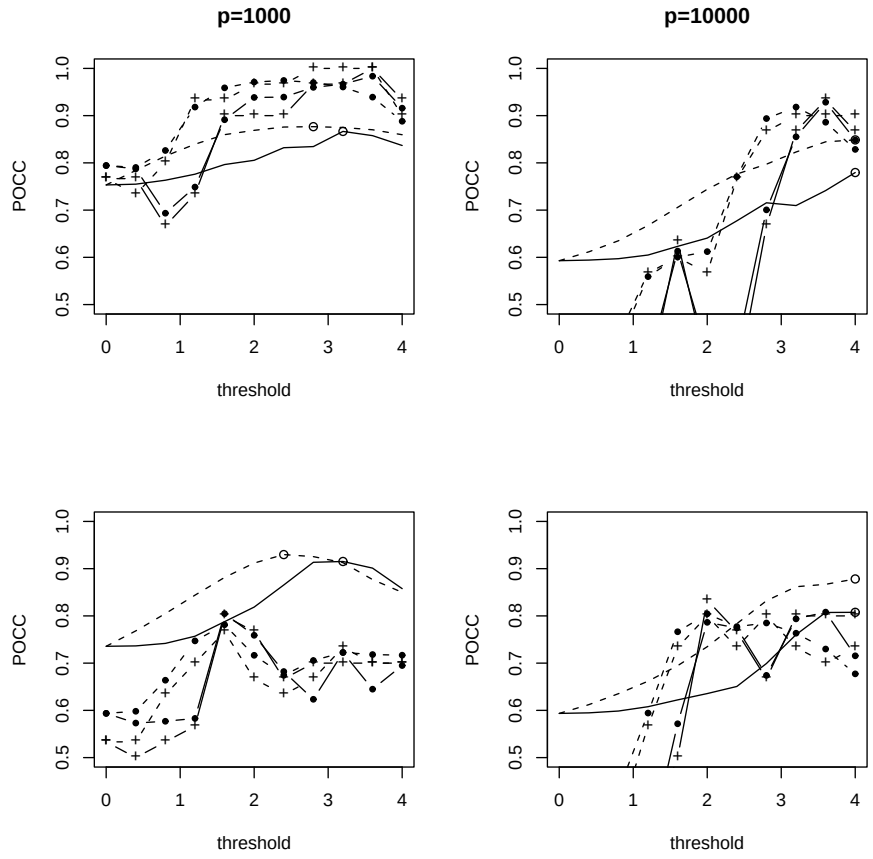

Figure 6: Examples of crossvalidation based estimates of the probability of correct classification (POCC). In all four subplots $n = m = 15$, there are 20 differentiable expressed variables, $\delta = 0.5$ for 10 variables, $\delta = 1.0$ for 7 variables, $\delta = 1.5$ for 2 variables, and one variable has $\delta = 2$. The total number of variables is $p = 1000$ in the first column and $p = 10000$ in the second column. The full drawn lines are for hard thresholding and the dashed lines are for soft thresholding. The lines with no marks (except for a circle showing the location of the maximum value) give the true probability of correct classification. The lines marked with + give the crossvalidation errors from (23), and the lines marked by bullets give (24).

soft threshold where an appreciable fraction of the cases give a lower probability of correct classification than the use of no threshold. Still, the median value is higher for the soft threshold than for the hard threshold, especially so for the case of $p = 10000$ variables. The column headed B in Table 5 will be explained in the next section.

		hard				soft			
p	no	max	cv-d	cv-e	B	max	cv-d	cv-e	B
1000	75	91	88 (0)	88 (0)	89 (0)	92	88 (18)	89 (18)	90 (17)
10000	59	80	75 (0)	75 (0)	79 (1)	85	80 (15)	80 (14)	80 (14)

Table 5: Median probability of correct classification in percent using data dependent threshold for the classification statistic D. Numbers are based on 1000 simulated values with sample sizes $n = m = 15$ and with 20 differentiable expressed variables as in Figure 6. The column *no* gives the probability of correct classification in the case of no thresholding, and the column headed *max* gives the result using the optimal threshold. The columns headed *cv-d* and *cv-e* are based on (23) and (24), respectively. Thus, in these cases the threshold parameter is found by maximizing (23) or (24). The column headed B is based on the Bayes modeling in Section 6. The numbers in parentheses are the percentages of the simulated cases for which the method used gives a lower probability than the use of no threshold.

5 Estimating POCC: crossvalidation

Above we used crossvalidation to choose a value of the threshold parameter Δ. The idea was that from the estimates (23) or (24) of the probability of correct classification we took the value $\hat{\Delta}$ of Δ giving the highest value. Intuitively, it is clear that the maximum of (24), say, will exaggerate the probability of correct classification when using $\hat{\Delta}$ for the threshold. The examples in Figure 6 illustrate this "overshoot".

Let us try to understand the problem through a very simple model. Let v_i, $i = 1, \ldots, k$, be a function that is estimated through $\hat{v}_i = v_i + u_i$. Let us assume that the u_is are independent and $N(0, \omega^2)$ distributed. Let $I = \arg\max_i \hat{v}_i$. Then the overshoot is

$$\max_i \hat{v}_i - v_I.$$

We consider the case where $v_i = \alpha + \beta i$. The mean of the overshoot becomes $\omega M_k(\beta/\omega)$, where M_k is given by

$$M_k(\eta) = \sum_{j=1}^{k} \int_{-\infty}^{\infty} z\varphi(z) \prod_{i \neq j} \Phi(z + \eta(j - i))dz. \tag{25}$$

Note that the value of α does not enter the mean of the overshoot and that β enters through β/ω. Partial integration shows that $M_2(\eta) = \exp(-\eta/4)/\sqrt{\pi}$. Intuitively, it is clear that for $\eta = \beta/\omega$ large the value of $M_k(\eta)$ will be close to $M_2(\eta)$ since with

a high probability $\max_i \hat{v}_i = \max\{\hat{v}_{k-1}, \hat{v}_k\}$ due to the the structure $v_i = \alpha + \beta i$. Values for $M_k(\eta)$, obtained by numerical integration, are given in Table 6.

η	0	0.5	1	2
$k=2$	0.54	0.53	0.44	0.21
$k=3$	0.85	0.75	0.54	0.21
$k=4$	1.03	0.85	0.56	0.21
$k=5$	1.16	0.89	0.56	0.21

Table 6: The mean overshoot $M_k(\eta)$ as given in 25.

Now, instead of (24) let us use $\Phi((-\hat{\xi}_N + \hat{\tilde{\xi}}_N)/(2\hat{\tau}_N))$ as our estimate of the probability of correct classification. Let $\hat{v}(\Delta) = (-\hat{\xi}_N + \hat{\tilde{\xi}}_N)/(2\hat{\tau}_N)$ be an estimate of $v(\Delta) = (-\xi_N + \tilde{\xi}_N)/(2\tau_N)$, where we have shown the dependency on the threshold explicitly in v and $\hat{v}$. Taking a discrete set of Δ values we use $\hat{v}_i = \hat{v}(\Delta_i)$. Let us now make the rough approximation that the terms entering $\hat{\xi}_N$ and $\hat{\tilde{\xi}}_N$ are independent and normally distributed with mean a and variance b^2. The mean and variance of $\hat{v}_i$ are then approximately

$$E(\hat{v}_i)) \approx \frac{a}{b}\frac{f-1/4}{f-1}, \qquad V(\hat{v}_i)) \approx \frac{f^2}{4nm(f-2)} + (\frac{a}{b})^2[\frac{f}{f-2} - (\frac{f-1/4}{f-1})^2].$$

If $\hat{v}_i$ is roughly unbiased we have $v_i \approx a/b$, and if these do not vary much we replace a/b by $\bar{v}$ in the variance formula. Taking $n = m$ we get approximately that the variance of $\hat{v}_i$ is

$$\omega^2 = \frac{1}{2n} + (\bar{v})^2\frac{1}{4n}. \tag{26}$$

Realistic and interesting values of $\bar{v}$ is in the range 1 to 2. To reduce the correlation between $\hat{v}_i$ (remember when deriving (25) the variables were assumed independent) we take $\hat{v}_i = \hat{v}(i)$, $i = 0, 1, 2, 3, 4$. A realistic value of β is then in the range 0 to 0.1. Combining this with (26) for moderate values of n we find from Table 6 that taking $M_k \approx 1$ gives a realistic value. Thus the overshoot is of the order

$$\omega \approx \frac{1}{\sqrt{n}},$$

for $\bar{v} \approx 2$.

There are many rough approximations in the above calculation so in Table 7 are some mean values of the overshoot based on simulations.

6 Bayes modeling

The classifier D_0 in (3) is obtained from the theoretical counterpart $\tilde{D}$ in (2) by inserting estimates for μ_j, $\mu_j + \delta_j\sigma_j$ and σ_j^2. Thus $3p$ parameters are being estimated freely. An alternative to this is to use a Bayes model to regularize the parameter estimation. One possibility is to take $(\sigma_j^2, \mu_j, \delta_j)$ independent and identically

p	$n=m$	k=0	k=20	k=80
1000	15	0.35 (0.33)	0.29 (0.42)	0.35 (0.50)
	30	0.30 (0.22)	0.17 (0.25)	0.25 (0.38)
10000	15	0.28 (0.34)	0.15 (0.44)	0.32 (0.56)
	30	0.35 (0.24)	0.20 (0.32)	0.27 (0.34)

Table 7: The mean overshoot based on 100 simulated values. The threshold is found by maximizing over the values $\Delta = 0.4, 0.8, \ldots, 4.0$. The values in parentheses are the standard deviations of the 100 simulated values.

distributed with

$$\begin{aligned} \frac{1}{\sigma_j^2} &\sim \frac{1}{s_0^2}\chi^2(d_0)/d_0, \\ \mu_j|\sigma_j^2 &\sim N(\mu_0, \tau^2\sigma_j^2), \\ \delta_j|\delta_j \neq 0, \sigma_j^2 &\sim N(0, v_0\sigma_j^2), \\ P(\delta_j \neq 0) &= p_0. \end{aligned}$$

This prior model has also been used in Smyth (2004) and previously in Lönnstedt and Speed (2002). When the hyper parameters $(d_0, s_0^2, \tau^2, v_0, p_0)$ have been estimated a full Bayesian classifier can be constructed. Details of this classifier are given in Appendix B. This approach is, however, not of main interest to us. Instead, we want to explore the use of the Bayesian model as a help to establish the optimal threshold in our classifier.

Let us first discuss how to estimate the hyper parameters. Estimates of d_0 and s_0^2 are obtained from the empirical variances s_j^2. Since $s_j^2|\sigma_j^2 \sim \sigma_j^2\chi^2(f)/f$ and $1/\sigma_j^2 \sim (1/s_0^2)\chi^2(d_0)/d_0$ one has $s_j^2 \sim s_0^2 F(f, d_0)$. Robust estimates of d_0 and s_0^2 are obtained from the equations

$$\frac{\mathrm{M}(F(f,d_0))}{\mathrm{IQR}(F(f,d_0))} = \frac{\mathrm{M}(s^2)}{\mathrm{IQR}(s^2)} \qquad \text{and} \qquad s_0^2\mathrm{M}(F(f,d_0)) = \mathrm{M}(s^2),$$

where M is the median and IQR is the inter quartile range. Next, consider estimation of the hyper parameters μ_0 and τ^2 in $\mu_j \sim N(\mu_0, \tau^2\sigma_j^2)$. Since $(n\bar{x}_j + m\bar{y}_j)/(n+m)$ has a symmetric distribution around μ_0, we take

$$\hat{\mu}_0 = \mathrm{M}((n\bar{x} + m\bar{y})/(n+m)).$$

Furthermore, it is easy to see that

$$\begin{aligned} \frac{n(\bar{x}_j - \mu_0) + m(\bar{y}_j - \mu_0)}{n+m}\Big|\sigma_j \sim{}& (1-p_0)\big[N\big(0, \sigma_j^2(1/(n+m) + \tau^2)\big)\big] \\ &+ p_0\big[N\big(0, \sigma_j^2(1/(n+m) + \tau^2 + (m/(n+m))^2 v_0)\big)\big], \end{aligned}$$

so that

$$\begin{aligned} \frac{n(\bar{x}_j - \mu_0) + m(\bar{y}_j - \mu_0)}{\sqrt{(n+m)s_j^2}} \sim{}& (1-p_0)\big[\{1 + (n+m)\tau^2\}^{1/2} t(f)\big] \\ &+ p_0\big[\{1 + (n+m)(\tau^2 + (m/(n+m))^2 v_0)\}^{1/2} t(f)\big], \end{aligned} \tag{27}$$

where $(1-p_0)[F]+p_0[F_2]$ denotes a mixture of the two distributions F_1 and F_2. Define

$$\tilde{t}_j = \frac{n(\bar{x}_j - \hat{\mu}_0) + m(\bar{y}_j - \hat{\mu}_0)}{\sqrt{(n+m)s_j^2}}.$$

Since in the applications we have in mind p_0 is very small we use the first term of (27) only and estimate τ^2 by solving

$$\{1+(n+m)\tau^2\}^{1/2}\,\mathrm{IQR}(t(f)) = \mathrm{IQR}(\tilde{t}_j).$$

Finally, consider the fraction of expressed variables p_0 and the scaled variance v_0 of the differential expression. We use the differences $\bar{y}_j - \bar{x}_j$ together with s_j^2 for the estimation. Let $t_j = (\bar{y}-\bar{x})/\sqrt{s_j^2(1/n+1/m)}$. Then

$$t_j \sim (1-p_0)[t(f)] + p_0[\{1+nmv_0/(n+m)\}^{1/2} t(f)]. \tag{28}$$

It is clear from this formula, that when p_0 is small there is little information in the data to estimate p_0 and v_0. To illustrate the lack of information let us consider an estimating equation obtained by equating the average of $|t_j|$ to the theoretical mean, that is,

$$E(|t(f)|)\big\{1+p_0[(1+nmv_0/(n+m))^{1/2}-1]\big\} = \frac{1}{p}\sum_{j=1}^{p}|t_j|. \tag{29}$$

The standard deviation of the average is roughly $\sqrt{0.4/p}$ and the term to be estimated is roughly $\eta = 0.8p_0[(1+nmv_0/(n+m)^{1/2}-1]$. Let the number of differentiable expressed variables be k so that $p_0 = k/p$, and let $v_0 = 1$. A few examples of the standard deviation and the value of η are:

		$p=1000$		$p=10000$	
$n=m$	k	sd	η	sd	η
10	20	0.02	0.02	0.006	0.002
10	80	0.02	0.09	0.006	0.009
20	20	0.02	0.04	0.006	0.004
20	80	0.02	0.15	0.006	0.015

If, instead, we look at the number of t_js with an absolute value greater than a chosen cutoff x, we compare below the expected numbers from a $t(f)$ distribution with the expected numbers from the mixture distribution (28) for the case $n=m=15$:

	$p=1000$			$p=10000$		
		mixture			mixture	
x	$t(f)$	$k=20$	$k=80$	$t(f)$	$k=20$	$k=80$
2	55	64	91	553	562	588
3	5.6	12	30	56	62	81
4	0.4	4.0	15	4.2	7.8	19

To illustrate the meaning of these numbers, assume that v_0 is known and only p_0 has to be estimated. Consider the entry in the table above with $x = 3$ and $k = 20$. The expected number from the mixture distribution is 12 and that from the $t(f)$ distribution is 5.6, which implies that in roughly 5% of the cases the estimate of p_0 becomes zero. The performance of the maximum likelihood estimate of p_0 with v_0 fixed is illustrated below. We have taken $n = m = 15$ and $v_0 = 1$ and simulated 100 samples with $p_0 = k/p$. The following table gives the 10%, 50%, and 90% quantiles of the estimated values of k:

	$p = 1000$			$p = 10000$		
k	10%	50%	90%	10%	50%	90%
20	13	24	35	10	31	47
80	74	90	106	74	104	132

These numbers show a bias towards larger values of p_0 as well as a large spread in the estimates. We can extend the maximum likelihood estimation to cover the estimation of both p_0 and v_0. However, as for example (29) shows, the data mostly supply information on a combination like $p_0[(1 + nmv_0/(n + m)^{1/2} - 1]$. When p_0 is very small it is probably better to fix v_0, say from previous experience, and estimate p_0 only.

Having estimated the hyper parameters we want to use the posterior means of μ_j and δ_j to calculate a posterior estimate of the probabilities of correct classification (14) and (15). Let I_j be one if variable j is differentiable expressed and zero if it is not differentiable expressed. The posterior means can be written as

$$\begin{aligned} &E(\mu|\bar{x}, \bar{y}, s^2) \\ &\quad = E(\mu|I = 0, \bar{x}, \bar{y}, s^2)P(I = 0|\bar{x}, \bar{y}, s^2) + E(\mu|I = 1, \bar{x}, \bar{y}, s^2)P(I = 1|\bar{x}, \bar{y}, s^2), \end{aligned} \tag{30}$$

and

$$E(\delta|\bar{x}, \bar{y}, s^2) = +E(\delta|I = 1, \bar{x}, \bar{y}, s^2)P(I = 1|\bar{x}, \bar{y}, s^2), \tag{31}$$

where subscript j has been left out. In the proposition below subscript j is not included either.

Proposition 6. *Let I be one if a variable is differentiable expressed and zero if it is not differentiable expressed. Define $q(n, m) = n + m + 1/\tau^2$,*

$$A = \frac{nm(\bar{y} - \bar{x})^2 + (n\bar{x}^2 + n\bar{y}^2)/\tau^2}{q(n, m)},$$

and

$$B = \frac{mn(\bar{y} - \bar{x})^2/v_0 + (n\bar{x}^2 + m\bar{y}^2)/(v_0\tau^2) + nm\bar{x}^2/\tau^2}{m(n + 1/\tau^2) + q(n, m)/v_0}.$$

We then have the following posterior statements

$$\begin{aligned} &P(I = 0|\bar{x}, \bar{y}, s^2) \\ &\quad = \Big\{1 + \frac{p_0}{1 - p_0}\Big[\frac{A + fs^2 + d_0s_0^2}{B + fs^2 + d_0s_0^2}\Big]^{(f+d_0+2)/2}\Big(\frac{q(n, m)/v_0}{m(n + 1/\tau^2) + q(n, m)/v_0}|\Big)^{1/2}\Big\}^{-1}, \end{aligned}$$

$$E(\sigma^2|s^2) = \frac{fs^2 + d_0 s_0^2}{f + d_0 - 2},$$
$$E(\mu|I=0,\bar{x},\bar{y},s^2) = \frac{n\bar{x} + m\bar{y}}{n + m + 1/\tau^2},$$
$$E(\mu|I=1,\bar{x},\bar{y},s^2) = \frac{n\bar{x} + \bar{y}/(v_0 + 1/m)}{n + 1/(v_0 + 1/m) + 1/\tau^2},$$

and

$$E(\delta|I=1,\bar{x},\bar{y},s^2) = \frac{nm(\bar{y}-\bar{x}) + m\bar{y}/\tau^2}{(n + m + 1/\tau^2)/v_0 + m(n + 1/\tau^2)}.$$

Proof. The joint density of $(\bar{x},\bar{y},s^2,I,\delta,\mu,v)$, where $v = 1/\sigma^2$, is

$$\begin{cases} (1-p_0)\frac{C}{\sqrt{2\pi\sigma^2}} v^{(f+d_0+2)/2-1} \exp\{-\frac{1}{2\sigma^2} f(\bar{x},\bar{y})\} & I=0, \\ p_0 \frac{C}{2\pi\sigma^2\sqrt{v_0}} v^{(f+d_0+2)/2-1} \exp\{-\frac{1}{2\sigma^2}[f(\bar{x},\bar{y}-\delta) + \delta^2/v_0]\} & I=1, \end{cases} \tag{32}$$

where

$$f(\bar{x},\bar{y}) = n(\bar{x}-\mu)^2 + m(\bar{y}-\mu)^2 + \mu^2/\tau^2 + fs^2 + d_0 s_0^2, \tag{33}$$

and

$$C = \frac{\sqrt{nm}(f/2)^{f/2}(d_0 s_0^2/2)^{d_0/2}(s^2)^{f/2-1}}{2\pi\tau\Gamma(f/2)\Gamma(d_0/2)}. \tag{34}$$

To find the posterior probability for I we integrate the first of (32) with respect to μ and then v. To this end we write

$$f(\bar{x},\bar{y}) = q(n,m)\Big(\mu - \frac{n\bar{x} + m\bar{y}}{q(n,m)}\Big)^2 + A + fs^2 + d_0 s_0^2,$$

where A is given in the proposition. This shows that

$$\mu|(I=0,\bar{x},\bar{y},s^2,\sigma^2) \sim N\Big(\frac{n\bar{x} + m\bar{y}}{q(n,m)}, \frac{\sigma^2}{q(n,m)}\Big). \tag{35}$$

After integrating over μ we get

$$(1-p_0)\frac{C}{\sqrt{q(n,m)}} v^{(f+d_0+2)/2-1} \exp\big\{-\frac{1}{2\sigma^2}[A + fs_0^2 + d_0 s_0^2]\big\},$$

and integration with respect to v gives

$$C_1(1-p_0)[A + fs_0^2 + d_0 s_0^2]^{-(f+d_0+2)/2}, \tag{36}$$

where

$$C_1 = C\frac{2^{(f+d_0+2)/2}}{\sqrt{q(n,m)}}. \tag{37}$$

In a similar fashion, looking at the second expression in (32), we see that

$$\begin{aligned} \mu|(I=1,\bar{x},\bar{y},s^2,\delta,\sigma^2) &\sim N\Big(\frac{n\bar{x}^+ m(\bar{y}-\delta)}{q(n,m)}, \frac{\sigma^2}{q(n,m)}\Big), \\ \delta|(I=1,\bar{x},\bar{y},s^2,\sigma^2) & \\ &\sim N\Big(\frac{m[n(\bar{y}-\bar{x}) + y/\tau^2]}{m(n + 1/\tau^2 + q(n,m)/v_0}, \frac{\sigma^2 q(n,m)}{m(n + 1/\tau^2 + q(n,m)/v_0}\Big), \end{aligned} \tag{38}$$

and integration with respect to μ, δ, and v, gives

$$C_1 p_0 [B + f s_0^2 + d_0 s_0^2]^{-(f+d_0+2)/2} [1 + v_0 m(n + 1/\tau^2)/q(n,m)]^{-1/2}, \tag{39}$$

where B is given in the proposition. Combining (36) and (39) the stated expression for $P(I = 0|\bar{x}, \bar{y}, s^2)$ is obtained.

From (35) follows

$$E(\mu|I = 0, \bar{x}, \bar{y}, s^2) = (n\bar{x} + m\bar{y})/q(n,m).$$

Equation (38) gives

$$E(\delta|I = 1, \bar{x}, \bar{y}, s^2) = \frac{m[n(\bar{y} - \bar{x}) + y/\tau^2]}{m(n + 1/\tau^2 + q(n,m)/v_0},$$

and

$$\begin{aligned} E(\mu|I = 1, \bar{x}, \bar{y}, s^2) &= (n\bar{x}^+ m\bar{y})/q(n,m) - \frac{m^2[n(\bar{y} - \bar{x}) + y/\tau^2]}{q(n,m)[m(n + 1/\tau^2 + q(n,m)/v_0]} \\ &= \frac{n\bar{x} + \bar{y}/(v_0 + 1/m)}{n + 1/(v_0 + 1/m) + 1/\tau^2}. \end{aligned}$$

Finally, the posterior distribution of $v = 1/\sigma^2$ given s^2 is a scaled χ^2-distribution with $f + d_0$ degrees of freedom and with scaling constant $(f + d_0)/(f s^2 + d_0 s_0^2)$. From this the conditional mean of $\sigma^2 = 1/v$ is easily obtained. □

Combining the results of Proposition 6 with (30) and (31) it is possible to calculate a posterior estimate of the probabilities of correct classification (14) and (15). We then consider these as a function of the threshold Δ and choose the value that maximizes the average of the two probabilities. The result of using this procedure can be seen in Table 5. Interestingly, it seems that this method improves the behaviour of the hard thresholding approach.

7 Three groups

For the case of three groups we consider instead of (1) the model

$$x_j \sim \begin{cases} N(\mu_j, \sigma_j^2) & \text{group 1,} \\ N(\mu_j + \delta_j \sigma_j, \sigma_j^2) & \text{group 2,} \\ N(\mu_j + \eta_j \sigma_j, \sigma_j^2) & \text{group 3.} \end{cases}$$

For three groups it is convenient to change the notation. Thus, we let x_i^r be the ith observation in group r, $i = 1, \ldots, n_r$, $r = 1, 2, 3$. For the average we use $\bar{x}^r$, and the variance estimate becomes $s_j^2 = (\sum_{r=1}^3 \sum_{i=1}^{n_r} (x_{ij}^r - \bar{x}_j^r)^2)/(n - 3)$, where $n = n_1 + n_2 + n_3$. The maximum likelihood classifier, with estimates inserted for the parameters, is

$$\arg\min_r \left\{ \sum_{j=1}^p \frac{(z_j - \bar{x}_j^r)^2}{s_j^2} \right\}, \tag{40}$$

where z is the new observation to be classified. Define for $r = 1, 2$

$$L_0^r(z) = \sum_{j=1}^{p} z_j(\bar{x}_j^3 - \bar{x}_j^r)/s_j^2 \quad \text{and} \quad \kappa_0^r(z) = \sum_{j=1}^{p}[(\bar{x}_j^3)^2 - (\bar{x}_j^r)^2]/s_j^2.$$

Then the classification (40) is equivalent to the rule

$$\begin{cases} \text{group 1} & \text{if } L_0^1 < \kappa_0^1,\ L_0^2 > L_0^1 + \kappa_0^2 - \kappa_0^1, \\ \text{group 2} & \text{if } L_0^2 < \kappa_0^2,\ L_0^2 < L_0^1 + \kappa_0^2 - \kappa_0^1, \\ \text{group 3} & \text{if } L_0^1 > \kappa_0^1,\ L_0^2 > \kappa_0^2. \end{cases}$$

For the case of unequal sample sizes n_1, n_2, n_3 we proceed as in Section 2.2 and define for $r = 1, 2$

$$L^r(z) = \frac{1}{2}\sum_j \left(\frac{1}{n_r}\sum_{i=1}^{n_r}(z_j - x_{ij}^r)\frac{\bar{x}_j^3 - \bar{x}_j^r(i)}{s_j^2(x_i^r)} + \frac{1}{n_3}\sum_{i=1}^{n_3}(z_j - \bar{x}_{ij}^3)\frac{\bar{x}_j^3(i) - \bar{x}_j^r}{s_j^2(x_i^3)} \right), \qquad (41)$$

which corresponds to (8). The classification rule with this choice becomes

$$\begin{cases} \text{group 1} & \text{if } L^1 < 0,\ L^2 > L^1, \\ \text{group 2} & \text{if } L^2 < 0,\ L^2 < L^1, \\ \text{group 3} & \text{if } L^1 > 0,\ L^2 > 0. \end{cases} \qquad (42)$$

Having decided on (L^1, L^2) as the basic statistic for classification there are of course other possibilities than the one in (42) for splitting the two-dimensional space into three regions.

For thresholding we use the three t statistics

$$t_j^{12} = \frac{\bar{x}_j^1 - \bar{x}_j^2}{\sqrt{s_j^2(1/n_1 + 1/n_2)}}, \quad t_j^{13} = \frac{\bar{x}_j^1 - \bar{x}_j^3}{\sqrt{s_j^2(1/n_1 + 1/n_3)}}, \quad t_j^{23} = \frac{\bar{x}_j^2 - \bar{x}_j^3}{\sqrt{s_j^2(1/n_2 + 1/n_3)}}.$$

This gives the possibility of using different thresholds as well as different combinations of these in L^1 and L^2. When used in (41) the appropriate observation is left out in the calculations of the t values. The simplest choice is to use the same genes in the two statistics and the same threshold, so that we have the weight function

$$w(t) = 1(\max\{|t^{12}|, |t^{13}|, |t^{23}|\} > \Delta)$$

in the case of hard thresholding.

8 Summary

In this paper I have illustrated that good classifiers can be build for high dimensional measurements where only a small number of the variables are differentiable expressed. In particular we have seen that thresholding can improve the performance of the classifier. For small samples the crossvalidation error cannot be trusted, the

overshoot can be large. If possible it would be good to have rules of thumb as to the size of the overshoot. Nevertheless, the crossvalidation can still be used to select a threshold in the classifier. Of course the best situation is to have a large independent test set!

The investigations have been made for a simple theoretical model. In a microarray setting there are many noise terms, some of which introduce bias in the measurements. Thus the investigations here show what can be achieved in an optimal setting. In real life situations we must expect the classifiers to have higher error rates than those obtained here.

A Variance of ξ_N when $n \neq m$

For the classifier $L(z)$ in (8) the variance of $\xi_N = E_N(L)$ is

$$V(\xi_N) = \frac{p}{n^2}c_{31} + \frac{1}{n}c_{32}\delta_\bullet^2 + c_{33}\delta_\bullet^4.$$

The coefficients are given by

$$\begin{aligned}
4c_{31} &= \frac{nf(f+1)(f-1)^2}{m(m-1)(n-1)(f-3)(f-5)} + (n-1)E[u_1u_2tu(1)tu(2)] \\
&\quad + \frac{n^2(m-1}{m}E[v_1v_2tv(1)tv(2)] - 2n^2E[u_1v_1tu(1)tv(1)], \\
4c_{32} &= \frac{(fm+2m-1)(f-1)^2}{m(m-1)(f-3)(f-5)} + (n-1)E\frac{u_1u_2}{s^2(u_1)s^2(u_2)} \\
&\quad + \frac{n(m-1)}{m}E\frac{2(v_1+v_2)+v_1v_2+(\bar v(1)-\bar u)(\bar v(2)-\bar u)}{s^2(v_1)s^2(v_2)} \\
&\quad - 2nE\frac{u1[v_1+(\bar v-\bar u(1))+(\bar v(1)-\bar u)]}{s^2(u_1)s^2(v_1)}, \\
4c_{33} &= \frac{(f-1)^2}{m(f-3)(f-5)} + \frac{m-1}{m}E\frac{1}{s^2(v_1)s^2(v_2)} \\
&\quad - \big[E\frac{1}{s^2(v_1)}\big]^2,
\end{aligned}$$

where the notation is as in Lemma 4.

B Appendix: Bayes classifier

Here we consider the Bayes model presented in Section 6 and establish the posterior probability for a new sample z to belong to a particular group. For convenience in the notation we take $\mu_0 = 0$. In the final formula z, $\bar x$, and $\bar y$ must then be replaced by $z-\mu_0$, $\bar x-\mu_0$, and $\bar y-\mu_0$. Also when considering one variable we leave out the subscript j.

Consider first the case where z belongs to group 1, that is, $z \sim N(\mu, \sigma^2)$. The full density for $(z, I, \bar x, \bar y, s^2, \mu, \delta, v)$, with $v = \frac{1}{\sigma^2}$, is similar to (32). The differences are that f in (33) contains the extra term $(z-\mu)^2$, C in (34) has 2π in the denominator

replaced by $(2\pi)^{3/2}$, and in (32) the exponent of v is $(f+d_0+3)/2-1$ instead of $(f+d_0+2)/2-1$. Integrating the density for $I=0$ with respect to μ and next with respect to v gives (36) with the exponent $(f+d_0+2)/2$ replaced by $(f+d_0+3)/2$, and where now $q(n,m)=1+n+m+1/\tau^2$,

$$A=\frac{n(z-\bar{x})^2+m(z-\bar{y})^2+nm(\bar{x}-\bar{y})^2+(z^2+n\bar{x}^2+m\bar{y}^2)/\tau^2}{q(n,m)},$$

and in the expression (37) for C_1 the exponent is $(f+d_0+3)/2$ instead of $(f+d_0+2)/2$.

When $I=1$ integrating with respect to μ, δ, and v gives (39) with

$$B=A-\frac{m^2[\bar{y}-(z+n\bar{x}+m\bar{y})/q(n,m)]^2}{1/v_0+m(1+n+1/\tau^2)/q(n,m)},$$

with the exponent $(f+d_0+2)/2$ replaced by $(f+d_0+3)/2$ as before and with $m(n+1/\tau^2)$ replaced by $m(1+n+1/\tau^2)$ in the last term of (39). The likelihood when z belongs to group 1 therefore becomes

$$\prod_{j=1}^{p} C_{1j}\frac{1-p_0}{[A_j+fs_j^2+d_0s_0^2]^{(f+d_0+3)/2}} \tag{43}$$
$$\times\Big\{1+\frac{1-p_0}{p_0}\Big(\frac{A_j+fs_j^2+d_0s_0^2}{B_j+fs_j^2+d_0s_0^2}\Big)^{(f+d_0+3)/2}\Big(\frac{q(n,m)/v_0}{q(n,m)/v_0+m(1+n+1/\tau^2)}\Big)^{1/2}\Big\}.$$

When z belongs to group 2 the case $I=0$, after integration over μ and v, gives the same result as when z belongs to group 1. The case $I=1$, after integration over μ, δ, and v, is as above with B replaced by

$$\tilde{B}=A-\frac{[m\bar{y}+z-(1+m)(z+n\bar{x}+m\bar{y})/q(n,m)]^2}{1/v_0+(1+m)(n+1/\tau^2)/q(n,m)},$$

and with $m(1+n+1/\tau^2)$ replaced by $(1+m)(1+n+1/\tau^2)$. The likelihood when z belongs to group 2 is then given by

$$\prod_{j=1}^{p} C_{1j}\frac{1-p_0}{[A_j+fs_j^2+d_0s_0^2]^{(f+d_0+3)/2}} \tag{44}$$
$$\times\Big\{1+\frac{1-p_0}{p_0}\Big(\frac{A_j+fs_j^2+d_0s_0^2}{\tilde{B}_j+fs_j^2+d_0s_0^2}\Big)^{(f+d_0+3)/2}\Big(\frac{q(n,m)/v_0}{q(n,m)/v_0+(1+m)(1+n+1/\tau^2)}\Big)^{1/2}.\Big\}$$

From (43) and (44) the posterior probability of belonging to one of the two groups is easily determined.

References

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

Andersen, L., T. Thyklær, M. Kruhøffer, J. Jensen, N. Marcussen, S. Dutoit, H. Wolf, and T. Ørntoft. (2003). Classification and characterization of bladder cancer stages using microarrays. stage and grade of bladder cancer defined by gene expression patterns. *Nature Genetics 33*(1), 90–96.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *SAGMB 3*, Article 33.

Ding, B. and R. Gentleman (2004). Classification using generalized partial least squares. Research Report 5, Bioconductor Project Working Papers.

Dudoit, S. and J. Fridlyand (2003). Classification in microarray experiments. In T. Speed (Ed.), *Statistical Analysis of Gene Expression Microarray Data*, pp. 93–158. Chapman & Hall/CRC, Boca Raton.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Lönnstedt, I. and T. Speed (2002). Replicated microarray data. *Statistica Sinica 12*, 31–46.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press, Inc., San Diego, CA.

Nguyen, D. and D. Roche (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics 18*, 39–50.

Pérez-Enciso, M. and M. Tenenhaus (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum Genet 112*, 581–592.

Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Gen. Mol. Biol. 3*(1), article 3.

Stone, M. and R. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression (with discussion). *J. Roy. Statist. Soc. B 52*, 237–269.

Tibshirani, R., T. Hastie, N. Balasubramanian, and G. Chu (2003). Class prdiction by nearest shrunken centroids, with application to DNA microarrays. *Statist. Sci. 18*, 104–117.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn (1984). The collinearity problem in linear regression. the pls approach to generalized inverses. *SIAM J. Sci. Comp. 5*, 734–743.