Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Classify strata

Note no. SAMBA/11/15

Authors Lars Holden

Date 16. mar. 2015

Norsk Regnesentral

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
office address
Gaustadalleen 23a
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Bankkonto · bank account
8200.01.48888
Org.nr. · enterprise no.
NO 952125001 VAT

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Classify strata**
Authors	**Lars Holden**
Date	16. mar. 2015
Year	2015
Publication number	SAMBA/11/15

Abstract
We assume there is a slight difference in the average value of the gene expression for a group of genes between two strata/time periods. There is no prior knowledge on which genes where the gene expression differs between the two strata. We test out the power of a large number of test statistics in order to identify whether the observed difference between the averages is significant. Secondly, we test out a large number of test statistics ability to classify new patients into the two strata based on the genes identified in the first analysis. This is performed both with data from one time-period and for cases where there is data from several time periods. Data from several time periods may be valuable if we expect a time development relative to time to diagnosis that is characteristic for the different strata.

Keywords	Gene expression data; Breast cancer; Development in time; Curve groups; Hypothesis testing using randomization; Prediction of diagnosis
Target group	Biostatisticians
Availability	Open
Project number	220633
Research field	Biostatistics
Number of pages	28

Table of Content

1 Introduction

This note analyses the possibility to separate between two different strata based on the gene expressions and our ability to classify patients between the two strata. The two strata may be case and control, patients with and without spread of cancer or two groups of patients that differ in time to prognosis. This is a theoretical study where we focus on how small differences/weak signals that may be detected and which methods and test statistics are best suited for this problem. Therefore, we use synthetic data where we know the exact properties of the data.

This problem is closely related to problems that consist of many parallel subproblems and where the problems consist of identifying the significant subproblems. This is the case when we test all genes in order to find the significant genes that may be used in differentiating between two strata. These problems are often characterized as "p>n" problems and the use of false discovery rate is central. Also in this paper we analyze p-values from a large number of subproblems, f.ex. for each gene. But the objective of our study is very different from a study with many subproblem since we focus on one problem, to differentiate between two strata, not to identify properties of each gene. Hence, we do not have the problem with many significant results due to the number of independent tests. We have only one test where we try to differentiate between the two strata. We analyze problems where the signal is so weak that we cannot expect to be able to identify significant genes. Since many genes have a weak signal, we may still be able to differentiate between the two strata.

The typical situation when separating between two strata is to focus on the gene expression in one or a few genes that already are identified. Here we assume a different situation. We assume:

1. There is a weak signal that separates the two strata in a large number of genes.
2. There is no prior information on which genes that may give a signal.
3. The signal is an additive term in some of the (log) genes expressions and there is no prior information on the distribution of this additive term.

The word "signal" should be interpreted that the distribution of the gene expression in the two different strata are not the same. This difference may be utilized to describe the gene expression and possibly classify patients in the two strata.

We are looking for methods using a large number of genes to separate between the strata. We calculate the t-statistics for all the genes and study methods based on the empirical distribution of the t-statistics. We focus on the t-statistics for the genes expressions where the t-statistics has the largest absolute values and compare several different methods based on the t-statistics. For each method we perform a hypothesis test with the same significance level. We compare the different methods by comparing the power of the hypothesis test for different assumptions for the gene expressions for the two strata. Similarly, when we classify patients in the two strata we set a fixed probability for a false true classification. Then we

compare the different methods by comparing the probability for a correct true classification for different assumptions on the signal. Some preliminary conclusions that are common for both one and several time periods:

1. The ability to differentiate between the strata depends obviously heavily on the strength of the signal relative to the natural variation of the gene expression and number of genes with a signal. Which method that is best to identify a signal depends on strength and type of signal.
2. It is possible to separate between strata and classify persons between strata also when no single gene is significant.
3. If the distribution of the signal has a normal distribution or has heavier tails, it seems optimal to have few elements in the sum. This is close to the situation where we should only focus on the one-three most extreme t-values. If the strength of the signal is the same in a large number of genes, it seems optimal to use many elements in the sum also larger than the number of genes with signal. Other distributions will vary between these two extremes.

When there is only one time period:

1. Number of genes in the analysis does not seem to be critical. To increase from 9.000 genes to 30.000 did only give a slight reduction on the strength of the tests.
2. The optimal method for differentiating between the two strata seems to be based on a weighted sum of the largest absolute values of the t-statistics. However, the weights and the optimal number of elements in the sum depend on the distribution of the signal.

Some preliminary conclusion when we study several time periods.

1. Which method that is best depends on the strength of the signal in the different time periods and the number of data in the different time periods.
2. Methods based on p-values from t-test in each time period seems better than methods based on curve groups.

So far we have focused on the t-statistics for each gene and the weighted sum of the t-statistics for several genes. T-statistics is the most common test statistics for these data and weighted sums is the most natural choice. If this is successful, it is possible to test out more complex models like the nearest neighbor in a high dimensional space with t-statics for each axis as is tested in a separate note (Holden and Holden, june 2014).

In all our models we assume the data is independent between the genes. We know there is a strong dependency between some of the genes. We have not included this in the models since it is difficult to find a good model that represents this complex dependency between the genes. We do not believe our results depend heavily on this assumption. But this is difficult to test without making assumptions on a particular joint statistical distribution. Our tests indicate that the result does not depend critically on the number of genes. Since many problems with a large number of correlated variables may be represented by a smaller number of independent variable, this may indicate that our assumption is correct. The distribution of extreme p-values

depends on the dependency between genes. However, our hypothesis tests and classification between the strata are based on randomization that maintains the joint distribution and hence should not be sensitive to this dependency. This implies that the result of our hypothesis test and classification are correct also with correlation. But when we compare the number of genes with a fixed strength of the signal, we may need more genes with this strength if the genes are correlated compared to when they are independent in order to get the same power in the hypothesis tests and classifications.

2 Model for two strata and one time-period

We have the (log) gene expressions $X_{i,j}$, where *i* is gene and *j* is patient. Each patient belong either to strata A or B and this is known in the control data set. We want to find out whether there is a difference in the gene expression between strata A and B and to classify persons into strata A or B for persons in a test set. A and B may be two strata or two time periods for same stratum.

For patient j in A we have $X_{i,j} = Y_{i,j} + a_i$, and for patient *j* in B we have $X_{i,j} = Y_{i,j}$, where $Y_{i,j}$ is $N(\mu_i, \sigma_i^2)$ and independent between different patients. We know very little about the variables a_i, denoted signal, except that we expect most of them to be zero. We want to find a good estimator for the classification A or B for different assumptions on a_i.

We have two alternative models for the a_i values when a_i is not identically equal to 0.

A. $a_i = \pm h_i$ with equal probability for a positive or negative value (h_i>0)
B. a_i as $N(0, u_i^2)$ Hence most of the a_i values are close to 0.

In order to compare these models we choose $u_i = h_i\sqrt{\pi/2}$ such that$E\{|a_i|\} = h_i$ in both cases. In this model only about 42% of the $a_i$values where $|a_i| > 0$ satisfies$|a_i| > h_i$.

2.1 Discussion of model and extensions

In this model there is no correlation between the gene expressions for the same patient except for some model of the a_i variable and no time development relative to time of diagnosis in the different strata. The preprocessing will remove a constant added to all gene expressions for a patient. There are many possible extensions of the model. One possibility is to divide genes into groups and assume correlation between genes expressions from same group and patient. If some of these groups are closely related to the group of genes with non-zero a_i this may make the test in this note much weaker. Otherwise, this correlation will not influence on the test described in this note.

2.2 Hypothesis test separating the two strata

For most genes it is natural to assume $a_i = 0$. If $|a_i|$ is large for one or a few genes, we should focus on these genes. Here we assume $|a_i|$ is different from 0 but quite small for a large number of genes. This makes it more natural to use a test statistics as follows:

c_i = average_{j in B} ($X_{i,j}$X_i,j) is an estimate for μ_i

s_i is an estimate for σ_i

$T_i = \frac{1}{d_i}(\frac{1}{n_A}\sum_{j\in A} X_{i,j} - \frac{1}{n_B}\sum_{j\in B} X_{i,j})$

is the t-estimator for a_i/σ_i. Here d_i is the estimate for the standard deviation of the numerator in the above expression and n_A and n_B are the number of patients in A and B respectively. Under the hypothesis $a_i = 0$ and we may simplify the expression in the denominator. However, we have chosen this expression in order to get an estimate for a_i/σ_i. This is probably not critical for the result. The nominator and the denominator in T_i are independent. Hence, these may be simulated independently in a simulation. We may simulate the nominator as a normal variable without simulating all the gene expressions in the two strata. The critical variation is in the nominator and we may use fewer samples for simulating the denominator. For the other variables, we define

$b_i = \frac{T_i}{|T_i|}$ i.e. an estimate for the sign of a_i

G_k is the set of genes with the k largest absolute values of T_i

We have two different test statistics comparing strata A and B:

Z^k is the k'th largest value of T_i and

$$H_k = \sum_{i=1}^{k} Z^k$$

The hypothesis is that strata A and B have the same properties. There are several possible test statistics $Z^k > z_k$ or $H_k > h_k$ for k=1,2,3….. We reject the hypothesis if the selected test statistics is above the threshold. Here we need to choose either Z or H and value of k in order to make as strong test as possible.

Figure 1-3 show the results from the tests. Figure 1 and 2 shows that we get highest power for small values of k when a_i is normally distributed while we for $|a_i|$ constant, we get highest power for larger values of k. If the distribution of the signal has heavier tails than the normal distribution, we should focus even more on the most extreme t-values, i.e. neglecting that there is a signal in many genes. Constant distribution of the signal in the selected genes is the most equal distribution of the signal. Here it seems optimal to use many t-values, even more t-values than there are genes with a signal. Notice that in figure 1A where $a_i$is normally distributed the critical change is for expected value changing from 0.4 to 0.6 increasing the power to about 0.9 while in Figure 1B where $|a_i|$ is constant, the critical change is when the expected value changes from 0.6 to 0.8 increasing the power to about 0.7. Increasing the number of patients in each stratum with 4 corresponds to doubling of the expected value of a_i. Figure 4 shows why we get highest power for small values of the tests for normally distributed $a_i$values.

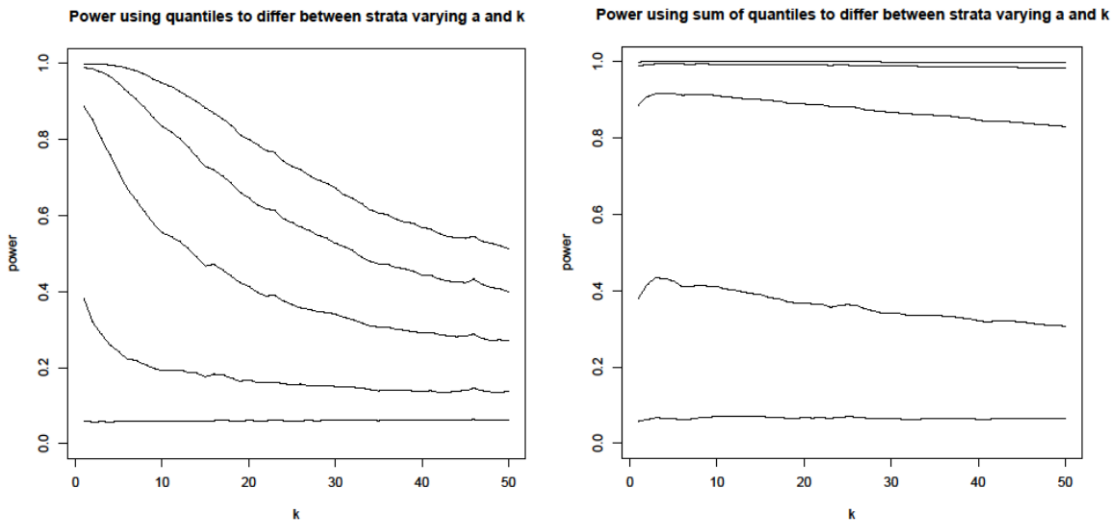

Figure 1A. Figure shows the power in the test separating different strata. In all cases there are 9.000 genes and 30 patients in each stratum and it is based on 10.000 simulations. In both figures there are 20 non-zero a_i values that are normally distributed with expectation equal 0 and expectation of the absolute value equal 0.2, 0.4, 0.6, 0.8 and 1 in the five lines. In the left figure we use the estimator Z^k and in the right figure we use the estimator H_k, which clearly gives higher power.

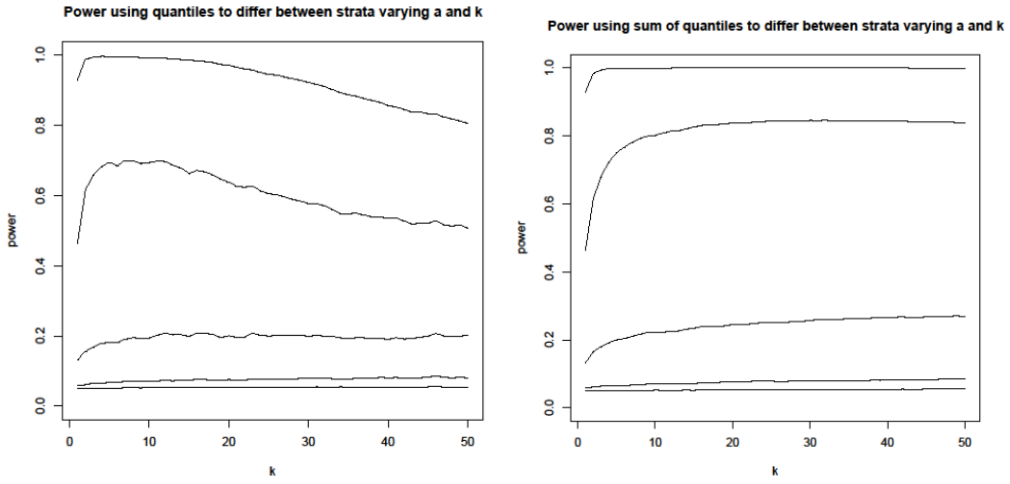

Figure 1B. Figure shows the power in the test separating different strata. In all cases there are 9.000 genes and 30 patients in each stratum and it is based on 10.000 simulations. In both figures there are 20 non-zero a_i values that are constant equal 0.2, 0.4, 0.6, 0.8 and 1 in the five lines. In the left figure we use the estimator Z^k and in the right figure we use the estimator H_k, which clearly gives higher power.

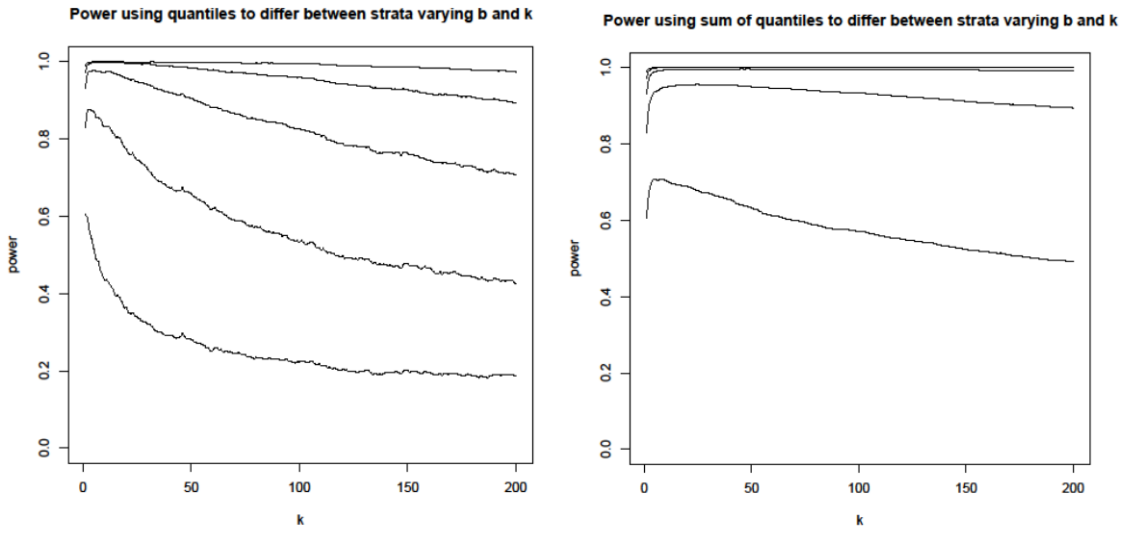

Figure 2A Figure shows the power in the test separating different strata. In all cases there are 9.000 genes and 30 patients in each stratum and it is based on 10.000 simulations. In both figures there are lines for 40, 80, 120, 160, and 200 respectively non-zero a_i values that are normally distributed expectation equal 0 and expectation of the absolute value equal 0.4. In the left figure we use the estimator Z^k and in the right figure we use the estimator H_k, which clearly gives higher power.

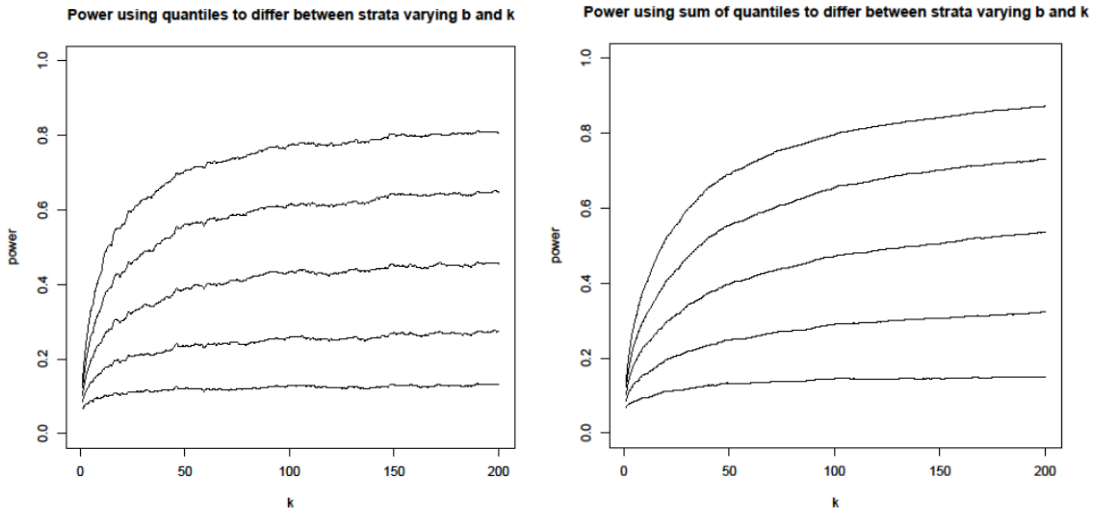

Figure 2B. Figure shows the power in the test separating different strata. In all cases there are 9.000 genes and 30 patients in each stratum and it is based on 10.000 simulations. In both figures there are lines for 40, 80, 120, 160, and 200 respectively non-zero a_i values that are constant equal 0.4. In the left figure we use the estimator Z^k and in the right figure we use the estimator H_k, which gives slightly higher power for large values of k. Notice that here the power increases with k in contrast to Figure 2A.

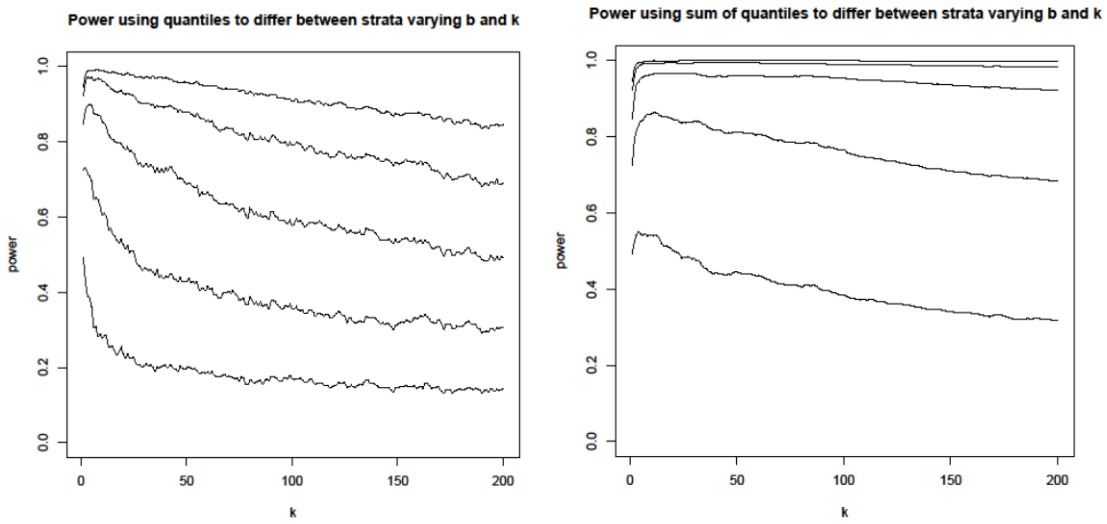

Figure 3. The figure shows the power in the test separating different strata. It is similar to Fig. 2A except that there are 30.000 genes instead of 9.000 and it is based on 2.000 simulations. In both figures there are lines for 40, 80, 120, 160, and 200 respectively non-zero a_i values that are normally distributed expectation equal 0 and expectation of the absolute value equal 0.4. In the left figure we use the estimator Z^k and in the right figure we use the estimator H_k, which clearly gives higher power. Notice that the power decreases slightly with increasing number of genes.

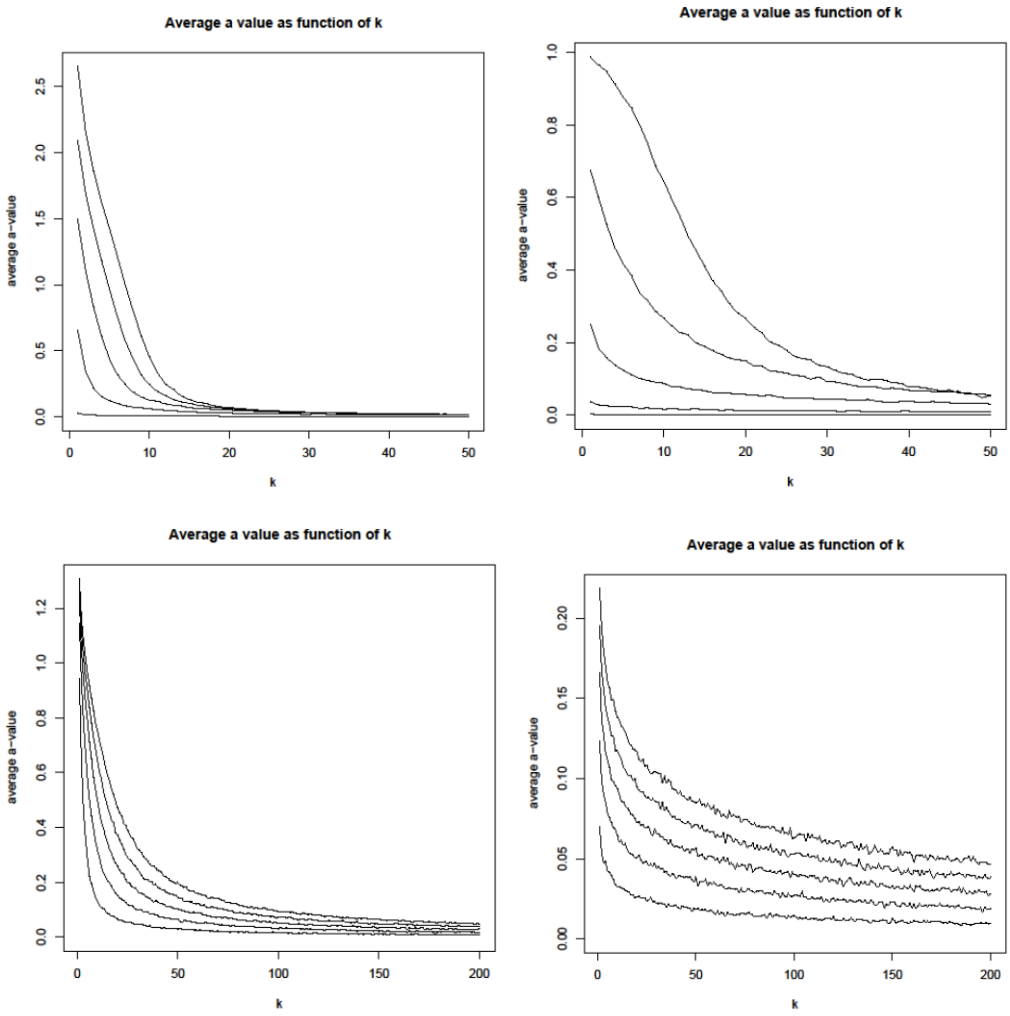

Figure 4. The average $a_i$values is largest for the most extreme quantiles of T_i. The figure shows how the average $a_i$value corresponding to Z^kdecreases with k. The upper left figure is for normally distributed $a_i$with expectation equal 0 and expectation of the absolute value equal 0.2, 0.4, 0.6, 0.8 and 1 respectively and the right figure is for a_i constant equal 0.2, 0.4, 0.6, 0.8 and 1 respectively (similar to Fig 1A and 1B). In the lower figures are for varying number of a_i values. (similar to Fig 2A and 2B respectively). These values are calculated for 20 non-zero $a_i$values. Notice that the values are smaller for constant $a_i$but the decrease is also smaller .This explains why we get strongest power in the tests for small values of k when a_i is normally distributed. Notice also that the larger the a-value is, the easier it is to identify the large values. If we divide the values in one curve with the values in the curve below, these values are increasing with k and for curves with smaller a-values. This shows that it is easier to identify a-values, the larger the a-values are.

2.3 Classification of stratum for new patients

If we find a significant difference between the strata, we would like to classify the stratum for a new patient. In the classification we set the probability for a false positive to 5%, i.e. the probability that a patient belonging to stratum B without a signal is classified to stratum A with a signal. We compare the different classification method on the probability for a correct

positive classification, i.e. to classify into stratum A when the patient belongs to this stratum. This formulation of the classification is relevant for patients that are classified with cancer (both strata A and B) and we will only apply an additional treatment with severe disadvantages if we are reasonably sure that that this treatment will help (i.e. belong to stratum A). The classification may be based on the estimator

$$S_{j,k} = \frac{1}{\sum_i w_i} \sum_{i \in G_k} \frac{w_i (X_{i,j} - c_i)}{s_i}$$

i.e. $S_{j,k}$ is the weighted average of $X_{i,j}$ for the genes *i* where we estimate for $|a_i|/\sigma_i$. to be largest. A possible choice is $w_i = b_i$ i.e. the estimated sign of a_i such that large values of $S_{j,k}$ indicates that patient j is in group A. We may also choose $w_i = Z_i$ such that genes with the expected largest values of $|a_i|/\sigma_i$ have largest weight.

If $|a_i|/\sigma_i$ is much larger for one or a few genes than for the other genes, our estimate should focus on these genes. We may then use the estimator $S_{j,k}$ for k=1,2,3. However, if many values of $|a_i|/\sigma_i$ have about the same size, we should probably use a larger value of n. Hence, we want to find the properties of $S_{j,k}$ and find the optimal value of k under for different assumptions on the distribution of $|a_i|/\sigma_i$.

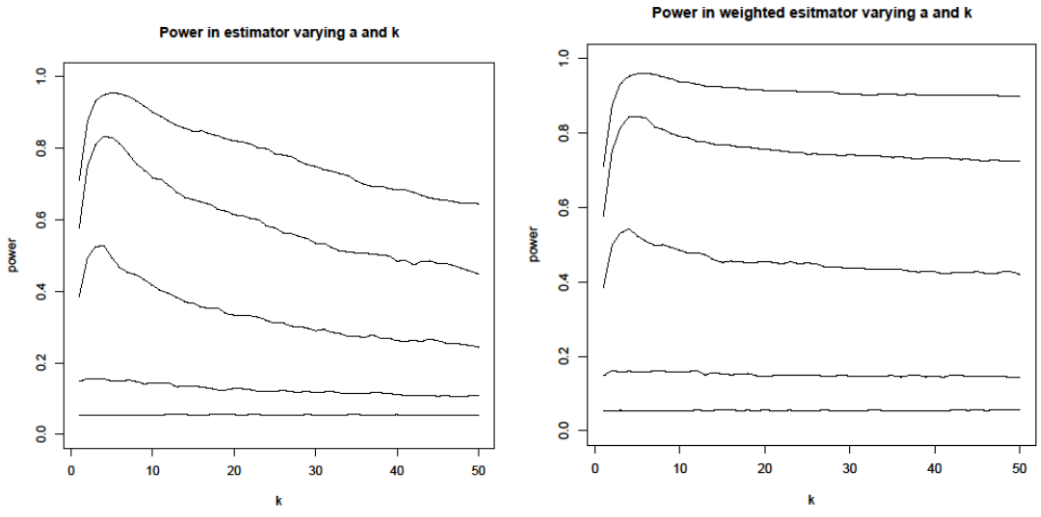

Figure 5. The figure shows probability for a true positive prediction, i.e. classify correctly in stratum A. In all cases there are 9000 genes and 30 patients in each stratum and based on 10.000 simulations. In both figures there are 20 non-zero a_i values that are normal distributed with expectation equal 0 and expectation of the absolute value equal 0.2, 0.4, 0.6, 0.8 and 1 in

the five lines. In the left figure we use weight equal $w_i = b_i$ and in the right figure we use weight $w_i = Z_i$which clearly gives higher power for k >5.

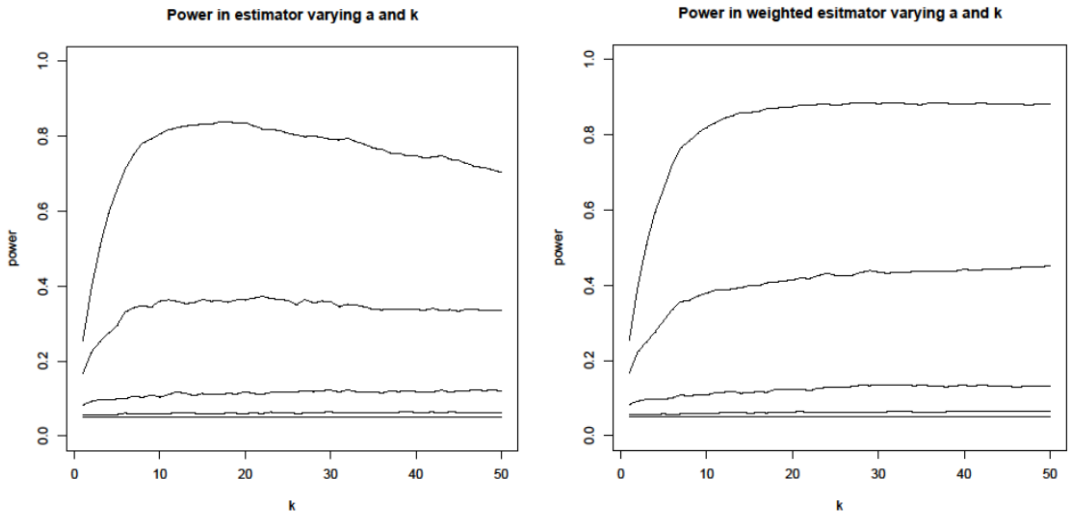

Figure 5B. The figure shows probability for a true positive prediction, i.e. classify correctly in stratum A. In all cases there are 9000 genes and 30 patients in each stratum and based on 10.000 simulations. In both figures there are 20 non-zero a_i values with constant value 0.2, 0.4, 0.6, 0.8 and 1 in the five lines. In the left figure we use weight equal $w_i = b_i$ and in the right figure we use weight $w_i = Z_i$ which gives higher power for k larger than 5-10. Notice that we here have optimal value for larger values of k compared to normally distributed a_i shown in figure 5.

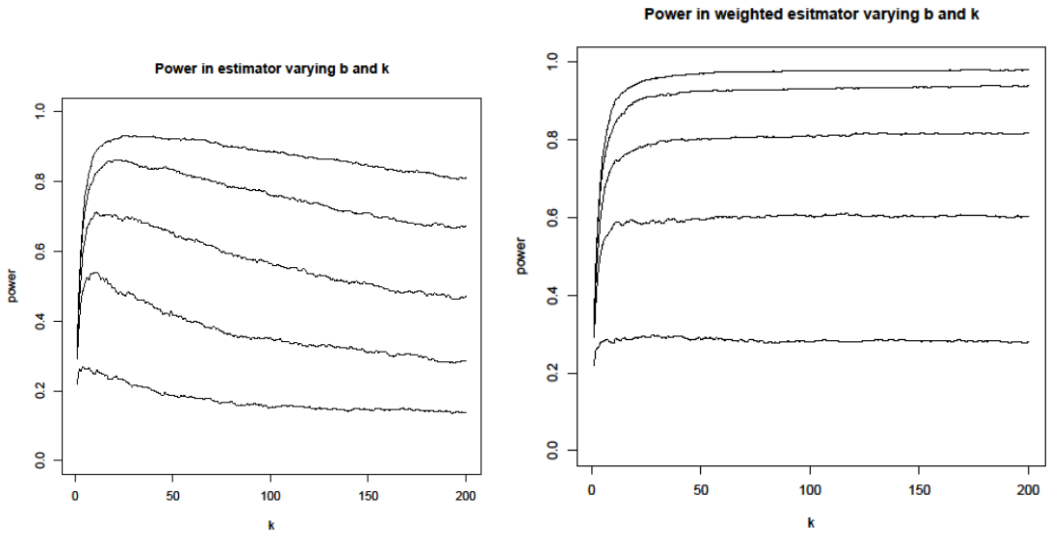

Figure 6. The figure shows probability for a true positive prediction, i.e. classify correctly in stratum A. In all cases there are 9000 genes and 30 patients in each stratum and based on 10.000 simulations. In both figures there are lines for 40, 80, 120, 160, and 200 non-zero a_i values respectively that are normal distributed with expectation equal 0 and expectation of the absolute value equal 0.4. In the left figure we use weight equal $w_i = b_i$ and in the right figure

we use weight $w_i = {}^{1}/_{(1+k)}$. *The weight seems to decrease so fast with k that there is almost no effect of k>25.*

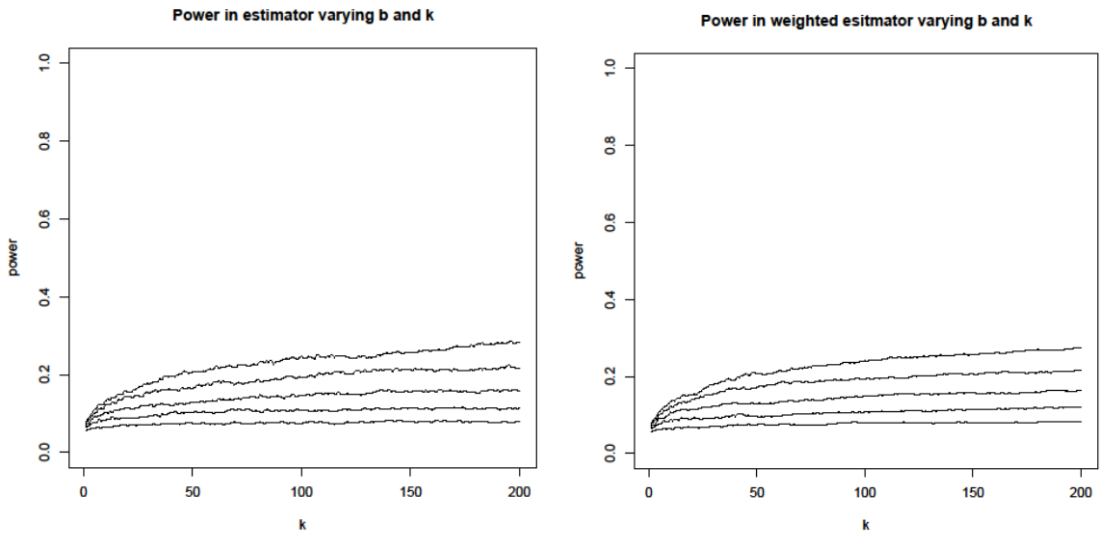

Figure 6B. The figure shows probability for a true positive prediction, i.e. classify correctly in stratum A. In all cases there are 9.000 genes and 30 patients in each stratum and based on 10.000 simulations. In both figures there are lines for 40, 80, 120, 160, and 200 non-zero a_i *values respectively that are constant equal 0.4. In the left figure we use weight equal* $w_i = b_i$ *and in the right figure we use weight* $w_i = Z_i$. *Hence, the lowest line in each of these figures corresponds to the second lowest lines in Figure 5B.*

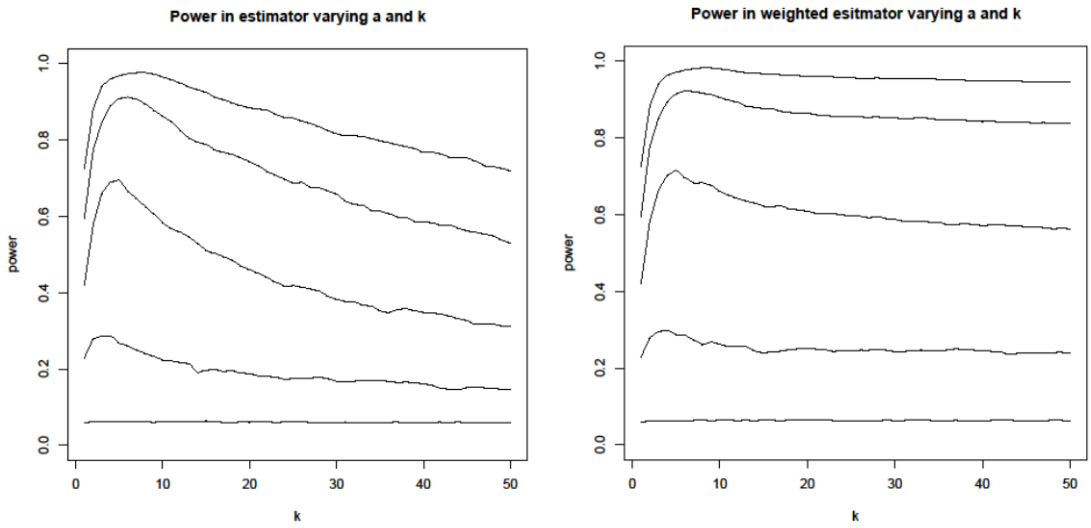

Figure 7. This figure corresponds to Figure 5 except that the number of patients is increased to 50 in each stratum. Otherwise, all parameters are as in Figure 5. There are 9.000 genes and it is based on 10.000 simulations. In both figures there are 20 non-zero a_i *values that are normal distributed with expectation equal 0 and expectation of the absolute value equal 0.2, 0.4, 0.6, 0.8, and 1 in the five lines. In the left figure we use weight equal* $w_i = b_i$ *and in the right figure we use the weight* $w_i = Z_i$. *There is a slight increase in power, but much smaller than in testing*

whether there is a difference between the two strata as commented in the legend to Figure 1. With more patients we are able to select better the genes with large values of a_i which improves the test.

3 Model for several time periods

Assume we want to differentiate between two strata where we assume the (log) gene expressions have a different time development relative to time to diagnosis. Also here we assume a weak signal in many genes that develop differently in the two strata, not a strong signal in one or a few genes. We want to find a strong test estimator that is able to identify this difference as good as possible.

Let the (log) gene expressions be denoted by $X_{i,j}$, where *i* is gene and *j* is patient. The function *s(j)* denote the strata for patient *j* and is equal A or B. The stratum is known in the control data set. The function *t(j)* denote the number of time periods before diagnosis for patient *j* and is equal *1,2,…,* n_t. We want to find out whether there is a time development in the gene expressions and a difference between strata A and B in the last time interval, ie. close to diagnosis. We also want to classify persons into strata A or B for persons in a test set in the last time interval.

The (log) gene expression is modelled as $X_{i,j} = Y_{i,j} + a_{i,s(j),t(j)}$ where $Y_{i,j}$ is $N(\mu_i, \sigma_i^2)$ and independent between different patients. We know very little about the variables $a_{i,s(j),t(j)}$, denoted the signal, except that it is equal 0 for most genes *i*. We also assume that for each gene the signal $a_{i,s(j),t(j)}$ is so weak that we are looking for methods based on many genes. We don't know which genes that gives a signal (i.e. $a_{i,s(j),t(j)}$ not vanishing) and all time periods give information on which genes that have a signal. We want to find as good estimator for the classification A or B for different assumptions on $a_{i,s(j),t(j)}$.

We have two alternative models for the $a_{i,s(j),1}$, i.e. values in the final time period for gene *i* where this variable is not identically equal to 0.

A. $a_{i,s(j),t(j)} = t(j)h_i > 0$
B. $a_{i,s(j),t(j)}$ as N(0, $(t(j)u_i)^2$)

Hence most of the $a_{i,s(j),t(j)}$ values are close to 0. In order to compare these models we choose $u_i = t(j)h_i\sqrt{\pi/2}$ such that E$|a_{i,s(j),t(j)}| = t(j)h_i$ in both cases. In this model only about 42% of the $a_{i,s(j),t(j)}$ values where $|a_{i,s(j),t(j)}|>0$ satisfies $|a_{i,s(j),t(j)}| > t(j)h_i$.

3.1 Hypothesis test for difference between strata

Similar to the situation for one time period, we make a hypothesis that there is no difference between the strata with the same significance level based on the different test statistics. Then

we find the power when using the different test statistics, i.e. the probability for rejecting the hypothesis, for different assumptions on the signal.

We compare two strata in four time periods where each measurement of the (log) gene expression is independent N(0,1) values in 1000 genes and assume there are 10 patients for each time period and each strata. One strata has no signal and one strata has the signal $a_{i,s(j),t(j)}$. In the test we assume the time development of the signal is $a_{i,s(j),t(j)} = a_{i,s(j),1}{}^{(n_t - t(j))}\big/_{(n_t - 1)}$ where n_t is the first time period. Hence, the signal is monotone in-/decreasing in time for each gene with the same increment between subsequent time periods, vanishing in the first time period and equal to $a_{i,s(j),1}$ at the time of diagnosis. In the test we vary the (expected) value of $a_{i,s(j),t_f}$ (values: 0.25, 0.5, 0.75) and the number of genes with a signal (values: 30, 100) and in addition whether $a_{i,s(j),1}$ is constant or normally distributed.

We compare 19 different test statistics for differentiating between the two strata. We make a hypothesis test that has a 5% significance level. Then we find the power of a hypothesis test with each test statistics, i.e. the probability for true positive conclusion for the given case. The data is tested in a simulation based on 10 sets of D_i 1000 data sets. The quantiles are found in each of the D_i data sets and then we take the average between theses 10 data sets. The same random variables are used for the different test statistics. The results are given in Table 1 with constant signal and Table 2 with normally distributed signal.

Test statistics for identify difference between two strata:

1. **p-value P1.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the last time period. Find the k'th smallest p-value/1-p-value and compare this with the corresponding p-value when the gene expressions are randomized between the two strata. We sort all the p-values and 1-p-values and base the method on the smallest of these values.
2. **p-value P12.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the two last time periods. Find a new value $s = qp_1 + (1 - q)p_2$ where p_1 and p_2 is the p-value from the two periods. Find the k'th smallest s/1-s value and compare this with the corresponding p-value when the gene expressions are randomized between the two strata.
3. **Sum p-value P1.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the last time period. Find the sum of the k'th smallest p-value and compare this with the corresponding value when the gene expressions are randomized between the two strata.
4. **Sum p-value P12.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the two last time periods. Find a new value $s = qp_3 + (1 - q)p_4$ where p_3 and p_4 is the p-value from the two periods. Find the sum of the k'th smallest s/1-s values and compare this with the corresponding s-value when the gene expressions are randomized between the two strata.

5. **Sum lg p-value P1.** Similar to 3 above except take the sum of the logarithm of the p-values
6. **Sum lg p-value P12.** Similar to 4 above except take the sum of the logarithm of the s-values
7. **Covariate P1.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the last time period. Let D_i denote the difference of sum of the covariates between the two strata. Find the sum of D_i for the k'th smallest p-value and compare this with the corresponding values when the gene expressions are randomized between the two strata.
8. **Covariate P12.** Find the p-value in a t-test for each gene comparing the gene expressions from the two strata from the two last time periods. Find a new value $s = qp_1 + (1-q)p_2$ where p_1 and p_2 is the p-value from the two periods. Find $D'' = qD'_i + (1-q)D_i$ where D'_i is the corresponding difference to D_i but in time period 2. Find the sum of D''_i for the k'th smallest q-value and compare this with the corresponding expression when the gene expressions are randomized between the two strata.
9. **NofGenes CG, all.** Classify all genes in groups based on the order of the average values of the gene expression in each time period. Group 1234 consist of genes where the average value of the gene expression increases for each time period. We compare the number of genes in the curve group 1234 with the number of genes in the same curve group when randomizing the gene expression between the two strata.
10. **NofGenes CG, 0.1/0.01** Similar to 9 above, we only count genes where the difference between average gene expression in the time period with smallest and largest gene expression is significant large, with a p-value less than 0.1 or 0.01 respectively. Hence we perform a t-test between the data in two time periods and only include genes where the p-value in this two sample t-test is less than 0.1 or 0.01 respectively.

These methods represent different ideas. It is possible to improve the methods by using other transformations than the logarithm, depend on p-value in more periods or use other combinations of two or more p-values than a linear combination. F.ex. use s=min($p_1, p_2, 1-p_1, 1-p_2$) + c max($p_1, p_2, 1-p_1, 1-p_2$) for a constant 0<c<1. This formula may be better if there is only one period with small p-value or one period has a low p-value and one period has a p-value close to 1. We have not tried to optimize the different test statistics since this will depend on the situation in each case.

The results are shown in Table 1-2. The best methods seem to sum of the logarithm of the 100 smallest/largest p-values, sum of the 100 or 200 covariates of the smallest/largest p-values where we in both cases uses data from the two last time periods. But this also depends on the type of signal. If we know that the trend is increasing, then the curve group method is equal good with the best methods. But it is more reasonable case where we don't know the direction of the signal, this method is not as good. As expected, the p-value method is best for low order p-values when the signal is normally distributed and for higher order when the signal is constant. Figure 8 and 9 indicate that if the distribution of the signal for the different genes has heavier tails than the normal distribution, it may be better to focus on the few genes with

a strong signal, while with a constant value of the signal, it is best to use many genes in the test statistics. This is similar to the result for one time period.

The Method denoted P1 only uses data from the last period. Notice that we are able to improve the predictions slightly when using data from the last two periods with methods denoted P12. The method based on curve groups uses data from all four periods. How much a method is improved by using data from more time periods depends on the strength of the signal in the different periods, the number of patients in each period and how we combine the results from the different periods. In our example, the strength of the signal increases linearly in time until time of diagnosis making data in the last time period most valuable. If the strength of the signal is the same in all periods, this will make the curve group method better. But it is also possible to use p-values from several time periods as illustrated above with the method using data from period 1 and 2, denoted P12. Figure 8 and 9 show that it is possible to analyze the strength of the signal in the different time periods.

The method based on p-values seems more flexible than the curve group method. If the two strata have a different time development in the time periods before diagnosis, then we would expect to get p-values close 0 or 1 in a t-test comparing the average values between the two strata in some of the time periods.

Genes with trend, constant		100	100	100	30	30	30
Diff. last period		0.25	0.5	0.75	0.25	0.50	0.75
Type of signal		Const.	Const.	Const.	Const.	Const.	Const.
Estimator	K						
p-value, P1	10	0.081	0.25	0.77	0.059	0.10	0.24
p-value, P1	100	0.10	0.41	0.88	0.062	0.11	0.23
p-value, P12	10	0.093	0.39	0.94	0.065	0.14	0.42
p-value, P12	100	0.12	0.53	0.94	0.065	0.12	0.27
Sum lg p-value P1	10	0.070	0.23	0.71	0.053	0.091	0.24
Sum lg p-value P1	100	0.10	0.46	0.97	0.064	0.13	0.35
Sum lg p-value P12	10	0.086	0.37	0.90	0.064	0.13	0.46
Sum lg p-value P12	100	0.13	0.66	0.996	0.070	0.18	0.48
Covariate, P1	10	0.075	0.26	0.75	0.058	0.11	0.33
Covariate, P1	50	0.098	0.45	0.96	0.060	0.13	0.41
Covariate, P1	100	0.11	0.53	0.98	0.066	0.14	0.41
Covariate, P1	200	0.12	0.57	0.98	0.069	0.15	0.38
Covariate, P12	10	0.096	0.39	0.89	0.069	0.16	0.54
Covariate, P12	50	0.12	0.62	0.996	0.065	0.18	0.62
Covariate, P12	100	0.13	0.70	0.998	0.069	0.19	0.60
Covariate, P12	200	0.15	0.74	0.999	0.071	0.19	0.53
NofGenes, CG,all		0.15	0.78	0.999	0.077	0.26	0.72
NofGenes, CG .01		0.13	0.73	0.999	0.074	0.27	0.76
NofGenes,CG.001		0.038	0.075	0.38	0.049	0.085	0.25

Table 1 comparing the power in 19 different test statistics when the signal is the same constant in all genes with signal.

Genes with trend, constant		100	100	100	30	30	30
Diff. last period		0.25	0.5	0.75	0.25	0.50	0.75
Type of signal		Normal	Normal	Normal	Normal	Normal	Normal
Estimator	K						
p-value, P1	10	0.055	0.12	0.85	0.054	0.071	0.28
p-value, P1	100	0.051	0.14	0.58	0.050	0.068	0.14
p-value, P12	10	0.051	0.17	0.93	0.048	0.072	0.37
p-value, P12	100	0.050	0.16	0.59	0.052	0.072	0.14
Sum p-value P1	10	0.048	0.11	0.90	0.056	0.072	0.42
Sum p-value P1	100	0.050	0.16	0.94	0.049	0.075	0.34
Sum p-value P12	10	0.057	0.19	0.97	0.050	0.081	0.62
Sum p-value P12	100	0.059	0.23	0.98	0.052	0.090	0.47
Covariate, P1	10	0.055	0.15	0.94	0.057	0.078	0.46
Covariate, P1	50	0.056	0.17	0.96	0.052	0.080	0.40
Covariate, P1	100	0.055	0.18	0.95	0.052	0.077	0.30
Covariate, P1	200	0.053	0.19	0.91	0.047	0.13	0.35
Covariate, P12	10	0.056	0.21	0.97	0.051	0.098	0.69
Covariate, P12	50	0.061	0.26	0.988	0.047	0.089	0.59
Covariate, P12	100	0.062	0.27	0.981	0.050	0.090	0.51
Covariate, P12	200	0.057	0.25	0.96	0.050	0.089	0.41
NofGenes, CG,all		0.059	0.22	0.94	0.049	0.090	0.42
NofGenes, CG .01		0.056	0.21	0.95	0.050	0.086	0.47
NofGenes,CG.001		0.055	0.10	0.76	0.046	0.060	0.34

Table 2 comparing the power in 19 different test statistics the signal is the normally distributed.

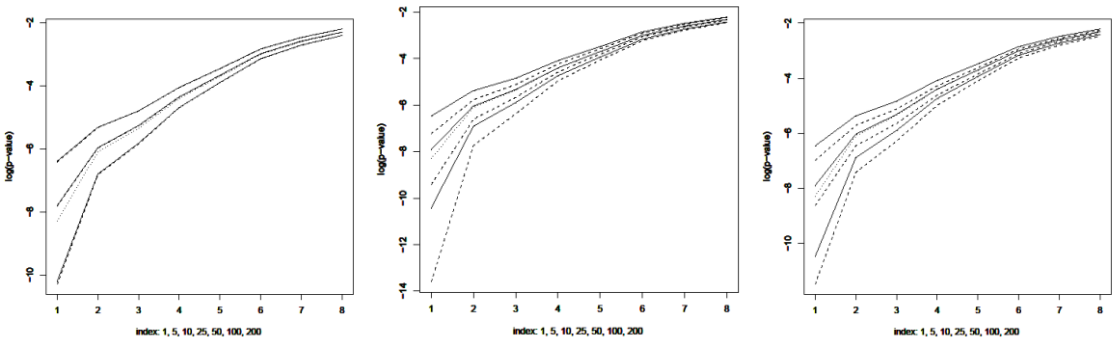

Figure 8. Quantiles of the extreme smallest/largest log(p-value) values with signal (dashed) and without signal (line) between the two strata. There are 30 genes with signal and the difference in the last period between the two strata is 0.25 in the left figure and 0.75 in two right figures. Normal signal in the two left figures and constant signal in the right hand figure.

Figure 8 and 9 show the quantiles of the log(p-value) for the extreme p-values in a test with 30 genes (figure 8) and 100 genes (figure 9) with signal and where the full difference of the constant signal is 0.25 and 0.75 respectively. It shows the quantiles of the extreme log(p-value) or log(1-p-value). For each of the 1,2,3,… values on the x-axis shows respectively quantiles of the 1st, 5th, 10th, 25th, 50th, 100th, and 200th most extreme log(p-values)/ log(1-p-value) along the horizontal axis. The three lines show the 0.05, 0.5 and 0.95 quantiles of the distribution

when there is not a signal and the dashed lines show the 0.05, 0.5 and 0.95 quantiles of the extreme p-values when there is a difference between the two strata. The dotted line is the expected value of the extreme p-values when there is no signal (k-0.5)/(2*number of genes) where k=1,5,10,…. These figures show that it is possible to analyze the strength of the signal in a group when there is sufficient number of patient data that it is possible to perform a t-test comparing to a reference population. Notice that the normally signal is more visible for low order p-values.

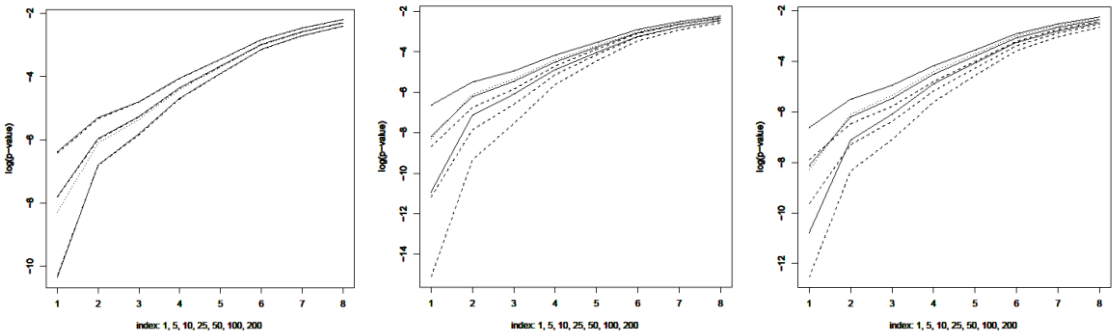

Figure 9. Quantiles of the extreme smallest/largest log(p-value) values with (dashed) and without (line) signal between the two strata. Similar to figure 8 except there are 100 genes with signal. Normal signal in the two left figures and constant signal in the right hand figure.

It is difficult to compare the difference when k is large since all the curves are close. In Figure 10 the values are scaled relative to the variability when there is no signal. Notice that the difference is not observable for 30 genes with 0.25 as expected difference since two of the dotted curve is almost identically equal 0 and 1 indicating that we have the same values for the quantiles as if there were no signal. When there are 100 genes with expected difference it is possible to notice a difference for the normal signal. When there are 100 genes with expected value 0.75, it is a large difference in the distribution. The difference is largest for p-value of order 5-10 for normally distributed signal and for p-values of order 25-100 for constant signal. This implies that also in the case when the signal is very weak, we will in some cases be able to state that there is a significant difference. Notice also that in the distribution of the ordered p-values are partly overlapping in the three cases with different signal (100 genes constant 0.75 difference, 100 genes constant 0.25 difference and 30 genes constant 0.75 difference). This means that we cannot expect to be able to identify the type of signal (e.g. number of genes, strength of signal in each gene) based on the distribution of the p-values in one data set. In some cases it may be possible and in general it will be possible to give an overall description of the strength of the signal.

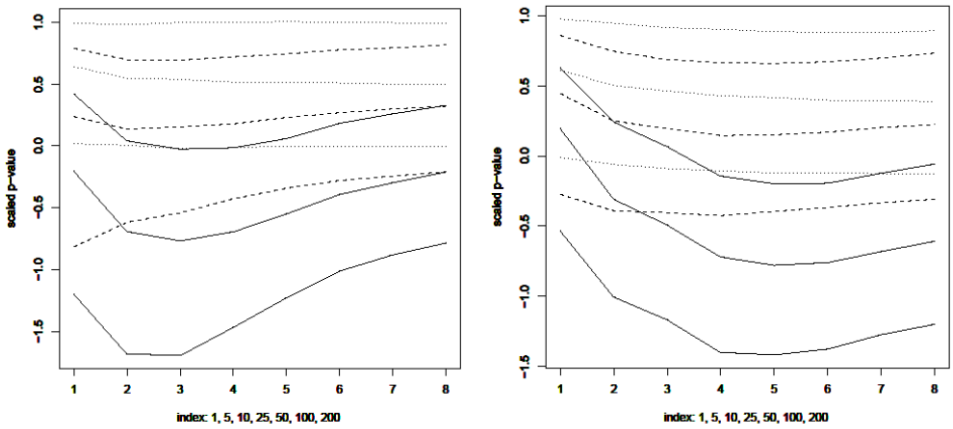

Figure 10 Quantiles of the extreme smallest/largest log(p-value) values when there is a signal between the two strata values scaled relative to similar quantiles when there is not a signal. The scale is such that (0,1) is similar to the variability between the 0.05 and 0.95 quantile of the p-value when there is no signal. The left figure is for signal that is normally distributed and the right figure is constant signal. In both figures we have three curves with lines (0.05, 0.5 and 0.95 quantiles when there are 100 genes with a constant 0.75 difference), three dashed curves (0.05, 0.5 and 0.95 quantiles when there are 100 genes with a constant 0.25 difference) and three dotted curves (0.05, 0.5 and 0.95 quantiles when there are 30 genes with a constant 0.75 difference). Horizontal axis is the order of the p-values 1, 5, 10, 25, 50, 100, 200 respectively.

3.2 Classification of strata for new patients

Assume that we have some data where we know the strata and performed the analysis in the previous section. If we receive new data where we don't know the strata, we may try to classify the strata for the new persons based on the gene expression for the genes that we identified in the analysis described in the previous section. Figure 11 compares how the curve group method and the extreme p-values are able to identify the genes with signal in an example with 30 out of 1000 genes have a signal and where the difference in the gene expression in the last time period is the constant value 0.75 relative to the standard deviation of the gene expression. The dotted line shows that the genes with smallest p-value/1-p-value has the probability 67% for being a gene with a signal and this is decreasing down to about 10% for the gene with 200th smallest p-value/1-p-value. In the curve group method the average result is that the 6 genes with the most significant increase in the trend has 54% probability to have a signal, decreasing to the 48 genes with a monotone increase in the trend has 22% probability to have a signal and the remaining genes have a 1.5% probability for a signal. Apriori all genes have a 3% probability for having a signal.

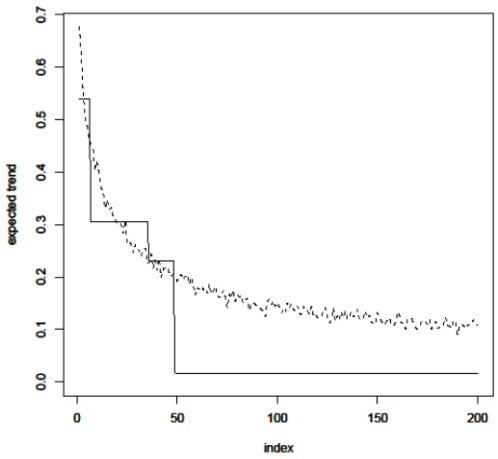

Figure 11. The probability for a the signal as a function of the order of genes sorted by the order of the extreme p-value (dashed line) and by curve group (line)

We will test different predictors for whether a new patient belongs to the strata with or without a signal. The predictors are based on the analysis in the previous section where we test different methods to separate between the two strata. Similar to the previous section we compare different test statistics for prediction whether a new patient belong to the strata with or without a signal. We make a classification that has a 5% probability for a false positive prediction, i.e. that a new patient that belong to the stratum without a signal wrongly is classified as the other strata. Then we find the probability for a correct positive classification, i.e. that a new patient that belong to the stratum with a signal is correctly classified. The method is tested on 10 data sets that each consists of 1000 synthetic genomes each with 1.000 genes. For each of the genomes we have classified 10 new patients. The same random variables are used for the different test statistics. The results are given in Table 3 with constant signal and Table 4 with normally distributed signal.

All the predictors have the form $Y_{j,k} = F(w) = \sum_i w_{i,k} X_{i,j}$. We set $G(w) = \sum_i w_{i,k}^2 \quad = 1$ in order that $Y_{j,k}$ is N(μ,1) where μ =0 in the case there is no signal. Then if $Y_{j,k}$>q_0.95 (the 0.95 quantile in the N(0,1) distribution) this gives a prediction that patient j belong to the strata with signal. This predictor has a 5% probability for a false positive prediction in the case the patient belongs to the stratum without a signal. From Figure 11 we see that we know the probability that the gene with the p-value of order q_i is a gene with a signal. We want to find the optimal vector w such that μ = EF(w) is as large as possible in the case the person belongs to the stratum with a signal. This is a linear optimization of $\sum_i w_{i,k} q_i$ under the nonlinear constraint $\sum_i w_{i,k}^2 \quad = 1$. The optimal solution of this is to set $w_i = {}^{q_i}/_{c}$ for a constant c since this gives in the optimal point ${}^{\partial F(w)}/_{\partial w_i} = q_i = c w_i = 2c \; {}^{\partial G}/_{\partial w_i}$

The expected value of the predictor is $\mu = \sum_i w_i q_i = c$.

The probability q_i shown in figure 11 is quite good approximated with q_i= 0.75exp(-i/65) +0.15-i/13000. This formula is used in the trends in Table 4. In Table 3 with constant signal we have used the formula q_i= 0.75exp(-i/15) +0.1-i/13000 which gives a better match for 100 genes with 0.75 signal.

We have three types of predictors. First the method is based only on the data in the last time period. It is performed a t-test based on the data from the two strata in the last time period and the genes are sorted in increasing p-value/1-p-value. We use the weight $w_{i,k} = c_k \exp(^{i}/_{h_k})$ if gene i has the i'th smallest p-value/1-p-value in the t-test. The variable h_k is a constant and h_k is set such that $\sum_i w_{i,k}^2 = 1$ and a sign such that a large value of $Y_{j,k}$ indicates that the new person is in the strata with signal. The second method use data from the two last periods and use the value $s = qp_1 + (1-q)p_2$ when identifying which genes that are most important. Here p_1 and p_2 are the p-value from the two periods. The weights $w_{i,k}$ is set similar to the first method but based on the value s for each gene instead of the p-value. The third method is based on the curve group classification which is estimated from data from all the periods. We find the genes that have a systematically increase in the average value of the four time periods. In the first of these methods we set $w_{i,k}$ equal the same value for all the genes where the average value increases in the four periods. In the second of these methods, we in addition require that the t-test comparing the values in the first and last time period is less than 0.1. It the third of these methods, we in addition require that the t-test comparing the values in the first and last time period is less than 0.01.

Genes with trend, constant		100	100	100	30	30	30
Diff. last period		0.25	0.5	0.75	0.25	0.50	0.75
Type of signal		const	Const	Const	Const	Const	Const
Predictor,data period	H_k						
P1	1	0.059	0.11	0.24	0.052	0.073	0.15
P1	5	0.062	0.13	0.33	0.053	0.077	0.18
P1	10	0.066	0.18	0.53	0.054	0.086	0.23
P1	25	0.072	0.23	0.70	0.056	0.095	0.27
P1	50	0.081	0.31	0.88	0.059	0.11	0.32
P1	100	0.090	0.38	0.94	0.061	0.12	0.34
P1	150	0.098	0.43	0.96	0.062	0.12	0.33
P1	200	0.0991	0.43	0.96	0.062	0.12	0.32
P1	250	0.0989	0.43	0.95	0.062	0.12	0.31
P1	Trend	0.091	0.39	0.94	0.061	0.12	0.35
P1-2	1	0.062	0.13	0.27	0.053	0.085	0.20
P1-2	5	0.064	0.16	0.40	0.055	0.094	0.26
P1-2	10	0.071	0.23	0.65	0.056	0.11	0.35
P1-2	25	0.077	0.31	0.84	0.057	0.12	0.42
P1-2	50	0.090	0.43	0.97	0.061	0.14	0.47
P1-2	100	0.10	0.51	0.988	0.063	0.15	0.47
P1-2	150	0.11	0.55	0.993	0.064	0.15	0.44

P1-2	200	0.11	0.55	0.991	0.065	0.14	0.41
P1-2	250	0.11	0.54	0.989	0.065	0.14	0.39
P1-2	Trend	0.10	0.51	0.985	0.065	0.15	0.49
P1-4, all genes		0.095	0.31	0.80	0.061	0.10	0.23
P1-4, p<0.1		0.089	0.32	0.85	0.060	0.11	0.27
P1-4, p<0.01		0.069	0.21	0.70	0.055	0.093	0.26
P1 mu	25	0.306	1.19	2.58	0.104	0.495	1.36

Table 3 comparing the power in 22 different predictions for whether a new patient belongs to the stratum with signal. Mu is the expected value of the predictor in case of a signal and when using data from the last period and weight with h=25.

Genes with trend, constant		100	100	100	30	30	30
Diff. last period		0.25	0.5	0.75	0.25	0.50	0.75
Type of signal		Normal	Normal	Normal	Normal	Normal	Normal
Predictor,data period	H_k						
P1	1	0.050	0.086	0.61	0.050	0.063	0.37
P1	5	0.051	0.095	0.75	0.051	0.066	0.42
P1	10	0.051	0.11	0.89	0.051	0.068	0.45
P1	25	0.051	0.12	0.94	0.051	0.070	0.45
P1	50	0.051	0.14	0.96	0.052	0.074	0.42
P1	100	0.053	0.15	0.96	0.051	0.074	0.37
P1	150	0.054	0.15	0.94	0.052	0.075	0.31
P1	200	0.054	0.15	0.92	0.051	0.074	0.28
P1	250	0.054	0.15	0.90	0.051	0.074	0.26
P1	Trend	0.053	0.15	0.97	0.052	0.076	0.42
P1-2	1	0.051	0.11	0.69	0.051	0.073	0.49
P1-2	5	0.052	0.12	0.84	0.051	0.076	0.56
P1-2	10	0.052	0.14	0.96	0.050	0.078	0.61
P1-2	25	0.053	0.16	0.98	0.050	0.079	0.60
P1-2	50	0.053	0.18	0.991	0.051	0.081	0.53
P1-2	100	0.054	0.19	0.988	0.050	0.081	0.45
P1-2	150	0.055	0.19	0.98	0.051	0.080	0.37
P1-2	200	0.056	0.19	0.96	0.051	0.078	0.32
P1-2	250	0.055	0.18	0.95	0.050	0.078	0.30
P1-2	Trend	0.055	0.20	0.992	0.051	0.084	0.54
P1-4, all genes		0.051	0.082	0.40	0.051	0.060	0.12
P1-4, p<0.1		0.051	0.088	0.50	0.052	0.060	0.15
P1-4, p<0.01		0.048	0.084	0.64	0.049	0.059	0.24
P1 mu	25	0.12	0.47	1.81	0.036	0.15	0.72

Table 4 comparing the power in 22 different predictions for whether a new patient belongs to the stratum with signal.

Notice that when the signal is constant it is best to have large h_k values that gives more weight to higher order p-values relative to when the signal is normally distributed. We are able to

increase the classification when we use weight that is closer to the curve shown in Figure 11. However, the optimal weights depend on the parameters of the signal. We see that p-value methods are better than curve groups and naturally it is better to use information from several time periods.

4 References

1. Marit Holden and Lars Holden: “Statistical analysis of gene expression data related to breast cancer diagnosis”. Norsk Regnesentral SAMBA/19/2014.