Nonparametic analysis of replicated microarray experiments

Ali Gannoun[1,2,3], Jérôme Saracco[1], Wolfgang Urfer[4] and George E. Bonney[2,3]

(1) Laboratoire de Probabilités et Statistique, Université Montpellier II, France

(2) Department of Microbiology, Howard University, 520 W Street NW, Rm 3019, Washington D.C. 20059, USA

(3) Statistical Genetics and Bioinformatics Unit, Howard University, National Human Genome Center, 2216 6th street, suite 206, Washington D.C. 20059, USA

(4) Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

Microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels of thousands of genes simultaneously. In microarray data analysis, the comparison of gene-expression profiles with respect to different conditions and the selection of biologically interesting genes are crucial tasks. Multivariate statistical methods have been applied to analyze these large data sets. In particular, Dudoit et al. [2002] developed methods using t-statistics with p-values calculated through permutations, and with the Westfall and Young [1993] step-down approach to correct for multiple testing. Thomas et al. [2001] developed a regression modelling approach. Following the idea of Efron et al. [2000] and Tusher et al. [2001], Pan [2002] proposed a Normal mixture modelling approach that relaxes many strong assumptions on the null distributions of the test statistics. This paper makes two contributions to the analysis of microarray data. The first is the introduction of a new method for the calculation of the cut-off point and the acceptance region, and the second is the replacement of the based Normal mixture density estimators proposed by Pan et al. [2002], with less restrictive kernel nonparametric ones. A useful modification is suggested in order to increase the performance of the kernel estimator on the tail of the distribution. We apply our approach to leukemia data of Golub et al. [1999] and compare our results to those of Pan [2002].

Keywords: Kernel estimator; Microarray; Mixture modeling; Regression modeling, t-test.

1 Introduction

Gene expression regulates the production of protein, the ultimate expression of the genetic information, which in turn governs many cellular processes in biological systems. The knowledge of gene expression has applications ranging from basic research on the mechanism of protein production to applications such as diagnosing, staging, treating and preventing of diseases. Microarray techniques provide a way of studying the RNA expression levels of thousands of genes simultaneously; see for example Brown and Botstein [1999], Lander [1999], Quackenbush [2001]. The identification of differentially expressed genes is a question which arises in a broad range of microarray experiments which produce enormous amounts of data, see Spellman et al. [1998], Galitski et al. [1999], Golub et al. [1999], Callow et al. [2000], Friddle et al. [2000], and Guimaraes and Urfer [2002], to name a few. The expression level can refer to summary measure of relative red to green channel intensities in a fluorescence-labeled complementary DNA or cDNA array, a radioactive intensity of a radiolabeled cDNA array, or summary difference of the perfect match (PM) and mis-match (MM) scores from an oligonucleotide array, see Li and Wong [2001]). Microarray experiments involve a number of distinct stages which are discussed in Smyth et al. [2002]. The expression levels may have been suitably preprocessed to acquire red and green foreground and background intensities for each spot of the arrays, including dimension reduction, data normalization and data transformation; see for example Dudoit et al. [2002], Efron et al. [2000], Li and Wong [2001]. We suppose here that all stages to get data are satisfied. For the purpose of the paper, let Rf and Gf (resp. *Rb* and *Gb*) be the foreground (resp. the background) red and green intensities for each spot. The *log*-differential expression ratio will be $Y = log_2(R/G)$ where usually $R = Rf - Rb$ and $G = Gf - Gb$, where $G > 0$. One of the core goals of microarray data is to compare, for example, the expression levels of genes in samples drawn from two different cell types, such as healthy versus diseased cells, and to identify which of the genes show good evidence of being differentially expressed. Statistical methods are very helpful to reach this goal. In the early days, many data analysis programs sort the genes according to the absolute level of $\overline{Y}$, where $\overline{Y}$ is Y -values for any particular gene across the replicate arrays, see Smyth et al. [2002] for more details. This is known to be unreliable (see Chen *et al.* [1997]) because statistical variability of the expression levels for each genes was not taken account. It has also been noted that data based on a single array may not be reliable and may contain high noises (see Lee et al. [2000]). Moreover, the need for independent replicates has been recognized (see Lee et al. [2000]), and several methods from combining information from several arrays have been proposed. These methods assign a test score to each of the genes and then select those that are "significant". The test statistics included the t-test (Zhang *et al.* [1997], Alon et al. [1999]), the ANOVA F-statistics (Kerr et al. 2000]) and the information theoric measure known as InfoScore (Hadenfalk et al. [2001]). Recently, Chilingaryan et al. [2002] used a multivariate approach based on Mahalanobis distance between vectors of gene expressions as a criterion for simultaneously comparing a set of genes, and developed an algorithm for maximizing it. A Similar vectorial approach, including principal components analysis, is also given by Kuruvilla et al. [2002]. Bayesian probabilistic framework for microarray data analysis are also developed by Friedman et al. [2000], Baldi et al. [2001], Imato et al. [2002] among others. In this paper we consider the detection of differentially expressed genes with replicated measurements of expression levels using Bayesian inference with the mixture model approach of Pan et al. [2002]. It is one of the three methods reviewed by Pan [2002]. In particular, we introduce a kernel estimator of density functions in order to form the test statistic in the Bayesian techniques.

The paper is organized as follows. In Section 2, we describe the statistical model and two existing testing methods: the t-test approach and the Normal mixture model approach. In Section 3, we propose a kernel estimation procedure, and we give a new method to determine the cut-off point and the acceptance region. This nonparametric approach is illustrated in Section 4 using the leukemia data of Golub et al. [1999], and is compared to the Normal mixture model approach of Pan et al. [2002]. Section 5 summarizes some concluding remarks and gives an outlook for further activities.

2 Statistical model and existing methods

In this section, we give a general statistical model from which we make the comparative studies. Then, we recall the construction of the t-test method and the mixture modeling approach.

2.1 The model

Various models are proposed to summarize multiple measurements of gene expression. A general survey is given by, for example, Thomas *et al.* [2001], Li and Wong [2001] and Sebastiani and Romani [2002]. We will focus on a simple model studied in particular by Pan et *al.* [2002].

Suppose that Y_{ij} is the expression level of gene i in array j, $i = 1, ..., n$ and $j = 1, ..., J$. We suppose that $J = J_1 + J_2$ and that the first J_1 and last J_2 arrays are obtained under the different conditions, say treatment and control, respectively.

We consider the following general statistical model:

$$Y_{ij} = \beta_i + \mu_i x_j + \varepsilon_{ij} \tag{1}$$

where $x_j = 1$ for $1 \leq j \leq J_1$ and $x_j = 0$ for $J_1 + 1 \leq j \leq J_1 + J_2$, and ε_{ij} are independent random errors with mean 0. Hence $\beta_i + \mu_i$ and β_i are the mean expression levels of gene i under the two conditions respectively.

Let H_{0i} denote the null hypothesis of equal treatment and control mean expression levels for gene i, $i = 1, ..., n$. Here we consider only two-sided alternative hypotheses; one-sided alternatives can be handled in similar manner. Then, determining whether a gene has differential expression is equivalent to testing the null hypothesis

$$H_{0i} : \mu_i = 0 \quad \text{against} \quad H_{1i} : \mu_i \neq 0.$$

A statistical test consists of two parts. The first is to construct a summary test statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence. The second is to choose a critical-value, or the significance level or p-value associated with the test statistic above which any value is considered to be significant. In many microarray studies the aim is to identify a number of candidate genes for confirmation and further study.

To focus on the main issue, we use $\alpha = 0.01$ as the genome-wide significance level. To account for multiple hypothesis testing, one may calculate adjusted p-values, see Shaffer ([1995] and Westfall and Young [1993]. According to Shaffer [1995], given any procedure, the adjusted p-value corresponding to the test of single hypothesis H_{0i} can be defined as the level of the entire test procedure at which H_{0i} would just be rejected, given the values of all test statistics involved. The Bonferroni method is perhaps the best known method with multiple testing (see Dudoit et al. [2002] and Thomas et al. [2001]). This method should be used here. Hence the gene-specific significance level (for a two-sided test) is $\alpha^* = \alpha/(2n)$.

In the following, we review two existing methods along the line.

2.2 The t-test

Let us recall that H_{0i} denote the null hypothesis of equal expression levels under the two different conditions (e.g. control and treatment) for gene i, $i = 1, ..., n$. As, we consider only two-sided alternative hypotheses, the t-statistic comparing gene expression for gene i is

$$Z_i = \frac{\overline{Y}_{i(1)} - \overline{Y}_{i(2)}}{\sqrt{\frac{s^2_{i(1)}}{J_1} + \frac{s^2_{i(2)}}{J_2}}}, \tag{2}$$

where $\overline{Y}_{i(1)}$ and $\overline{Y}_{i(2)}$ denote the average expression level of gene i in the J_1 treatment and J_2 control hybridizations, respectively. Similarly, $s^2_{i(1)}$ and $s^2_{i(2)}$ denote the sample variances of gene i's expression level in the treatment and control hybridizations, respectively.

Large absolute t-statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. Note that the replication is essential for such an analysis because of the need for assessing the variability of gene expression levels in the treatment and control groups.

Under the Normality assumption of Y_{ij}, Z_i approximately has a t-distribution with degree of freedom

$$d_i = \frac{(s^2_{i(1)}/J_1 + s^2_{i(2)}/J_2)^2}{(s^2_{i(1)}/J_1)^2/(J_1-1) + (s^2_{i(2)}/J_2)^2/(J_2-1)}.$$

A classical approximation of d_i is given by $J_1 + J_2 - 1$, see for example Scheffé [1970] and Best and Rayner [1987]. If we do not assume the t-distribution, we use permutation to estimate their distribution, see Dudoit et al. [2002] for more details. Wastfall and Young [1993] suggest approximating the p-values using asymptotic theory, see also Dudoit et al. [2002] for a computational algorithm.

2.3 The mixture model approach

The ordinary t-statistic is not ideal because of its restrictive assumptions. Strong assumptions (e.g. normality, equality of variances) are needed for the null distribution of the test statistics. To estimate the null distribution, Pan [2002] and Pan et al. [2002] constructed the following null statistics

$$z_i = \frac{Y_{i(1)}u_i/J_1 - Y_{i(2)}v_i/J_2}{\sqrt{\frac{s^2_{i(1)}}{J_1} + \frac{s^2_{i(2)}}{J_2}}} \tag{3}$$

where $Y_{i(1)} = (Y_{i1}, Y_{i2}, ..., Y_{iJ1})$, $Y_{i(2)} = (Y_{iJ_1+1}, Y_{iJ_1+2}, ..., Y_{iJ_1+J_2})$, u_i is a random permutation of column vector containing $J_1/2$ 1's and -1's respectively, and v_i is a random permutation of column vector containing $J_2/2$ 1's and -1's respectively.

Let f and f_0 be the distribution densities of Z_i and z_i.

If there is no expression change for gene i , then Z_i should have the same distribution as that of z_i. Under the weak assumption that the random variable ε_{ij} in (1) has a distribution symmetric about its mean 0, then under H_{0i}, $f = f_0$.

If we assume that the distribution of Z_i's for genes that are differentially expressed is f_1, f can be expressed as a mixture of f_0 and f_1, that is

$$f = p_0 f_0 + p_1 f_1$$

where p_1 is an unknown proportion of the genes that are differentially expressed and $p_0 = 1 - p_1$.

For any given Z, if we know the densities f_0 and f, we use the likelihood ratio test statistic:

$$LR(Z) = f_0(Z)/f(Z) \tag{4}$$

to test for H_0. Then, by the optimal Neyman-Pearson test, a small value of $LR(Z)$, say $LR(Z) < c$, provides evidence to reject H_0. The cut-off point c is determined such that the type I error kind is

$$\frac{\alpha}{n} = \int_{LR(z)<c} f_0(z)dz \tag{5}$$

where α is the genome wide significance level.

In the absence of strong parametric assumptions, the parameters p_0, f_0 and f_1 are not identifiable, see Efron et al. [2000]. By assuming a normal distribution for Z_i, for each i, one can estimate the components of the mixture by using for example the EM algorithm (see Dempster et al. [1977]). Lee et al. [2000] and Newton et al. [2001] considered parametric approaches by assuming Normal or Gamma distributions for f_0 and f_1 respectively. Efron et al. [2000] avoided such parametric assumptions and considered a nonparametric empirical Bayesian approach.

Practical determination of the cut-off point c and decision rule.
From Efron and Tibshirani [1993], Pan [2002], Pan et *al.* [2002], Efron et al. [2000, 2001], a parametric bootstrap approach proceeds as follows. We draw B random samples from f_0: $z^{(1)}, z^{(2)}, ..., z^{(B)}$, where $z^{(b)} = \{z_1^{(b)}, z_2^{(b)}, ..., z_N^{(b)}\}$ for $b = 1, ..., B$. Then for a possible cut-off point c, the average of false rejections can be calculated by:

$$False(c) = \frac{1}{B}\sum_{b=1}^{B} \#\{i : LR(z_i^{(b)}) < c\}.$$

Based on a desired false rejection number, we can choose the corresponding c. This cut-off point c yields the corresponding rejection region for H_0 which is given by

$$\{Z: \ LR(Z) < c\}.$$

Remark. With the Normal mixture model in Pan *et al.* [2002], it is possible to numerically solve the equation (5) in order to determine the cut-off point c by using the bisection method (Press et al.[1992]).

3 A fully nonparametric approach

Based on the likelihood ratio test approach of Pan et al. [2002] and using z_i's and Z_i's defined respectively in (2) and (3), we will nonparametrically estimate f_0 and f by a kernel method and develop a procedure to determine the rejection region from an approximation of (5).

3.1 Kernel estimation of f_0 and f

The construction of a kernel estimator of the density functions f and f_0 requires a choice of a real (density) function K (called kernel), and bandwidths h_n and h_{0n} which are sequences of positive numbers tending to 0 as n tends to infinity. From $\{Z_i,\ i = 1, ..., n\}$ and $\{z_i,\ i = 1, ..., n\}$, f and f_0 can be estimated nonparametrically by

$$f_n(z) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h_n}\right) \tag{6}$$

and

$$f_{0n}(z) = \frac{1}{nh_{0n}} \sum_{i=1}^{n} K\left(\frac{z - z_i}{h_{0n}}\right). \tag{7}$$

Well known theoretical results show that the choice of a reasonable K does not seriously affect the quality of the estimators (6) and (7). In order to get smoother estimation, one can use a kernel K which is bounded, symmetric and satisfying $|z|\, K(z) \to 0$ as $|z| \to \infty$ and $\int z^2 K(z)dz < \infty$. Some special kernel functions are given in Table 1.

Kernel	$K(z)$
Uniform	$\frac{1}{2}1(\lvert z\rvert \leq 1)$
Triangle	$(1 - \lvert z\rvert)1(\lvert z\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - z^2)1(\lvert z\rvert \leq 1)$
Quartic	$\frac{15}{16}(1 - z^2)^2 1(\lvert z\rvert \leq 1)$
Triweight	$\frac{35}{36}(1 - z^2)^3 1(\lvert z\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}$
Cosines	$\frac{\pi}{4}\cos(\frac{\pi}{4}z)1(\lvert z\rvert \leq 1)$

Table 1: *Examples of density kernel functions.*

On the contrary the choice of the bandwidths h_n and h_{0n} turns to be crucial for the accuracy of the estimators (6) and (7). Some indications about this choice are given in Bosq and Lecoutre [1987]. For example, one can use

$$h_n = \widehat{\sigma}_n n^{-1/5} \quad \text{and} \quad h_{0n} = \widehat{\sigma}_{0n} n^{-1/5}, \tag{8}$$

where $\widehat{\sigma}_n$ and $\widehat{\sigma}_{0n}$ denote the empirical standard deviation of the Z_i's and the z_i's. From a theoritical point of view, this choice minimizes some asymptotic mean square error (see Deheuvels [1977]). In practice, this choice gives an idea of the amount of smoothing needed for the estimator. For the graphical aspect of the corresponding estimated density function curve, the user can choose to increase or decrease the value of the bandwidth in order to obtain the desired smoothing of the density estimators.

Note that it is well-known that the kernel density estimator does not perform well on the support edges of the distribution. In the following, we suggest a very simple and practical method for overcoming edge effect problems, and by the way for giving more efficient estimator of the LR function.

3.2 The reflection approach in kernel estimation

Reflection principles in density estimation have been described and studied by Schuster [1985], Silverman [1986], and Cline and Hart [1991]. Here we present a slighty different version of the geometric approach for removing the edge effects proposed by Hall and Wehrly [1991].

Let $x_{(1)}, \ldots, x_{(n)}$ be the initial ordered data from which we will determine the estimator of the density function, say g. We add $\beta\%$ artificial observations in the tail of the distribution using the following principle.

- In the left tail, the "new" observations are $\tilde{x}_{(i+1)} = x_{(1)} - \left(x_{(i+1)} - x_{(1)}\right)$, $i = 1, \ldots, [\beta n/2]$, where $[m]$ is the integer part of m.
- In the right tail, the "new" observations are $\hat{x}_{(i+1)} = x_{(n)} + \left(x_{(n)} - x_{(n-i)}\right)$, $i = 1, \ldots, [\beta n/2]$.

Finally we estimate g from the overall data set (i.e. from the union of the original data x_i and the pseudo-data $\tilde{x}_i$ and $\hat{x}_i$.

Remark 1. When the number n of observations is large, the adjusted estimator is very sensitive to the percentage β of artificial observations. Generally, it suffices to take it very small (around 0.5%) to get a reasonable estimator.

Remark 2. If there are not enough observations close to the extreme values $x_{(1)}$ and $x_{(n)}$, we can adapt the same outline described previously by replacing $x_{(1)}$ and $x_{(n)}$ by some extreme empirical quantiles, such as the 1th and 99th centiles of the data.

3.3 Implementation of the nonparametric method

Here we propose an empirical method to solve (5). This method works even in Pan's approach and with any estimator of f and f_0.

For the purpose of this paper, the densities f and f_0 are replaced by their kernel estimators f_{0n} and f_n given in (6) and (7). We solve the modified equation

$$\frac{\alpha}{n} = \int\limits_{\widehat{LR}(z)<c} f_{0n}(z)dz, \tag{9}$$

where $\widehat{LR}(z) = f_{0n}(z)/f_n(z)$.

For a fixed value $c > 0$, let $A_c = \{z : T < c\}$ where $T = LR(z)$. We generate an ordered grid of N points $\{\tilde{z}_k,\ k = 1, \ldots, N\}$ covering the support of the Z_i's. Let $\widehat{T}_k = \widehat{LR}(\tilde{z}_k),\ k = 1, \ldots, N$. Let us define $\widehat{A}_c = \left\{\tilde{z}_k : \widehat{T}_k < c,\ k = 1, \ldots, N\right\}$ and $\overline{\widehat{A}_c} = \left\{\tilde{z}_k : \widehat{T}_k \geq c,\ k = 1, \ldots, N\right\}$, the complementary of $\widehat{A}_c$. We assume now that $\overline{\widehat{A}_c}$ is a connex set (that is an interval). Let $\tilde{z}_{c,(1)}, \tilde{z}_{c,(2)}, \ldots, \tilde{z}_{c,(q)}$ be the q ordered values of $\overline{\widehat{A}_c}$. Then

$$\begin{aligned}\int\limits_{A_c} f_0(z)dz \approx \int\limits_{\widehat{A}_c} f_{0n}(z)dz &\approx \int_{-\infty}^{\tilde{z}_{c,(1)}} f_{0n}(z)dz + \int_{\tilde{z}_{c,(q)}}^{+\infty} f_{0n}(z)dz \\ &\approx \int_{\tilde{z}_1}^{\tilde{z}_{c,(1)}} f_{0n}(z)dz + \int_{\tilde{z}_{c,(q)}}^{\tilde{z}_N} f_{0n}(z)dz.\end{aligned}$$

The left hand side integral can be evaluated by classical numerical integration method (trapezoidal quadrature). Now, the approximate cut-off point is the value c^* of the set $\{\frac{l}{\mathcal{N}},\ l = 0, 1, \ldots, \mathcal{N}\}$ where $\mathcal{N}$ is chosen as large as possible, such that

$$\frac{\alpha}{n} \approx \int\limits_{\widehat{A}_{c^*}} f_{0n}(z)dz.$$

From this cut-off point c^*, we can easily deduce the rejection region which is given by

$$\{Z : Z < \tilde{z}_{c^*,(1)} \ \text{ or } \ Z > \tilde{z}_{c^*,(q)}\}.$$

4 Data analysis

This section is devoted to the application of our proposed method. We describe the data and present the results on expression level study of genes. Then, using simulation study, we check the efficiency of the kernel method against the true Normal Mixture model.

4.1 The data

We apply the methods on the leukemia data of Golub et al. [1999]. Data have been generated for leukemic myeloid (AML) and lymphoblastic (ALL) cells taken from different individuals. There are 27 ALL samples and 11 AML samples. Here our goal is to find genes with differential expression between ALL and AML. This data set was analyzed by Thomas et al. [2001], Pan [2002], Grant et al. [2001] among others. There are $n = 7129$ genes in each sample.

We take the genome-wide significance level at the usual $\alpha = 0.01$. Data preprocessing and normalization are accomplished by Pan [2002].

4.2 Summary of the results obtained with the Normal mixture model

The density function f_0 and f estimated by Pan [2002] are

$$f_{0m}(z) = 0.479\phi(z, -0.746, 0.697) + 0.521\phi(z, 0.739, 0.641) \tag{10}$$

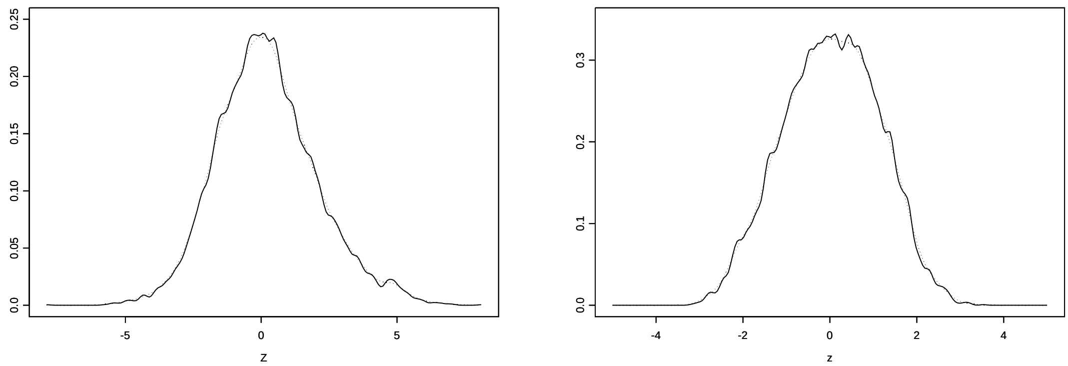
and

$$f_m(z) = 0.518\phi(z, -0.318, 1.803) + 0.482\phi(z, 0.7781, 4.501), \tag{11}$$

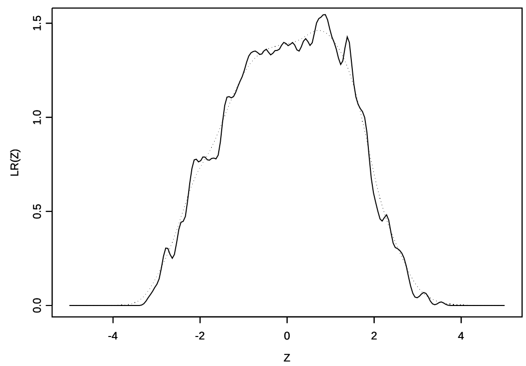
where $\phi(z, a, b)$ denotes the normal density function with mean a and variance b. Using the bisection method (Press et al. [1992], p.353), the cut-off point obtained by Pan [2002] is $c = 0.0003437$. The corresponding rejection region for H_0 is $\{Z : Z < -4.8877 \text{ or } Z > 4.4019\}$, which gives 187 genes with significant expression changes. More details are in Pan [2002].

4.3 Results obtained with our nonparametric approach

In order to implement our nonparametric modelling approach, we have to choose the kernel and the bandwidth. To estimate nonparametrically f_0 and f, we used the Gaussian density as kernel K. Other kernel functions did not yield noticeably different results and will not be presented here. Concerning the choice of the bandwidths h_n and h_{0n}, we first used the formulae given in (8). We obtained the following values: $h_n = 0.313$ and $h_{0n} = 0.187$. The estimated densities f_n and f_{0n} defined respectively in (6) and (7) are plotted on Figures 1(a). For comparison, the density functions f_{0m} and f_m of Pan [2002] given in (10) and (11) are also plotted with dotted lines. One can observe that with this choice of bandwidths these curves are clearly not sufficiently smoothed. The same is true for the corresponding likelihood ratio curve presented in Figure 1(b). The deviations from the smooth curves are due



(a) Kernel density estimation of the Z_i's with the bandwith h_n (on the left hand side) and of the z_i's (on the right hand side)

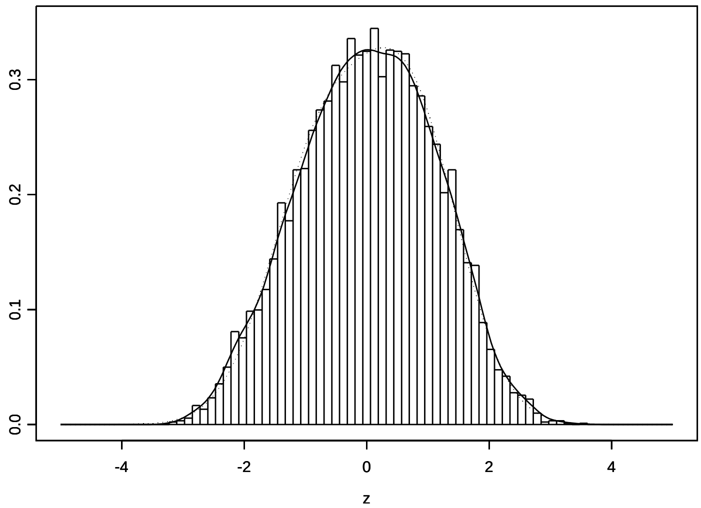


(b) Corresponding estimated LR function

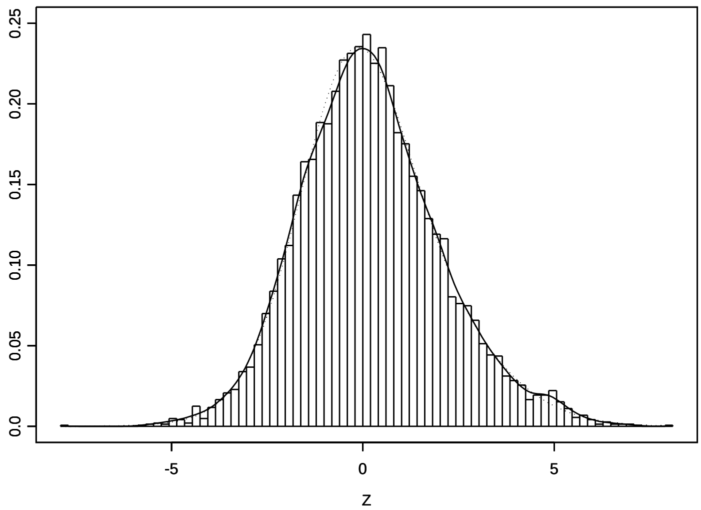
Figure 1: *Estimation of the density functions f and f_0 and of the LR function without over-smoothing. (The dotted lines are the corresponding curves obtained by Pan [2002].)*

to background noises which are not informative. Smoother curves can be obtained by broadening the bandwidths.

From graphical point of view, in order to obtain a reasonably smoothed estimator of f, f_0 and LR, we need to increase the values of the bandwidths. This is done by multiplying them by a factor 2.5 which is the "optimal value" obtained from our computational study of this data. The corresponding bandwidths are $h_n^* = 0.782$ and $h_{0n}^* = 0.468$. Figure 2(a) and 2(b) present the histogram of the z_i's and the Z_i's, and the estimated densities f_{0n} and f_n. Again for comparison, the density functions f_{0m} and f_m given in (10) and (11)) are also plotted in Figure 2(a) and 2(b). The corresponding LR function is depicted in Figure 3.



(a) Histogram of the z_i's, Pan estimated density (dotted line) and over-smoothed kernel estimated density with bandwidth h^*_{0n} (solid line)



(b) Histogram of the Z_i's, Pan estimated density (dotted line) and over-smoothed kernel estimated density with bandwidth h^*_n (solid line)

Figure 2: *Estimation of the z_i's and Z_i's densities.*

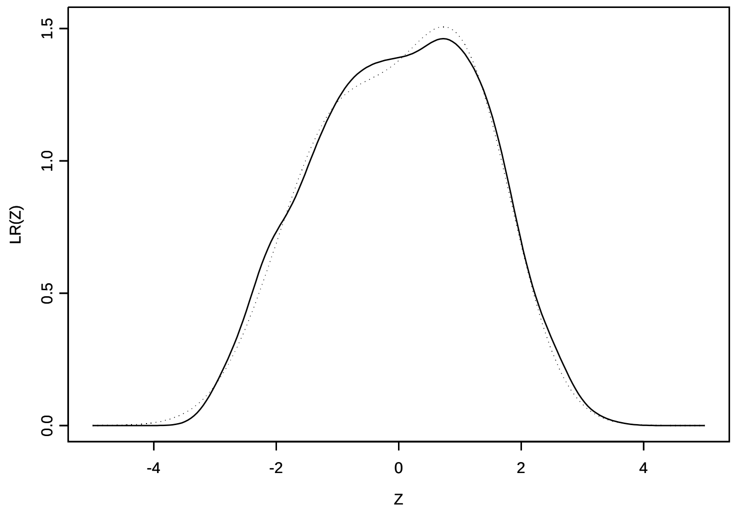


Figure 3: *Estimation of the LR function (Pan estimator (dotted line) and over-smoothed kernel estimator (solid line).*

To solve the equation (5), we use the approximation presented in (9) and the implementation procedure described in Section 3.3. Then we get the cut-off point $c = 0.00002$, yielding a rejection region of $\{Z : Z < -4.158 \text{ or } Z > 4.485\}$ for H_0. It gives 211 genes with significant expression changes comparing to the 187 obtained with the Normal mixture model of Pan [2002]. Note that the common rejection region between kernel and Normal approaches is $\{Z : Z < -4.887$ or $Z > 4.485\}$, and the common number of genes with significant expression changes is 178. With our approach, we obtain 33 differentially expressed genes not detected by Pan's approach; similarly 9 differentially expressed genes have been detected by the Normal mixture model but not with our nonparametric method.

As we pointed out in Section 3, the kernel estimation method is not very efficient in the distribution edges. It may be one of the reasons why a greater number of differentially expressed genes was detected by this nonparametric method compared to the Normal mixture model in Table 2. To improve our estimator, we use the reflection method described in Section 3.2. The percentage β varies between 0% and 0.5%. Results are summarized in table 2.

We obtained, for all the β values, a cut-off point close to 0.00002. The rejection region and the corresponding number of differentially expressed genes decrease as β increases. This phenomenom can be easily explained by the fact that the rejection techniques artificially inflate the tail of the distribution if β is too large. In all the cases, we studied there were some differentially expressed genes detected by our

β	Rejection region	Number of differentially expressed genes (number of differentially expressed genes in common with the Normal mixture model of Pan [2002])
0%	$\{Z : Z < -4.158 \text{ or } Z > 4.485\}$	211 (178)
0.10%	$\{Z : Z < -4.325 \text{ or } Z > 4.843\}$	147 (125)
0.20%	$\{Z : Z < -4.472 \text{ or } Z > 4.991\}$	108 (98)
0.25%	$\{Z : Z < -4.498 \text{ or } Z > 5.055\}$	102 (92)
0.3%	$\{Z : Z < -4.549 \text{ or } Z > 5.189\}$	85 (75)
0.4%	$\{Z : Z < -4.607 \text{ or } Z > 5.298\}$	76 (68)
0.5%	$\{Z : Z < -4.645 \text{ or } Z > 5.349\}$	71 (64)

Table 2: *Results obtained with reflection in extreme observations.*

kernel approach which were not found by the Normal mixture model of Pan [2002], and vice versa.

4.4 A simulation study

The aim of the simulation study is to validate our nonparametric computational approach to find the rejection region by solving the equation (5).

We consider the Normal mixture model defined in (10) and (11) as the "true" model for f_0 and f.

First, using our knowledge of f and f_0, we evaluate the "true" cut-off point and the corresponding "true" rejection region for H_0 by numerically solving (5) with $n = 7129$ (the sample size of our real data). We obtain $c = 0.00036$ and the rejection region $\{Z : Z < -4.876 \text{ or } Z > 4.395\}$, which are very close to those obtained by Pan [2002] with the bisection method.

Then, we generate $N = 100$ samples of size $n = 7129$ from this Normal mixture model. For each simulated sample, we estimate the cut-off point and the corresponding rejection region for H_0 by our method described in Section 3.3, using the Gaussian kernel and the choice of the bandwidths described in Section 4.3. For each simulated sample, the value of c is close to 0.00002 and the lower and upper bounds of the rejection region are also close to the "true" ones. Figure 4 shows the boxplots of these lower and upper bounds. The variations in the estimated bounds are due to the sampling fluctuations of the simulations, in particular those of the edges distribution. Nevertheless, in all cases our computational approach yielded essentially the same true rejection region.

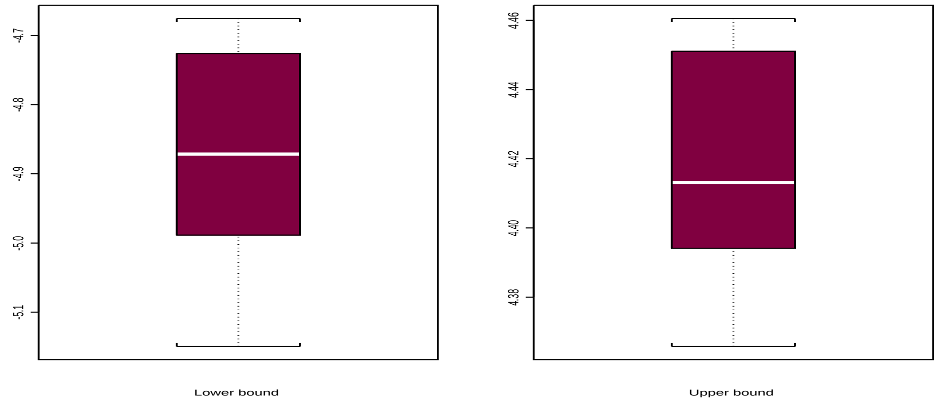


Figure 4: *Boxplots of the lower and upper bounds of the rejection region for H_0, for 100 simulated samples*

5 Discussion and concluding remarks

We have reviewed and extented methods for the analysis of microarray experiments. Following the principle of "letting the data speak about themselves", we have introduced a nonparametric kernel estimation into mixture models. Our method has three principal advantages.

1) An assumption of normality is not needed.

2) The estimation of the degrees of freedom in the existing t-test is avoided.

3) We need not use bootstrap to estimate the cut-off point and the corresponding rejection region.

4) The reflection method is proposed to overcome the edge effect of the kernel estimators.

For microarray data, small sample sizes are very common. Thus the asymptotic justification for the t-test is not applicable, and its validity depends on normality assumptions. Alternatives have been proposed in the literature. For example Baldi and Lang [2001], Dudoit et al. [2000], Kerr et al. [2000] and Thomas et al. [2001] proposed parametric or partiallu nonparametric methods. In this paper we have considered an alternative that is totally nonparametric. Furthermore, our simulation studies show that, if the true state of the nature is the Normal mixture, our methods yield the expected results. However, as in all kernel estimators, our approach is sensitive to distributional edge effects. We adapted the reflection method to study this problem and found a practical optimal solution to mimize the edge effects.

Acknowledgments.

The work was supported by the United States Public Service Grant No AG 16996 from the National Institutes of Health, and the Genome Research Council through the Collaborative Research Center (SFB 475) at the University of Dortmund (Germany). The authors thank Dr. Wei Pan from University of Minnesota for hepful discussion.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine, AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci., USA, 96, 6745-6750.

Baldi P, Long AD. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-Test and Statistical Inferences of Gene Changes. Bioinformatics, 17, 509-519.

Beran R. 1981. Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

Best DJ, Rayner JCW. 1987. Welsh approximate solution for the Behrens-Fisher problem. Technometrics, 29, 205-210.

Bosq D, Lecoutre JP. 1987. Théorie de l'estimation fonctionnelle. Economica.

Brown PO, Botstein D. 1999. Exploring the New World of the genome with DNA microarrays. Nature Genetics, 21, 33-37.

Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. 2000. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Genome Research, 10, 2022-2029.

Chen Y, Dougherty ER, Bittner M. 1997. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. Biomedical Optics, 2, 364-374.

Chilingaryan A, Gevorgyan N, Vardanyan A, Jones D, Szabo A. 2002. Multivariate approach for selecting sets of differentially expressed genes. Mathematical Biosciences, 176, 59-69.

Cline DBH, Hart JD. 1991. Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics*, 22, 69-84.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihoud estimation from incomplete data, via the EM algorithm (with discussion). J .R. Statist Soc. B, 39, 1-38.

Deheuvels P. 1977. Estimation non parametrique de la densité par histogramme generalise. Rev. Stat. Apppliquée. 35F42.

Dudoit S, Yang YH, Speed TP, Callow MJ. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, 111–139.

Efron B, Tibshirani RJ. 1993. An introduction to the Bootsrap. Chapman & Hall: London.

Efron B, Tibshirani R, Goss V, Chu G. 2000. Microarrays and Their Use in a Comparative . Technical report, Stanford University.

Efron B, Storey J, Tibshirani R. 2001. Microarrays, Empirical Bayes Methods, and False Discovery Rates. Technical report, Univ. California, Berkeley.

Friddle C, Koga T, Rubin E, Bristow J. 2000. Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. Proc. Nat. Acad. Sci., USA, 97, 6745-50.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology, 7, 601–620.

Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. 1999. Ploidy regulation of gene expression. Science, 285, 251-254.

Golub TR , Slonim DK, Tamayo P, Huard C , Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531-537.

Grant G, Manduch, E, Stoeckert C. 2002. Using non-parametric methods in the context of multiple testing to identify differentially expressed genes. Methods of microarray data analysis. Editors S.M. Lin and K.F. Johnson, Kluwer Academic Publishers, 37-55.

Guimaraes G, Urfer W. 2002. Self-organizing maps and its applications in sleep apnea research and molecular genetics. In: O.Opitz and M. Schwaiger (eds) Exploratory data analysis in empirical research. Studies in classification, data analysis and knowledge organization. Springer Verlag, Heidelberg, 332-345.

Hall P, Wehrly TE. 1991. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86, 665-672.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Wilfond B, Sauter G, Kallioniemi OP, Borg Å, Jeffrey T. 2001. Gene Expression Profiles in Hereditary Breast Cancer. New England Journal of Medicine, 244, 539-548.

Imoto S, Goto T, Miyano S. 2002. Estimation of Genetic Networks and Functional Structures between Genes by using Bayesian Network and Nonparametric Regression. Proc. Pacific Symposium on Biocomputing, 7, 175-186

Kerr MK, Martin M, Churchill GA. 2000. Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7, 819-837.

Kuruvilla FG, Park PJ, Schreiber SL. 2002. Vector algebra in the analysis of genome-wide expression data. Genome Biology, 3, 1-11.

Lander ES. 1999. Array of hope. Nature Genetics, 21, 3–4.

Lee MLT, Kuo FC, Whitmore GA, Sklar J. 2000. Importance of microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations", Proc. Nat. Acad. Sci., USA, 97, 18, 9834-9839.

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, PNAS. 98, 31-36.

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology, 8, 37-52.

Pan W. 2002. A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments". Bioinformatics, 12, 546-554.

Pan W, Lin J, Le CT. 2002. A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data. GenomeBiology (To appear).

Press WH, Teukolsky CM, Vetterling WT, Flannery BP. 1992. Numerical recepes in C, The Art of Scientific Computing. 2nd ed. Combridge: New York.

Quackenbush J. 2001. Computational analysis of microarray data. Nat. Rev. Genet, 2, 418-27.

Scheffé H. 1970. Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Sociaty, 65, 1501-1508.

Schuster EF. 1985. Incorporating support constraints into nonparametric estimation of densities. *Communications in Statistics, Theory and Methods*, 14, 1123-1136.

Sebastiani P, Ramoni M. 2002. Statistical Challenges in Functional Genomics. Technical report, Massachustts University.

Shaffer JP. 1995. Multiple hypothesis testing. Annu. Rev. Psychol., 46, 561-584.

Silverman BW. 1986. Density estimation for statistics and data analysis. London : Chapman & Hall.

Smyth GK, Yang YH, Speed TP. 2002. Statistical issues in microarray data analysis. In: Functional Genomics, Methods and Protocols, M. J. Brownstein and A. B. Khodursky (eds.), Methods in Molecular Biology series, Humana Press, Totowa, NJ.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9, 3273-3297.

Thomas JG, Olson JM, Tapscott SJ, Zhao LP. 2001. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res, 11, 1227-1236.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci, U S A, 98, 5116-5121.

Westfall PH, Young SS. 1993. Resampling-based multiple testing: Examples and Methods for p-value adjustement. Wiley series in probability and mathematical statistics. Wiley, 1993.

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. 1997. Gene Expression Profiles in Normal and Cancer Cells. Science, 276, 1268-1272.