Journal of Statistical Computation & Simulation
Vol. 00, No. 00, Month 200x, 1–10

RESEARCH ARTICLE

Searching for differential expression: a non parametric approach

I. Ortega-Serrano[a], M.C. Ruiz de Villa [b*] and A. Miñarro[b]

[a]*Institut Recerca Hospital Universitari Vall Hebron , Barcelona, Spain*; [b]*Department of Statistics. University of Barcelona, Barcelona, Spain*

(*released January 2011*)

Microarray experiments are being widely used in medical and biological research. The main features of these studies are the large number of variables (genes) involved and the low number of replicates (arrays). It seems clear that the most appropriate models, when looking for detecting differences in gene expression are those who exploit the most useful information to compensate for the lack of replicates. On the other hand, the control of the error in the decision process plays an important role for the high number of simultaneous statistical tests (one for each gene), so that concepts such as the False Discovery Rate (FDR [3]) take a special importance.

One of the alternatives for the analysis of the data in these experiments is based on the calculation of statistics derived from modifications of the classical methods used in this type of problems (moderated-t, B-statistic [11]). Nonparametric techniques have been also proposed ([5], [6]), allowing the analysis without assuming any prior condition about the distribution of the data, which make them especially suitable in such situations. This paper presents a new method to detect differentially expressed genes based on nonparametric density estimation by a class of functions that allow us to define a distance between individuals in the sample (characterized by the coordinates of the individual (gene) in the dual space tangent to the manifold of parameters) [2]. From these distances, we designed the test to determine the rejection region based on the control of FDR.

Keywords: Microarrays; Differential expression; Non-parametric density estimation

AMS Subject Classification: 62-07;62C12;62F03;62F40;62G07

1. Introduction

Since DNA microarrays appeared in the late 90s they have become a powerful tool in molecular biology research. They provide a way to obtain expression measurements of thousands of genes at the same time, giving a picture of the interactions among genes in biological processes. Nowadays it is already assumed the need of replicating microarrays in each experimental condition in order to achieve reliability in the conclusions.

An important goal in microarrays studies is to detect differentially expressed genes between two or more different experimental conditions. Since distributional assumptions in the expression levels may be sometimes violated, several nonparametric methods have been proposed to deal with that issue (e.g. [5], [4], [6]).

Among the nonparametric methods, several works studying the use of mixture models to estimate densities of the data have been published ([6], [7], [9], [8]). These

[*] Corresponding author. Email: mruiz_de_villa@ub.edu

ISSN: 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/0094965YYxxxxxxxx
http://www.informaworld.com

methods have the advantage of using information from the large amount of genes for inferential purposes.

As pointed out in [10], a mixture model can be thought as a smooth of a kernel estimate, providing a powerful density estimation with the advantage over the kernel method that it gives more stable estimates of tail probabilities, which play a critical role to determine a rejection region in a statistical test [6]. In the other hand, this smoothness could lead to a loss of accuracy in the estimate. The aim of this paper is to propose an alternative method for selecting differentially expressed genes based on orthogonal series to perform a density estimation step, as well as compare it with the MMM method presented in [6], which is based on normal mixtures.

2. Methods

2.1. *Preliminars*

Let $X_{1i},...,X_{ni},Y_{1i},...,Y_{mi}$ be the expression levels from n and m microarrays under two different experimental conditions for the gene $i = 1,...,N$ (after any possible normalization and (log)transformation of the original array signals).The goal is to find which genes modify their expression among the different groups.

If we only have two groups we can calculate a t-type score as the test statistic:

$$Z_i = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{s_{(1),i}/n + s_{(2),i}/m + a_0}}, \tag{1}$$

with $s_{(1),i}, s_{(2),i}$ being the sample variances for $X_{1i},...,X_{ni}$ and $Y_{1i},...,Y_{mi}$ respectively, and a_0 the 90-th percentile of $\left\{s_{(1),i}/n + s_{(2),i}/m\right\}$, proposed by ([5], [6])to stabilize the variances of the Z_i avoiding its dependency from the specific gene.

When searching for differential expression among different conditions it arises in a natural way a two-component mixture model, as has been considered in [5], [6], [9] or [12]. Let Z_i be the test statistic for the gene i , and let f be the density function of Z_i. Consider G_0 and G_1 the set of genes non-differentially and differentially expressed, respectively. Suppose that f_0 is the density for Z_i, when the gene i belongs to G_0 and f_1 when it belongs to G_1. Thus, the density of Z_i overall the genes can be written as:

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{2}$$

where p_0 and p_1 are the proportion of genes with no change in their expression and those changing, respectively.

As microarray data is usually composed by a high number of expression genes (say N), the values of $Z_i, i = 1,...,N$ can be used to obtain an accurate estimate of f and f_0 using non-parametric techniques. Several methods have been proposed using normal mixture models to fit the estimates of f and f_0 in order to construct methods to detect differential expression ([6], [9]).

In the next section we explain the details about fitting normal mixture models.

2.2. *Density estimation using normal mixtures*

Density estimation of $f(z)$ using gaussian mixtures assumes that

$$f(z;\Psi_g) = \sum_{i=1}^{g} \pi_i \phi(z;\mu_i,\sigma_i^2),$$

where g is the number of components in the mixture, $\phi(z;\mu_i,\sigma_i^2)$ is the gaussian density function with mean μ_i and variance σ_i^2, π_i is the mixing proportion of the component i, and $\Psi_g = \{(\pi_i,\mu_i,\sigma_i^2), i = 1,\dots,g\}$ denotes the set of all the parameters in the model.

In order to fit the model, following [10], the EM-algorithm has been used to obtain the maximum likelihood estimate $\hat{\Psi}_g$. Given N observations $z_1,\dots,z_N$, the EM-algorithm gives $\hat{\Psi}_g$ as follows:

- If $\Psi_g^{(k)} = \{(\pi_i^{(k)},\mu_i^{(k)},\sigma_i^{2(k)}), i = 1,\dots,g\}$ are the parameter estimates at step k, the new estimations at step $k+1$ are:

$$\begin{aligned}
\pi_i^{(k+1)} &= \frac{1}{N}\sum_{j=1}^{N} \tau_{ij}^{(k)} \\
\mu_i^{(k+1)} &= \sum_{j=1}^{N} \tau_{ij}^{(k)} z_j / \sum_{j=1}^{N} \tau_{ij}^{(k)} \\
\sigma_i^{2(k+1)} &= \sum_{j=1}^{N} \tau_{ij}^{(k)} (z_j - \mu_i^{(k+1)})^2 / \sum_{j=1}^{N} \tau_{ij}^{(k)}
\end{aligned}$$

  where

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)}\phi(z_j;\mu_i^{(k)},\sigma_i^{2(k)})}{f(z_j;{\Psi_g}^{(k)})}$$

- Iterating the previous step while $\left\|{\Psi_g}^{(k+1)} - {\Psi_g}^{(k)}\right\| \geq c$, for a fixed cutoff c, $\hat{\Psi}_g$ is obtained.

The Bayesian Information Criterion (BIC) has been used to select the number of components in the normal mixture [6],

$$BIC = -2logL(\hat{\Psi}_g) + \nu_g log(N), \qquad (3)$$

where ν_g is the number of free parameters in Ψ_g, and $logL$ indicates the log-likelihood function. In order to decide the number of components g in the mixture model, different values of g has been used to fit the model, and then that g corresponding to the first local minimum for the BIC has been selected.

2.3. *Density estimation using orthogonal functions*

In this section we propose an alternative density estimation approach, based on orthogonal series estimators ([13]).

If f is a density then $\sqrt{f} \in \mathcal{L}^2(\mathbb{R})$, so $\sqrt{f}$ belongs to the set of square-integrable real functions. Following the classical functional analysis theory, as $\mathcal{L}^2(\mathbb{R})$ is a Hilbert space, f can be expressed as:

$$f(z;\Theta) = \left[\sum_{i=0}^{\infty} \theta_i h_i(z)\right]^2,$$

where $\{h_i(z)\}_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{L}^2(\mathbb{R})$, Θ denotes the set of all unknown parameters, $\Theta = \{\theta_i, i = 1,\ldots,\infty\}$, and θ_i are the Fourier coefficients of $\sqrt{f}$ in this basis, and $\sum_{i=0}^{\infty}\theta_i^2 = 1$.

An orthonormal basis of $\mathcal{L}^2(\mathbb{R})$ is a set $\{h_i(z)\}_{i\in\mathbb{N}}$ verifying:

$$\int_{\mathbb{R}} h_i(z)\cdot h_j(z)dz = \begin{cases} 1 \; \mathit{if} \quad \text{i=j} \\ 0 \; \mathit{otherwise} \end{cases}.$$

As functions in $\mathcal{L}^2(\mathbb{R})$ are real-valued, a natural basis for that space are the family of normalized Hermite polynomials, defined as:

$$h_i(z) = \frac{1}{(i!2^i\sqrt{\pi})^{1/2}} e^{-z^2/2} H_i(z),$$

$$\text{where} \quad H_i(z) = (-1)^i e^{z^2} \frac{d^i}{dz^i} e^{-z^2}.$$

To estimate f we take:

$$\hat{f}(z;\hat{\Theta}) = \left[\sum_{i=0}^{N} \hat{\theta}_i h_i(z)\right]^2, \tag{4}$$

with $\sum_{i=0}^{N}\hat{\theta}_i^{\,2} = 1$.

The estimation of θ_i's presents some problems [13], in particular , the difficulties appear when N tends to be large. In order to overcome this handicap [1] proposed an estimation algorithm that we have slightly modified in order to improve the execution time.

Essentially, the algorithm presented is a forward procedure that begins taking the single normalized Hermite polynomial $h_{i_1}(z)$ that has the highest likelihood based on the sample data. At step k, the strategy is to add the polynomial $h_{i_k}(z)$ that maximizes the likelihood. Statistical significance is checked by means of a likelihood ratio test. This algorithm is here detailed:

(1) Transform the data, z to z^*, to have mean 0 and sample variance 1/2.
(2) Let's define $\hat{l}_{i_0,\ldots,i_k} = \sup_{\{\theta_{i_0},\ldots,\theta_{i_k} \mid \sum\theta_{i_r}=1\}} L(\theta_{i_0},\ldots,\theta_{i_k})$ where $L(\theta_{i_0},\ldots,\theta_{i_k})$ is the likelihood function, and $i_0,\ldots,i_k \in \mathbb{N}$.
(3) Select $i_0 \in I = \{0,\ldots,K\}$ such that $\hat{l}_{i_0} \geq \hat{l}_i, \forall i \in I \setminus \{i_0\}$, where K is the number of Hermite functions to take into account.
(4) (i) For $k = 1$ to K, being fixed $i_0,\ldots,i_{k-1}$, we choose $i_k \in I \setminus \{i_0,\ldots,i_{k-1}\}$ such that $\hat{l}_{i0,\ldots,i_k} \geq \hat{l}_{i_0,\ldots,i_{k-1},i}, \forall i \in I \setminus \{i_0,\ldots,i_{k-1}\}$
    (ii) A hypothesis test is then performed to detect significant differences in the likelihood, due to the adding of a new term in the estimation of

f. We use the statistic:

$$\Lambda = \frac{\hat{l}_{i_0,\ldots,i_{k-1}}}{\hat{l}_{i_0,\ldots,i_{k-1},i_k}}$$

(5) Wilks theorem ensures convergence of $-2ln\Lambda$ to a χ^2 distribution, under the null hypothesis of non significant differences between their likelihoods. Rejecting the test implies to return to **(i)**. Otherwise, the estimation of f is:

$$\hat{f}(z;\hat{\Theta}) = \frac{1}{b}\left[\sum_{r=0}^{k} \theta_{i_r} h_{i_r}(z^*)\right]^2$$

where $i_r \in \{0,\ldots,N\}$, $r = 0,\ldots,k$

It is interesting to remark that in the case when only one component is chosen by the algorithm the density function estimated is a normal with sample mean and sample variance parameters.

2.4. *Searching for differential expression*

2.4.1. STP Method

One of the advantages of using the previous density estimation method is that it allows for the use of a distance that can be used to search for differential expression.

Let $\hat{f}$ be a density function estimated from (4), for $k > 1$. Following Miñarro and Oller [2], a distance between the points in the sampling space Θ can be defined using the coordinates of the subjects in the dual space tangent to Θ, where $\Theta \subset \mathbb{R}^{k-1}$ is open and $\theta = (\theta_1, \cdots, \theta_{k-1}) \in \Theta$, by:

$$d^2(x_1,x_2) = \cdots = (\partial_\theta log\hat{f}(x_1|\theta) - \partial_\theta log\hat{f}(x_2|\theta))^t G^{-1}(\theta)(\partial_\theta log\hat{f}(x_1|\theta) - \partial_\theta log\hat{f}(x_2|\theta)) \quad (5)$$

where $\partial_\theta logf(x|\theta) = \left(\frac{\partial logf(x|\theta)}{\partial\theta_1}, \cdots, \frac{\partial logf(x|\theta)}{\partial\theta_{k-1}}\right)$, and $G^{-1}(\theta) = \frac{1}{4}(I_{k-1} - \theta\theta^t)$ is the inverse of the Fisher information matrix, with I_{k-1} the identity matrix.

The decision rule to select a gene i as being expressed will be that Z_i is, in some way, far away from f_0.

Based on the previous distance we propose the following steps to select differentially expressed genes:

(1) For each gene i compute the score Z_i following (1), $i = 1,\ldots,N$.
(2) For each gene i compute a set of B z_i's using (1) but randomly permuting the sample labels. Here B is the number of permutations.
(3) Estimate the null distribution f_0 using the set of $B \cdot N$ z_i's scores generated from all the genes, as explained in section **2.3**.
(4) Compute the test statistic $dZ_i = d(Z_i, 0)$, and $dz_i = d(z_i, 0)$ as in (5), taking $\hat{f}_0$ as the reference distribution.
(5) We decide that the gene i is expressed if $dZ_i >= c$, for some c.
(6) For this c, we can estimate the False Discovery Rate (FDR) using:

$$F\hat{D}R(c) = \frac{\frac{1}{B}\#dz_i \geq c}{\#dZ_i \geq c}$$

The correct choose of c will give a value of the FDR under some previously set value.

2.4.2. MMM Method

Based on [5], [6] suggests the LR statistic:

$$LR(Z) = \frac{f_0(Z)}{f(Z)} \tag{6}$$

And two alternatives to set the cut-off for gene selection:

- Directly (*lr*): A large value of $LR(Z)$ gives no evidence against H_0, thus a too small value leads to rejecting H_0. The rejection region RR(α) is obtained as: $\alpha = \int_{LR(Z)<s} f_0(Z)dz$
- Using only the f_0 distribution (*tail*): Choosing the rejection region as the two tails of f_0. The rejection region RR(α) is obtained as: $\alpha = \int_{|Z|>t} f_0(Z)dz$

3. Results

3.1. *Simulation results*

In this section we use simulations to assess the performance of the STP method in comparison to the MMM approach using the EM-algorithm in both *tail* and *lr* alternatives to set the cut-off for gene selection. The distribution of all Z statistics has been modeled through

$$f(z) = (1-p)f_0(z) + pf_1(z), \tag{7}$$

where $f_0(z)$ stands for the standard normal density and

$$f_1(z) = 0.5f_{m-}(z) + 0.5f_{m+}, \tag{8}$$

where f_{m+} and f_{m-} denote the normal densities of mean m and $-m$ respectively and with a standard deviation of 1. So we are supposing that the distribution of f is symmetrical around 0 with heavy tails for the f_1 distribution.

Those Z generated under f_1 represent genes differentially expressed among the conditions, whereas Z generated from f_0 are genes with no changes in their expression.

We have considered several values for m and p, in particular $m = 1.5, 2, 2.5$ and p ranging from 0.1 to 0.5. For each combination of parameters 1000 samples were generated, each one consisting of N=5000 values (genes) of Z and for each sample all three estimations: STP, MMM-tail and MMM-lr have been computed.

Once differential expression has been accepted or rejected for each sample following Section 2, we have assess the performance of each estimation method by computing the false discovery rate (FDR) and the power. Results are shown in Figure 1 and Figure 2 respectively.

From the results we conclude that STP and MMM-lr show a similar behavior in terms of false discovery rate and power. MMM-tail shows a quite different result with a higher FDR and also a higher power under all situations. As expected, increasing the distance between f_0 and f_1 (that is increasing the absolute value of m) causes a decrease in the FDR and an increase of the power in all the simulations.

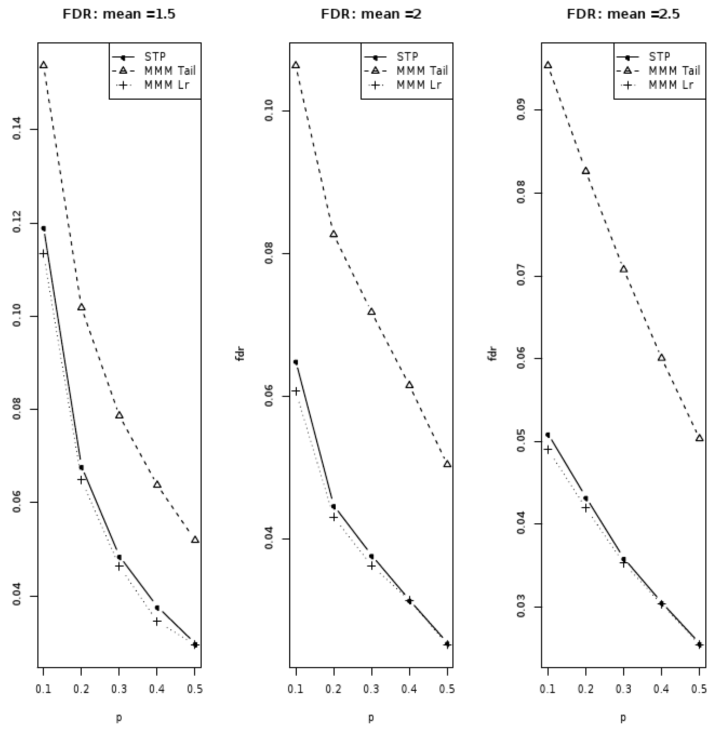

Figure 1. FDR results for the three estimations.

It is also interesting to note that as the percentage of differentially expressed genes (p) increases, FDR decreases and power increases in all three methods. The results show that MMM-tail method do not achieve good results in the control of FDR in all situations studied, and if we have to choose between STP and MMM-lr our preference goes to STP since in the process of estimation it is not necessary to fix a determined number of mixture terms and also the selection of the final model is based on a likelihood test rather than a BIC criterion. STP method also offers the possibility of defining a natural distance between objects (genes) that we consider is a real advantage in front of other methods.

3.2. *Real data*

We have used the proposed method with data from an experiment involving the study of a treatment applied to mice. Nine Affymetrix chips were made from treated mice and nine from control samples. After appropriate preprocessing, normalization and filtering of data, we worked with approximately 2500 genes. For each of these genes Z_i is estimated according to the formula (1). From $B = 100$ permutations a set of B z_i 's was obtained for each gene to estimate $\hat{f}_0(x)$. Once obtained the estimation we computed the distances dZ_i and dz_i. Figure 3 shows a diagram of the method. Table 1 shows the estimation of f_0 and in Figure 4 we show the boxplots of the obtained distances.

As the method allows the calculation of a distance between two points of the sample space, we have calculated the matrix of distances and we have obtained a two dimensional representation by classical multidimensional scaling. The result is shown in Figure 5. In red are shown the points corresponding to the 20 genes with greater differential expression according to the method proposed. The first component explains 99 % of the total variability.

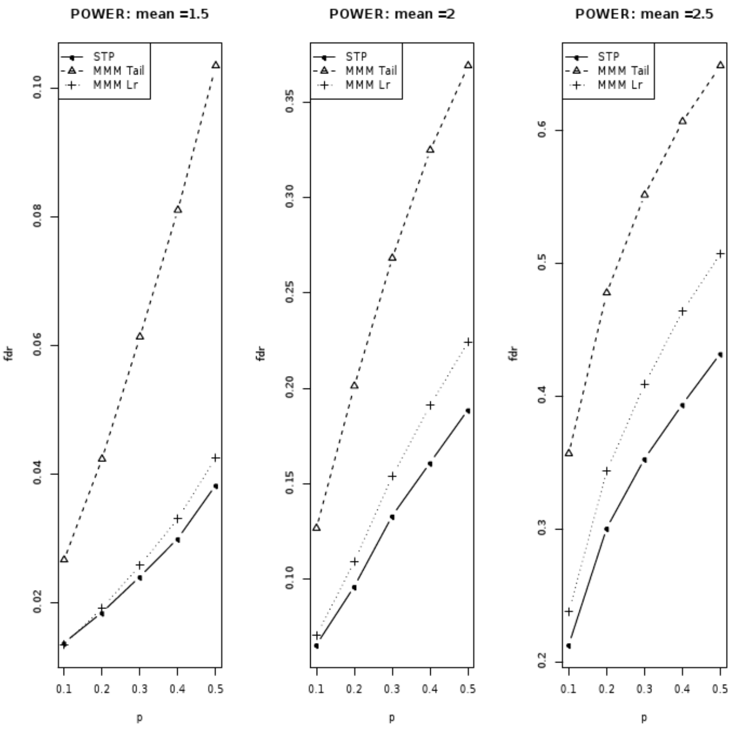

Figure 2. Power results for the three estimations.

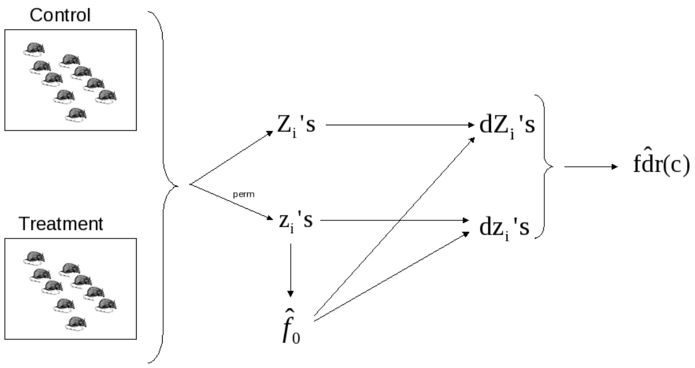

Figure 3. Flowchart of the process.

4. Conclusions

- MMM-lr and STP have a similar behavior in terms of FDR and POWER.
- MMM-tail performs better in POWER but not in FDR.
- As the percentage of differentially expressed genes increases FDR decreases and POWER increases in all three methods.
- MMM-lr could be a good approach since it uses f and f_0 but the problem is that its behavior is very unstable, giving computing problems in many simulations.

Table 1. Coefficients of $\hat{f}_0$ in $\hat{f}_0 = \left[\sum_{i=0}^{10} \theta_i h_i(x)\right]^2$.

i	θ_i
0	0.99937
1	0
2	-0.00499
3	0.01005
4	0.03064
5	0.00973
6	0.00814
7	0.00248
8	0.00470
9	0.00251
10	0

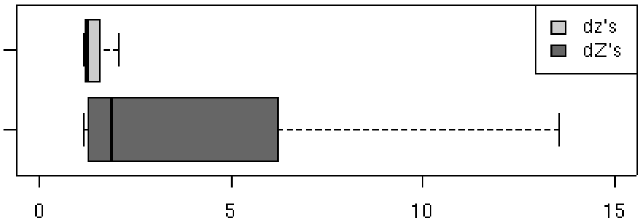

Figure 4. Boxplots of distances.

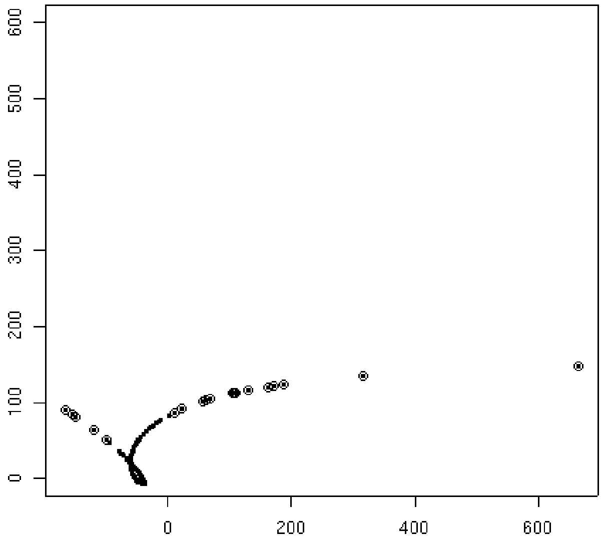

Figure 5. Multidimensional scaling from the distance matrix between the Z_i's. Circled points indicate the points corresponding to differentially expressed genes.

- STP estimation does not need to set a determined number of terms of the model, the selection of the final model is based on a likelihood test.
- STP method offers the possibility of defining a natural distance between objects (genes) that we consider is a real advantage in front of other methods.
- The possibility of calculating a distance allows the selection of differentially expressed genes based on it and a graphical representation of the genes in a

reduced dimension.

References

[1] Miñarro, A., Oller, J.M. *On a class of probability density functions and their information metric.* Sankhya, SERIES A, 55(2)(1993),pp. 214–225.

[2] Miñarro, A., Oller, J.M. . Some remarks on the individuals-score distance and its applications to statistical inference. Qüestiió, 16(1992), pp. 43–57.

[3] Benjamini, Y., Hochberg, Y. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser B. 57(1995), pp. 289–300

[4] Tusher, V., Tibshirani, R., Chu G. *Significance analysis of microarrays applied to the ionizing radiaton response.* Proc. Natl. Acad. Sci. 98(2001), pp. 5116–5121

[5] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V. *Empirical Bayes Analysis of a Microarray Experiment.* Journal of the American Statistical Association 96(2001), pp. 1151–1160

[6] Pan, W., Lin, J., Le, C.T. *A mixture model approach to detecting differentially expressed genes with microarray data.* Functional and Integrative Genomics 3(2003),pp. 117–124

[7] Najarian,K., Zaheri,M., Rad, A., Najarian, S., Dargahi, J. *A novel Mixture Model Method for identification of differentially expressed genes from DNA microarray data.* BMC Bioinformatics 5(2004), pp. 201.

[8] Jiao, S., Zhang, S. *The t-mixture model approach for detecting differentially expressed genes in microarrays.* Functional and Integrative Genomics, 8(2008), pp. 181–186.

[9] McLachlan, G.J., Bean,R.W., Ben-Tovim Jones, L. *A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays* Bioinformatics 22,13(2006), pp. 1608–1615

[10] McLachlan, G.J., Peel, D., and Bean, R.W. *Modelling high-dimensional data by mixtures of factor analyzers.* Computational Statistics and Data Analysis 41(2003), pp. 379–388.

[11] SmythG, K. *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Statistical Applications in Genetics and Molecular Biology, 3(2004) Article 3.

[12] Rossell,D., Guerra, R., Scott, C. *Semi-Parametric Differential Expression Analysis via Partial Mixture Estimation* The Berkeley Electronic Press 7(2008), Issue 1.

[13] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, England, 1986, 23–25.