UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 10-44
Statistics and Econometrics Series 27
November 2010

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

MULTIPLE HYPOTHESIS TESTING AND CUSTERING WITH MIXTURES OF NON-CENTRAL T-DISTRIBUTIONS APPLIED IN MICROARRAY DATA ANALYSIS

J. Miguel Marín[1] and M. Teresa Rodríguez-Bernal[2]

Abstract

Multiple testing analysis, based on clustering methodologies, is usually applied in Microarray Data Analysis for comparisons between pair of groups. In this paper, we generalize this methodology to deal with multiple comparisons among more than two groups obtained from microarray expressions of genes. Assuming normal data, we define a statistic which depends on sample means and sample variances, distributed as a non-central t-distribution. As we consider multiple comparisons among groups, a mixture of non-central t-distributions is derived. The estimation of the components of mixtures is obtained via a Bayesian approach, and the model is applied in a multiple comparison problem from a microarray experiment obtained from *gorilla*, *bonobo* and *human* cultured fibroblasts.

Keywords: Clustering, MCMC computation, Microarray analysis, Mixture distributions, Multiple hypothesis testing, non-central t-distribution.

[1] Dep. de Estadística, U. Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: jmmarin@est-econ.uc3m.es
[2] Dep. de Estadística e IO, U. Complutense de Madrid, Plaza de Ciencias, 3 Ciudad Universitaria 28040 (Madrid), e-mail: mayter@mat.ucm.es

Multiple hypothesis testing and clustering with mixtures of non-central t-distributions applied in microarray data analysis

J.M. Marín
Dep. Estadística
U. Carlos III, 28903 Getafe (Madrid), Spain

M.T. Rodríguez-Bernal
Dep. Estadística e I.O.
Fac. Matemáticas, U. Complutense, 28040 Madrid, Spain

Abstract

Multiple testing analysis, based on clustering methodologies, is usually applied in Microarray Data Analysis for comparisons between pair of groups. In this paper, we generalize this methodology to deal with multiple comparisons among more than two groups obtained from microarray expressions of genes. Assuming normal data, we define a statistic which depends on sample means and sample variances, distributed as a non-central t-distribution. As we consider multiple comparisons among groups, a mixture of non-central t-distributions is derived. The estimation of the components of mixtures is obtained via a Bayesian approach, and the model is applied in a multiple comparison problem from a microarray experiment obtained from gorilla, bonobo and human cultured fibroblasts.

Keywords: Clustering; MCMC computation; Microarray analysis; Mixture distributions; Multiple hypothesis testing; non-central t-distribution.
AMS 2000 subject classifications. Primary 62-02; secondary 62E10, 62F15.

1 Introduction

Nowadays, in Bioinformatics the analysis of expression data from microarray technologies is one of the main tasks. Specifically, in Genomics, the analysis of gene expression from microarrays is undertaken with two main techniques: multiple hypothesis testing and cluster analysis (e.g. Medvedovic and Sivaganesan (2002), Dudoit et al. (2003) and McLachlan et al. (2002)). Clustering techniques are widely used in Bioinformatics because genes usually present high correlations and they can be grouped together; this correlation may reflect underlying biological factors of interest, such as regulation by common transcription factors. On other hand, multiple testing methods aim to detect differences in the expressions of genes under different treatment conditions.

Although both techniques has been treated independently, recently they have been joined in a common framework (see Yuan and Kendziorski (2006), Dahl and Newton (2007), and Dahl et al. (2008)). They consider a mixture of both techniques where the multiple testing procedure are highly improved taking into account clusters that share similar parameter values.

In this paper, we consider a methodology to analyse differences among mean expressions of genes in microarray data, under different treatment conditions. We will follow the mixed approach of multiple testing and cluster analysis. Usually, multiple testing procedures only deal with comparison between pairs of groups, and with a cluster approach not only the process of multiple comparisons is improved but many groups can be compared at the same time.

In the gene expression context, after normalization and data cleaning, it can be assumed that expressions are normally distributed. But as we deal with different sample sizes and sample estimates, it is convenient to define statistics which take account of them. Therefore, we will consider a statistic distributed with a non-central t-distribution (see Johnson et al. (1995)).

We take groups of statistics, distributed as non-central t-distributions, related for each treatment condition and possibly similar expression profiles among different genes. In this way, we can take a mixture of non-central t-distributions approach, to tackle with different groups of expressions along the different treatment conditions where gene expressions are considered.

As the non-central t-distribution has cumbersome expressions for its density function (see Johnson et al. (1995)), we take an alternative parametrization in terms of scale mixtures of normal distributions (see Tsionas (2002)). In this way, the model can be seen as a discrete mixture of scale mixtures of normal distributions and the estimation process may be cumbersome. Nevertheless a Bayesian approach presents a good performance with not huge amount of data.

We have used, as software to program the algorithms, `Jags` (see Plummer (2003)). It seems to be adapted to mixture modelling and it permits sorting the parameters for avoiding problems about identifiability of the components of the mixtures.

The paper is organized as follows. In section 2 we describe the theoretical model. In section 3 we show the prior distributions and compute the posterior distributions of the parameters of the model. In section 4 we consider first some simulated data to check the procedure, and then an analysis of a microarray data set described in Karaman et al. (2003) with three related species (gorillas, bonobos and humans). Finally, in section 5, we present some conclusions and hints about the methodology and results shown in the paper.

2 Model for the distribution of microarray expressions

We consider a known set of g genes ($g = 1, \ldots, G$) under a treatment conditions whose expressions are measured by microarray techniques. The respective expressions are modelized by $g \times a$ random variables: $X_1^g, \ldots, X_a^g$. Let us denote, $\left(x_{i1}^g, \ldots, x_{in_{ig}}^g\right)$ a sample from a random variable $X_i^g \sim N\left(\mu_{Xi}^g, \sigma^g\right)$, where $i = 1, \ldots, a$ treatment conditions and $g = 1, \ldots, G$ genes.

Each element g is the expression of a given gene that can be measured under a different conditions and we want to compare all groups of expressions by using a multiple tests procedure:

H_0^g : $\mu_{X_1}^g = \mu_{X_2}^g = \cdots = \mu_{X_a}^g$

H_1^g : at least one $\mu_{X_i}^g \neq \mu_{X_j}^g$

for $g = 1, \ldots, G$

For each gene (g) we define the following statistics

$$T_{X_i}^g = \frac{\overline{x}_{ig}}{S_{x_i}^g / \sqrt{n_{ig}}},$$

for $i = 1, \ldots, a$ treatment conditions. As $\left(x_{i1}^g, \ldots, x_{in_{ig}}^g\right)$ is a random sample from a $N\left(\mu_{X_i}^g, \sigma^g\right)$, then $T_{X_i}^g$ is distributed as a *non-central* t-distribution for $i = 1, \ldots, a$ (see chap. 31 of Johnson et al. (1995)). We will denote

$$T_{X_i}^g \sim t_{n_{ig}-1}(\delta_{ig}),$$

where $(n_{ig} - 1)$ and $\delta_{ig} = \sqrt{n_{ig}}\mu_{X_i}^g / \sigma^g$ are the respective grades of freedom and centrality parameters of $T_{X_i}^g$.

Let us denote as T the vector of all possible $T_{X_i}^g$ statistics with dimension aG; if we consider a multiple test procedure, via a cluster approach, there may be k different groups, and the distribution of this vector can be modelized as a mixture of k non-central t-distributions,

$$f\left(t|\nu, \delta, \alpha\right) = \sum_{j=1}^{k} \alpha_j f\left(t|\nu_j, \delta_j\right)$$

where $\nu = (\nu_1, \ldots, \nu_k)$ are the grades of freedom, $(\delta_1, \ldots, \delta_k)$, are the centrality parameters, $\alpha = (\alpha_1, \ldots, \alpha_k)$ are the weights of the mixture and $f\left(t|\nu_j, \delta_j\right)$ are the density functions of the non-central t-distributions.

The null hypothesis H_0^g is not rejected when all groups in which a gene g is considered, belong to the same component of the mixture of non-central t-distributions.

2.1 Parametrization of the non-central t-distribution as a scale mixture of distributions

There are several ways to write the density function of the non-central t-distribution, but they usually involve complex expressions in terms of integrals, e.g. as

$$f(x) = \frac{\nu^{\frac{\nu}{2}\exp\left[-\frac{\nu\mu^2}{2(x^2+\nu)}\right]}}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)2^{\frac{\nu-1}{2}}\left(x^2+\nu\right)^{\frac{(\nu+1)}{2}}}\int_0^\infty y^\nu \exp\left[-\frac{1}{2}\left(y - \frac{\mu x}{\sqrt{x^2+\nu}}\right)^2\right]dy.$$

Expressions like this are fairly cumbersome to use in computational tasks. But in Tsionas (2002) the non-central t-distribution is showed to be a scale mixture of normal distributions, and it is introduced in a regression problem with a Bayesian approach.

In this work, we will use this parametrization in terms of scale mixture of normal distributions. Therefore, by definition, the non-central t-distribution is

$$T = \omega^{1/2}(Z + \delta) \sim t_\nu(\delta),$$

where $Z \sim N\left(0,1\right)$ and $\frac{\nu}{\omega} \sim \chi^2_\nu$ independently, and $\delta \in \mathbb{R}$.

Then, it follows, if we write previous expressions in terms of a hierarchical model that

$$\begin{aligned} X|\omega &\sim N\left(\omega^{1/2}\delta,\ \omega^{1/2}\right) \\ \frac{\nu}{\omega} &\sim \chi^2_\nu. \end{aligned}$$

Or, equivalently, the density function can be written as

$$f(x|\omega,\delta) = \frac{1}{\sqrt{2\pi}}\frac{\nu^{\frac{\nu}{2}}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty \exp\left(-\frac{1}{2\omega}\left((x-\omega^{1/2}\delta)^2+\nu\right)\right)\omega^{-\frac{\nu}{2}-\frac{3}{2}}d\omega.$$

The expression of a mixture of normal distributions in terms of a hierarchical framework permits to apply a MCMC simulation-based procedure under a Bayesian approach (see Choy and Smith (1997)).

Likelihood of the model

As we consider a discrete mixture of t-distributions, for $t = (t_1, \ldots, t_{aG})$, the complete likelihood is

$$L\left(\nu,\delta,\alpha|\mathbf{t}\right) = \prod_{i=1}^{aG} f\left(t_i|\nu,\delta,\alpha\right) = \prod_{i=1}^{aG}\sum_{j=1}^{k}\alpha_j f\left(t_i|\nu_j,\delta_j\right).$$

As this expression in practice is unwieldy, we can introduce (see e.g. Marin and Robert (2007)) index variables Z_i, for $i = 1, \ldots, aG$ that point out the element of the mixture which each t_i belongs to.

In this way,

$$P\left(Z_i = j|\alpha\right) = \alpha_j,$$

for $i = 1, \ldots, aG$ and $j = 1, \ldots, k$. Then, the conditional distribution of t_i for each fixed $Z_i = j$ is

$$f\left(t_i | Z_i = j, \nu, \delta\right) = f\left(t_i | Z_i = j, \nu_j, \delta_j\right),$$

where

$$\left(t_i | Z_i = j, \nu_j, \delta_j\right) \sim t_{\nu_j}\left(\delta_j\right),$$

and the joint distribution of t_i and Z_i is

$$f\left(t_i, Z_i | \nu, \delta, \alpha\right) = f\left(t_i | Z_i, \nu, \delta\right) \cdot P\left(Z_i | \alpha\right).$$

As we consider the expression of the non-central t-distribution as a mixture of normals distributions, the likelihood can be expressed as a three-stage hierarchical model,

$$\begin{array}{lll} t_i | Z_i, \nu, \delta, \omega \sim N\left(\omega_j^{1/2}\delta_j,\ \omega_j^{1/2}\right) & \text{for} & i = 1, \ldots, aG \quad j = 1, \ldots, k \\ \omega_j | \nu_j, \delta \sim IChiq(\nu_j) & \text{for} & j = 1, \ldots, k \\ P\left(Z_i = j\right) = \alpha_j & \text{for} & i = 1, \ldots, aG \quad j = 1, \ldots, k \end{array}$$

Or, equivalently,

$$L\left(\nu, \delta, \alpha, \omega | \mathbf{t}, Z\right) = \prod_{i=1}^{aG} f\left(t_i, \omega | Z_i, \nu, \delta\right) \cdot P\left(Z_i | \alpha\right) =$$

$$\prod_{\{i:Z_i=1\}} \alpha_1 f\left(t_i, \omega_1 | \nu_1, \delta_1\right) \cdot \cdots \cdot \prod_{\{i:Z_i=k\}} \alpha_k f\left(t_i, \omega_k | \nu_k, \delta_k\right) =$$

$$\prod_{j=1}^{k} \left\{ \left(\alpha_j \frac{1}{\sqrt{2\pi}} \frac{\nu_j^{\frac{\nu_j}{2}}}{2^{\nu_j/2}\Gamma\left(\frac{\nu_j}{2}\right)} \omega_j^{-\frac{\nu_j}{2}-\frac{3}{2}} \right)^{n_j} \prod_{\{i:Z_i=j\}} \exp\left(-\frac{1}{2\omega_j} \left((t_i - \omega_j^{1/2}\delta_j)^2 + \nu_j \right) \right) \right\}$$

where $n_j = \#\{i : Z_i = j\}$.

3 Posterior distributions of parameters

The posterior distribution of the parameters $(\nu, \delta, \alpha, \omega)$ of the mixture of non-central t-distributions is the product of the likelihood function by the prior distribution of the parameters,

$$\pi\left(\nu, \delta, \alpha, \omega | \mathbf{t}, Z\right) \propto L\left(\nu, \delta, \alpha, \omega | \mathbf{t}, Z\right) \cdot \pi(\nu) \cdot \pi(\delta) \cdot \pi(\alpha) \cdot \pi(\omega)$$

We consider the prior distributions for the parameters, introducing vague or diffuse information.

(i) For parameter α we consider a non-informative Dirichlet distribution: $\alpha \sim Dirichlet(1, \ldots, 1)$, namely, $\pi(\alpha) \propto 1$.

(ii) For parameters ν_j we consider truncated Poisson distributions $\nu_j \sim TruncPoisson(\lambda)$, where $\lambda > 1$, namely, $P(\nu_j = l) = \frac{1}{l!}\lambda^l e^{-\lambda} \cdot \frac{1}{(1-e^{-\lambda}-\lambda e^{-\lambda})}$, for $l = 2, \ldots$

(iii) For vector of parameters δ we assume $\pi(\delta) \propto \prod_{j=1}^{k} \pi(\delta_j)$, where each δ_j is distributed as truncated normal distributions in $(0, \infty)$, $\delta_j \sim TruncN(\mu, \sigma)$, namely, $\pi(\delta_j) = \frac{1}{1-\Phi\left(-\frac{\mu}{\sigma}\right)} \frac{1}{\sigma} \phi\left(\frac{\delta_j-\mu}{\sigma}\right)$, where ϕ is the density function and Φ the distribution function of a standard normal distribution.

(iv) $\omega_j \sim IG(\alpha_0, \beta_0)$, namely, $\pi(\omega_j) \propto \left(\frac{1}{\omega_j}\right)^{\alpha_0+1} \exp\left(-\frac{\beta_0}{\omega_j}\right)$.

Then the full posterior distribution can be written as

$$\pi\left(\nu, \delta, \alpha, \omega|\mathbf{t}, Z\right) \propto L\left(\nu, \delta, \alpha, \omega|\mathbf{t}, Z\right) \cdot \prod_{j=1}^{k} \left[\frac{1}{\nu_j!}\lambda^{\nu_j} \cdot \left(\frac{1}{\omega_j}\right)^{\alpha_0+1} \cdot \exp\left(-\frac{\beta_0}{\omega_j}\right) \cdot \phi\left(\frac{\delta_j - \mu}{\sigma}\right)\right] \propto$$

$$\propto \prod_{j=1}^{k} \left\{ \left(\alpha_j \frac{1}{\sqrt{2\pi}} \frac{\nu_j^{\frac{\nu_j}{2}}}{2^{\nu_j/2}\Gamma\left(\frac{\nu_j}{2}\right)}\right)^{n_j} \frac{1}{\nu_j!}\lambda^{\nu_j} \cdot \phi\left(\frac{\delta_j - \mu}{\sigma}\right) \cdot \omega_j^{\left(-\frac{\nu_j}{2}-\frac{3}{2}\right)n_j - \alpha_0 - 1} \cdot \right.$$

$$\left. \exp\left(-\frac{\beta_0}{\omega_j}\right) \cdot \prod_{\{i:Z_i=j\}} \exp\left(-\frac{1}{2\omega_j}\left((t_i - \omega_j^{1/2}\delta_j)^2 + \nu_j\right)\right) \right\}$$

The corresponding conditional posterior distributions of the parameters are,

(i) $\alpha|\nu, \delta, \omega, \mathbf{t}, Z \sim \mathit{Dirichlet}\left(n_1 + 1, \ldots, n_k + 1\right)$, namely,

$$\pi\left(\alpha|\nu, \delta, \omega, \mathbf{t}, Z\right) \propto \alpha_1^{n_1} \cdot \cdots \cdot \alpha_k^{n_k}$$

(ii)

$$\pi\left(\nu_j = l_j|\delta, \omega, \mathbf{t}, Z, \delta_j\right) \propto$$
$$\left(\frac{l_j^{\frac{l_j}{2}}}{2^{l_j/2}\Gamma\left(\frac{l_j}{2}\right)}\right)^{n_j} \frac{1}{l_j!}\lambda^{l_j} \cdot \omega_j^{\left(-\frac{l_j}{2}-\frac{3}{2}\right)n_j - \alpha_0 - 1}$$
$$\prod_{\{i:Z_i=j\}} \exp\left(-\frac{1}{2\omega_j}\left((t_i - \omega_j^{1/2}\delta_j)^2 + l_j\right)\right)$$

for $j = 1, \ldots, k$.

(iii)

$$\pi\left(\delta_j|\mathbf{t}, Z, \nu_j, \alpha, \omega\right) \propto \phi\left(\frac{\delta_j - \mu}{\sigma}\right) \cdot \prod_{\{i:Z_i=j\}} \exp\left(-\frac{1}{2\omega_j}\left((t_i - \omega_j^{1/2}\delta_j)^2 + \nu_j\right)\right)$$

for $j = 1, \ldots, k$.

(iv)

$$P\left(Z_i = j|\alpha, \nu, \delta, t_i, \omega\right) \propto$$
$$\left(\alpha_j \frac{\nu_j^{\frac{\nu_j}{2}}}{2^{\nu_j/2}\Gamma\left(\frac{\nu_j}{2}\right)}\right)^{n_j} \cdot \frac{1}{\nu_j!}\lambda^{\nu_j} \cdot \phi\left(\frac{\delta_j - \mu}{\sigma}\right) \cdot \omega_j^{\left(-\frac{\nu_j}{2}-\frac{3}{2}\right)n_j - \alpha_0 - 1} \cdot \exp\left(-\frac{\beta_0}{\omega_j}\right) \cdot$$
$$\exp\left(-\frac{1}{2\omega_j}\left((t_i - \omega_j^{1/2}\delta_j)^2 + \nu_j\right)\right)$$

for $i = 1, \ldots, aG$.

(v)

$$\pi\left(\omega_j|\mathbf{t}, Z, \nu_j, \alpha, \delta\right) \propto \omega_j^{\left(-\frac{\nu_j}{2}-\frac{3}{2}\right)n_j - \alpha_0 - 1} \cdot \exp\left(-\frac{\beta_0}{\omega_j}\right) \cdot \prod_{\{i:Z_i=j\}} \exp\left(-\frac{1}{2\omega_j}\left((t_i - \omega_j^{1/2}\delta_j)^2 + \nu_j\right)\right)$$

for $j = 1, \ldots, k$.

Classification rule In order to classify the observation t_i into a given component, we compute for all $j = 1, \ldots, k$,

$$j_i = \max_j \left\{ \#\{t : Z_i^{(t)} = j\} \right\}$$

and then we classify t_i in the corresponding component j_i where the maximum is attained.

In terms of the multiple hypothesis testing procedure, for $g = 1, \ldots, G$, the hypothesis H_0^g is rejected if at least one $T_{X_i}^g$ (where $i = 1, \ldots, a$ treatment conditions) is located in a different component of the mixture of t-distributions to the other $T_{X_j}^g$ (where $j \neq i$ and $j = 1, \ldots, a$).

4 Applications

We show an application of the previous theoretical results, by analysing the expressions from a microarray experiment of human, bonobo and gorilla cultured fibroblasts rendered by Karaman et al. (2003). As a preliminary step we test the procedure with simulated data. In all cases we have programmed the algorithms with `Jags` (see Plummer (2003)). One advantage of using `Jags` is that not only it constructs the full conditional distributions and it carries out the Gibbs sampling from the model specifications, but it allows to sort the parameters for avoiding problems about identifiability of the components of the mixtures. All codes are available from the authors, upon request.

4.1 Synthetic data

In order to check the procedure, we simulate expressions of 50 synthetic genes under three given conditions from normal distributions with different means $(10, 40, 100, 150, 200, 250, 300)$ and same variance equal to 10. As sampling sizes for each group of conditions, we take $n_1 = 20$, $n_2 = 30$ and $n_3 = 20$. In this case, 40 genes were simulated as having the same means over the three groups and 10 genes had different means for each group.

We take as prior distributions those shown in section 3, and we consider a mixture of non central t-distributions with an unknown number of components. Observe that the number of components may be considered as a parameter, and a reversible jump methodology can be applied to explore among spaces of different dimensions (see e.g. Green (1995) and Richardson and Green (1997)). Nevertheless, in the context of multiple testing associated with cluster problems it is better, from a practical point of view, to use a criterion of optimal measure of complexity and fit of models, to determine the number of components of a mixture of t-distributions. We use a modified version of the standard DIC coefficient, because in mixture models problems of identifiability can appear. This modification of the DIC coefficient was pointed out by Richardson (2002) and revised by Celeux et al. (2006), who named that version as DIC$_3$. The expression of this coefficient is

$$\text{DIC}_3 = -4E_{\theta|\mathbf{y}}[\log f(\mathbf{y} \mid \theta)] + 2\log \hat{f}(\mathbf{y}),$$

where $\hat{f}(\mathbf{y}) = \prod_{i=1}^n \hat{f}(y_i)$, and $\hat{f}(y_i) = E_{\theta|\mathbf{y}}[f(y_i \mid \theta)]$. We use this criterion to compare among models with different number of components of mixtures, and we take the model with the smallest DIC$_3$.

In table 1 we show the DIC$_3$ values for different number of components of the mixtures (k); we take for each computation of DIC$_3$ 20000 iterations with 10000 iterations for burn-in. A *post hoc* analysis of chains did not show a significant departure from convergence. The lowest values are found for $k = 4$ and $k = 7$ with not many differences among them; in this case a slightly smaller value, 613.90, is obtained for $k = 7$. Observe that this result is fairly closed to the desired objective, as the simulation was based in 7 groups with different means and same variance.

k	2	3	4	5	6	7	8	9	10
DIC$_3$	692.35	621.67	614.04	625.24	625.55	613.90	617.51	618.79	622.24

Table 1: Values of DIC$_3$ for k components of mixtures

We consider, therefore, the analysis of a mixture model with $k = 7$ elements. Results show a good behaviour, as the estimated probability of type I error is only 0.02, namely, only a 2% of all hypotheses are rejected being true. By the other hand, we have an estimated value of the false discovery rate (FDR) of 0.025, namely, only a 2.5% of those hypotheses that are true, are rejected. We also have obtained that no one of hypotheses are accepted being false, but among them, 4% of hypotheses being false are classified with a different mean than those that were simulated from. The mixture of non-central t-distributions, which corresponds to the distribution of the statistics that we defined, fits correctly the simulated data. Moreover, with this approach we can consider at the same time different groups of genes in order to make multiple comparisons.

4.2 An application in Bioinformatics

Once we have checked the procedure with simulated data, we consider an application in Bioinformatics. We take data from a microarray experiment with human (*Homo sapiens*), bonobo (*Pan paniscus*) and gorilla (*Gorilla gorilla*) cultured fibroblasts done by Karaman et al. (2003). Expressions profiles are obtained by means of Affymetrix HG U95Av2 chips for 12625 genes in 46 samples (23 humans, 11 bonobos and 12 gorillas), and they are available in the Bioconductor library `fibroEset` (see Gentleman et al. (2004)). In order to study genes with relevant effects, we select 95 genes whose expression scores are greater or equal than 6000.

As a first step, we consider a glance at the data by means of the package `made4` (Culhane et al. (2005)), from the `Bioconductor` bundle. We show a dendrogram with a hierarchical cluster analysis (based on a Pearson correlation distance metric with average linkage) of the 46 individuals, a boxplot and a histogram of the data.

In figure 1, red color corresponds to bonobos, green color corresponds to humans and blue color to gorillas. Apparently there is a slightly more proximity between bonobos and humans than in the case of gorillas. This is coherent with the evolutionary origin of the three species. Having a look to the boxplot it shows that sample variances are roughly equal for the three species, although distribution of data are far from a single distribution. The histogram shows an asymmetric distribution of data; therefore, as the original data were originally re-normalized by Karaman et al. (2003), they correspond to the three population normal distributions of the expressions of genes.

In order to have a look to the number of possible groups included in data, we apply a *hybrid* clustering method, *HOPACH* (Hierarchical Ordered Partitioning And Collapsing Hybrid), which is a popular technique used in Bioinformatics. It builds a hierarchical tree of clusters using partitioning and agglomerative methods (see van der Laan and Pollard (2003)). We obtain the following dendrogram, that suggests about 7 main clusters.

Now, we consider a mixture of non-central t-distributions with an unknown number of components. Although in figure 2 it is suggested about 7 clusters we compute also the DIC$_3$ values for different number of components k, which are shown in table 2. Results are obtained after a total of 20000 iterations with 10000 iterations for burn-in. The lowest value, 1438.83, is obtained for $k = 6$. We also considered a *post hoc* analysis of chains which did not show a significant departure from convergence.

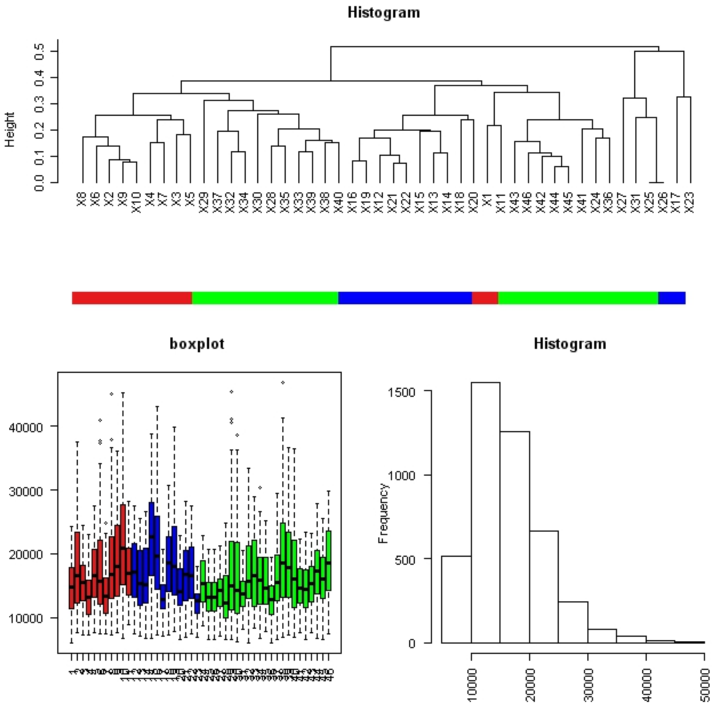

Figure 1: Overview of data: dendrogram with average linkage clustering, boxplot and histogram.

k	2	3	4	5
DIC$_3$	1602.50	1536.20	1478.37	1465.82
k	6	7	8	9
DIC$_3$	1438.83	1454.75	1457.89	1450.32
k	10			
DIC$_3$	1468.37			

Table 2: Values of DIC$_3$ with respect to k

We consider, therefore, the analysis of a mixture model with $k = 6$ elements. The posterior distribution of the means (δ_j, $j = 1, \dots 6$) of each component of the mixture are shown in table 3.

	mean	**sd**	**HPD 2.5%**	**Median**	**HPD 97.5%**
μ_1	2872.6	620	880	2930	3880
μ_2	3754.5	700	2700	3670	5230
μ_3	4534.1	780	3270	4420	6540
μ_4	5539.3	980	4050	5460	7200
μ_5	7070.0	2820	4650	6600	12870
μ_6	19884.4	30250	5880	10060	79020

Table 3: Posterior distribution of means of the six components of the mixture

Based on previous clusters, we can propose the classification of each T statistic, based on the criterion presented in section 3 as the rule of classification. Results are shown in table 4.

We deal, as an indirect way to validate the results, with 95 independent ANOVAs which are computed gene by gene. Direct comparison of results is not possible as ANOVAs do not take into

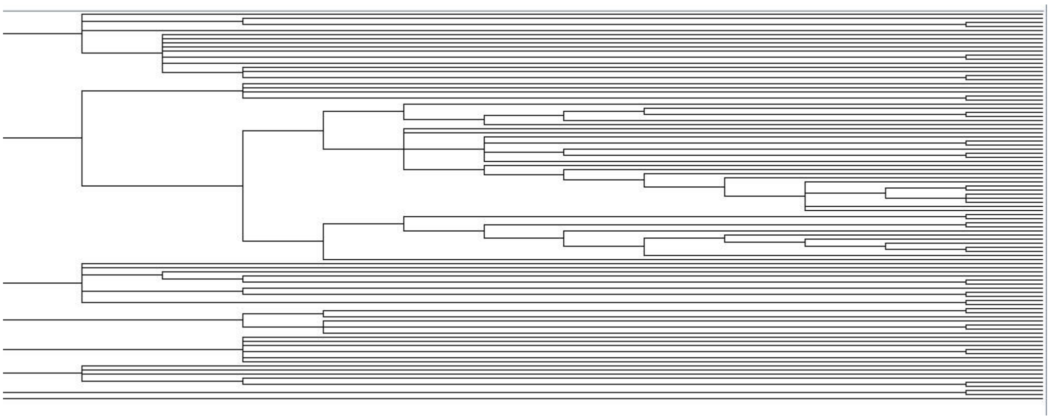

Figure 2: Dendogram using HOPACH.

account the conjoint relations among genes. We show the corresponding *p-values* of the tests in the last column of table 4. In terms of hypothesis testing, the decisions with respect to equality of the 95 genes are similar between ANOVAs and mixture of t-distributions method, although the independent ANOVAs tend to accept the hypotheses of equality among genes more frequently than the mixture of t-distribution method.

It is observed that the statistics show a closer relationship between human and bonobos than respective to gorillas; this fact is supported by the evolutionary relationships among the three species. Anyway, deeper insights on a pure genetic interpretation of results obtained in these particular data are beyond the scope of this paper.

	Gor	Bon	Hum	*p-val*
1	5	5	3	0.267
2	5	5	3	0.001
3	1	1	1	0.197
4	3	2	1	0.000
5	3	5	2	0.068
6	2	1	1	0.000
7	3	3	1	0.115
8	5	5	3	0.035
9	1	1	1	0.535
10	5	3	2	0.007
11	5	5	2	0.012
12	3	3	1	0.213
13	3	3	2	0.506
14	5	3	3	0.002
15	1	5	1	0.000
16	2	1	1	0.017
17	5	5	1	0.000
18	3	3	1	0.000
19	5	5	2	0.012
20	3	5	1	0.024
21	3	3	1	0.003
22	3	5	3	0.261
23	5	2	2	0.000
24	3	3	1	0.082
25	1	5	2	0.000
26	3	5	3	0.093
27	3	3	1	0.115
28	3	5	1	0.000
29	3	3	1	0.000
30	4	3	3	0.352
31	3	2	1	0.010
32	2	1	1	0.001
33	5	3	2	0.065
34	3	3	1	0.285
35	3	2	1	0.015
36	3	3	1	0.014
37	3	3	1	0.054
38	5	5	3	0.000
39	5	5	3	0.111
40	3	3	1	0.000
41	2	2	1	0.012
42	3	3	1	0.129
43	5	1	1	0.000
44	5	5	3	0.272
45	5	5	3	0.059
46	5	5	2	0.004
47	2	2	1	0.781
48	3	3	1	0.072
49	2	3	1	0.591
50	3	5	1	0.000
51	1	2	1	0.005
52	5	5	3	0.272
53	5	3	3	0.000
54	2	3	1	0.321
55	1	1	1	0.091
56	5	3	2	0.063
57	5	5	3	0.345
58	3	3	2	0.150
59	3	3	1	0.117
60	3	3	1	0.032
61	5	4	1	0.000
62	3	5	3	0.051
63	1	1	1	0.003
64	5	5	3	0.395
65	3	3	1	0.970
66	1	2	1	0.000
67	5	3	3	0.012
68	2	1	1	0.000
69	2	3	1	0.000
70	1	1	1	0.462
71	3	3	1	0.000
72	3	3	1	0.465
73	1	1	1	0.647
74	3	3	1	0.015
75	1	1	2	0.001
76	1	3	1	0.010
77	1	1	1	0.000
78	2	3	1	0.194
79	5	5	2	0.100
80	2	3	1	0.220
81	1	1	1	0.003
82	3	5	1	0.000
83	2	2	1	0.522
84	1	1	1	0.596
85	3	3	1	0.502
86	1	1	1	0.987
87	5	5	4	0.209
88	1	1	1	0.000
89	5	5	3	0.916
90	5	4	1	0.012
91	5	5	3	0.563
92	5	5	3	0.361
93	5	5	3	0.491
94	4	5	3	0.020
95	3	3	1	0.788

Table 4: Classification of statistics in the defined groups

5 Conclussions

In this paper we have shown a methodology that is well adapted to multiple hypothesis testing problems, based on a clustering methodology in Bioinformatics. With this approach, it is possible to tackle with multiple comparisons among more than two groups of different microarray expressions of genes. All groups can be analysed at the same time, and it is not necessary to compare them by pairs, as it is done in the standard post hoc analysis of multiple comparisons analysis.

Observe that original data are assumed to be normally distributed, although this is an usual assumption in microarray analysis, after standard procedures of normalization and data cleaning. Then, a mixture of non-central t-distributions is straightforwardly obtained when we define a standard statistic (denoted by T). This is related only with the sample means and variances of groups of data; namely, there are not other guesses in our model but normality of data, and computation of sample means and sample variances.

By the other hand, the results obtained with simulated data are adequate, and the analysis of real data renders sensible conclusions in the line of other standard methods of clustering or multiple testing techniques used in Bioinformatics.

References

G. Celeux , F. Forbes, C. P. Robert and D. M. Titterington (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* **1(4)**, 651–674.

S. T. B. Choy and A. E. M. Smith (1997). Hierarchical Models with Scale Mixtures of Normal Distributions. *TEST* **6(1)**, 205–221.

A. C. Culhane, J. Thioulouse, G. Perriere and D. G. Higgins (2005). MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* **21(11)**, 2789–2790.

D. B. Dahl and M. A. Newton (2007). Multiple Hypothesis Testing by Clustering Treatment Effects. *JASA* **102** 478, 517–526.

D. B. Dahl, Q. Mo and M. Vannucci (2008). Simultaneous inference for multiple testing and clustering via a Dirichlet process mixture model. *Statistical Modelling* **8(1)**, 23–39.

S. Dudoit, J. P. Shaffer and J. C. Boldrick (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* **18(1)**, 71–103.

R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarryand, F. Leisch, C. Li, M. Maechler, A. Rossiniand, G. Sawitzki, C. Smith, G. Smyth, L. Tierneyand, J. Yang and J. Zhang, (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.
`http://www.bioconductor.org`.

P. Green (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Y. Ji, Y. Lu and G. B. Mills (2008). Bayesian models based on test statistics for multiple hypothesis testing problems. *Bioinformatics* **24(7)**, 943–949.

N. L. Johnson, S. Kotz and N. Balakrishnan (1995). *Continuous Univariate Distributions Volume 2*. Wiley & Sons, New York.

M. W. Karaman, M. L. Houck, L. G. Chemnick, S. Nagpal, D. Chawannakul, D. Sudano, B. L. Pike, V. V. Ho, O. A. Ryder, and J. G. Hacia (2003). Comparative Analysis of Gene-Expression Patterns in Human and African Great Ape Cultured Fibroblasts. *Genome Research* **13**, 1619–1630.

M. van der Laan and K. Pollard (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J. of Statistical Planning and Inference* **117**, 275–303.

J. M. Marin and C. P. Robert (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* Springer, New York.

G. J. McLachlan, R. W. Bean and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18(3)**, 413–422.

M. Medvedovic and S. Sivaganesan (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18(9)**, 1194–1206.

M. A. Newton A. Noueiry D. Sarkar and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5(2)**, 155–176.

M. Plummer (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *DSC 2003 Working Papers*
`http://www-fis.iarc.fr/∼martyn/software/jags/`

S. Richardson and P. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Royal Statist. Soc. B* **59**, 731–792.

S. Richardson (2002). Discussion of Spiegelhalter et al. *J. of the Royal Statist. Soc. B* 631.

E. G. Tsionas (2002). Bayesian Inference in the Noncentral Student-t Model. *J. of Computational and Graphical Statistics* **11(1)**, 208–221.

M. Yuan and C. Kendziorski (2006). A Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification. *Biometrics* **62**, 1089–1098.