Brazilian Journal of Probability and Statistics
2015, Vol. 29, No. 1, 145–171
DOI: 10.1214/13-BJPS233

A gene-by-gene multiple comparison analysis: A predictive Bayesian approach

Erlandson F. Saraiva[a] **and Francisco Louzada**[b]

[a]*Universidade Federal de Mato Grosso do Sul*
[b]*Universidade de São Paulo*

Abstract. In this paper, we propose a hierarchical Bayesian framework with a prior Dirichlet process for gene-by-gene multiple comparison analysis. The comparison among experimental conditions are made using the posterior probability for hypothesis of equality or inequality. To calculate the posterior probabilities, we use the Polya urn scheme through latent variables and the Bayes factor. The performance of the proposed method, as well as a comparison with usual Tukey-test, are evaluated on artificial data and on a shotgun proteomics data set. The results reveal a better performance of the proposed methodology in identification of difference of means and/or variance.

1 Introduction

A common interest in gene expression data analysis is to identify genes that present significant changes in expression levels among biological experimental condition. The identification of these genes is important because it may allow biologists and geneticists to study possible relationships among genes, among genes and proteins, which genes may be involved in the origin and/or evolution of same disease with genetic origin, or which genes react to a drug stimulus, and so on. For further discussion and additional references on DNA array technology, see Schena et al. (1995), DeRisi et al. (1997), Arfin et al. (2000), Lonnstedt and Speed (2001), Wu (2001), Hatifield et al. (2003). Here, we restrict our discussion to DNA microarray data sets from oligonucleotide arrays (Irizarry et al., 2003). We assume that data consists of a set of replicate measurements for each gene.

Under the first level of analysis, where each gene is analyzed separately, the identification of genes differentially expressed, usually, is made by using a statistic and a cutoff value to separate the genes differentially expressed from the non differentially expressed ones. The literature on statistical methods to identify genes differentially expressed is extensive. We can cite the usual two-sample t-test (TT), the Cyber-t (CT) proposed by Baldi and Long (2001), the Bayesian t-test (BTT) proposed by Fox and Dimmic (2006) and the predictive Bayes Factor (PBF) proposed by Louzada et al. (2014). The CT and BTT are developed through modifications of the standard error estimate of the two sample difference present in the

Key words and phrases. Gene expression, multiple comparison, Bayesian Inference, Bayes factor, predictive density.

Received September 2012; accepted October 2013.

denominator of the standard t statistics. The PBF compare observed gene expression level from treatment and control using the posterior probability of the difference which is calculated using the Bayes factor. A large simulation study revealed a better performance of the PBF in identification of difference of means and/or variance in small sized samples, usually present in gene expression data analysis (Louzada et al., 2014).

The methods presented above can be applied only to compare a treatment with a control. This can be seen as a drawback to be overcome, since in practice we often find the need for multiple comparisons. For instance, consider the shotgun proteomics data set, extracted from the site http://cybert.ics.uci.edu/anova (Baldi and Long, 2001). The data set is composed by proteins from a control and two treatment conditions.

In this paper, we extend the Bayesian approach proposed by Louzada et al. (2014), by making a comparison gene-by-gene in the multiple comparison case, that is, from a control and more than one treatment experimental condition. The proposed approach is within a Bayesian framework with a Dirichlet process prior. The advantage of using the Dirichlet process prior is its discreteness which allows the parameters to be coincident with positive probability. Using this fact, we develop a multiple comparison approach using the posterior probabilities for hypothesis of equality or inequality among of experimental conditions. The posterior probabilities are calculated based on the Polya urn scheme (Blackwell and MacQueen, 1973) using latent variables and the Bayes factor. The advantage of using the Bayes factor is that it allows for compare the observed expression levels from treatments as well as the distributions associated to different treatment experimental conditions (Louzada et al., 2014).

The proposed method performance is verified in a generated and in a real dataset, where also it is compared to an analysis of variance (ANOVA) followed by a Tukey-test (Cox and Reid, 2000). The ANOVA is applied to identify genes which show significant difference among experimental conditions. But, it does not identify which experimental conditions show the difference. Thus, we apply the Tukey-test to selected genes from the ANOVA in order to identify which experimental conditions show significant difference. The choice of the Tukey-test is based on the fact that it is a commonly used post hoc test, see for example Pavlids (2003), Parkitna et al. (2006), Goeman and Bühlmann (2007).

The comparison between methods is made in terms of the true positive rate, true discovery rate and false discovery rate. The simulation results reveal a better performance of the proposed method. We also illustrate the performance of the proposed method using a real data set. The real data set is a shotgun proteomics experiment extracted from the site http://cybert.ics.uci.edu/anova (Baldi and Long, 2001).

The remainder of the paper is structured as follows. In Section 2, we describe the Bayesian model for gene expression data analysis and the Polya urn scheme using latent variables. In Section 3, we calculate the posterior probabilities for hypothesis of equality or inequality among experimental conditions using the Bayes factor. The performance of the proposed approach as well as a comparison with Tukey-test is presented in Section 4. In Section 5, the paper is concluded with final remarks.

2 Bayesian model for gene expression data analysis

Consider a DNA array experiment with N genes performed for experimental conditions $E_1, \dots, E_M$, where E_1 represents the control, E_2 represent the first treatment and successively until E_M, the last treatment. Assume that each experimental condition is replicated n times. Denote by x_{ig_m} the ith observed expression level (or its logarithm), for gene g, in experimental condition m, for $m = 1, \dots, M$, $i = 1, \dots, n$ and $g = 1, \dots, N$. Let $\mathbf{X}_g = \{\mathbf{X}_{g_1}, \dots, \mathbf{X}_{g_M}\}$ be all observed expression levels for gene g in M experimental conditions, where $\mathbf{X}_{g_m} = (x_{1g_m}, \dots, x_{ng_M})'$ is a $n \times 1$ vector of conditionally independent observations for gene g on treatment m, for $g = 1, \dots, N$ and $m \in \{1, \dots, M\}$.

Assume that data have already been preprocessed with appropriate normalization. For further discussion and additional references on normalization methods, see Yang et al. (2002), Huber et al. (2002), Bolstad et al. (2003), Smyth and Speed (2003), Chen et al. (2004). The real data set used in the paper is normalized according to Variance Stabilization and Normalization (VSN) method (Huber et al., 2002), as described in the site http://cybert.ics.uci.edu/anova from where the data set was downloaded.

Consider the logarithm of the observed gene expression levels in control and treatments are generated from normal distributions with mean μ_{g_m} and variance $\sigma^2_{g_m}$, $X_{ig_m} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{g_m}, \sigma^2_{g_m})$, for $i = 1, \dots, n$, $g = 1, \dots, n$ and $m = 1, \dots, M$.

In order to simplify the notation hereafter we omit the index g in next expressions. Denote parameters by $\theta_m = (\mu_m, \sigma^2_m)$ and by $\Theta = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_M);\ \theta_m \in \mathbb{R} \times \mathbb{R}^+\}$ the parametric space, for $m = 1, \dots, M$.

The interest here is to verify whether gene g presents different gene expression levels in different experimental conditions, i.e., if $\theta_m = \theta_j$ or $\theta_m \neq \theta_j$, for all $m \in \{1, \dots, M\}$, $j \in \{1, \dots, M\}$ and $m \neq j$. This leads to the following multiple hypotheses testing

$$H_0 : \Theta_0 = (\boldsymbol{\theta}; \theta_1 = \cdots = \theta_M),$$
$$H_1 : \Theta_1 = (\boldsymbol{\theta}; \theta_1 \neq \theta_2, \theta_2 = \theta_3 = \cdots = \theta_M),$$

successively for all combinations of inequality 2 to 2, 3 to 3 (see Apendix A), until the last one hypothesis, that is,

$$H_T : \Theta_T = (\boldsymbol{\theta}; \theta_1 \neq \cdots \neq \theta_M).$$

The equality (or not) between θ_m's in the hypotheses above, determines partitions on the parameter space, that is, the hypotheses $H_t : \Theta_t$, $t = 0, 1, \ldots, T$, are disjoint and $\bigcup_{t=0}^{T} \Theta_t = \Theta$. This allow us to develop a hierarchical Bayesian approach with a prior Dirichlet process on $\theta_1, \ldots, \theta_M$ in order to make simultaneous comparisons among θ_m's (Gopalan and Berry, 1998, Neal, 2000).

2.1 Prior Dirichlet process

Assume that prior distributions for $\theta_1, \ldots, \theta_M$ are sampled from a unknown distribution G and that G follows a prior Dirichlet process (Ferguson, 1973, Antoniak, 1974) with baseline distribution G_0 and mass parameter $\alpha > 0$, that is,

$$\begin{aligned} \theta_1, \theta_2, \ldots, \theta_M | G &\overset{\text{i.i.d.}}{\sim} G, \\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha G_0). \end{aligned} \tag{1}$$

The main advantage of using the prior framework in (1) is the discreteness of the prior distribution G, given the assumption of a Dirichlet process. Under such assumption, the parameters θ_m's are coincident with positive probability. This fact is discussed by Blackwell and MacQueen (1973), which show that integrating G over its prior distribution in (1), $\theta_1, \ldots, \theta_M$ follows a Polya urn scheme, which can be written as

$$\begin{aligned} \theta_1 &\sim G_0, \\ \theta_m | \theta_1, \ldots, \theta_{m-1} &\sim \frac{\alpha}{\alpha + m - 1} G_0 + \frac{1}{\alpha + m - 1} \sum_{j=1}^{m-1} \mathcal{I}_{\theta_m}(\theta_j), \end{aligned} \tag{2}$$

where $\mathcal{I}_{\theta_m}(\theta_j) = 1$ if $\theta_m = \theta_j$ and $\mathcal{I}_{\theta_m}(\theta_j) = 0$ otherwise, for $j \in \{1, \ldots, m-1\}$ and $m \in \{2, \ldots, M\}$.

Note that, at each step of the sample procedure defined in (2), θ_m may replicate one of the previous θ_j's, with probability $\frac{1}{\alpha+m-1} \sum_{j=1}^{m-1} \mathcal{I}_{\theta_m}(\theta_j)$, or may assume a new value, generated from baseline distribution G_0, with probability $\frac{\alpha}{\alpha+m-1}$. Thus, a sample from joint distribution of $\theta_1, \ldots, \theta_M$ yields k groups ($1 \leq k \leq M$) of θ_m's with distinct values given by $\phi_1, \ldots, \phi_k$, generated from the baseline distribution G_0. In the next section, we explore this fact using latent variables in order to develop the proposed multiple comparison.

2.2 Prior Dirichlet process via latent variables

In order to represent the k groups of θ_m's consider the latent variables $\mathbf{Z} = (Z_1, \ldots, Z_M)$ in a way that Z_m is paired with θ_m and $Z_m = j$ indicates that $\theta_m = \phi_j$, $\phi_j \sim G_0$ for $m = 1, \ldots, M$ and $j = 1, \ldots, k$. The configuration of $\mathbf{Z}$ defines the groups and the group formed by the subset of index m, so that, $Z_m = Z_1$

define the group of experimental conditions that present no evidence for difference in relation to the control experimental condition.

Moreover, by introducing the latent variables $\mathbf{Z}$ we obtain a partition of all observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ in k groups, $\{D_1, \ldots, D_k\}$, where $D_j = \{\mathbf{x}_m; Z_m = j\}$, with $\bigcup_{j=1}^{k} D_j = \mathbf{x}$. The set $\{D_1, \ldots, D_k\}$ is paired with the set $\{\phi_1, \ldots, \phi_k\}$, i.e., the observations in D_j are modeled by the same distribution $F(\phi_j)$. The likelihood function for $\mathbf{Z}$ is given by

$$L(\mathbf{Z}|\mathbf{x}) = \prod_{j=1}^{k} \mathbf{I}(D_j), \tag{3}$$

in which

$$\mathbf{I}(D_j) = \int \left[\prod_{\mathbf{x}_m \in D_j} f(\mathbf{x}_m|\phi_j) \right] \pi_{G_0}(\phi_j)\, d\phi_j, \tag{4}$$

where $\pi_{G_0}(\cdot)$ and $f(\cdot)$ represent the densities of the baseline distribution G_0 and of the normal distribution, respectively, for $m = 1, \ldots, M$ and $j = 1, \ldots, k$.

Considering n_j the number of observations in D_j given the configuration $Z_1, \ldots, Z_{m-1}$, the Polya urn scheme in (2) can be replicated by the following steps:

(i) Initialize $Z_1 = 1$, $k = 1$, $D_1 = \{\mathbf{x}_1\}$ and generate ϕ_1 from the baseline distribution, $\phi_1 \sim G_0$.

(ii) For $m = 2, \ldots, M$ sample Z_m with probabilities given by

$$P(Z_m = j|Z_1, \ldots, Z_{m-1}) = \frac{n_j}{\alpha + m - 1}, \tag{5}$$

$$P(Z_m \neq Z_j, \forall j < m|Z_1, \ldots, Z_{m-1}) = \frac{\alpha}{\alpha + m - 1}, \tag{6}$$

for $j = 1, \ldots, k$. For the case in (6) we consider that Z_m assumes a new value $j^* = \max(Z_1, \ldots, Z_{m-1}) + 1 = k + 1$;

(a) If $Z_m = j$ for some $j \in \{1, \ldots, k\}$, do $D_j = D_j \cup \mathbf{x}_m$ and $n_j = n_j + 1$;

(b) If $Z_m = j^*$, does $D_{j^*} = \{\mathbf{x}_m\}$ and generate ϕ_{j^*} from the baseline distribution G_0, $\phi_{j^*} \sim G_0$. The number of groups increases by one unit, $k = k + 1$.

(iii) Conditional on $\mathbf{Z} = (Z_1, \ldots, Z_M)$, set $\theta_m = \phi_j$ for all $Z_m = j$, $j = 1, \ldots, k$.

2.2.1 *Choice of* G_0. It is now necessary to specify the prior mean G_0 of G. Following, Escobar and West (1995) and Casella et al. (2000) we assume that under G_0

$$\mu_m|\sigma_m^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma_m^2}{\lambda}\right) \quad \text{and} \quad \sigma_m^2 \sim \mathcal{IG}\left(\frac{\tau}{2}, \frac{\beta}{2}\right),$$

for $m = 1, 2, \dots, M$, where μ_0, λ, τ and β are hyperparameters and $\mathcal{IG}(\cdot)$ represents the inverse gamma distribution with parametrization so that the mean is given by $\tau/(\beta - 2)$. The choices of the hyperparameters will generally depend upon the application at hand. At this moment, we leave them unspecified.

Thus, from (4)

$$\mathbf{I}(D_j) = \left[\frac{1}{\beta\pi}\right]^{n_j/2} \lambda^* \Gamma^* \left[1 + \frac{\sum_{\mathbf{x}_m \in D_j} \mathbf{x}_m^2 + \lambda\mu_0^2}{\beta} - \frac{(\sum_{\mathbf{x}_m \in D_j} \mathbf{x}_m + \lambda\mu_0)^2}{\beta(n_j + \lambda)}\right]^{-((\tau+n_j)/2)}, \tag{7}$$

where $\lambda^* = [\frac{\lambda}{n_j+\lambda}]^{1/2}$ and $\Gamma^* = \Gamma(\frac{\tau+n_j}{2})/\Gamma(\frac{\tau}{2})$, for $j = 1, \dots, k$.

2.2.2 *Choice of* α. It is also necessary to either specify a value for α or put a prior distribution on it. Escobar (1994) and Bhattacharya (2008) assume for α a Gamma(a_α, b_α) prior distribution and develop a Gibbs sampler algorithm in order to estimate a vector of normal means and α. On the other hand, Escobar and West (1995), Medvedovic and Sivaganesan (2002), Jain and Neal (2004) and Jain and Neal (2007), fix α equals 1, $\alpha = 1$. This value of α is a natural choice due to the way of the Polya urn scheme in (2). Gopalan and Berry (1998) propose a elicitation procedure to fix a value for α using probabilities $P(H_0) = \alpha(M-1)!/\prod_{m=1}^{M}(\alpha + m - 1)$ and $P(H_T) = \alpha^M/\prod_{m=1}^{M}(\alpha + m - 1)$. Thus, setting up $P(H_0)/P(H_T) = 1$ we obtain $\alpha = \sqrt[M-1]{(M-1)!}$.

As the proposed procedure does not need MCMC methods to calculate the posterior probabilities described in Section 3, we opt to follow Escobar (1994), Medvedovic and Sivaganesan (2002), Jain and Neal (2004, 2007) and Gopalan and Berry (1998), fixing $\alpha = 1$ and $\alpha = \sqrt[M-1]{(M-1)!}$. In our experience, these both values of α, worked well. However, it does not restrict the method for being applicable in cases where the interest also lies in estimation of α, as in approach of Escobar (1994) and Bhattacharya (2008).

3 Multiple comparison via posterior probability for Z

In this section, we describe the multiple comparison approach using the posterior probabilities for the latent variables $\mathbf{Z}$.

From Bayes theorem, updating the prior probabilities in (5) and (6) via likelihood function in (3), the conditional posterior probabilities are

$$P(Z_m = j | Z_1, \dots, Z_{m-1}, \mathbf{x}) = b\frac{n_j}{\alpha + m - 1} \int f(\mathbf{x}_m | \phi_j)\pi(\phi_j | D_j)\, d\phi_j \tag{8}$$

and

$$P\big(Z_m = j^* | Z_1, \dots, Z_{m-1}, \mathbf{x}\big) = b\frac{\alpha}{\alpha + m - 1} \int f(\mathbf{x}_m | \phi_{j^*})\pi_{G_0}(\phi_{j^*})\, d\phi_{j^*}, \tag{9}$$

where $\pi(\phi_j|D_j)$ is the density of the posterior distribution for ϕ_j given the set $D_j = \{\mathbf{x}_{m'}; Z_{m'} = j\ \forall m' < m\}$, $j^* = k+1$ and b is the normalizing constant in order the probabilities sum up one.

Using (7), the probabilities in (8) and (9) are given by

$$P(Z_m = j|Z_1,\ldots,Z_{m-1},\mathbf{x}) = b\frac{n_j}{\alpha+m-1}\frac{\mathbf{I}(D_j \cup \mathbf{x}_m)}{\mathbf{I}(D_j)} \tag{10}$$

and

$$P\big(Z_m = j^*|Z_1,\ldots,Z_{m-1},\mathbf{x}\big) = b\frac{\alpha}{\alpha+m-1}\mathbf{I}(\mathbf{x}_m). \tag{11}$$

We describe bellow the probabilities in (10) and (11) in terms of the Bayes factor for some particular cases.

3.1 A control and a treatment condition

For this case, $m = 1, 2$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. Initialize with $Z_1 = 1$ and let $D_1 = \{\mathbf{x}_1\}$.

Thus, from (10) and (11), respectively,

$$\begin{aligned} P(Z_2 = 1|Z_1 = 1,\mathbf{x}) &= \frac{1}{1+\alpha B_{21}} \quad \text{and} \\ P(Z_2 = 2|Z_1 = 1,\mathbf{x}) &= \frac{\alpha B_{21}}{1+\alpha B_{21}}, \end{aligned} \tag{12}$$

where $B_{21} = \frac{\mathbf{I}(D_1)\mathbf{I}(\mathbf{x}_2)}{\mathbf{I}(D_1 \cup \mathbf{x}_2)}$ is the Bayes factor (Kass and Raftery, 1995) of the model which assume $\mathbf{x}_1 \sim F(\phi_1)$ and $\mathbf{x}_2 \sim F(\phi_2)$ for $\phi_1 \neq \phi_2$ related to a model which assume $\mathbf{x}_1, \mathbf{x}_2 \sim F(\phi_1)$. We calculate B_{21} according to proposal of Louzada et al. (2014).

For $\alpha = 1$, probabilities in (12) are the probabilities for models M_0 and M_1 in proposal of Louzada et al. (2014). Moreover, following Louzada et al. (2014), if $P(Z_2 \neq 1|Z_1 = 1,\mathbf{x}) > P(Z_2 = 1|Z_1 = 1,\mathbf{x})$ we set up $Z_2 = 2$. In this case, the gene presents evidence for difference between treatment and control. Otherwise, we do $Z_2 = Z_1 = 1$. The gene does not have evidence for difference.

3.2 A control and two treatment conditions

For this case, $m = 1, 2, 3$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$. Initialize by applying the procedure described in Section 3.1, in order to compare treatment condition 1 with the control condition.

(a) Given that $Z_2 = Z_1 = 1$, do $D_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$. The posterior probabilities for Z_3 are given by

$$P(Z_3 = j|Z_1 = 1, Z_2 = 1,\mathbf{x}) = \begin{cases} \dfrac{2}{2+\alpha B_{31}}, & \text{for } j = 1, \\ \dfrac{\alpha B_{31}}{2+\alpha B_{31}}, & \text{for } j = 2, \text{ i.e., } j \neq 1, \end{cases}$$

for $B_{31} = \frac{\mathbf{I}(D_1)\mathbf{I}(\mathbf{x}_3)}{\mathbf{I}(D_1 \cup \mathbf{x}_3)}$. If $P(Z_3 \neq 1|Z_1 = 1, Z_2 = 1, \mathbf{x}) > P(Z_3 = 1|Z_1 = 1, Z_2 = 1, \mathbf{x})$ do $Z_3 = 2$. Otherwise, do $Z_3 = Z_2 = Z_1 = 1$.

(b) Given that $Z_2 \neq Z_1$, do $D_1 = \{\mathbf{x}_1\}$ and $D_2 = \{\mathbf{x}_2\}$. The posterior probabilities for Z_3 are given by

$$P(Z_3 = j|Z_1 = 1, Z_2 = 2, \mathbf{x}) = \begin{cases} \dfrac{B_{32}}{B_{31} + B_{32} + \alpha B_{31} B_{32}}, & \text{for } j = 1, \\ \dfrac{B_{31}}{B_{31} + B_{32} + \alpha B_{31} B_{32}}, & \text{for } j = 2, \\ \dfrac{\alpha B_{31} B_{32}}{B_{31} + B_{32} + \alpha B_{31} B_{32}}, & \text{for } j = 3, \text{ i.e., } j \neq 1, 2, \end{cases}$$

where $B_{3j} = \frac{\mathbf{I}(D_j)\mathbf{I}(\mathbf{x}_3)}{\mathbf{I}(D_j \cup \mathbf{x}_3)}$ for $j = 1, 2$. If $P(Z_3 = j|\cdot) = \max_{i=1,2,3}(P(Z_3 = i|\cdot))$ do $Z_3 = j$. At this point, another possibility would be randomly generate $Z_3 = j$ with probability $P(Z_3 = j|\cdot)$ or, following Shapiro (1977), to consider the maximum posterior probability, which we therefore prefer.

In the Appendix B, we present the posterior probabilities for the case with a control and three treatments.

3.3 Algorithm for the general case

For the general case, the probabilities can be calculated by the following steps:

(i) Initialize with $Z_1 = 1$, $D_1 = \{\mathbf{x}_1\}$ and $k = 1$;

(ii) for $m = 2, \ldots, M$ do the following:
- (a) Calculate $\mathbf{I}(D_j)$, $\mathbf{I}(D_j \cup \mathbf{x}_m)$ and $\mathbf{I}(\mathbf{x}_m)$ according to (7), for $j = 1, \ldots, k$;
- (b) From (10), calculate $P(Z_m = j|Z_1, \ldots, Z_{m-1}, \mathbf{x}) \propto \frac{n_j}{\alpha+m-1} \frac{\mathbf{I}(D_j \cup \mathbf{x}_m)}{\mathbf{I}(D_j)}$;
- (c) From (11), calculate $P(Z_m = j^*|Z_1, \ldots, Z_m, \mathbf{x}) \propto \frac{\alpha}{\alpha+m-1}\mathbf{I}(\mathbf{x}_m)$, for $j^* = k + 1$;
- (d) If $P(Z_j = j|\cdot) = \max_{j=1,\ldots,k}(P(Z_m = j|\cdot), P(Z_m = j^*|\cdot))$, do $D_j = D_j \cup \mathbf{y}_m$ and $n_j = n_j + 1$;
- (e) If $P(Z_j = j^*|\cdot) = \max_{j=1,\ldots,k}(P(Z_m = j|\cdot), P(Z_m = j^*|\cdot))$, do $D_{k+1} = \{\mathbf{y}_m\}$, $n_{j^*} = 1$ and $k = k + 1$.

Given $\mathbf{Z} = (Z_1, \ldots, Z_M)$, the set $D_1 = \{\mathbf{x}_m; Z_m = 1\}$ is composite by the treatment conditions which does not have evidence for difference related to the control, for $m \in \{2, \ldots, M\}$.

Hereafter, we refer to our approach as Bayesian multiple comparison via Bayes factor (MCBF).

4 Data analysis

In this section, the proposed MCBF approach is applied to artificial and a real datasets. The artificial data sets were generated as a mix of both differentially and non-differentially expressed genes where the fraction of differentially expressed genes is small.

To evaluate the performance of the MCBF and to compare with the ANOVA followed by a Tukey-test (Tuk), we consider the true positive rate, the true discovery rate and the false discovery rate.

Following Louzada et al. (2014) to specify the hyperparameters values, we set up $\mu_0 = [\min(\mathbf{x}) + \max(\mathbf{x})]/2$, $\lambda = 10^{-2}$, $\tau = 3$ and $\beta = (\tau - 2)R$, where $R = \max(\mathbf{x}) - \min(\mathbf{x})$ is the length of the interval of variation of the observed data $\mathbf{x}$.

4.1 Artificial data set

Here we present the performance of MCBF for the case with a control and two treatment conditions. The five hypothesis written in terms of latent variables are: $H_0 : Z_1 = Z_2 = Z_3$, $H_1 : Z_1 = Z_2 \neq Z_3$, $H_2 : Z_1 = Z_3 \neq Z_2$, $H_3 : Z_1 \neq Z_2 = Z_3$ and $H_4 : Z_1 \neq Z_2 \neq Z_3$.

To generate data sets, we follow Louzada et al. (2014) fixing the control parameters as $\mu_1 = -14$ and $\sigma_1^2 = 0.8$. For this case, $M = 3$, we obtain from Gopalan and Berry's (1998) procedure, $\alpha = \sqrt{2}$. The sample size n was fixed at $n = 5$, based on the real data set discussed in the next section. We also fix $N = 1000$ and proportions generated from each hypothesis as 0.80 from H_0 and 0.05 from H_j, $j = 1, \ldots, 4$. To verify how the method behaves when treatment parameters (μ_j, σ_j), $j = 2, 3$, moves away from control parameters (μ_1, σ_1), we fix parameters values for hypothesis H_j as follows:

- for H_1 we fix $(\mu_2, \sigma_2) = (\mu_1, \sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$;
- for H_2 we fix $(\mu_3, \sigma_3) = (\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$;
- for H_3 we fix $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2, \sigma_2)$;
- for H_4 we fix $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2 + \delta\sigma_2, \gamma\sigma_2)$,

for $\delta = \{0, 0.50, 1, 1.50, 2, 2.50, 3, 3.50, 4\}$ and $\gamma = \{1, 2, 3\}$.

Thus, the generation of the simulated data sets is as follows. For $g = 1, \ldots, N$, generate u_g from $U \sim \mathcal{U}(0, 1)$;

(i) if $u_g \leq 0.80$, fix parameters values according to H_0. Let the index vector $\mathbb{G}_g = (1, 1, 1)$ to indicate that case g is generated from H_0;
(ii) if $0.80 < u_g \leq 0.85$, fix parameters values according to H_1 and set $\mathbb{G}_g = (1, 1, 2)$;
(iii) if $0.85 < u_g \leq 0.90$, fix parameters values according to H_2 and set $\mathbb{G}_g = (1, 2, 1)$;
(iv) if $0.90 < u_g \leq 0.95$, fix parameters values according to H_3 and set $\mathbb{G}_g = (1, 2, 2)$;

(v) if $u_g > 0.95$, fix parameters values according to H_4 and set $\mathbb{G}_g = (1, 2, 3)$;
(vi) fixed parameters according to one the steps above, generate $\mathbf{X}_j = (X_{j1}, \ldots, X_{jn}) \sim \mathcal{N}(\mu_j, \sigma_j^2)$, for $j = 1, 2, 3$.

We apply the MCBF and the Tuk (with significance level at 0.05) to the generated the data sets. To record the configuration obtained by the MCBF and the Tuk, we consider the index vector $\mathbb{Z}_g^{\text{method}}$, where $\mathbb{Z}_g^{\text{method}}$ assume one of the following configurations: $(1, 1, 1)$, $(1, 1, 2)$, $(1, 2, 1)$, $(1, 2, 2)$ or $(1, 2, 3)$, for method $= \{\text{MCBF}, \text{Tuk}\}$. So, we compare performance of the methods by using the true positive rate (TPR), the true discovery rate (TDR) and the false discovery rate (FDR), as presented in the Appendix C.

Moreover, for each pair (δ, γ) we generate $L = 100$ different artificial data sets according to steps (i) to (vi) described above and present the results using the mean of the TPR, TDR and FDR. For instance, the mean of TPR is given by $\overline{\text{TPR}} = \sum_{l=1}^{L} \text{TPR}^{(l)}/L$, where $\text{TPR}^{(l)}$ is the TPR calculated for lth generated data set.

The plots in Figures 1 and 2 show the performances of both methods, for Tuk with significance level at 0.05 and MCBF with $\alpha = 1$ and $\alpha = \sqrt{2}$, respectively. We observe the MCBF performs better than Tuk, by presenting higher TPR and TDR and smaller FDR. Besides, increasing the variance of the treatment ($\gamma = \{2, 3\}$) better is the performance of MCBF in relation to the Tuk.

The plots in Figures 3 and 4 show the performances of both methods for $n = 10$. The MCBF also presents better performance by presenting higher TPR and TDR and smaller FDR than Tuk.

The plots in Figures 5 and 6 show the performances of both methods, but now for Tuk with significance level at 0.10. The MCBF also presents better performance.

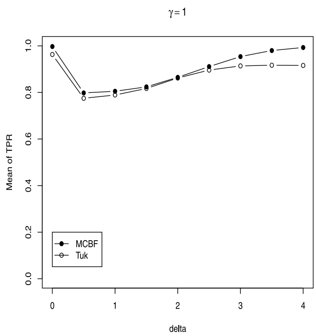
From the biological practical point of view, it indicates the MCBF may identify gene differences which are not identified by Tuk, specially, genes with differences in means and variances.

In the Appendix D, one can find the comparison of the performance of methods for $M = 4$. For such case, the MCBF also presents higher TPR and TDR and smaller FDR.

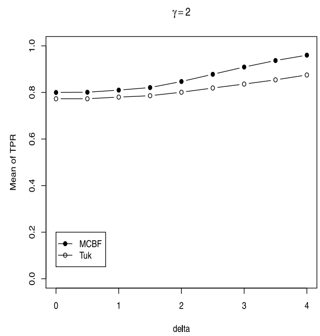
4.2 Real data set

Now recall the shotgun proteomics data set mentioned in the introduction, extracted from the site cybert.ics.uci.edu/anova/ (Baldi and Long, 2001). The data set is composed by $N = 1088$ proteins from a control and two treatment conditions. The sample size from each experimental condition is $n = 5$.

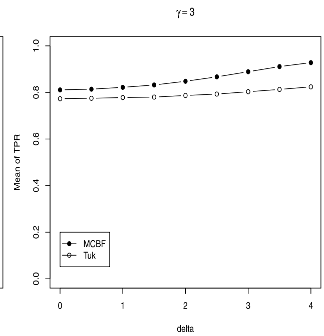
Results from MCBF are the same for $\alpha = 1$ and $\alpha = \sqrt{2}$. The MCBF identified 12 cases under H_1, 70 under H_2, 22 under H_3 and none under H_4. While, the Tuk identifies 3, 60, 6 and 27 cases under H_1, H_2, H_3 and H_4, respectively. Out of



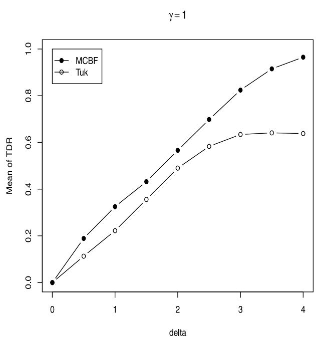
(a) $\overline{TPR}$



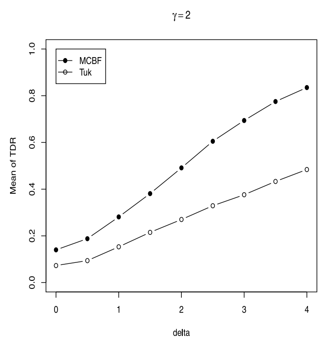
(b) $\overline{TPR}$



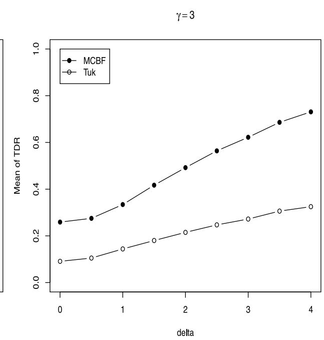
(c) $\overline{TPR}$



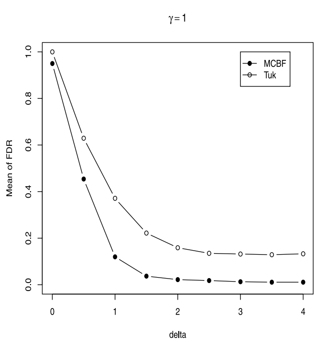
(d) $\overline{TDR}$



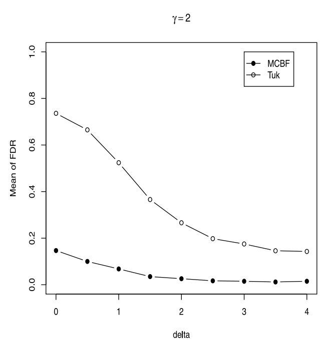
(e) $\overline{TDR}$



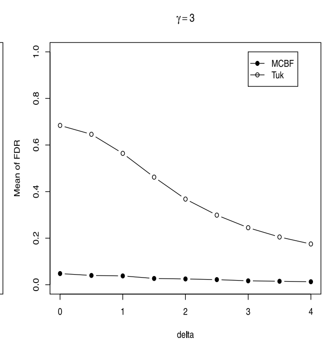
(f) $\overline{TDR}$



(g) $\overline{FDR}$



(h) $\overline{FDR}$



(i) $\overline{FDR}$

Figure 1 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 3$ *and* $n = 5$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = 1$.

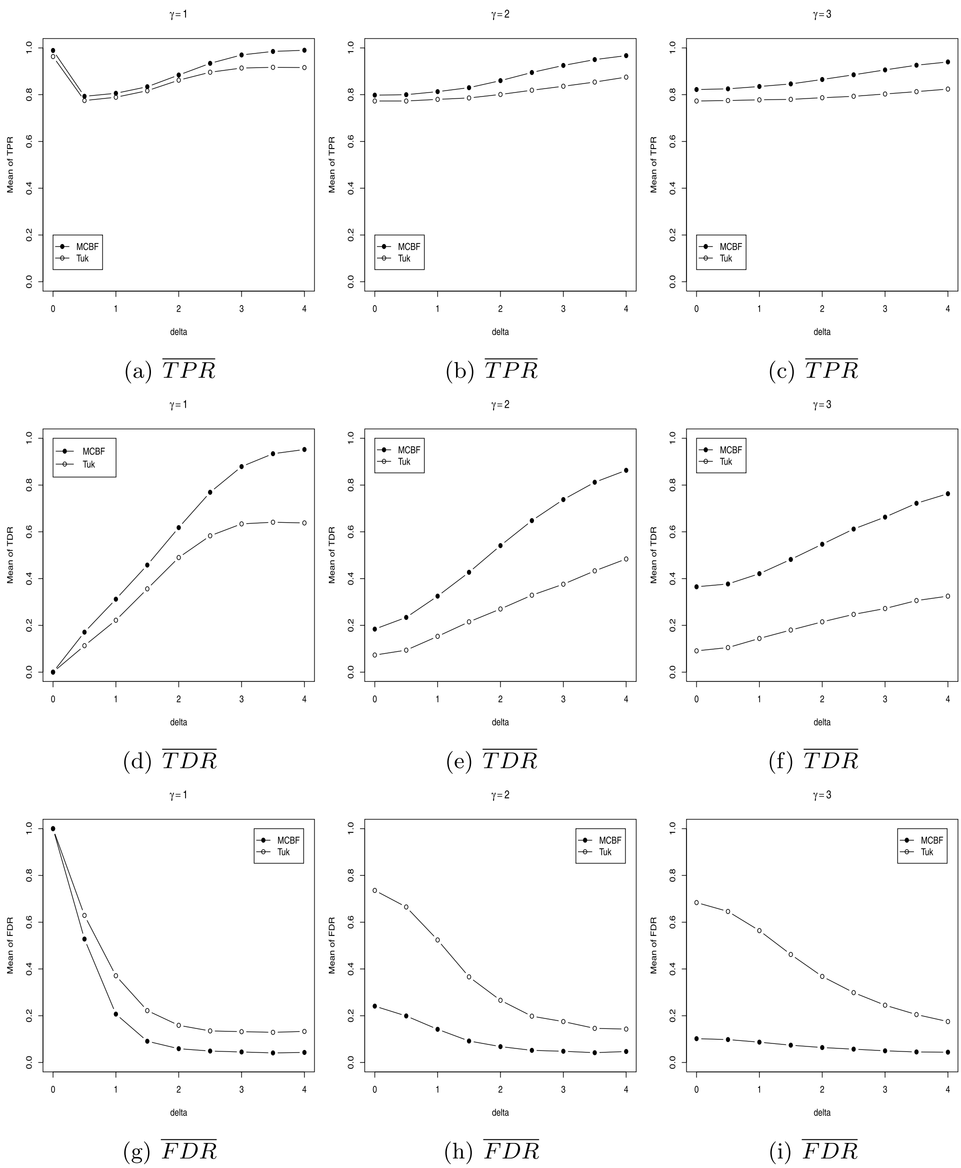


Figure 2 $\overline{\mathrm{TPR}}$, $\overline{\mathrm{TDR}}$ *and* $\overline{\mathrm{FDR}}$ *by method*, $M = 3$ *and* $n = 5$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = \sqrt{2}$.

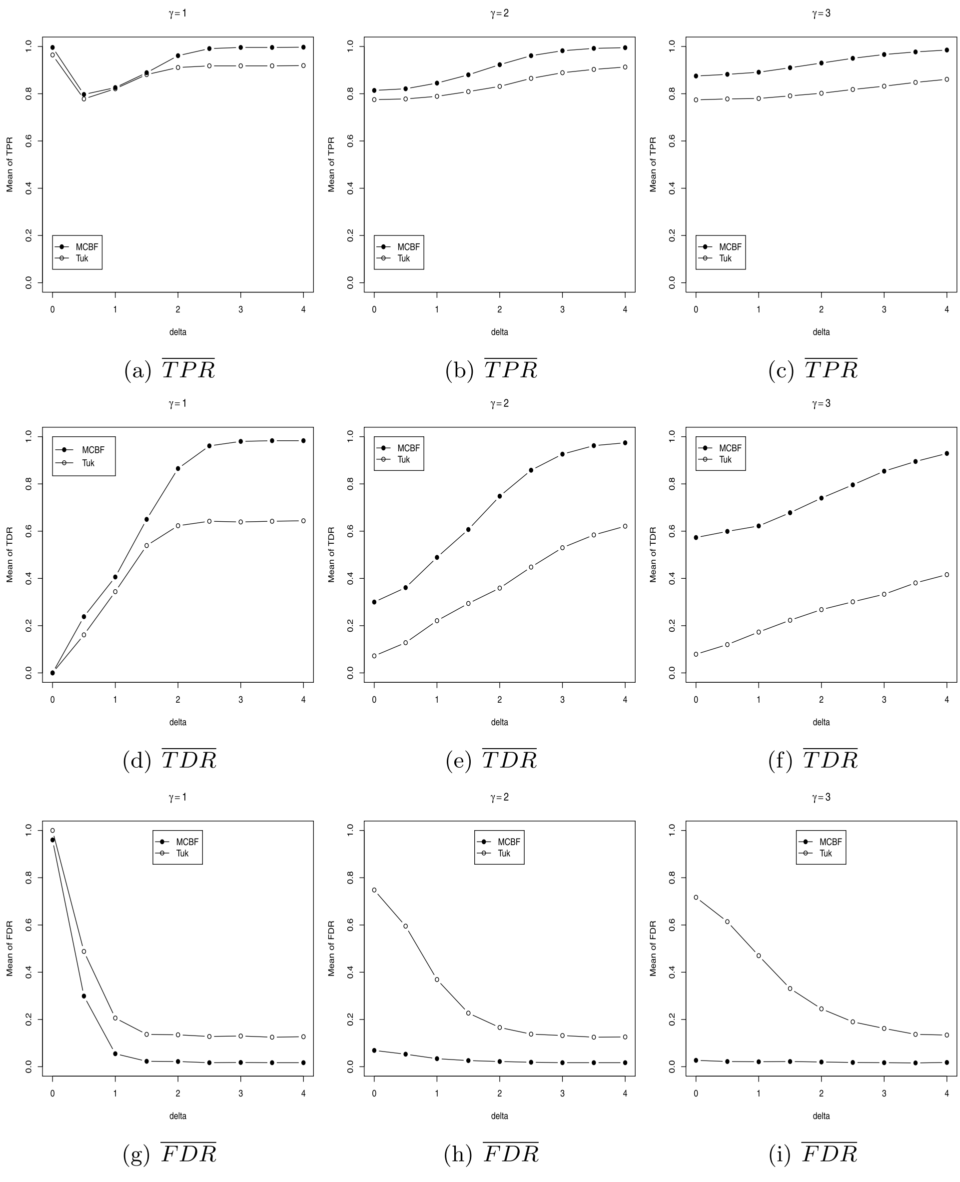
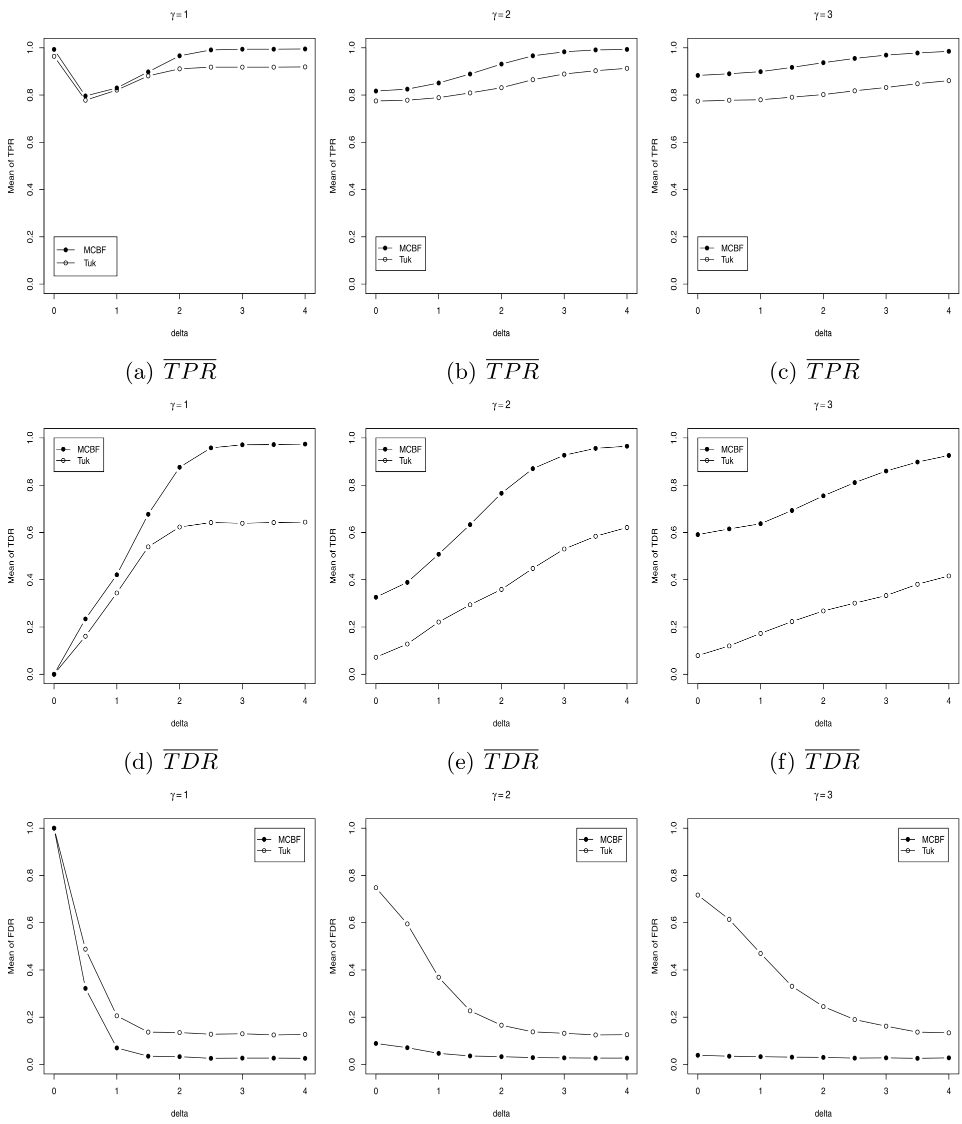


Figure 3 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 3$ *and* $n = 10$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = 1$.



(a) $\overline{TPR}$ (b) $\overline{TPR}$ (c) $\overline{TPR}$

(d) $\overline{TDR}$ (e) $\overline{TDR}$ (f) $\overline{TDR}$

(g) $\overline{FDR}$ (h) $\overline{FDR}$ (i) $\overline{FDR}$

Figure 4 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 3$ *and* $n = 10$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = \sqrt{2}$.

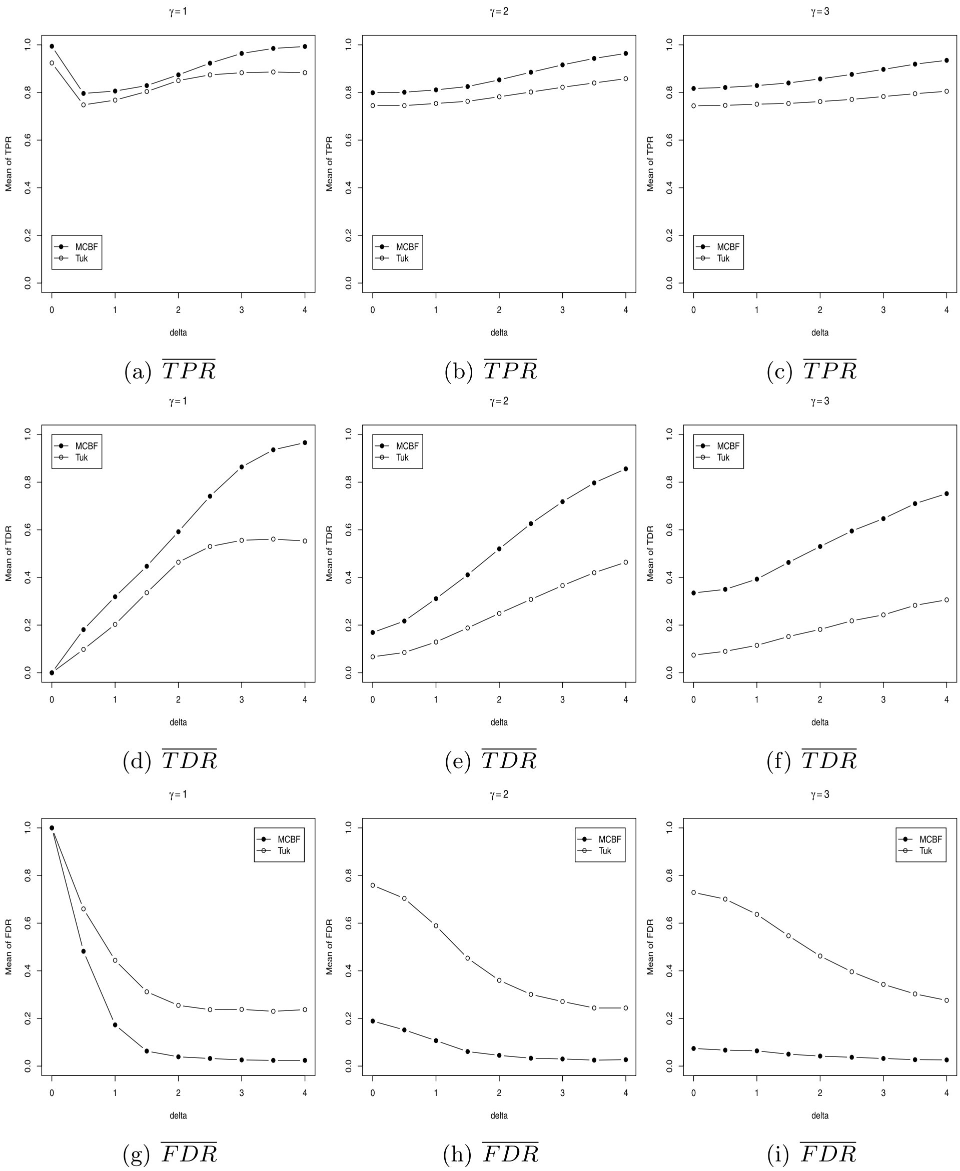
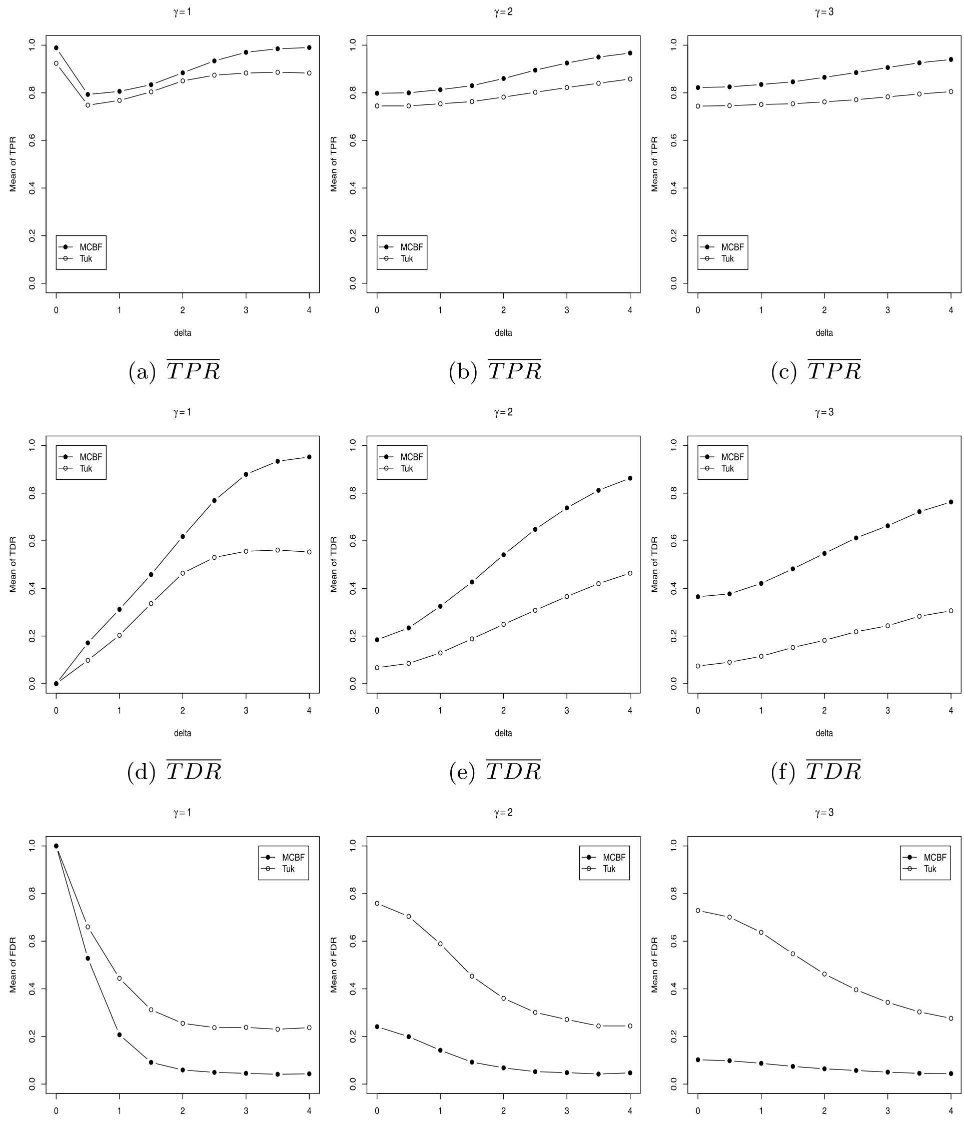


Figure 5 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 3$ *and* $n = 5$. *Tuk with significance level at* 0.10 *and MCBF with* $\alpha = 1$.



(a) $\overline{TPR}$ (b) $\overline{TPR}$ (c) $\overline{TPR}$

(d) $\overline{TDR}$ (e) $\overline{TDR}$ (f) $\overline{TDR}$

(g) $\overline{FDR}$ (h) $\overline{FDR}$ (i) $\overline{FDR}$

Figure 6 $\overline{\mathrm{TPR}}$, $\overline{\mathrm{TDR}}$ *and* $\overline{\mathrm{FDR}}$ *by method*, $M = 3$ *and* $n = 5$. *Tuk with significance level at* 0.10 *and MCBF with* $\alpha = \sqrt{2}$.

96 rejected null hypothesis by the ANOVA, 71 (73, 96%) were also rejected by the MCBF. Under H_1, out of three cases identified by the Tuk, two were also identified by the MCBF. Under H_2, out of the 60 cases identified by the Tuk, 45 were also identified by the MCBF. Under H_3, the six cases identified by the Tuk were also identified by the MCBF.

Tables 1 and 2 show the ten most significantly cases identified by MCBF and Tuk, respectively. In these tables, column 1 shows the number of the protein in the data set; columns 2, 3, 4 and 5, 6, 7 show the sample mean and standard-deviation (s.d.) from control, treatment 1 and treatment 2, respectively; columns 8 and 9 show the configuration identified; column 10 show the posterior probability for configuration identified by MCBF and column 11 show the p-value from the ANOVA.

Table 1 *Ten most significantly cases identified by MCBF*

	Sample mean			Sample s.d.			Configuration		Posterior	
Number	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_2$	$\bar{\mathbf{x}}_3$	s_1	s_2	s_3	MCBF	Tuk	probability	p-value
690	11.603	13.780	13.536	0.742	0.505	0.553	(1, 2, 2)	(1, 2, 2)	0.920	<0.001
666	11.885	14.087	12.304	0.643	0.617	0.650	(1, 2, 1)	(1, 2, 3)	0.896	<0.001
932	11.732	14.123	12.259	0.806	0.537	0.839	(1, 2, 1)	(1, 2, 3)	0.893	<0.001
661	15.975	15.982	15.095	0.020	0.005	1.813	(1, 1, 2)	(1, 1, 1)	0.842	0.339
847	12.042	13.188	10.087	0.788	0.780	3.411	(1, 1, 2)	(1, 1, 1)	0.810	0.095
557	8.740	13.114	12.675	5.156	1.384	0.909	(1, 2, 2)	(1, 1, 1)	0.778	0.090
936	10.942	12.778	11.289	0.449	1.073	0.463	(1, 2, 1)	(1, 2, 3)	0.773	0.004
1024	12.042	13.896	12.061	0.702	0.805	0.691	(1, 2, 1)	(1, 2, 3)	0.763	0.002
625	10.898	12.660	11.216	0.558	0.741	0.856	(1, 2, 1)	(1, 2, 3)	0.745	0.005
1012	11.550	13.185	11.552	0.613	0.578	0.702	(1, 2, 1)	(1, 2, 3)	0.742	0.002

Table 2 *Ten most significantly cases identified by Tuk*

	Sample mean			Sample s.d.			Configuration		Posterior	
Number	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_2$	$\bar{\mathbf{x}}_3$	s_1	s_2	s_3	MCBF	Tuk	probability	p-value
690	11.603	13.780	13.536	0.742	0.505	0.553	(1, 2, 2)	(1, 2, 2)	0.920	<0.001
666	11.885	14.087	12.304	0.643	0.617	0.650	(1, 2, 1)	(1, 2, 3)	0.896	<0.001
932	11.732	14.123	12.259	0.806	0.537	0.839	(1, 2, 1)	(1, 2, 3)	0.893	<0.001
649	12.152	12.985	11.168	0.623	0.342	0.703	(1, 1, 2)	(1, 1, 2)	0.557	0.001
1012	11.550	13.185	11.552	0.613	0.578	0.702	(1, 2, 1)	(1, 2, 3)	0.742	0.002
730	12.095	13.908	11.708	0.800	0.557	0.989	(1, 2, 1)	(1, 2, 3)	0.725	0.002
1024	12.042	13.896	12.061	0.702	0.805	0.691	(1, 2, 1)	(1, 2, 3)	0.763	0.002
1020	11.798	13.240	10.899	1.213	0.492	0.643	(1, 2, 1)	(1, 2, 3)	0.430	0.003
936	10.942	12.778	11.289	0.449	1.073	0.462	(1, 2, 1)	(1, 2, 3)	0.773	0.004
132	11.574	13.109	11.202	0.673	0.420	1.031	(1, 2, 1)	(1, 2, 3)	0.531	0.004

Note from Tables 1 and 2 that cases with variances well apart are not identified by the Tuk, they are identified by the MCBF. Examples are cases 661 and 847 (see Table 1). In accordance with our simulation results, here the MCBF is capable of identify differentially expressed cases which are not identified by the Tuk, specially, genes with differences in means and/or variances.

5 Discussion

In this paper, we propose a hierarchical Bayesian approach via Dirichlet process prior to develop a gene-by-gene multiple comparison analysis. The proposed approach is a semi-parametric Bayesian model with priors on the parameters $\theta_1, \theta_2, \ldots, \theta_M$ being non-parametric, sampled from the Dirichlet process. But, the distribution of $\mathbf{X}_m$ given $\theta_m = (\mu_m, \sigma_m^2)$ has a parametric form, given by $\mathbf{X}_m | \mu_m, \sigma_m^2 \sim \mathcal{N}(\mu_m, \sigma_m^2)$, for $m = 1, \ldots, M$.

The comparison among experimental conditions are made by using the posterior probability for hypothesis, which are calculated through the Polya urn scheme using latent variables to indicate the equality or inequality among the experimental conditions. For some particular cases, we described the posterior probabilities in terms of the Bayes factor.

The performance of the proposed MCBF method as well as its comparison with the Tuk was verified on an artificial data sets and on a real data set. Results from the artificial data sets show a better performance of MCBF in relation to Tuk.

From the biological point of view the MCBF may bring to light cases not identified when Tuk is considered. We can observe this fact comparing the results obtained when both methods are applied to the real data set (please see Tables 1 and 2). Moreover, the MCBF can be easily implemented in usual softwares. The source code used in data set analysis was implemented in software R (the Comprehensive R Archive Network, http://cran.r-project.org) and can be obtained by email the authors.

In section data set analysis, we apply the MCBF fixing the mass parameters α equal to 1 and $\sqrt[M-1]{(M-1)!}$. Results for these two values of α are similar and lead to a better performance than Tuk. But, the posterior probabilities can depend greatly on mass parameter α, so careful assessment of α is important. A further development is to consider the proposed approach with one more hierarchical level and to specify a prior distribution on α and its estimation.

Appendix A: Hypothesi with inequality 3 to 3

An way to write the hypothesi with inequality 3 to 3 is

$$H_{t'''} : \Theta_{t'''} = \left(\boldsymbol{\theta}; \theta'_m \neq \theta_{m''} \neq \theta_{m'''} \text{ and } \theta_i = \theta_j, \forall i, j \in \{1, \ldots, M\} \setminus \{m', m'', m'''\}\right)$$

for $m', m'', m''' \in \{1, \ldots, M\}$ and $m' \neq m'' \neq m'''$ and $t''' \in \{((\binom{M}{2}+1)+1, \ldots, T\}$, where $\binom{M}{2}+1$ is the number of hypothesis with inequality 2 to 2 more the null hypothesi.

Appendix B: A control and three treatments

In this case, we have $m = 1, 2, 3, 4$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$. We first apply procedures described in Sections 3.1 and 3.2, to compare control and treatments 1 and 2 and then include treatment 3.

The posterior probabilities for Z_4 are as follows:

(a) If $Z_3 = Z_2 = Z_1$ ($Z_1 = 1, Z_2 = 1, Z_3 = 1$), do

$$P(Z_4 = j|Z_1, Z_2, Z_3, \mathbf{x}) = \begin{cases} \dfrac{3}{3+\alpha B_{41}}, & \text{for } j = 1, \\ \dfrac{\alpha B_{41}}{3+\alpha B_{41}}, & \text{for } j = 2, \text{ i.e., } j \neq 1, \end{cases}$$

where $B_{41} = \frac{\mathbf{I}(D_1)\mathbf{I}(\mathbf{x}_4)}{\mathbf{I}(D_1 \cup \mathbf{x}_4)}$ for $D_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$;

(b) If (b_1) $Z_3 \neq Z_2 = Z_1$ ($Z_1 = 1, Z_2 = 1, Z_3 = 2$) or ($b_2$) $Z_2 \neq Z_3 = Z_1$ ($Z_1 = 1, Z_2 = 2, Z_3 = 1$), then

$$P(Z_4 = j|Z_1, Z_2, Z_3, \mathbf{y}) = \begin{cases} \dfrac{2B_{42}}{B_{41}+2B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 1, \\ \dfrac{B_{41}}{B_{41}+2B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 2, \\ \dfrac{\alpha B_{41}B_{42}}{B_{41}+2B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 3, \text{ i.e., } j \neq 1, 2, \end{cases}$$

where $B_{4j} = \frac{\mathbf{I}(D_j)\mathbf{I}(\mathbf{x}_4)}{\mathbf{I}(D_j \cup \mathbf{x}_4)}$ for $j = 1, 2$, and (b_1) $D_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$ and $D_2 = \{\mathbf{x}_3\}$; (b_2) $D_1 = \{\mathbf{x}_1, \mathbf{x}_3\}$ and $D_2\{\mathbf{x}_2\}$;

(c) If $Z_3 = Z_2 \neq Z_1$ ($Z_1 = 1, Z_2 = 2, Z_3 = 2$), then

$$P(Z_4 = j|Z_1, Z_2, Z_3, \mathbf{x}) = \begin{cases} \dfrac{B_{42}}{2B_{41}+B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 1, \\ \dfrac{2B_{41}}{2B_{41}+B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 2, \\ \dfrac{\alpha B_{41}B_{42}}{2B_{41}+B_{42}+\alpha B_{41}B_{42}}, & \text{for } j = 3, \text{ i.e., } j \neq 1, 2, \end{cases}$$

where $B_{4j} = \frac{\mathbf{I}(D_j)\mathbf{I}(\mathbf{x}_4)}{\mathbf{I}(D_j \cup \mathbf{x}_4)}$, for $j = 1, 2$, $D_1 = \{\mathbf{x}_1\}$ and $D_2 = \{\mathbf{x}_2, \mathbf{x}_3\}$;

(d) If $Z_3 \neq Z_2 \neq Z_1$ ($Z_1 = 1, Z_2 = 2, Z_3 = 3$), then do

$$P(Z_4 = j | Z_1, Z_2, Z_3, \mathbf{x}) = \begin{cases} \dfrac{B_{42}B_{43}}{B_{41}B_{42} + B_{41}B_{43} + B_{42}B_{43} + \alpha B_{41}B_{42}B_{43}}, \\ \quad \text{for } j = 1, \\ \dfrac{B_{41}B_{43}}{B_{41}B_{42} + B_{41}B_{43} + B_{42}B_{43} + \alpha B_{41}B_{42}B_{43}}, \\ \quad \text{for } j = 2, \\ \dfrac{B_{41}B_{42}}{B_{41}B_{42} + B_{41}B_{43} + B_{42}B_{43} + \alpha B_{41}B_{42}B_{43}}, \\ \quad \text{for } j = 3, \\ \dfrac{\alpha B_{41}B_{42}B_{43}}{B_{41}B_{42} + B_{41}B_{43} + B_{42}B_{43} + \alpha B_{41}B_{42}B_{43}}, \\ \quad \text{for } j = 4, \text{ i.e., } j \neq 1, 2, 3, \end{cases}$$

where $B_{4j} = \frac{\mathbf{I}(D_j)\mathbf{I}(\mathbf{x}_4)}{\mathbf{I}(D_j \cup \mathbf{x}_4)}$, $D_j = \{\mathbf{x}_j\}$ for $j = 1, 2, 3$.

Appendix C: TPR, TDR and FDR

(i) The true positive rate (TPR) is given by the number of hypothesis correctly identified divided by N, i.e.,

$$\text{TPR} = \frac{\sum_{g=1}^{n} \mathbb{I}_{Z_g^{\text{method}}}(\mathbb{G}_g)}{N}, \tag{13}$$

where $\mathbb{I}_{Z_g^{\text{MC}}}(\mathbb{G}_g) = 1$ if configuration identified by the method is equal to $\mathbb{G}_g$ and $\mathbb{I}_{Z_g^{\text{MC}}}(\mathbb{G}_g) = 0$ otherwise, for method $=$ {MCBF, Tuk};

(ii) The true discovery rate (TDR) is given by the number of true positives (number of hypothesis H_j, $j = 1, 2, 3, 4$, correctly identified) divided by the number of rejected null hypothesis, i.e.,

$$\text{TDR} = \frac{\sum_{g=1}^{n} \mathbb{I}_{Z_g^{\text{method}}}(\mathbb{G}_g) \cdot (1 - \mathbb{I}_{\mathbb{Z}_g^{\text{method}}}(Z_0))}{N - \sum_{g=1}^{n} \mathbb{I}_{\mathbb{Z}_g^{\text{method}}}(Z_0)}, \tag{14}$$

where $\mathbb{I}_{\mathbb{Z}_g^{\text{method}}}(Z_0) = 1$ if configuration identified is equal to configuration Z_0 of the null hypothesi H_0 and $\mathbb{I}_{\mathbb{Z}_g^{\text{method}}}(Z_0) = 0$ otherwise, for method $=$ {MCBF, Tuk};

(iii) The false discovery rate is given by the number of false positives (number of null hypothesi incorrectly rejected) divided by the number of rejected null hypothesi, given by

$$\mathrm{FDR} = \frac{\sum_{g=1}^{n}(1 - \mathbb{I}_{Z_g^{\mathrm{method}}}(\mathbb{G}_g)) \cdot \mathbb{I}_{\mathbb{G}_g}(Z_0)}{N - \sum_{g=1}^{n} \mathbb{I}_{\mathbb{Z}_g^{\mathrm{method}}}(Z_0)}, \tag{15}$$

where $\mathbb{I}_{\mathbb{G}_g}(Z_0) = 1$ if case g ($\mathbb{G}_g$) is generate according to configuration Z_0 of the null hypothesi H_0 and $\mathbb{I}_{\mathbb{G}_g}(Z_0) = 0$ otherwise, for method = {MCBF, Tuk}.

Appendix D: Results for $M = 4$

For this case, all 15 hypothesis are described in Table 3.

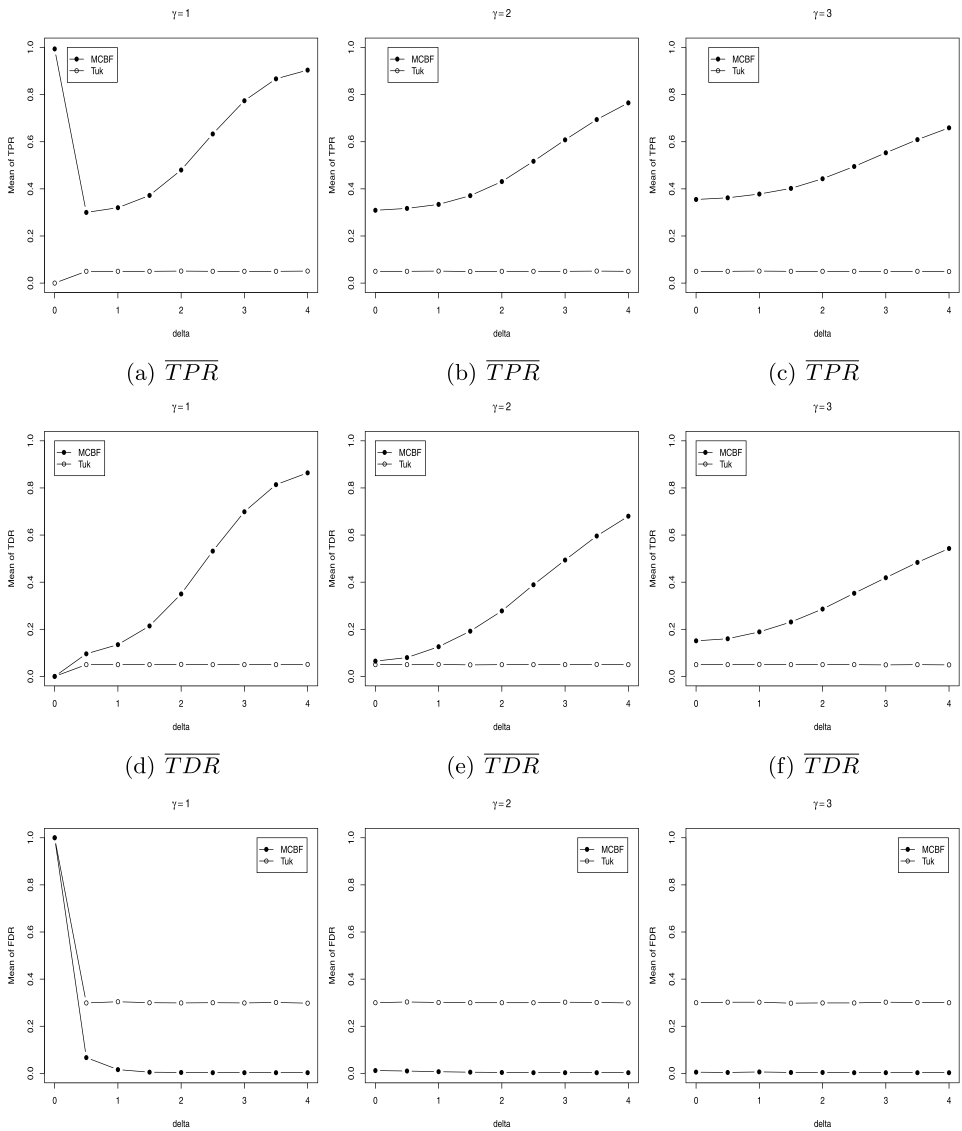
We fix proportions generated from each hypothesis as 0.30 from H_0 and 0.05 from H_j, $j = 1, \ldots, 14$. The data are generate in a similar way as made for $M = 3$. For example, if $u_g \leq 0.30$ we fix parameters according to H_0 and we set up $\mathbb{G} = (1, 1, 1, 1)$; if $0.30 < u_g \leq 0.35$ we fix parameters according to H_1 and we set up $\mathbb{G} = (1, 1, 1, 2)$. So, we generate $\mathbf{X}_j = (X_{j1}, \ldots, X_{jn}) \sim \mathcal{N}(\mu_j, \sigma_j^2)$, $j = 1, 2, 3, 4$.

For this case, $M = 4$, we obtain from Gopalan and Berry's (1998) elicitation procedure, $\alpha = \sqrt[3]{6}$.

Graphics in Figures 7 and 8 show performance of both methods for $n = 5$, for Tuk with significance level at 0.05 and MCBF with $\alpha = 1$ and $\alpha = \sqrt[3]{6}$, respectively. Graphics in Figures 9 and 10 show performance of both methods for $n = 10$. As for $M = 3$, the proposed MCF present higher TPR and TDR and smaller FDR.

Table 3 *Hypothesis for a control and three treatment experimental conditions*

Hypothesis	Hypothesis	Hypothesis
$H_0 : Z_1 = Z_2 = Z_3 = Z_3$	$H_5 : Z_1 = Z_2 \neq Z_3 = Z_4$	$H_{10} : Z_1 = Z_4 \neq Z_2 \neq Z_3$
$H_1 : Z_1 = Z_2 = Z_3 \neq Z_4$	$H_6 : Z_1 = Z_3 \neq Z_2 = Z_4$	$H_{11} : Z_1 \neq Z_2 = Z_3 \neq Z_4$
$H_2 : Z_1 = Z_2 = Z_4 \neq Z_3$	$H_7 : Z_1 = Z_4 \neq Z_2 = Z_3$	$H_{12} : Z_1 \neq Z_2 = Z_4 \neq Z_3$
$H_3 : Z_1 = Z_3 = Z_4 \neq Z_2$	$H_8 : Z_1 = Z_2 \neq Z_3 \neq Z_4$	$H_{13} : Z_1 \neq Z_2 \neq Z_3 = Z_4$
$H_4 : Z_1 \neq Z_2 = Z_3 = Z_4$	$H_9 : Z_1 = Z_3 \neq Z_2 \neq Z_4$	$H_{14} : Z_1 \neq Z_2 \neq Z_3 \neq Z_4$

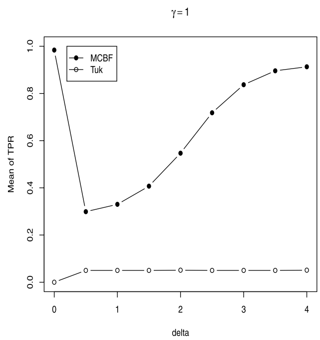


(a) $\overline{TPR}$ (b) $\overline{TPR}$ (c) $\overline{TPR}$

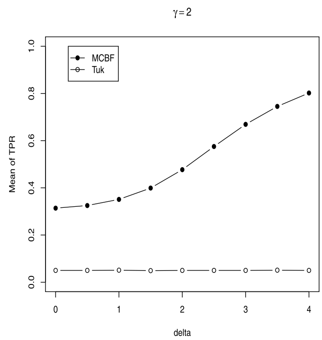
(d) $\overline{TDR}$ (e) $\overline{TDR}$ (f) $\overline{TDR}$

(g) $\overline{FDR}$ (h) $\overline{FDR}$ (i) $\overline{FDR}$

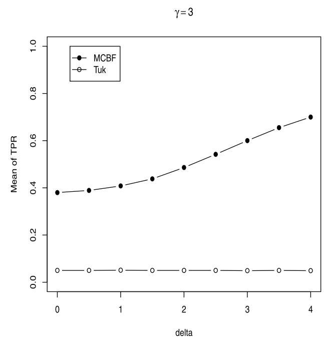
Figure 7 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 4$ *and* $n = 5$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = 1$.



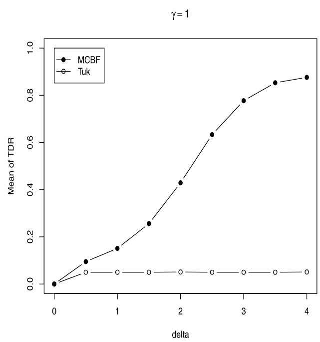
(a) $\overline{TPR}$



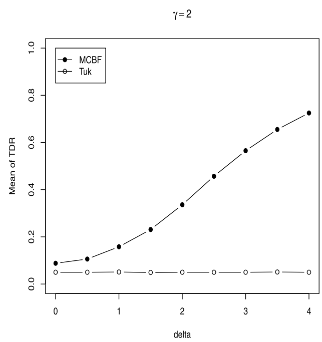
(b) $\overline{TPR}$



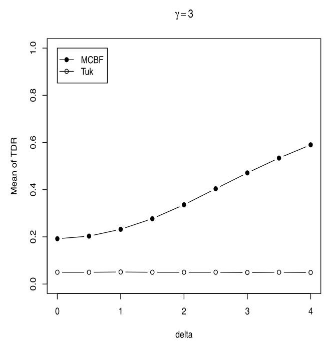
(c) $\overline{TPR}$



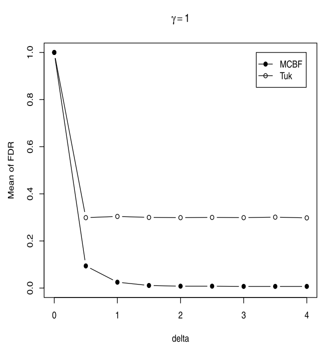
(d) $\overline{TDR}$



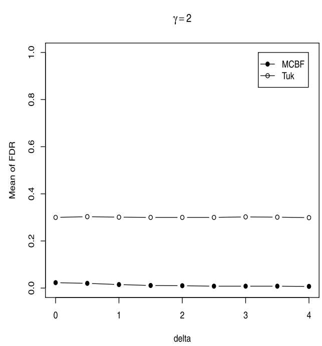
(e) $\overline{TDR}$



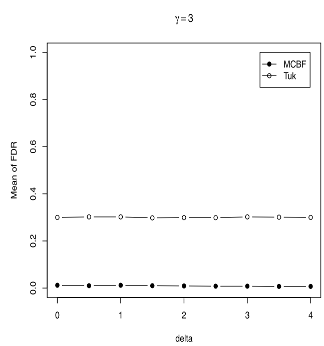
(f) $\overline{TDR}$



(g) $\overline{FDR}$



(h) $\overline{FDR}$



(i) $\overline{FDR}$

Figure 8 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 4$ *and* $n = 5$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = \sqrt[3]{6}$.

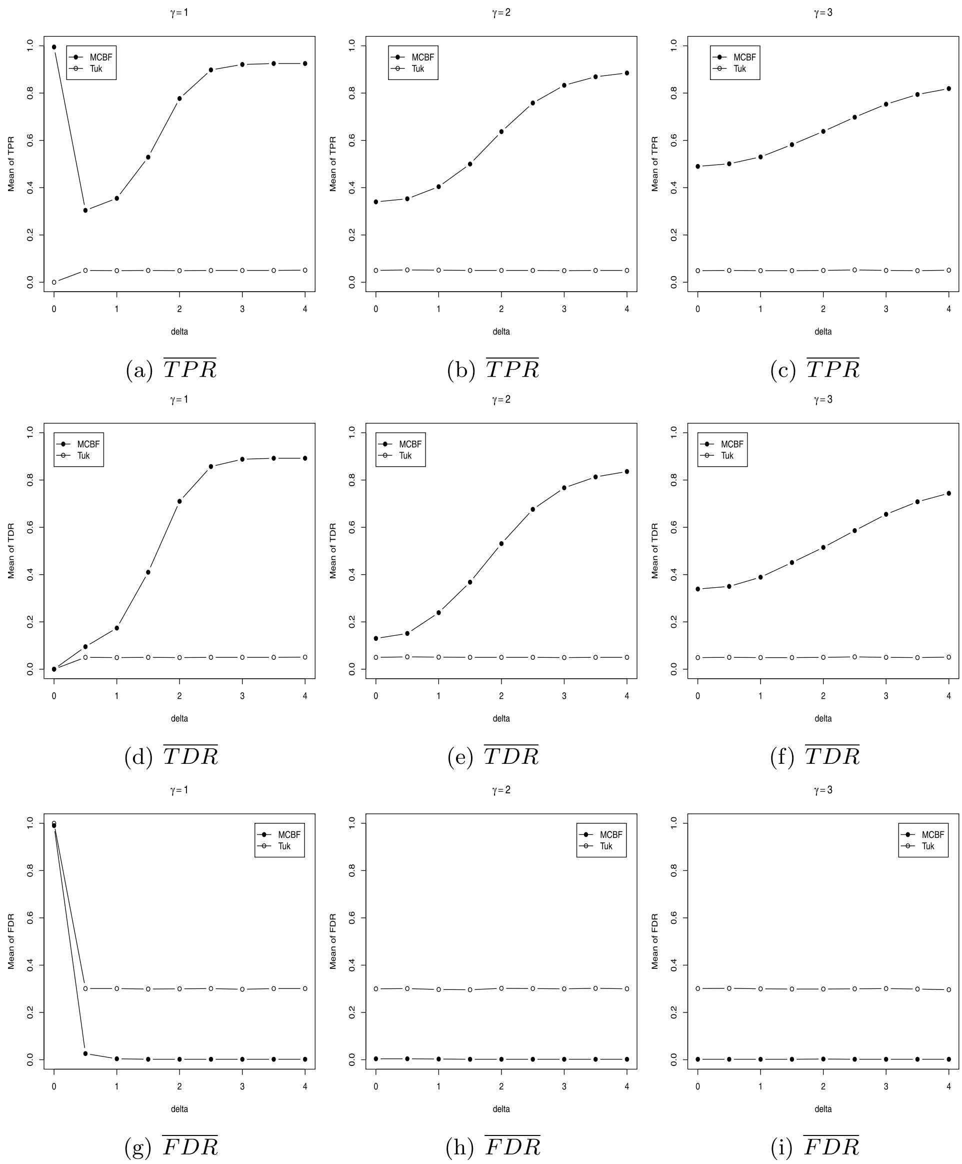
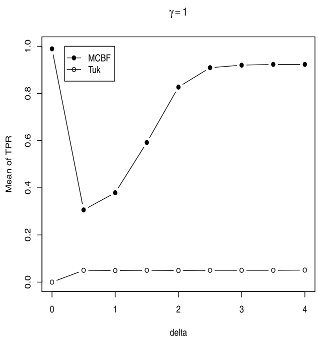
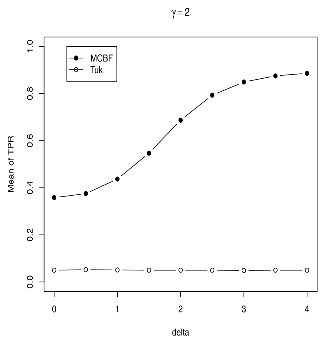


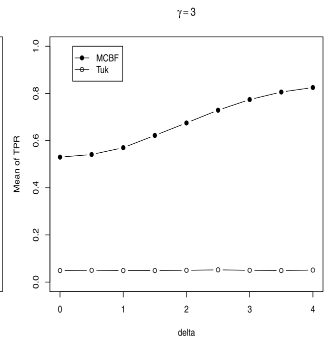
Figure 9 $\overline{\mathrm{TPR}}$, $\overline{\mathrm{TDR}}$ *and* $\overline{\mathrm{FDR}}$ *by method*, $M = 4$ *and* $n = 10$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = 1$.



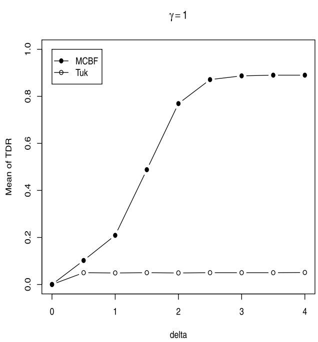
(a) $\overline{TPR}$



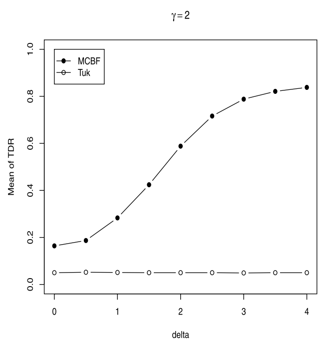
(b) $\overline{TPR}$



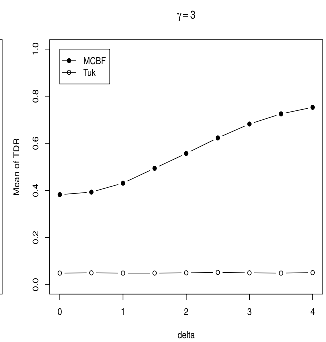
(c) $\overline{TPR}$



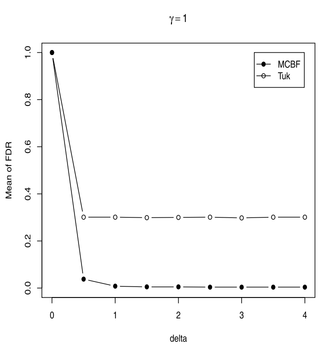
(d) $\overline{TDR}$



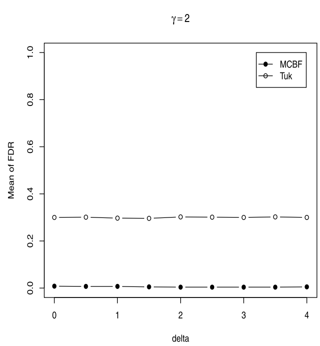
(e) $\overline{TDR}$



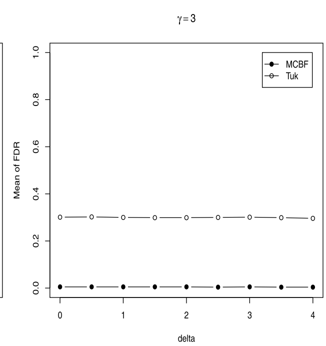
(f) $\overline{TDR}$



(g) $\overline{FDR}$



(h) $\overline{FDR}$



(i) $\overline{FDR}$

Figure 10 $\overline{\text{TPR}}$, $\overline{\text{TDR}}$ *and* $\overline{\text{FDR}}$ *by method*, $M = 4$ *and* $n = 10$. *Tuk with significance level at* 0.05 *and MCBF with* $\alpha = \sqrt[3]{6}$.

Acknowledgments

We thank the Editor and the referees for their comments, suggestions and criticisms which have led to improvements of this article. The researches of Francisco Louzada is supported by the Brazilian organization CNPq.

References

Antoniak, C. E. (1974). Mixture of processes Dirichlet with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174. MR0365969

Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000). Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor. *The Journal of Biological Chemistry* **275**, 29672–29684.

Baldi, P. and Long, D. A. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bhattacharya, S. (2008). Gibbs sampling based Bayesian analysis of mixtures with unknown number of components. *Sankhyā* **70**, 133–155. MR2507480

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distribution via Polya urn schemes. *The Annals of Statistics* **1**, 353–355. MR0362614

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003). A comparison of normalization methods for high dendity oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Casella, G., Robert, C. and Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology* **1**, 1–18.

Chen, J. J., Delongchamp, R. R., Tsai, C.-A., Hsueh, H.-m., Sistare, F., Thompson, K. L., Desai, V. G. and Fuscoe, J. C. (2004). Analysis of variance components in gene expression data. *Bioinformatics* **20**, 1436–1446.

Cox, D. R. and Reid, N. M. (2000). *The Theory of Design of Experiments*. London: Chapman & Hall/CRC.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277. MR1266299

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588. MR1340510

Ferguson, S. T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. MR0350949

Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics* **7**, 126.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene set: Methodological issues. *Bioinformatics* **23**, 980–987.

Gopalan, R. and Berry, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association* **93**, 1130–1139. MR1649207

Hatifield, G. W., Hung, S. and Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. *Molecular Microbiology* **47**, 871–877.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probel level data. *Nucleic Acids Research* **31**, e15.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **13**, 158–182. MR2044876

Jain, S. and Neal, R. (2007). Splitting and merging components of a nonconjugated Dirichlet process mixture model. *Bayesian Analysis* **2**, 445–472. MR2342168

Kass, R. and Raftery, A. (1995). Bayes factor. *Journal of the American Statistical Association* **90**, 773–795.

Lonnstedt, I. and Speed, T. P. (2001). Replicated microarray data. *Statistical Sinica* **12**, 31–46. MR1894187

Louzada, F., Saraiva, E. F., Milan, L. A. and Cobre, J. (2014). A predictive Bayes factor approach to identify genes differentially expressed: An application to *Escherichia coli* bacterium data. *Brazilian Journal of Probability an Statistics* **28**, 167–189.

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

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265. MR1823804

Parkitna, J. R., Korostynski, M., Kaminska-Chowaniec, D., Obara, I., Mika, J., Przewlocka, B. and Przewlocki, R. (2006). Comparison of gene expression profiles in neuropathic and inflammatory pain. *Journal of Physiology and Pharmacology* **57**, 401–414.

Pavlids, P. (2003). Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* **31**, 282–289.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Shapiro, C. P. (1977). Classification by maximum posterior probability. *The Annals of Statistics* **5**, 185–190. MR0431530

Smyth, G. K. and Speed, T. P. (2003). Normalization of cDNA microarray data. *Methods* **31**, 265–273.

Wu, T. D. (2001). Analyzing gene expression data from DNA microarray to identify candidates genes. *Journal of Pathology* **195**, 53–65.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

INMA
Universidade Federal de Mato Grosso do Sul
Campo Grande, MS
Brazil
E-mail: erlandson.saraiva@ufms.br

ICMC
Universidade de São Paulo
São Carlos, SP
Brazil
E-mail: louzada@icmc.usp.br