Journal of Data Science 14(2016), 491-508

# A Bayesian multiple comparison approach for gene expression data analysis

Erlandson F. Saraiva[1*], Lu´ıs A. Milan[2]

[1]*Institute of Mathematics, Federal University of Mato Grosso do Sul*

[2]*Department of Statistcs, Federal University of Sao Carlos*

*Abstract:* Methods used to detect differentially expressed genes in situations with one control and one treatment are t-tests. These methods do not per- form well when control and treatment variances are different. In situations with a control and more than one treatment, it is common to apply analysis of variance followed by a Tukey and/or Duncan test to identify which treat- ment caused the difference. We propose a Bayesian approach for multiple comparison analysis which is very useful in the context of DNA microarray experiments. It uses a priori Dirichlet process and Polya urn scheme. It is a unified procedure (for cases with one or more treatments) which detects differentially expressed genes and identify treatments causing the difference. We use simulations to verify the performance of the proposed method and compare it with usual methods. In cases with control and one treatment and control and more than one treatment followed by Tukey and Duncan tests, the method presents better performance when variances are different. The method is applied to two real data sets. In these cases, genes not detected by usual methods are identified by the proposed method.

*Key words*: Gene Expression, Bayesian approach, Prior Dirichlet process, Polya urn, Multiple comparison.

## 1. Introduction

A common interest in gene expression data analysis is to identify genes with different expression levels. Identifying these genes allows us to detect relationships between genes and between genes and proteins; also, it allows us to identify which genes are involved in the origin and/or evolution of diseases with genetic origin, or which genes react to a drug stimulus (Schena et al., 1995; Allison et al., 2006; DeRisi et al., 1997; Arfin et al., 2000; Lonnstedt and Speed, 2001; Rosenfeld, 2007; Wu, 2001).

Gene expression data can be analyzed in at least three levels of increasing complexity (Baldi and Long, 2001). In the first level, each gene is analyzed separately and the purpose is to verify whether the expression levels are different in treatment and control conditions. In the second

level, clusters of genes are analyzed in terms of common functionalities and interactions. In the third level, the purpose is to understand the relationship between genes and proteins.

Here we focus on identifying differentially expressed genes. One of the first proposed approaches was the fold-change (Schena et al., 1995; Allison et al., 2006), where a gene is considered differentially expressed if the average of the logarithm of the observed expression levels in treatment and control differ by more than a cutoff value, Rc, which is previously prefixed.

Another method used for gene expression data analysis is the two-sample t- test (T T ), see Baldi and Long (2001) and Hatfield et al. (2003). A limitation for applying T T to gene expression data is the usual small sample size, which may lead to underestimated variances and low power of test. To avoid such limitations, some T T modifications were proposed, such as the Cyber-t (CT) proposed by Baldi and Long (2001) and the Bayesian t-test (BTT) proposed by Fox and Dimmic (2006). Basically, these methods modify the standard error estimate of the two sample differences found in the denominator of the standard t statistic.

These methods can be applied to only two experimental conditions (control and treatment). This is a drawback since in many microarray experiments, the gene expression response is monitored under M , M > 2, treatment conditions.

The interest is to consider all M (M − 1)/2 pairwise treatments, (Dudoit et al., 2003).

We propose a hierarchical Bayesian approach with a priori Dirichlet process, which compares two or more experimental conditions. The comparison is made using the a posteriori probabilities for models in a model selection procedure. The a posteriori probabilities are calculated using the Polya urn scheme (Blackwell and MacQueen, 1973).

In order to verify the performance of the proposed method (denoted by PU ), we carried out a simulation study with small sample sizes, which is usual in gene expression data analysis. In situations with a control and a treatment, we compared the performance of PU with T T , CT and BT T . We also considered situations with a control and two and three treatments. In these cases, we com- pared the performance of PU with analysis of variance (ANOVA) followed by the Tukey-test (Cox and Reid, 2000) and Duncan-test (Duncan, 1955). ANOVA was applied to identify differentially expressed genes. As it does not discrimi- nate which treatments differ from one another, we applied the Tukey-test and Duncan-test as a post hoc test.

The comparison among the methods was made in terms of the true positive rate and the false discovery rate. The simulation study showed a better perfor- mance for PU , i.e., the greater true positive rate and the smaller false discovery rate, for cases with different mean and variance. We also applied these meth- ods to two real data sets. The first is from an experiment with Escherichia coli bacterium with a control and a treatment condition (Arfin et al., 2000). The second refers to Plasmodium falciparum protein microarray with a control and two treatments, obtained from the website cybert.ics.uci.edu (Baldi and Long, 2001).

This paper is structured using the hierarchical Bayesian model for the gene expression data analysis, described in Section 2. The a posteriori probabilities are calculated using the Polya urn scheme and the Bayes factor in Section 3. In Section 4, we compare the performance of the

proposed approach and methods usually considered. Section 5 concludes the paper with final remarks.

## 2. Bayesian Model

Consider an experiment with N genes under experimental conditions $E_1, \ldots, E_M$ and $E_1$ as the control. For each gene g, g = 1, . . . , N , let $\{y_{1m}, \ldots, y_{nmm}\}$be the set of measurements of log-expression levels in experimental condition m, where nm is the sample size, for m = 1, . . . , M . Although this is not necessary for the proposed approach, to simplify notation we assume a balanced design common to all genes so that $n_m = n$, for m = 1, . . . , M .

Let $Y = \{Y_1, \ldots, Y_M\}$ be the set of all observed expression levels for gene g, where $Ym = (y_{1m}, \ldots, y_{nm})'$ is a $n \times 1$ vector of conditionally independent observations for treatment m. Assume $Y_{im} \sim F(\theta_m)$, where $F(\theta_m)$ is a parametric distribution indexed by unknown parameters θm. Denote the parametric space by $\Theta = \{\theta = (\theta_1, \ldots, \theta_M); \theta_m \in R^d$, where d is the dimension of $\theta_m$, m = 1, . . . , M }.

Our interest is to verify whether a gene g is differentially expressed in different experimental conditions, i.e., we search for a model which best fits the data and meets these conditions. These models can be described as $M_0 : \Theta_0 = \{\theta; \theta_1 = \ldots = \theta_M\})$; $M_1 : \Theta 1 = \{\theta; \theta_1 \neq \theta_2, \theta_2 = \theta_3 = \ldots = \theta_M\}$, or $M_2 : \Theta_2 = \{\theta; \theta_1 = \theta_2, \theta_2 \neq \theta_3, = \theta_3 = \ldots = \theta_M\}$, and successively for all combinations until $M_T : \Theta_T = \{\theta; \theta_1 \neq \ldots \neq \theta_M\}$.

The equality (or not) of $\theta_m$'s determines partitions in parameter space Θ, i.e., $\Theta_0, \Theta_1, \ldots, \Theta_T$ are disjointed and $\Theta_0 \cup \Theta_1 \cup \ldots \cup \Theta_T = \Theta$. This allows us to develop a hierarchical Bayesian approach using an a priori Dirichlet process (DP ) on $\theta_1, \ldots, \theta_M$ in order to make simultaneous comparisons of $\theta_m$'s (Gopalan and Berry, 1998; Neal, 2000). This exploits the discreteness of the Dirichlet process that allows parameters to be coincident with positive probability.

We assume the following semi-parametric Bayesian model (see Ferguson (1973) and Antoniak (1974)),

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\theta} &\sim F(\boldsymbol{\theta}) \\ \boldsymbol{\theta}|G &\sim G \\ G \mid \alpha, G_0 &\sim DP(\alpha G_0). \end{aligned} \tag{1}$$

Integrating tt over its a priori distribution in (1), θ follows the Polya urn scheme (Blackwell and MacQueen, 1973), and can be written as

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

where $\theta_{m-1} = (\theta_1, \ldots, \theta_{m-1})$, $I_{\theta m}(\theta_j) = 1$ if $\theta_m = \theta_j$ and $I_{\theta m}(\theta_j) = 0$ otherwise, for $j \in \{1, \ldots, m-1\}$ and $m \in \{2, \ldots, M\}$.

Note that at each step of the sampling procedure defined in (2), $\theta_m$ replicates one of the previous $\theta_j$'s, with probability $\frac{1}{a+m-1}\sum_{j=1}^{m-1} I_{\theta m}(\theta_j)$, or it assumes a new value, generated from the base distribution $G_0$, with probability $\frac{1}{a+m-1}$ Thus, a sample from the joint distribution of $\theta_1, \ldots, \theta_M$ yields $k$ groups ($1 \leqq k \leqq M$) of $\theta_m$'s with distinct values, $\emptyset\ 1, \ldots, \emptyset\ k$, generated from the base distribution $G_0$.

## 2.1 A priori Dirichlet process via latent variables

Consider the latent variables $c = (c1, \ldots, cM)$ in a way that $cm = j$ indicates that $\theta m = \varphi j$, $\varphi j \sim G0$ for $m = 1, \ldots, M$ and $j = 1, \ldots, k$. c classifies the observed data $y = (y1, \ldots, yM)$ in k groups, $\{D1, \ldots, Dk\}$, where $Dj = \{ym; cm = j\}$ with $\bigcup_{j=1}^{k} Dj = y$.

The likelihood function for c is

$$L(\mathbf{c}/\mathbf{y}) = \prod_{j=1}^{k} P(D_j) \tag{3}$$

where

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

and $\pi G0(\cdot)$ is the density of the base distribution G0.

Letting nj be the number of observations in Dj given the configuration $cm-1 = (c1, \ldots, cm-1)$, the Polya urn scheme in (2) can be described by

(i) Initialize $c_1 = 1$, $k = 1$, $D_1 = \{\mathbf{y}_1\}$ and generate $\phi_1 \sim G_0$.

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

$$P(c_m = j|\mathbf{c}_{m-1}) = \frac{n_j}{\alpha + m - 1} \tag{5}$$

$$P(c_m = j^*|\mathbf{c}_{m-1}) = \frac{\alpha}{\alpha + m - 1}, \tag{6}$$

where $j^* = k + 1$ and $j = 1, \ldots, k$.

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

(b) If $c_m = j^*$, set $D_{j^*} = \{\mathbf{y}_m\}$ and generates $\phi_{j^*}$ from base distribution $G_0$, $\phi_{j^*} \sim G_0$. The number of groups increases by one unit, $k = k + 1$.

(iii) Conditional on $\mathbf{c} = (c_1, \ldots, c_M)$, set $\theta_m = \phi_j$ for all $c_m = j$, $j = 1, \ldots, k$.

## 3. Multiple comparison

Updating the a priori probabilities in (5) and (6) via the likelihood function in (3), we obtain the conditional a posteriori probabilities

$$P(c_m = j|\mathbf{c}_{m-1}, \mathbf{y}) = b\frac{n_j}{\alpha + m - 1}\frac{P(D_j \cup \mathbf{y}_m)}{P(D_j)}, \quad (7)$$

and

$$P(c_m = j^*|\mathbf{c}_{m-1}, \mathbf{y}) = b\frac{\alpha}{\alpha + m - 1}P(\mathbf{y}_m), \quad (8)$$

where b is the normalizing constant and P ( ·) is given by (4).

In order to specify the mass parameter α, from (5) and (6), we define

$$P(M_0) = \frac{\alpha(M-1)!}{\prod_{m=1}^{M}(\alpha + m - 1)} \quad \text{and} \quad P(M_T) = \frac{\alpha^M}{\prod_{m=1}^{M}(\alpha + m - 1)},$$

See Gopalan and Berry (1998). Setting P (M0)/P (MT ) = 1, we obtain

$$\alpha = \begin{cases} 1 & \text{for } M = 2 \\ \sqrt[M-1]{(M-1)!} & \text{for } M \geq 3. \end{cases} \quad (9)$$

### 3.1 Particular cases

Now we show some particular cases of (7) and (8).

#### 3.1.1 Control and one treatment

In this case, we have M = 2, y = ($y_1$, $y_2$) and α = 1. Initialize with $c_1 = 1$ and $D_1 = \{y_1\}$.

Thus, from (7) and (8), we have

$$P(c_2 = 1|c_1 = 1, \mathbf{y}) = \frac{P(D_1 \cup \mathbf{y}_2)}{P(D_1 \cup \mathbf{y}_2) + \alpha P(D_1)P(\mathbf{y}_2)}$$

and

$$P(c_2 = 2|c_1 = 1, \mathbf{y}) = \frac{\alpha P(D_1)P(\mathbf{y}_2)}{P(D_1 \cup \mathbf{y}_2) + \alpha P(D_1)P(\mathbf{y}_2)}.$$

Let B21 = $\frac{P(D1)P(y2)}{P(D1 \cup y2)}$ be the Bayes factor (Kass and Raftery, 1995). Compare models with the first assuming $Y_1 \sim F(\varphi_1)$ and $Y_2 \sim F(\varphi_2)$, for $\varphi_1 \neq \varphi_2$, and the second assuming $Y_1, Y_2 \sim F(\varphi_1)$. Thus

$P(c_2 = 1/c_1 = 1, y) = \frac{1}{1+\alpha B_{21}}$ and $P(c = 2|c = 1, y) = \frac{\alpha B_{21}}{1+\alpha B_{21}}$

If $P(c_2 = 2|c_1 = 1, y) > P(c_2 = 1|c_1 = 1, y)$ do $c_2 = 2$. In this case, there is evidence for a difference between the control and the treatment. Otherwise, do c2 = c1 = 1 considering that there is not enough evidence for the difference.

### 3.1.2 Control and two treatment

We now have M = 3, y = (y1, y2, y3) and $\alpha = \sqrt{2}$. Apply the procedure in 3.1.1 to classify treatment 1 and define c2, then do the following procedure.

(i)If c2 = c1 = 1, i.e., treatment 1 does not differ in relation to the control,do $D_1 = \{y_1, y_2\}$.

The a posteriori probabilities for $c_3$ are given by

$$P(c_3 = j \mid c_2, y) = \begin{cases} \frac{2}{2+\alpha B_{31}}, & for\ j = 1 \\ \frac{\alpha B_{31}}{2+\alpha B_{31}}, & for\ j = 2 \end{cases}$$

and $B_{31} = \frac{P(D1)P(y3)}{P(D1 \cup y3)}$, where $c_2 = (c_1 = 1, c_2 = 1)$.

If $P(c_3 = 2 \mid c_2, y) > P(c_3 = 1 \mid c_2, y)$ do $c_3 = 2$, otherwise do $c_3 = 1$.

(ii) If $c_2 \neq c_1$ ($c_1 = 1$ and $c_2 = 2$) then $D_1 = \{y_1\}$ and $D_2 = \{y_2\}$.

The a posteriori probabilities for c3 are

$$P(c_3 = j|\mathbf{c}_2, \mathbf{y}) = \begin{cases} \frac{B_{32}}{B_{31}+B_{32}+\alpha B_{31}B_{32}}, & \text{for } j = 1 \\ \frac{B_{31}}{B_{31}+B_{32}+\alpha B_{31}B_{32}}, & \text{for } j = 2 \\ \frac{\alpha B_{31}B_{32}}{B_{31}+B_{32}+\alpha B_{31}B_{32}}, & \text{for } j = 3, \end{cases}$$

$$\text{where } B_{3j} = \frac{P(D_j)P(\mathbf{y}_3)}{P(D_j \cup \mathbf{y}_3)}, \text{ for } j = 1, 2.$$

Do $c_3 = \underset{j=1,2,3}{\text{argmax}}(P(c3 = j| \cdot))$.

In Appendix 1 of the additional matter (AM), we present a posteriori probabilities for the case with a control and three treatments.

### 3.1.3 Algorithm for the general case

The proposed method can be expressed as:

(i) Initialize with $c_1 = 1$, $D_1 = \{y_1\}$, k = 1 and fix α according to (9);

(ii) For $m = 2, \ldots, M$ do

(a) Calculate $P(Dj)$, $P(Dj \cup \mathbf{y}m)$ and $P(\mathbf{y}m)$ according to (4), for $j = 1,\ldots,k$;

(b) Calculate $P(c_m = j|c_{m-1}, y) \propto \frac{n_j}{\alpha+m-1}\frac{P(D_j \cup y_m)}{P(Dj)}$, as in (7); $P(c_m = k+1|c_{m-1}, y) \propto \frac{n_j}{\alpha+m-1} P(Ym)$, as in (8);

(c) If $P(c_m = j| \cdot) = \max_{j=1,..,k}(P(cm = j| \cdot), P(c_m = k+1| \cdot))$, for $j \in \{1,\ldots k\}$, do $D_j = D_j \cup y_m$ and $n_{j+1}$. Else do $D_{k+1} = \{y_m\}$, $n_{k+1} = 1$ and k = k+1;

Given the configuration $c = (c_1 \ldots, c_M)$, if $cm = 1$ for all $m = 1, \ldots, M$, we select model $M_0$, otherwise, if at least one $c_m \neq 1$, we select another model.Choosing $M_0$ means no differentially expressed gene for all treatments while in any other model it would imply at least one treatment with a differentially expressed gene.

## 4. Data Analysis

We observed the performance of PU and compared it with standard methods using the simulation.

Gene expression levels in control and treatments are generated from normal distribution, $Y_{im}|\mu m, \sigma^2 \sim N(\mu m, \sigma^2)$, for $i = 1, \ldots, n$ and $m = 1, \ldots, M$ .The normal assumption for expression data (log-transformed) is usual in gene expression data analysis, see for example Baldi and Long (2001); Hatfield et al.(2003); Fox and Dimmic (2006); Saraiva and Milan (2012); Louzada et al. (2014).

Assume that

$$\mu_m|\sigma_m^2, \lambda \sim \mathcal{N}\left(\mu_0, \frac{\sigma_m^2}{\lambda}\right) \quad \text{and} \quad \sigma_m^2|\tau, \beta \sim \mathcal{IG}\left(\frac{\tau}{2}, \frac{\beta}{2}\right), \tag{10}$$

for $m = 1, \ldots, M$, where $\mu_0$, $\lambda$, $\tau$ and $\beta$ are known hyperparameters and $IG(\cdot)$ is the inverse gamma distribution with mean $\beta/(\tau - 2)$, see Escobar and West (1995) and Casella et al., (2000).

Thus, from (4)

$$P(D_j) = \beta^*\lambda^*\Gamma^*\left[1 + \frac{\sum_{\mathbf{y}_m \in D_j} \mathbf{y}_m^2 + \lambda\mu_0^2}{\beta} - \frac{\left(\sum_{\mathbf{y}_m \in D_j} \mathbf{y}_m + \lambda\mu_0\right)^2}{\beta(n_j + \lambda)}\right]^{-\tau^*}$$

where $\beta^* = \left(\frac{1}{\beta\pi}\right)^{\frac{n_j}{2}}$, $\lambda^* = \left(\frac{\lambda}{n_j+\lambda}\right)^{\frac{1}{2}}$, $\Gamma^* = \frac{\Gamma\left(\frac{\tau+n_j}{2}\right)}{\Gamma\left(\frac{\tau}{2}\right)}$ and $\tau^* = \left(\frac{\tau+n_j}{2}\right)$, for $j = 1, \ldots, k$.

Consider (a, b) as roughly the interval which would include all observations produced by the experiment. We defined the a priori distributions choosing $\tau$ and $\beta$ such that $E[\sigma_m^2] = \frac{\beta}{\tau-2} = R$, where R is range of the interval R = b-a.Thus , we obtain $\beta=(\tau-2) \cdot R$ and we set $\tau$ =3. The hyperparameter $\mu_0$ was chosen to be the middle point of the interval $\mu_0$ = (a+b) /2 .We also set $\lambda$=0.01.

## 4.1 Control and one treatment

For this case, we compared the performance of PU with T T , CT (Baldi and Long, 2001) and the BT T (Fox and Dimmic, 2006).

### 4.1.1 Simulated data sets

We used $\mu_1 = -14$ and $\sigma_1^2 = 0.8$ to simulate the data sets.These values are the mean and variance of the expression levels (log transformed) from the control group of the Escherichia coli bacterium data set. The sample sizes used are 4 and 8.

To verify how the method performs when treatment parameters $\theta_2 = (\mu_2, \sigma_2^2)$ move away from the control parameters $\theta 1 = (\mu_1, \sigma_1^2)$, we simulate this using $\mu 2 = \mu 1 \pm \delta\sigma_1$ and $\sigma 2 = \gamma\sigma_1$, for $\delta \in \{0.0, 0.5, 1.0, 1.5, 2, 2.5, 3\}$ and $\gamma \in \{1, 2, 3\}$.

Data sets were generated to mimic a mix of both differentially and non- differentially expressed genes where the proportion of differentially expressed genes is small. We fixed the proportion of differentially expressed genes at 5%, being 3% and 2% for situations over and under expressed, respectively.

The data sets were simulated following the steps. For g = 1, . . . , N , N = 1000, simulate $u_g \sim U(0, 1)$:

(i) If $u_g \leq 0.95$ fix $\mu_2 = \mu_1$ and $\sigma_2 = \sigma_1$. Let the index variable IIg = 0 to indicate that case g is generated under $M_0$;

(ii) If $0.95 < u_g \leq 0.98$ fix $\mu_2 = \mu_1 + \delta\sigma 1$ and $\sigma 2 = \gamma\sigma 1$. Set IIg = 1 to indicate that case g is generated under M1;

(iii) If $u_g > 0.98$ fix $\mu_2 = \mu_1 - \delta\sigma_1$ and $\sigma_2 = \gamma\sigma_1$. Set IIg = 1 to indicate that case g is generated under $M_1$;

(iv) Simulate Yim $\sim N(\mu_m, \sigma_m^2)$, for *m* = 1, 2 and *i* = 1, . . . , n.

After generating the data sets, we apply PU and t-tests to identify the cases with a difference. To record cases identified with a difference by PU , we consider an index variable IP U = 1 if P (c2 = 2|c1 = 1, y) > 0.5 and IP U = 0 otherwise.Similarly, for T T , CT and BT T , we consider Imethod = 1 (where the method is TT or CT or BTT ) for cases with $p - value < 0.05$ and $II_g^{method}$ = 0 otherwise.

We define as performance indicators the true positive rate, TPr, and the false discovery rate, FDr, as

$$TPr^{method} = \frac{\sum_{g=1}^{n} \mathbb{I}_g \cdot \mathbb{I}_g^{method}}{\sum_{g=1}^{n} \mathbb{I}_g} \quad \text{and} \quad FDr^{method} = \frac{\sum_{g=1}^{n} (1 - \mathbb{I}_g) \cdot \mathbb{I}_g^{method}}{\sum_{g=1}^{n} \mathbb{I}_g^{method}}.$$

When $\delta = 0$ and $\gamma = 1$ no case is generated under the alternative model, for this case $TP_r^{method} = 0$.

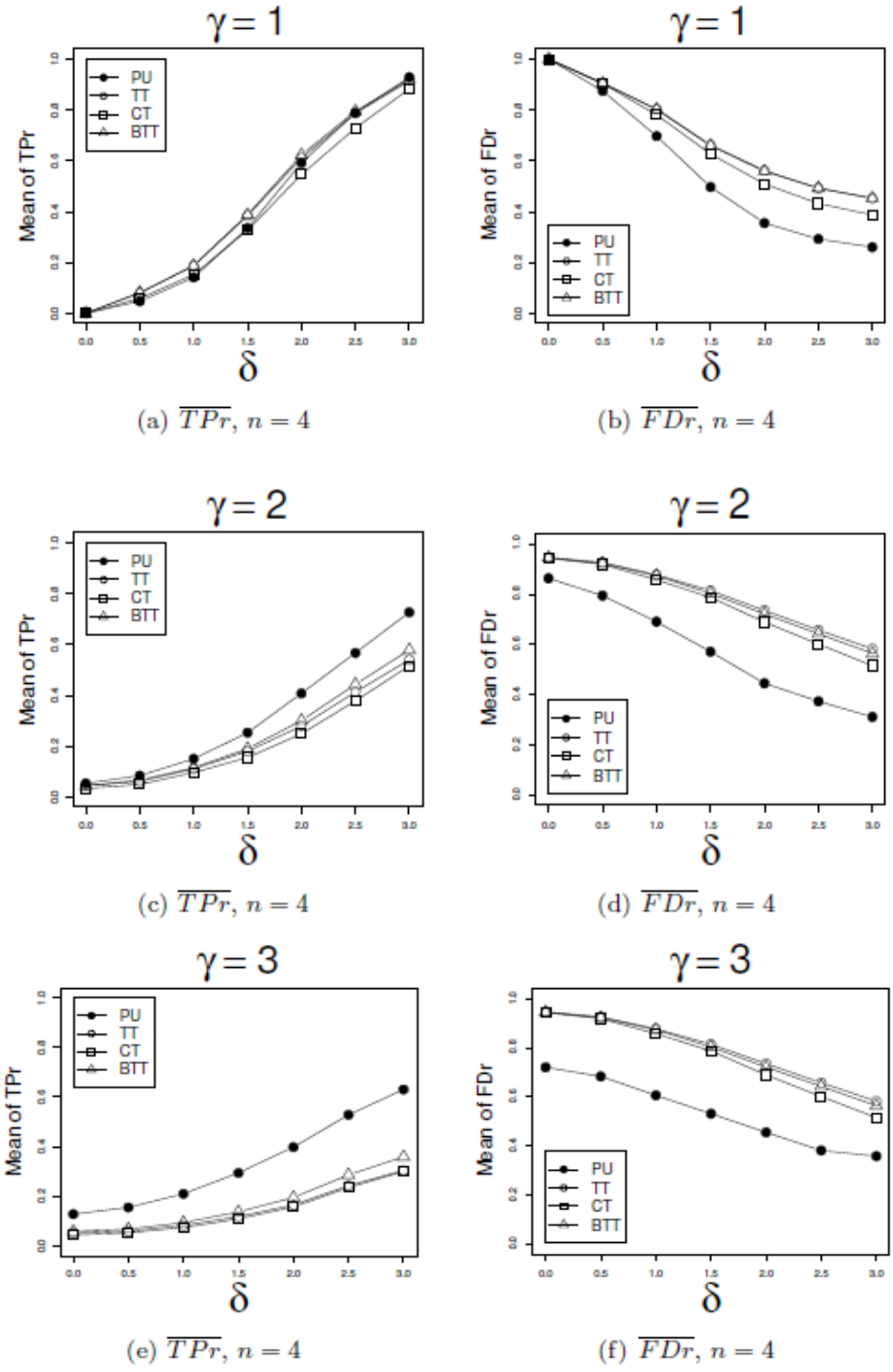
We generate L = 100 different artificial data sets for each pair $(\delta, \gamma)$, as described above. The average values of $TP_r$ and $FD_r$, denoted by $\overline{TPr}^{method}$ and $\overline{FDr}^{method}$, for n = 4, are presented in Figure 1. Figure 1 in Appendix 2 of the additional matter (AM) shows the $\overline{TPr}$ and $\overline{FDr}$ for n = 8.

Average true positive curves are presented in Figure 1-a,c,e. As can be ob- served in Figure 1-a, the performance is similar for n = 4 and all values of δ. For

sample size n = 8 (Figure 1-a of AM), equal variance, $\gamma = 1$, and small changes in mean ($0.5 \leq \delta \leq 2.0$), T T , CT and BT T show a better performance than PU, while for $\delta \geq 2.5$ all methods are similar.

As the difference in variance increases, $\gamma = 2$ and $\gamma = 3$, PU performs better than all methods tested, and the performance improves as the difference increases, as can be observed in Figures 1-c and 1-e.

Figure 1-b,d,f shows the average false positive (also see Figure 1-b,d,f of the AM). Considering the condition of equal variances, Figure 1-b, all methods are similar for equal means and as the difference between the means increases, the performance of PU improves more rapidly than in the other methods. For dif- ferent variances, Figures 1-d and 1-f, PU is better in all tested situations. This good performance of PU for small differences in means and large differences in variance is especially interesting for detecting differentially expressed genes.




(a) $\overline{TPr}$, $n = 4$ (b) $\overline{FDr}$, $n = 4$

(c) $\overline{TPr}$, $n = 4$ (d) $\overline{FDr}$, $n = 4$

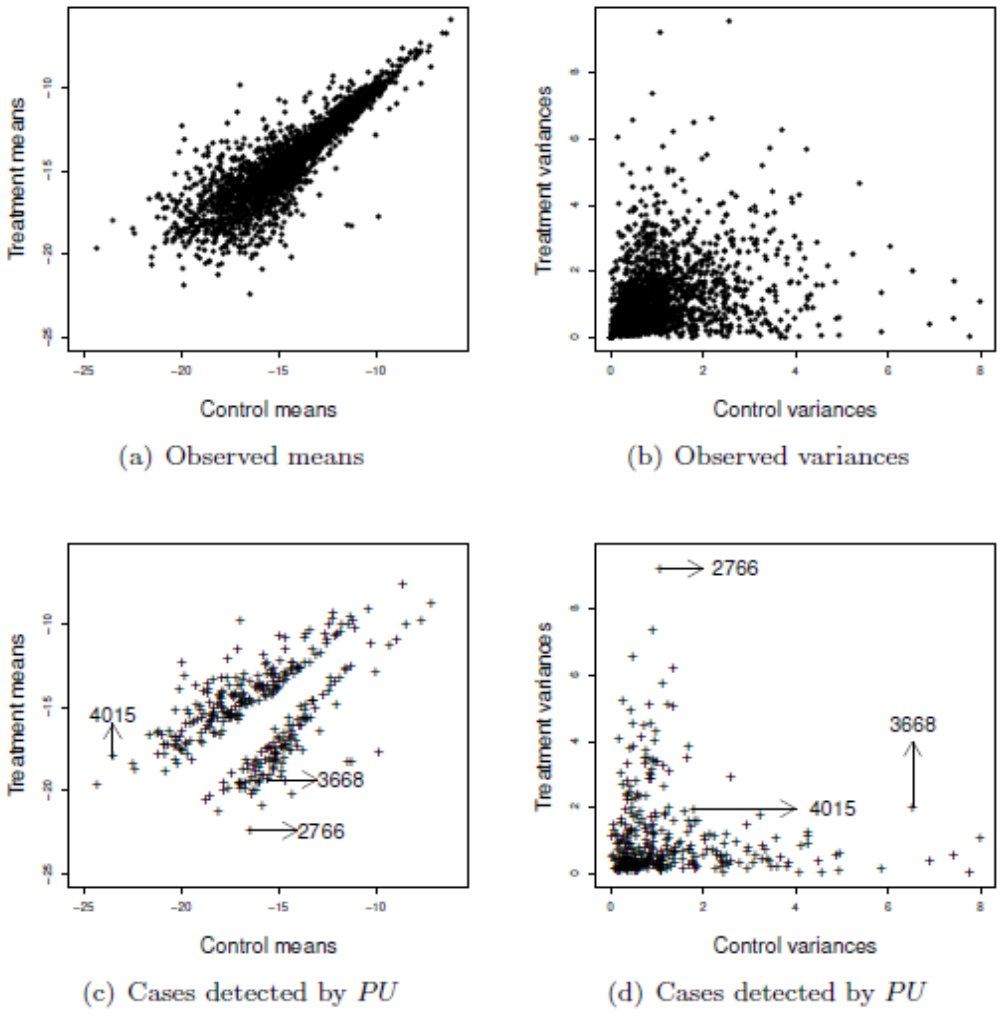
(e) $\overline{TPr}$, $n = 4$ (f) $\overline{FDr}$, $n = 4$

Figure 1: Average of true positive and false discovery rate for M = 2.

### 4.1.2 Escherichia coli bacterium data set.

Now consider the gene expression data for the Escherichia Coli bacterium described in Arfin et al. (2000), which presents N = 4, 290 and n = 4.

Figure 2-a,b show the observed control means and variances versus observed treatment means and variances for all genes of this dataset. Figure 2-c,d highlight the cases identified with a difference by PU . Cases identified with a difference by TT , CT and BTT are presented in Figure 3.



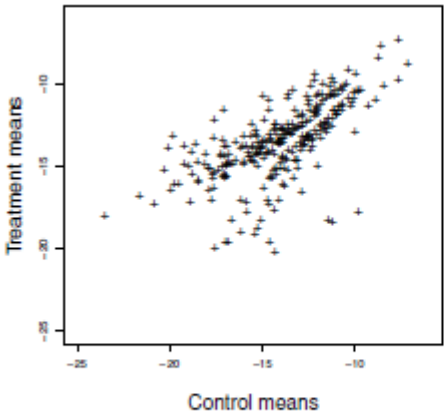

(a) Observed means (b) Observed variances

(c) Cases detected by *PU* (d) Cases detected by *PU*

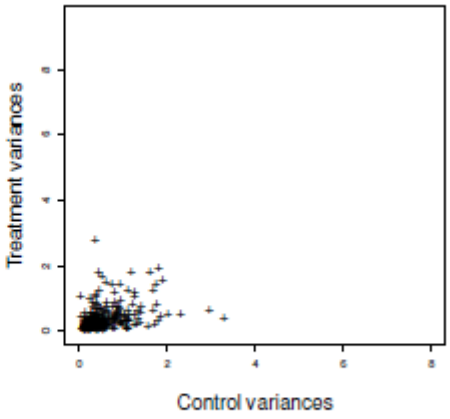
Figure 2: Control and treatment observed means and variances for PU

PU identifies 327 genes as differentially expressed while TT identifies 287, CT 219 and BTT 288 genes. Out of 287 genes identified by TT , 159 (55.40%) were also identified by PU ; out of

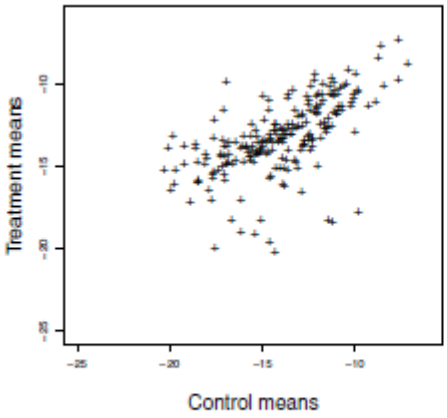
219 identified by CT , 141 (64.38%) were identified by PU and out of 288 identified by BTT , 164 (56, 94%) were also identified by PU . 133 genes were identified by all four methods.



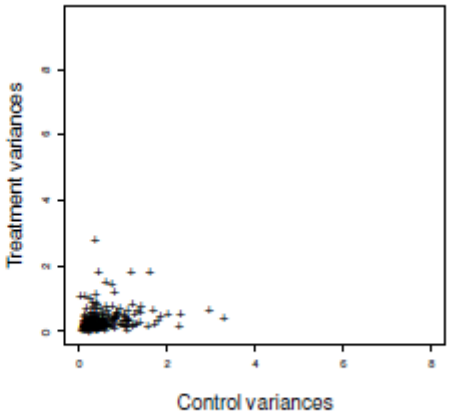

(a) TT



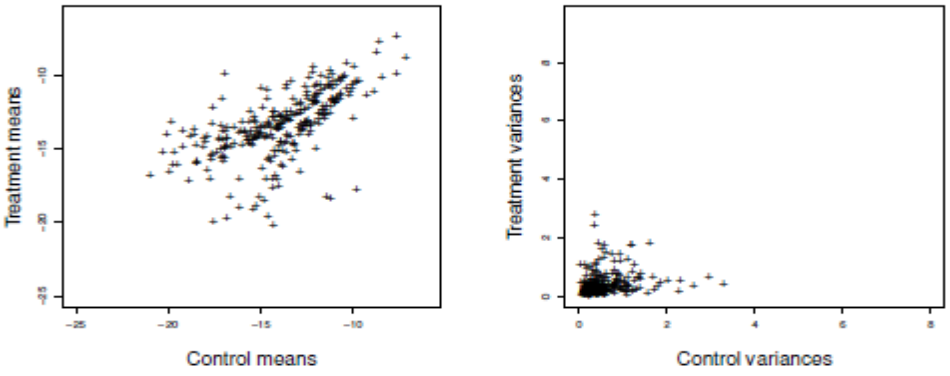

(b) TT




(c) CT




(d) CT




(e) BTT (f) BTT

Figure 3: Cases identified by TT , CT and BTT

For this example, 152 genes identified by PU are not detected by the other methods. These cases are shown in Figure 2 in Appendix 3 of the AM.

Note that genes with means well apart from eachother are better identified by PU than by t-tests. Examples are genes 2766 (b1326(f262)) and 3668 (pnuC), which are highlighted in Figure 2(c,d), and have P ($c_2 = 2|c_1 = 1$, y) equal to 0.81 and 0.53, respectively. These two genes are not identified by t-tests. For gene 2766, the p-values obtained by TT , CT and BTT are, respectively, 0.81, 0.86 and 0.87; and for gene 3668 are 0.53, 0.72 and 0.80. The reason for this error is the low performance of t-tests in situations with differences between means and variances, as observed in the simulations. Genes with similar variances and different means are well identified by t-tests. An example is the gene 4015 (yehP) which has p-value=0.01 for TT , CT and BTT and P ($c_2 = 2|c_1 = 1$, y) = 0.90 for PU

### 4.2 Control and two treatments

For this case, we compare the performance of PU with ANOVA followed by the Tukey-test (denoted by AT ) and Duncan-test (denoted by AD) with a significance level of 0.05.

The five models written in terms of latent variables are: $M_0$ : $c_0 = (c1 = c2 = c3)$ *versus* $M_t$: for one of $c_t$'s: $c_1 = (c_1 = c_2 \neq c_3)$, $c2 = (c_1 = c_3 \neq c_2)$, $c_3 = (c_1 \neq c_2 = c_3)$ and $c_4 = (c_1 \neq c_2 \neq c_3)$.

#### 4.2.1 Simulated data sets

In order to generate the data sets, we fix control parameters as $\mu_1 = -14$ and $\sigma_1^2 = 0.8$ and set the proportion of cases generated from each model to 0.80 for $M_0$ and 0.05 for $M_t$, for t = 1, . . . , 4.

The values of the parameters for each configuration are

- $(\mu_3, \sigma_3) = (\mu_2, \sigma_2) = (\mu_1, \sigma_1)$ for c0;
- $(\mu_2, \sigma_2) = (\mu_1, \sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ for $c_1$;
- $(\mu_3, \sigma_3) = (\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ for $c_2$;

– $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2, \sigma_2)$ for $c_3$;
– $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2 + \delta\sigma_2, \gamma\sigma_2)$ for $c_4$,

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

The generation of a simulated data set is as follows. For g = 1, . . . , N , generate $u_g$ from $U \sim U(0, 1)$;

(i) If $u_g \leq 0.80$, fix parameters values according to $c_0$. Let the index vector Gg = (1, 1, 1) to indicate that case g is generated from $M_0$;
(ii) If $0.80 < u_g \leq 0.85$, fix parameters values according to c1 and set $G_g$ = (1, 1, 2);
(iii) If $0.85 < u_g \leq 0.90$, fix parameters values according to c2 and set $G_g$ = (1, 2, 1);
(iv) If $0.90 < u_g \leq 0.95$, fix parameters values according to c3 and set $G_g$ = (1, 2, 2);
(v) If $u_g > 0.95$, fix parameters values according to c4 and set $G_g$ = (1, 2, 3);
(vi) Generate $Y_{im} \sim N(\mu_m, \sigma_m^2)$, for m = 1, 2, 3 and i = 1, . . . , n.

For each pair $(\delta, \gamma)$, we generate L = 100 data sets according to steps (i) to (vi) described above and the results are presented using $\overline{TPr}$ and $\overline{FDr}$, see Appendix 4 of AM.

Figure 4-a,c,e shows the true positive rate, $\overline{TPr}$. For equal variance, $\gamma = 1$, and n = 4, Figure 4-a, the performance of PU is similar to AT and slightly worse than AD. Increasing the sample size, n = 8 (Figure 3-a in Appendix 5 of AM), AD is better than PU and AT . For different variances, Figures 4-c and 4-e (Figures 3-c and 3-e in Appendix 5 of AM), PU is better than both AT and AD.

The graphs in Figure 4-b,d,f (Figure 3-b,d,f in Appendix 5 of AM) show the false discovery rate. In all tested situations, PU presented better results, especially in cases where the variances are different, as shown in Figures 4-d and 4-f (Figures 3-d,f of AM), with the performance increasing as the difference in variance increases and for a small difference in means.

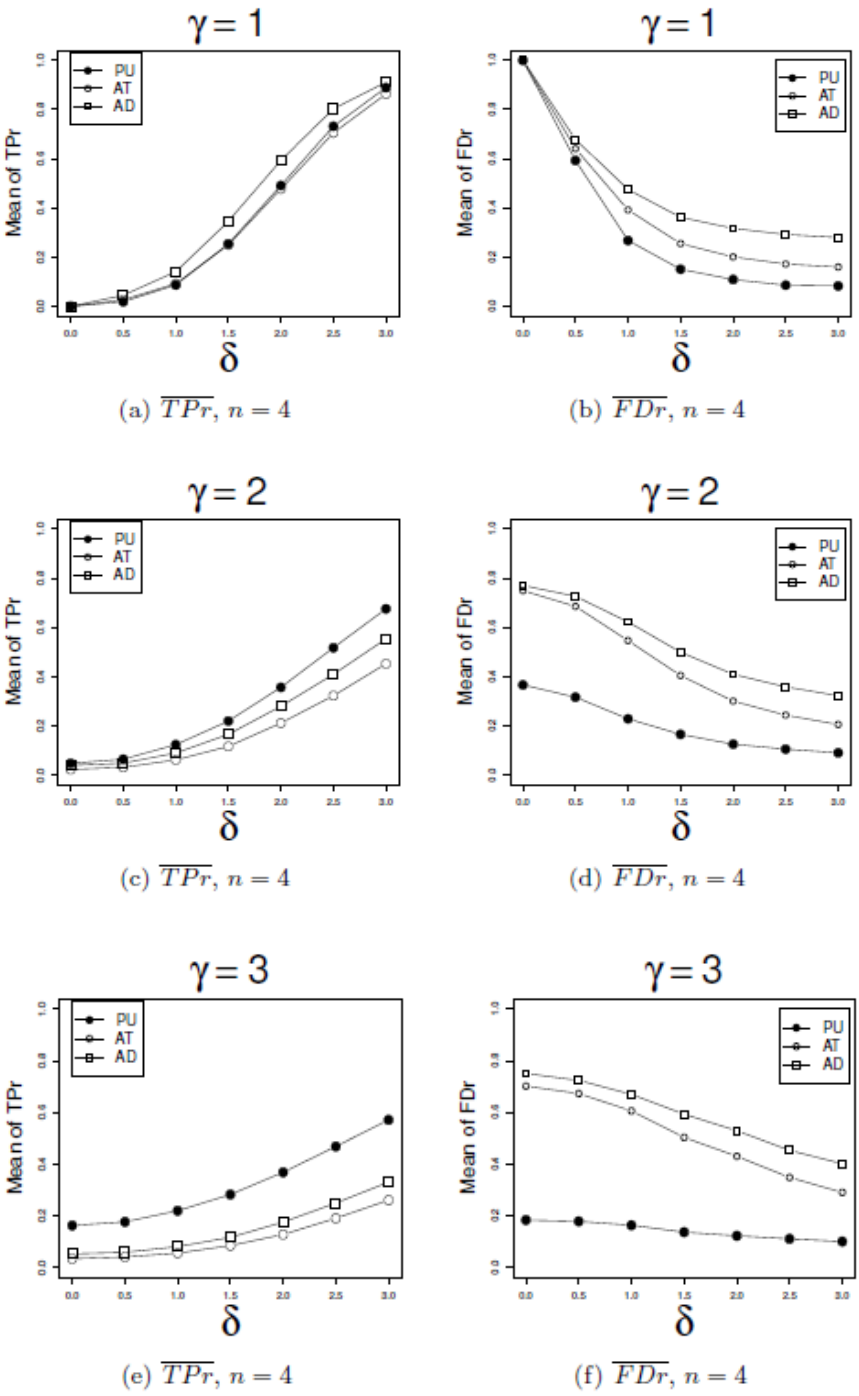
Appendix 6 of AM shows a comparison of performance of the methods for M = 4. For this case, the PU also presents higher $\overline{TPr}$ and smaller $\overline{FDr}$.

### 4.2.2 Proteomics data set

Consider the shotgun proteomics microarray data set, taken from the website http:// cybert.ics.uci.edu (Baldi and Long, 2001). The data set consists of N = 1, 088 proteins, with a control and two treatments and sample size n = 5.

Table 1 shows the number of cases identified for each model by each method. means.

| Method | Model | | | | | Total of diff. expressed genes identified |
|---|---|---|---|---|---|---|
| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | |
| PU | 925 | 24 | 106 | 33 | 0 | 163 |
| AT | 948 | 46 | 27 | 67 | 0 | 140 |
| AD | 932 | 32 | 69 | 53 | 2 | 156 |




(a) $\overline{TPr}$, $n = 4$ (b) $\overline{FDr}$, $n = 4$

(c) $\overline{TPr}$, $n = 4$ (d) $\overline{FDr}$, $n = 4$

(e) $\overline{TPr}$, $n = 4$ (f) $\overline{FDr}$, $n = 4$

Figure 4: Average of true positive and false discovery rate for M = 3.

Out of the 140 rejected null models by AT , 105 (75%) were also rejected by PU ; out of the 46 cases identified by AT as M1, 15 (32.61%) were also identified by PU ; all 27 cases identified as $M_2$ by AT , were also identified by PU ; and out of the 67 cases identified under $M_3$ by AT , 17 (25.37%) were also identified by PU .

Out of the 156 rejected null models by AD, 112 (71.79%) were also rejected by PU ; out of the 32 cases identified by AD as $M_1$, 15 (46.88%) were also identified by PU ; out of the 69 cases identified as $M_2$ by AD, 49 (71.01%) were also identified by PU ; out of the 53 cases identified as $M_3$ by AD, 18 (33.96%) were also identified by PU ; and the two cases identified as $M_4$ by AD (proteins 60 and 649) were not identified by PU as $M_4$, but as $M_1$.

Tables 2 and 3 in Appendix 7 of AM show the ten most evident cases identified by PU and AT -AD, respectively.

For this proteomics data, 51 null hypothesis rejected by PU were not rejected by any of the other methods. These cases are shown in Table 4 in Appendix 8 of the AM.

## 5. Discussion

Results from simulations showed a better performance for PU than TT , CT and BTT , for experiments with control and one treatment. This was clearer in situations with different variances. For experiments with control and two or three treatments, we compared the performance of PU with ANOVA followed by the Tukey-test (AT ) or Duncan-test (AD). Once more, PU presented a better performance than AT and AD emphasizing itself in cases with different variances. Methods were also applied to real data sets with control and one treatment and control and two treatments. In both cases, PU showed a better performance. From a biological point of view, the main interest is that PU brings to light genes that are not identified when using the other methods considered in compar- isons, (TT , AT and AD), AT or AD. This suggests an eventual complementarity

of the methods.

Additional points in favour of PU are: (1) It is easier to use, especially when $M > 2$ and (2) it performs well in situations with small sample sizes which are common in gene expression data analysis. Besides this, the PU can be easily implemented in usual software such as the software R. The code used for computing is in the R language and can be obtained by e-mail from the first author.

## Acknowledgment

We thank the Editor and the referees for their comments, suggestions and criticisms which have led to improvements of this article. The first author ac- knowledges the Brazilian institution CNPq.

## References


[1] Allison, D. B., Cui, X., Page, G. P. and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus, *Nat. Rev. Genet.*, **7**, 55-65.

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

[3] Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S., Hatfield, G. W. (2000). Global gene expression profiling in *Escherichia Coli* K12. *J. Biol. Chem*, **275**, 29672-29684.

[4] Baldi, P., Long, D. A. (2001). A Bayesian framework for the analysis of mi- croarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509-519.

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

[6] Casella, G., Robert, C., and Wells, M. (2000). Mixture models, latent variables and partitioned importance sampling. *Technical Report 2000-03*, CREST, INSEE, Paris.

[7] Cox, D. R. and Reid, N. M. (2000). The theory of design of experiments. Chapman-Hall/CRC.

[8] 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**, 680-68.

[9] Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18(1)**, 71-103.

[10] Duncan, D B. (1955). Multiple range and multiple F tests. *Biometrics*, **11**, 1-42.

[11] Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference using Mixtures. *Journal of the American Statistical Association*, **90**, 577– 588.

[12] Ferguson, S. T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **2**, 209-230.

[13] Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for mi- croarray data. *BMC Bioinformatics*, **7**:126.

[14] Gopalan, R.; Berry, D. A. (1998). Bayesian multiple comparisons using Dirich- let process priors. *Journal of the American Statistical Association*, vol.93, No.**443**, 1130-1139.

[15] Hatifield, G. W., Hung, S. and Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. *Molecular Microbiology*, **47**(4), 871-877.

[16] Kass, R., and Raftery, A. (1995). Bayes Factor. *Journal of the American Statistical Association*, **90**, 773-795.

[17] Lonnstedt, I. Speed, T. P. (2001). Replicated microarray data. *Statistica Sinica*, **12**, 31-46.

[18] 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 Statistics.*, **28**, 167-189.

[19] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mix- ture models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[20] Rosenfeld, S.(2007) Detection of Differentially Expressed Genes In Small Sets of cDNA Microarrays. *Journal of Data Science*, **5**, 00-00(JDS-341).

[21] Saraiva, E. F. and Milan, L. A. (2012). Clustering Gene Expression Data using a Posterior Split-Merge-Birth Procedure. *Scandinavian Journal of Statistics*, **39**, 399-415.

[22] Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantita- tive monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

[23] Wu, T. D. (2001). Analyzing gene expression data from DNA microarray to identify candidates genes. *Journal of Pathology*, **195**(1), 53-65.

Received June 25, 2014; accepted September 28, 2014

Erlandson F. Saraiva
Institute of Mathematics

Federal University of Mato Grosso do Sul Campo
Grande, MS, Brazil

Lu´ıs A. Milan Department of
Statistics

Federal University of Sao Carlos Sao
Carlos, SP, Brazil