Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 627–640

https://doi.org/10.29220/CSAM.2017.24.6.627
Print ISSN 2287-7843 / Online ISSN 2383-4757

Identifying differentially expressed genes using the Polya urn scheme

Erlandson Ferreira Saraiva[1,a], Adriano Kamimura Suzuki[b], Luís Aparecido Milan[c]

[a]Institute of Mathematics, Federal University of Mato Grosso do Sul, Brazil;
[b]Department of Applied Mathematics and Statistics, University of São Paulo, Brazil;
[c]Department of Statistics, Federal University of São Carlos, Brazil

Abstract

A common interest in gene expression data analysis is to identify genes that present significant changes in expression levels among biological experimental conditions. In this paper, we develop a Bayesian approach to make a gene-by-gene comparison in the case with a control and more than one treatment experimental condition. The proposed approach is within a Bayesian framework with a Dirichlet process prior. The comparison procedure is based on a model selection procedure developed using the discreteness of the Dirichlet process and its representation via Polya urn scheme. The posterior probabilities for models considered are calculated using a Gibbs sampling algorithm. A numerical simulation study is conducted to understand and compare the performance of the proposed method in relation to usual methods based on analysis of variance (ANOVA) followed by a Tukey test. The comparison among methods is made in terms of a true positive rate and false discovery rate. We find that proposed method outperforms the other methods based on ANOVA followed by a Tukey test. We also apply the methodologies to a publicly available data set on Plasmodium falciparum protein.

Keywords: gene expression, Bayesian approach, prior Dirichlet process, Polya urn scheme, Gibbs sampling

1. Introduction

DNA array technology is capable of providing simultaneous gene expression level measurements for thousands of genes under different biological experimental conditions. Once the expression levels have been obtained one of the objectives is to identify genes that present significant changes in the expression levels among experimental conditions. The identification of these genes is important because it may bring to reveal new biological discoveries such as which genes are involved in the origin and/or evolution of some genetic disease or which genes react to a drug stimulus. For further discussion and additional references on DNA array technology, see DeRisi *et al.* (1997), Arfin *et al.* (2000), Wu (2001), Hatfield *et al.* (2003), and their references.

According to Baldi and Long (2001) the first level of gene expression data analysis is the identification of the genes with expression levels different in a treatment condition in relation to a control condition. For this case, the usual methods used to identify differentially expressed genes are based on t-tests such as usual t-test for unequal variances, the Cyber-t (CT) proposed by Baldi and Long (2001) and the Bayesian t-test (BTT) proposed by Fox and Dimmic (2006). CT, and BTT modify

[1] Corresponding author: Institute of Mathematics, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Brazil.
E-mail: erlandson.saraiva@ufms.br

Published 30 November 2017 / journal homepage: http://csam.or.kr

the standard error estimate of the two-sample differences found in the denominator of the standard t-statistic. Under the Bayesian approach, Oh and Yang (2006) developed a two-sample comparison considering a multiple test scenery. The authors assume a conditionally prior distribution for each pair of comparing means. In order to estimate parameters of interest from posterior distribution the authors propose an importance sampling method. But, in situations with a control and more than one treatment, remains common to apply analysis of variance (ANOVA) followed by a Tukey test to identify which treatment caused the difference; see for example Pavlidis (2003), Parkitna *et al.* (2006), and Goeman and Bühlmann (2007).

In this paper, we consider gene expression data analysis from experimental conditions with a control and more than one treatment. We model each one of the hypothesis of equality or inequality among experimental conditions by a model. In this way our interest is to search for a model which best fits the data and meets conditions of inequality among the experimental conditions. We use a hierarchical Bayesian approach in order to select one of the models considered. This approach uses the Dirichlet process prior and its representation via Polya urn scheme (Blackwell and MacQueen, 1973). The advantage of using the Dirichlet process prior is its discreteness that allows the parameters to be coincident with positive probability. To calculate the posterior probability for each model we implement a Gibbs sampling (GS) algorithm considering the Polya urn scheme written through latent variables.

We develop a simulation study to verify the performance of the proposed method and compare it with the usual method based on ANOVA followed by Tukey test. The simulation study was implemented considering the cases with a control and two- and three- treatment experimental conditions. The ANOVA is applied to identify genes which show a significant difference among experimental conditions. ANOVA does not identify which experimental conditions show the difference; therefore, we also apply the Tukey test as a post-hoc test to identify which experimental conditions show significant difference, see for example Pavlidis (2003), Parkitna *et al.* (2006), Goeman and Bühlmann (2007), and Zollanvari *et al.* (2009). As comparison criterion between methods we consider the true positive rate (TPR) and false discovery rate (FDR).

The simulation results show a better performance of the proposed method. We also apply both methods to a real data set downloaded from http://cybert.ics.uci.edu/anova that concerns a proteomics experiment (Baldi and Long, 2001).

The pioneering paper using the Dirichlet process prior for multiple-comparison was by Gopalan and Berry (1998). The paper considers a hierarchical Bayesian approach with the Dirichlet process prior and develop a GS algorithm to make a comparison among various hypotheses. The posterior probability for the hypotheses are estimated using the Rao-Blackwellization method proposed by Gelfand and Smith (1990). Guindani *et al.* (2009) recently developed a semi-parametric Bayesian model with the Dirichlet process prior for multiple-comparison problems. Guindani *et al.* (2009) consider a loss function based on positive and false positive counts to proposes a decision rule that is based on a threshold of the posterior probability. Zou *et al.* (2010) consider a two-sample comparison and model the t-statistics using a hierarchical Bayesian approach with a Dirichlet process prior on the non-centrality parameter. Estimates for parameters of interest and false discovery rate are estimated via a GS algorithm. The distinction among the previous approaches and ours is that here the Dirichlet process is used jointly with a model selection procedure to identify the cases differentially expressed. The discreteness of the Dirichlet process is used to identify equality (or not) among parameters of the models considered. In this way, the procedure allow the source for a model that best fits the data and the identification of cases with difference in relation to mean e/or variance.

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 compare the performance of the proposed method with the usual approach. In Section 4, both methods are applied to a real dataset. Section 5 concludes the paper with final remarks.

2. Hierarchical Bayesian model

Consider a DNA array experiment with N genes performed for experimental conditions $E_1, \ldots, E_T$, where E_1 represent the control experimental condition, E_2 represent the first treatment experimental condition and successively until the last one treatment experimental condition E_T. Suppose that each one of experimental conditions are replicated n times. Denote by y_{ig_t} the i^{th} observed expression level for gene g in experimental condition $t \in \{1, \ldots, T\}$, for $i = 1, \ldots, n$ and $g = 1, \ldots, N$.

We omit the index g in the next expressions in order to simplify the notation hereafter. Thus, let $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$ be the set of all observed expression levels for gene g in T experimental conditions, where $\mathbf{Y}_t = (y_{1t}, y_{2t}, \ldots, y_{nt})'$ is a $n \times 1$ vector of conditionally independent observations for gene g on treatment t, for $g = 1, \ldots, N$ and $t = 1, \ldots, T$.

As is usual in gene expression data analysis, consider that the logarithm of the observed gene expression levels in control and treatment conditions are generated from normal distributions with mean μ_t and variance σ_t^2, $Y_{it} \sim \mathcal{N}(\mu_t, \sigma_t^2)$, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. See for example, Baldi and Long (2001), Fox and Dimmic (2006), Kim *et al.* (2013), Saraiva and Milan (2012), Louzada *et al.* (2014), and Oh (2015), among others.

Denote parameters of the experimental condition t by $\theta_t = (\mu_t, \sigma_t^2)$ and let $\Theta = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T);\ \theta_t \in \mathbb{R} \times \mathbb{R}^+\}$ be the parametric space, for $t = 1, \ldots, T$. The likelihood function for $\boldsymbol{\theta}$ given $\mathbf{y}$ is given by

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{t=1}^{T} \prod_{i=1}^{n} f(y_i|\theta_t) \propto \prod_{t=1}^{T} \left(\sigma_t^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_t^2} \sum_{i=1}^{n} (y_{it} - \mu_t)^2\right\},$$

where $f(y_i|\theta_t)$ is the probability density function of the normal distribution with parameters $\theta_t = (\mu_t, \sigma_t^2)$, for $t = 1, \ldots, T$.

Our interest is to verify whether a gene g is differentially expressed among the different experimental conditions, i.e., if $\theta_t = \theta_j$ or $\theta_t \neq \theta_j$, for all $t, j = 1, \ldots, T$ and $t \neq j$. This equality or inequality between θ_t and θ_j can be represented by the following models:

$$\begin{aligned} M_0 &: \theta_1 = \theta_2 = \cdots = \cdots = \theta_T, \\ M_1 &: \theta_1 \neq \theta_2, \theta_1 = \theta_3 = \cdots = \theta_T, \\ M_2 &: \theta_1 \neq \theta_3, \theta_1 = \theta_2 = \theta_4 = \cdots = \theta_T, \end{aligned}$$

successively for all combinations of inequality 2 to 2, 3 to 3, ..., until the last one model

$$M_Q : \theta_1 \neq \theta_2 \neq \cdots \neq \theta_T.$$

In this way our interest is to search for a model which best fits the data and meets conditions defined by models M_q, $q = 0, 1, \ldots, Q$. For each model M_q, the equality or inequality among θ_t's determine a particularly partition on parameter space Θ, for $t \in \{1, \ldots, T\}$. This then allow us to use a hierarchical Bayesian approach with the Dirichlet process prior on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ in order to make simultaneous comparisons among θ_t's. This is possible due the discreteness of the Dirichlet process that allows parameters from distinct experimental conditions to be coincident with positive probability. For more details on discreteness of the Dirichlet process see Gopalan and Berry (1998), Neal (1998), and references.

Thus, consider the following hierarchical Bayesian model with the Dirichlet process prior

$$\begin{aligned} \mathbf{Y}_t \mid \theta_t = \left(\mu_t, \sigma_t^2\right) &\sim \mathcal{N}\left(\mu_t, \sigma_t^2\right) \\ \theta_t \mid G &\sim G \\ G \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha G_0), \end{aligned} \tag{2.1}$$

for $t = 1, \ldots, T$, where G is a unknown distribution function with a prior distribution given by a Dirichlet process with concentration parameter α and baseline distribution G_0 (Antoniak, 1974; Ferguson, 1973). The model (2.1) is denominated in the literature by Dirichlet process mixture model (DPMM).

In order to complete the specification of the model (2.1) we fix $\alpha = 1$, as done by Escobar and West (1995), Medvedovic and Sivaganesan (2002), and Jain and Neal (2004). Besides, in order to explore the complete conjugacy of the model we set up G_0 as

$$G_0 \equiv \mathcal{N}\left(\mu_0, \frac{\sigma_t^2}{\lambda}\right) \mathcal{IG}\left(\frac{\tau}{2}, \frac{\beta}{2}\right),$$

where μ_0, λ, α, andβ are known hyperparameters and $IG(\cdot)$ represents the inverse gamma distribution in a parametrization so that the expected value is $\beta/(\tau - 2)$. The choice of the hyperparameters values will generally depend upon the application at hand. At this moment, we leave them unspecified.

Using the result of Blackwell and MacQueen (1973), we have that integrating G over its prior distribution in model (2.1) $\boldsymbol{\theta}$ follows the Polya urn scheme and can be written as

$$\theta_1 \sim G_0 \tag{2.2}$$

$$\theta_t | \theta_1, \ldots, \theta_{t-1} \sim \frac{\alpha}{\alpha + t - 1} G_0 + \frac{1}{\alpha + t - 1} \sum_{j=1}^{t-1} \mathbb{I}_{\theta_t}(\theta_j), \tag{2.3}$$

where $\mathbb{I}_{\theta_t}(\theta_j) = 1$ if $\theta_t = \theta_j$ and $\mathbb{I}_{\theta_t}(\theta_j) = 0$ otherwise, for $t = 1, \ldots, T$.

Note that, at each step of the sample procedure defined by (2.3), θ_t may assume a new value generated from baseline distribution G_0 with probability $\alpha/(\alpha + t - 1)$ or may assume the value of one of previous θ_j's with probability $\{1/(\alpha + t - 1)\}\sum_{j=0}^{t-1} \mathbb{I}_{\theta_t}(\theta_j)$. Thus, a sample from joint distribution of $\theta_1, \ldots, \theta_T$ yields k groups $(1 \leq k \leq T)$ of θ_t's with distinct values $\phi_1, \ldots, \phi_k$ generated from baseline distribution G_0. Using this fact we proposed a MCMC procedure to estimate the posterior probabilities for models M_q through the Polya urn scheme, given by expressions (2.2) and (2.3), written in terms of latent indicator variables, for $q = 1, \ldots, Q$.

2.1. Polya urn scheme via latent variables

Let $\mathbf{c} = (c_1 \ldots, c_T)$ be a vector of latent indicator variables, so that, $c_t = j$ if $\theta_t = \phi_j$, for $t = 1, \ldots, T$ and $j = 1, \ldots, k$. Consider $D_j = \{y_t; c_t = j\}$ be the cluster (set) of observations with identical configuration indicators c_t, where $D_1, \ldots, D_k$ are paired with $\phi_1, \ldots, \phi_k$, respectively, for $t = 1, \ldots, T$.

Assume that clusters are numbered consecutively as they arise; therefore, the sampling procedure defined by (2.2) and (2.3) can be replicated by the following procedure:

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

(ii) for $t = 2, \ldots, T$ calculate the probabilities

$$P(c_t = j|c_1, \ldots, c_{t-1}) = \frac{n_j}{\alpha + t - 1} \quad \text{and} \quad P(c_t = j^*|c_1, \ldots, c_{t-1}) = \frac{\alpha}{\alpha + t - 1}, \tag{2.4}$$

for $j = 1, \ldots, k_{(t)}$, $j^* = k_{(t)} + 1$, where $k_{(t)}$ is the number of different values in $\mathbf{c}_{t-1} = (c_1, \ldots, c_{t-1})$ and n_j is the number of $c_{t'} = j$ in $\mathbf{c}_{t-1}$, for $t' = 1, \ldots, t-1$ and $j = 1, \ldots, k_{(t)}$.

(ii-a) Generate $\mathbf{Z}_t \sim \text{Multinom}(1, \mathbf{P}_t)$, where $\mathbf{P}_t = (P(c_t = 1|\cdot), \ldots, P(c_t = k_{(t)}|\cdot), P(c_t = j^*|\cdot))$ and $\text{Multinom}(1, \mathbf{P}_t)$ is the multinomial distribution with $k_{(t)} + 1$ modalities and a single observation;

(ii-b) If $Z_{tj} = 1$, for some $j \in \{1, \ldots, k_{(t)}\}$, then do $c_t = j$, $D_j = \{D_j\} \cup \{y_t\}$ and $n_j = n_j + 1$;

(ii-c) If $Z_{tj^*} = 1$, then do $c_t = j^*$, $D_{j^*} = \{y_t\}$, $n_{j^*} = 1$ and generate $\phi_{j^*} \sim G_0$.

(iii) Given $\mathbf{c}$, set $\theta_t = \phi_j$ for all $c_t = j$, $t = 1, \ldots, T$ and $j = 1, \ldots, k$.

Note from this procedure, that D_1 is the set composite by the Treatments conditions that do not have difference in relation to control experimental condition.

As a sample from a Dirichlet process is exchangeable (Antoniak, 1974; Jain and Neal, 2004; MacEachern, 2016), so from (2.4) we have conditional probabilities given by

$$P(C_t = j|\mathbf{c_{-t}}) = \frac{n_{j,-t}}{\alpha + T - 1} \quad \text{and} \quad P(C_t = j^*|\mathbf{c}_{-t}) = \frac{\alpha}{\alpha + T - 1},$$

where $\mathbf{c}_{-t} = (c_1, \ldots, c_{t-1}, c_{t+1}, \ldots, c_T)$ and $n_{j,-t}$ is the number of $c_t = j$ in $\mathbf{c}_{-t}$, for $j = 1, \ldots, k_{(t)}$. Here, $k_{(t)}$ is the number of different values in configuration $\mathbf{c}_{-t}$. The ϕ_j's remains drawn independently from baseline distribution G_0.

Using the Bayes rule, the conditional posterior probabilities for latent indicator variable are given by

$$P(C_t = j|\mathbf{c_{-t}}, \mathbf{y}_t, \phi_j) = b_t \frac{n_{j,-t}}{\alpha + T - 1} f(\mathbf{y}_t|\phi_j) \quad \text{and} \quad P(C_t = j^*|\mathbf{c}_{-t}, \mathbf{y}_t) = b_t \frac{\alpha}{\alpha + T - 1} q(\mathbf{y_t}), \tag{2.5}$$

where b_t is the appropriate normalizing constant for those probabilities sum to one, $f(\mathbf{y}_t|\phi_j)$ is the joint probability for $\mathbf{y}_t$ given ϕ_j and

$$q(\mathbf{y}_t) = \int f\left(\mathbf{y}_t|\phi_{j^*}\right) \pi_{G_0}\left(\phi_{j^*}\right) d\phi_{j^*} = \beta^* \lambda^* \Gamma^* \left[1 + \frac{\sum_{\mathbf{y}_m \in D_j} \mathbf{y}_m^2 + \lambda\mu_0^2}{\beta} - \frac{\left(\sum_{\mathbf{y}_m \in D_j} \mathbf{y}_m + \lambda\mu_0\right)^2}{\beta(n_j + \lambda)}\right]^{-\tau^*},$$

in which,

$$\beta^* = \left(\frac{1}{\beta\pi}\right)^{\frac{n_j}{2}}, \quad \lambda^* = \left(\frac{\lambda}{n_j + \lambda}\right)^{\frac{1}{2}}, \quad \Gamma^* = \frac{\Gamma\left(\frac{\tau + n_j}{2}\right)}{\Gamma\left(\frac{\tau}{2}\right)} \quad \text{and} \quad \tau^* = \left(\frac{\tau + n_j}{2}\right),$$

for $t = 0, 1, \ldots, T$.

Given a configuration $\mathbf{c}$, the full conditionals for μ_j and σ_j^{-2} are, respectively,

$$\mu_j|\sigma_j^2, \lambda, \mathbf{y} \sim \mathcal{N}\left(\frac{n_j}{n_j + \lambda}\bar{y}_j + \frac{\lambda}{n_j + \lambda}\mu_0, \frac{\sigma_j^2}{n_j + \lambda}\right) \tag{2.6}$$

and

$$\sigma_j^{-2}|\alpha,\beta,\mathbf{y} \sim \mathcal{IG}\left(\frac{\alpha+n_j+1}{2}, \beta+(n_j-1)s_j^2+\frac{n_j\lambda\left(\bar{y}_j-\mu_0\right)^2}{n_j+\lambda}\right), \qquad (2.7)$$

where $\bar{y}_j$ and s_j^2 are the sample mean and variance, respectively, of the cluster D_j, for $j = 1, \ldots, k$.

2.1.1. Gibbs sampling algorithm

In order to estimate the posterior probabilities for models M_q we use a GS algorithm by iterating between aligns (2.5)–(2.7).

- **Gibbs sampling algorithm**: Let the current state of the Markov chain consist of $(\mathbf{c}^{(l-1)}, \boldsymbol{\phi}^{(l-1)})$, where l is the l^{th} iteration of the algorithm, for $l = 1, \ldots, L$, where $\mathbf{c}^{(0)} = (c_0^{(0)}, c_1^{(0)}, \ldots, c_T^{(0)})$ and $\boldsymbol{\phi}^{(0)} = (\phi_1^{(0)}, \ldots, \phi_k^{(0)})$ are the initial values. So, do the following:

  (1) For $t = 1, \ldots, T$:

    (a) Calculate $\mathbf{P}_t = (P(c_t = 1|\cdot), \ldots, P(c_t = k_{(t)}|\cdot), P(c_t = j^*|\cdot))$ as in (2.5);

    (b) Generate $\mathbf{Z}_t \sim \text{Multinom}(1, \mathbf{P}_t)$. If $z_{t,j} = 1$, for some $j \in \{1, \ldots, k_{(t)}\}$, do $c_t^{(l)} = j$. Otherwise, if $z_{j,k_{(t)}+1} = 1$, do $c_t^{(l)} = k_{(t)} + 1$;

  (2) Conditional on $\mathbf{c}^{(l)}$ update parameters ϕ_j from posterior distribution in (2.6) and (2.7);

  (3) Let $\mathbb{I}_q^{(l)}$ be an indicator variable, so that, $\mathbb{I}_q^{(l)} = 1$ if configuration $\mathbf{c}^{(l)}$ defines model M_q; and $\mathbb{I}_q = 0$ otherwise, for $q = 0, 1, \ldots, Q$.

We discard the first B values of the generated chains as a *burn in*. The posterior estimates for probabilities of models M_q are given by

$$\tilde{P}_q = \frac{1}{L-B}\sum_{l=B+1}^{L-B} \mathbb{I}_q^{(l)},$$

for $q = 0, 1, \ldots, Q$. If $P_q = \max_{1\leq q'\leq Q} P_{q'}$, then M_q is the selected model, for $q \in \{1, \ldots, Q\}$.

3. Data analysis

In this section we verify performance of the proposed method called Polya urn within Gibbs sampling (PUGS) and compare it with a standard method based on ANOVA followed by Tukey test (ATUK) using simulated data sets and a real data set.

In order to establish the hyperparameters values for the PUGS we consider the following procedure. Let (a, b) be the roughly interval which would include all observations produced by the experiment. Then, the hyperparameter μ_0 was chosen to be the middle point of the interval $\mu_0 = (a + b)/2$ and we set up $\lambda = 0.1$. Besides, we choose τ and β in a way that $E[\sigma_t^2] = \beta/(\tau - 2) = R$, where R is the range of the interval $R = b - a$. Thus, we obtain $\beta = (\tau - 2)R$ and we set $\tau = 3$. For ATUK we consider a significance level at 0.05.

3.1. Simulated data set

Consider a DNA experiment with a control and two experimental conditions. The five possible models written in terms of latent variables are:

$$
\begin{aligned}
M_1 &: \mathbf{c}_1 = (c_1 = c_2 \neq c_3);\\
M_2 &: \mathbf{c}_2 = (c_1 = c_3 \neq c_2);\\
M_3 &: \mathbf{c}_3 = (c_1 \neq c_2 = c_3);\\
M_4 &: \mathbf{c}_4 = (c_1 \neq c_2 \neq c_3). \qquad (3.1)
\end{aligned}
$$

In order to generate the data sets, we fix control parameters as $\mu_0 = 8.44$ and $\sigma_0^2 = 0.67$. These values are the average of the observed expression levels in control experiment carried through with the proteomics data set (Baldi and Long, 2001). The parameters values for each configuration are given by

- $(\mu_3, \sigma_3) = (\mu_2, \sigma_2) = (\mu_1, \sigma_1)$ for $\mathbf{c}_0$;

- $(\mu_2, \sigma_2) = (\mu_1, \sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ for $\mathbf{c}_1$;

- $(\mu_3, \sigma_3) = (\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ for $\mathbf{c}_2$;

- $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2, \sigma_2)$ for $\mathbf{c}_3$;

- $(\mu_2, \sigma_2) = (\mu_1 + \delta\sigma_1, \gamma\sigma_1)$ and $(\mu_3, \sigma_3) = (\mu_2 + \delta\sigma_2, \gamma\sigma_2)$ for $\mathbf{c}_4$,

for $\delta \in \{0.50, 1.00, 1.50, 2.00, 2.50, 3.00, 3.50, 4.00\}$ and $\gamma \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. We consider the proportion of cases generated from each model are $(0.70, 0.10, 0.10, 0.05, 0.05)$ for $(M_0, M_1, M_2, M_3, M_4)$, respectively. Besides, we set up $n = 5$ and $N = 1,000$.

The generation of a simulated data set is given by the following steps. For $g = 1, \ldots, N$, generate U_g from a uniform distribution on interval $(0, 1)$, $U_g \sim \mathcal{U}(0, 1)$, and consider $\mathbb{G}_g$ an indicator vector of dimension 1×3 that record from which model data are generated from:

(i) If $u_g \leq 0.70$, fix parameters values according to $\mathbf{c}_0$. Let the index vector $\mathbb{G}_g = (1, 1, 1)$ to indicate that case g is generated from M_0;

(ii) If $0.70 < u_g \leq 0.80$, fix parameters values according to $\mathbf{c}_1$ and set $\mathbb{G}_g = (1, 1, 2)$;

(iii) If $0.80 < u_g \leq 0.90$, fix parameters values according to $\mathbf{c}_2$ and set $\mathbb{G}_g = (1, 2, 1)$;

(iv) If $0.90 < u_g \leq 0.95$, fix parameters values according to $\mathbf{c}_3$ and set $\mathbb{G}_g = (1, 2, 2)$;

(v) If $u_g > 0.95$, fix parameters values according to $\mathbf{c}_4$ and set $\mathbb{G}_g = (1, 2, 3)$;

(vi) Generate $Y_{it} \sim \mathcal{N}(\mu_t, \sigma_t^2)$, for $t = 1, 2, 3$ and $i = 1, \ldots, n$.

Generated the data set, we apply PUGS and ATUK to identify the cases with a difference. We apply the PUGS fixing L = 33,000 iterations and burn-in B = 3,000; in addition, Besides, out of 30,000 iterations, we consider jumps of size 10, i.e., only 1 draw from every 10 was kept, in order to construct a sample of size of 3,000 to make inferences.

To record the configuration obtained by the PUGS and the ATUK, 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 = {PUGS, ATUK}. So, we compare performance of the methods by using the TPR (number of models correctly identified divided by N) and the FDR (number of models M_0 incorrectly selected divided by the number of rejected models M_0) given by

$$\mathrm{TPR} = \frac{\sum_{g=1}^{n} \mathbb{I}_{Z_g^{\mathrm{method}}}(\mathbb{G}_g)}{N}, \tag{3.2}$$

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

$$\mathrm{FDR} = \frac{\sum_{g=1}^{n} \left(1 - \mathbb{I}_{Z_g^{\mathrm{method}}}(\mathbb{G}_g)\right) \cdot \mathbb{I}_{\mathbb{G}_g}(\mathrm{Z}_0)}{N - \sum_{g=1}^{n} \mathbb{I}_{\mathbb{Z}_g^{\mathrm{method}}}(\mathrm{Z}_0)}, \tag{3.3}$$

where $\mathbb{I}_{\mathbb{G}_g}(\mathrm{Z}_0) = 1$ if case g $(\mathbb{G}_g)$ is generate according to configuration Z_0 of the null model M_0 and $\mathbb{I}_{\mathbb{G}_g}(\mathrm{Z}_0) = 0$ otherwise, for method = {PUGS, ATUK};

Moreover, for each pair (δ, γ) considered, we generate $L = 100$ different artificial data sets according to steps (i) to (vi) described above and present results using the mean of the TPR and FDR,

$$\overline{\mathrm{TPR}} = \frac{1}{L}\sum_{l=1}^{L} \mathrm{TPR}^{(l)} \quad \text{and} \quad \overline{\mathrm{FDR}} = \frac{1}{L}\sum_{l=1}^{L} \mathrm{FDR}^{(l)},$$

where $\mathrm{TPR}^{(l)}$ and $\mathrm{FDR}^{(l)}$ is the TPR and FDR calculated for l^{th} generated data set, respectively.

Figure 1 show the $\overline{\mathrm{TPR}}$ and $\overline{\mathrm{FDR}}$ for both methods. Note that, PUGS present better performance than ATUK for all simulated cases, i.e., PUGS has higher $\overline{\mathrm{TPR}}$ and smaller $\overline{\mathrm{FDR}}$. These results mean that PUGS has a better performance in correctly classified the cases; and the errors are smaller than the ATUK method. Particularly, this better performance occurs for cases with differences in variances, $\gamma = 2$ and $\gamma = 3$, as can be viewed in Figures 1(b), (c), (e), (f). This fact is especially interesting from the biological point of view because PUGS may show cases identified when the usual method ATUK is considered.

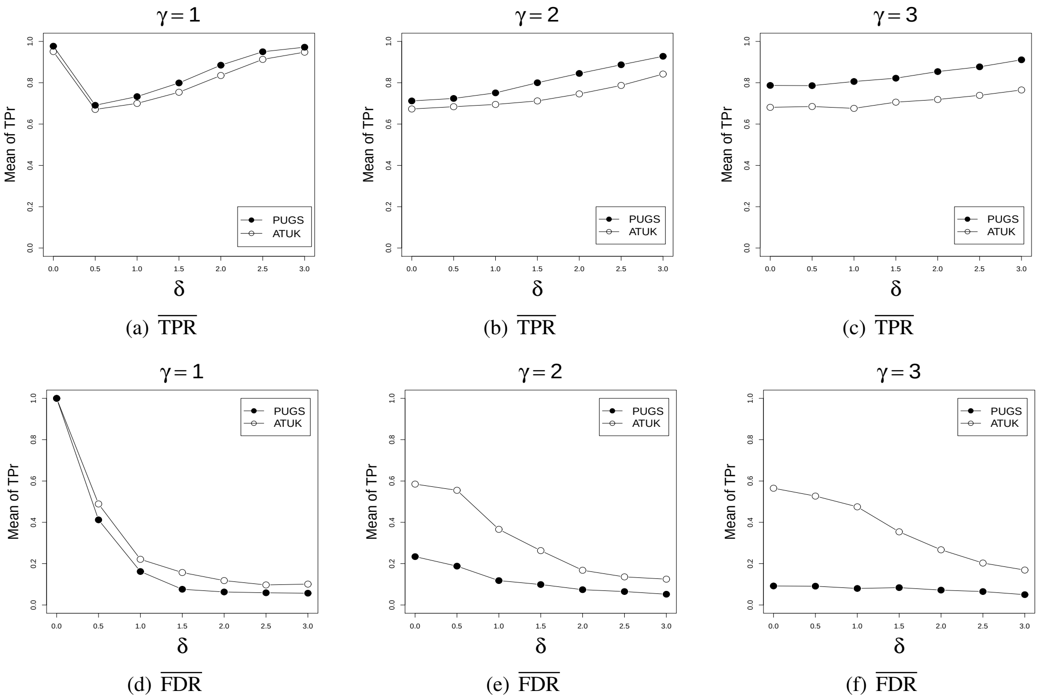
As in application presented in the next Section no case was identified under model M_4, so we present a simulation study in Appendix A similar to the presented case above, but considering only the configurations $\mathbf{c}_j$, $j = 0, 1, 2, 3$, from (3.1). In the Appendix B we present a simulation study conducted for a situation with a control and three treatment experimental conditions. Analogously to the results describe in this section, PUGS also present better performance than ATUK, specially for cases with difference in variances.

4. Application

Now consider the proteomics data set mentioned in the introduction. This dataset was extracted from the website cybert.ics.uci.edu/anova/. The data set is composed by $N = 1{,}088$ proteins from a control and two treatment conditions. The sample size from each experimental condition is $n = 5$.

For application of the PUGS we consider the same number of iteration, burn-in and hyperparameters values used in the simulation section. Table 1 shows the number of cases identified under each model by method. The last column of this table shows the number of cases in which model M_0 was not selected, i.e., the number of cases with difference identified by each method.

ATUK identify 140 case with evidence for difference while PUGS identify 127. Out of case identified with difference, 92 were identified by both methods. No cases were identified under M_4 in either method.



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

(d) $\overline{\text{FDR}}$ (e) $\overline{\text{FDR}}$ (f) $\overline{\text{FDR}}$

Figure 1: *Average of TPR and FDR for PUGS and ATUK. TPR = number of models correctly identified divided by N; FDR = number of models M_0 incorrectly selected divided by the number of rejected models M_0; ATUK = analysis of variance followed by Tukey test; PUGS = Polya urn within Gibbs sampling.*

Table 1: Number of cases identified by method

Method	Model					Total of cases with difference
	M_0	M_1	M_2	M_3	M_4	
PUGS	961	21	89	17	0	127
ATUK	948	46	27	67	0	140

PUGS = Polya urn within Gibbs sampling; ATUK = analysis of variance followed by Tukey test.

Tables 2 and 3 show the ten most evident cases identified by PUGS and ATUK, respectively. These tables show the number of the protein in the dataset, the sample mean and the standard deviation (SD) for control and two treatments, the configuration identified by PUGS and ATUK, the posterior probability obtained by PUGS and the p-value from ANOVA in ATUK.

The 10 most evident cases identified by ATUK were also identified by PUGS. Out of the 10 most evident cases identified with difference by PUGS, two were not identified by ATUK. These both cases are highlighted with $*$ in Table 2. Note that these both cases have higher differences in control SD (highlighted in bold) in relation to treatment SDs. As in the simulation study, this result indicates that PUGS has ability to identify cases not identified by the usual method ATUK if the difference is in variances.

5. Final remarks

In this paper, we develop a gene-by-gene multiple comparison analysis using a semi-parametric

Table 2: Ten most evident cases identified by PUGS, where $\mathbf{c}_j$ is given in (3.1), $j = 0, \ldots, 4$

Number	Sample mean			Sample standard deviation			Configuration		Posterior probability	p-value
	$\bar{\mathbf{y}}_1$	$\bar{\mathbf{y}}_2$	$\bar{\mathbf{y}}_3$	s_1	s_2	s_3	PUGS	ATUK		
557*	6.0579	9.0897	8.7857	**3.5739**	**0.9594**	**0.6304**	$\mathbf{c}_3$	$\mathbf{c}_0$	0.8938	0.0897
690	8.0427	9.5514	9.3821	0.5140	0.3500	0.3832	$\mathbf{c}_3$	$\mathbf{c}_3$	0.8854	0.0002
526*	8.8327	8.1400	8.7649	**0.9355**	**4.6101**	**0.5055**	$\mathbf{c}_2$	$\mathbf{c}_0$	0.8682	0.9076
666	8.2380	9.7643	8.5284	0.4457	0.4276	0.4502	$\mathbf{c}_2$	$\mathbf{c}_2$	0.8331	0.0003
1069	7.3039	8.8703	7.3194	0.6188	1.0238	0.3377	$\mathbf{c}_2$	$\mathbf{c}_2$	0.8248	0.0065
1024	8.3471	9.6321	8.3603	0.4866	0.5582	0.4792	$\mathbf{c}_2$	$\mathbf{c}_2$	0.8105	0.0023
936	7.5842	8.8570	7.8247	0.3111	0.7440	0.3207	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7926	0.0039
932	8.1317	9.7893	8.4973	0.5588	0.3725	0.5818	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7925	0.0006
730	8.3837	9.6405	8.1156	0.5543	0.3864	0.6852	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7637	0.0021
1012	8.0056	9.1393	8.0072	0.4247	0.4005	0.4868	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7633	0.0019

PUGS = Polya urn within Gibbs sampling; ATUK = analysis of variance followed by Tukey test.

Table 3: Ten most evident cases identified by ATUK, where $\mathbf{c}_j$ is given in (3.1), $j = 0, \ldots, 4$.

Number	Sample mean			Sample standard deviation			Configuration		Posterior probability	p-value
	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_2$	$\bar{\mathbf{x}}_3$	s_1	s_2	s_3	PUGS	ATUK		
690	11.6032	13.7804	13.5362	0.7423	0.5050	0.5537	$\mathbf{c}_3$	$\mathbf{c}_3$	0.8854	<0.001
666	11.8855	14.0871	12.3043	0.6432	0.6170	0.6496	$\mathbf{c}_2$	$\mathbf{c}_2$	0.8331	<0.001
932	11.7326	14.1231	12.2591	0.8065	0.5371	0.8393	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7925	0.001
60	12.6841	13.7752	11.6590	0.8529	0.6342	0.3794	$\mathbf{c}_2$	$\mathbf{c}_1$	0.3671	0.001
649	12.1523	12.9857	11.1680	0.6238	0.3427	0.7030	$\mathbf{c}_1$	$\mathbf{c}_1$	0.4103	0.001
1012	11.5504	13.1859	11.5521	0.6131	0.5783	0.7021	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7633	0.002
730	12.0950	13.9087	11.7082	0.8001	0.5572	0.9892	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7637	0.002
1024	12.0420	13.8963	12.0614	0.7023	0.8051	0.6912	$\mathbf{c}_2$	$\mathbf{c}_2$	0.8105	0.002
1020	11.7981	13.2402	10.8990	1.2130	0.4921	0.6438	$\mathbf{c}_2$	$\mathbf{c}_2$	0.5424	0.003
936	10.9425	12.7781	11.2891	0.4494	1.0730	0.4630	$\mathbf{c}_2$	$\mathbf{c}_2$	0.7926	0.004

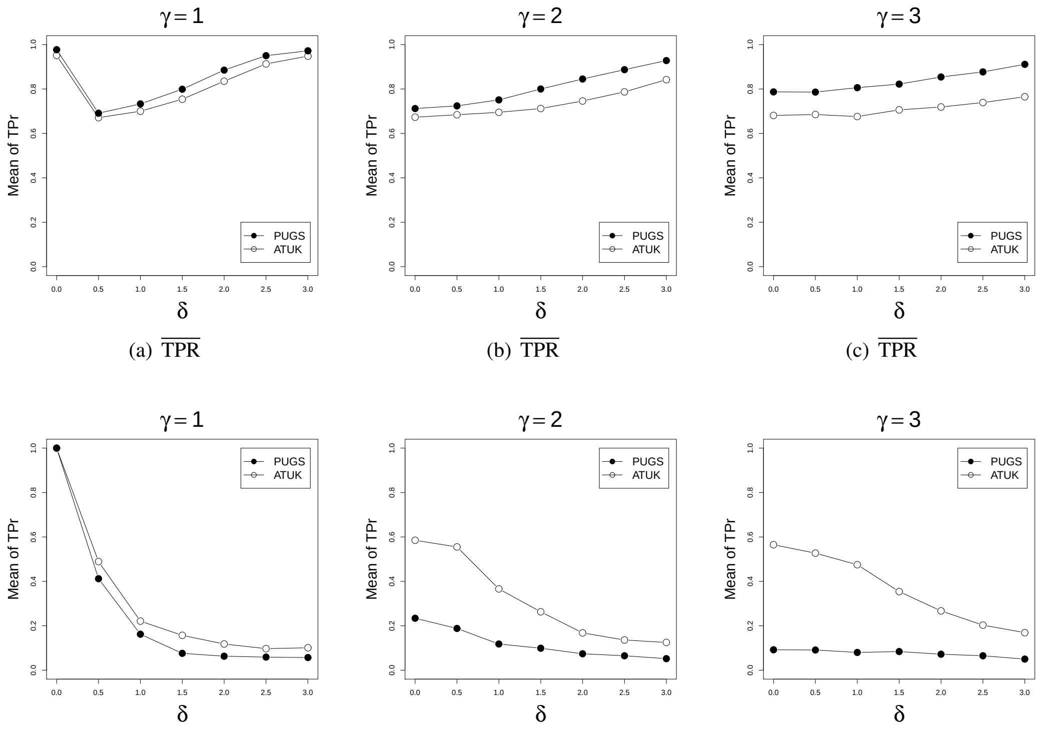
PUGS = Polya urn within Gibbs sampling; ATUK = analysis of variance followed by Tukey test.

Bayesian model with prior given by a Dirichlet process. The comparison among experimental conditions were made using the discreteness of the Dirichlet process within a model selection framework. The posterior probability for models were calculated through a GS algorithm that was implemented using the Polya urn scheme written in terms of latent variables to indicate equality or inequality among the experimental conditions.

The performance of the PUGS as well as its comparison with the ATUK was verified on artificial data sets and on a real dataset. Results show a better performance of PUGS for cases with differences in variances. Two examples of the better performance of the PUGS are cases 557 and 526 presented in Table 2. These two cases present a clear difference in standard deviation among control and treatments conditions, but are not identified as a case with evidence for difference among control and treatment conditions by ATUK. However, PUGS consider these both cases as being from a model that considers differences between the control and the two-treatment conditions.

From the biological point of view the results shows that PUGS may illustrate cases not identified when the usual method ATUK is considered. From the statistical point of view the proposed method may be viewed as an effective Bayesian alternative to solve problems of multiple comparison. PUGS can also be easily implemented in usual software such as the R software (The Comprehensive R Archive Network, http://cran.r-project.org). The code used for computing is in the R language and can be obtained by e-mail from the first author.

Here we provide some additional results from simulation study for PUGS and ATUK.



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

(d) $\overline{\text{FDR}}$ (e) $\overline{\text{FDR}}$ (f) $\overline{\text{FDR}}$

Figure A.1: *Average of TPR and FDR for PUGS and ATUK. TPR = number of models correctly identified divided by N; FDR = number of models M_0 incorrectly selected divided by the number of rejected models M_0; ATUK = analysis of variance followed by Tukey test; PUGS = Polya urn within Gibbs sampling.*

Appendix A: Simulated dataset for $T = 3$ without M_4

In this Appendix we present a simulation considering a situation with a control and two experimental condition, similar to described in Section 3.1. However, here we do not generate data from model M_4 which consider $c_1 \neq c_2 \neq c_3$ (see expression in (3.1)). This situation is similar to the real dataset.

The parameters values and sample size are the same used in Section 3.1. We consider the proportion of cases generated from each model are $(0.70, 0.10, 0.10, 0.10)$ for (M_0, M_1, M_2, M_3), respectively.

Figure A.1 show the $\overline{\text{TPR}}$ and $\overline{\text{FDR}}$ for both methods. As we can note, PUGS also present better performance than ATUK for this situation, i.e., higher $\overline{\text{TPR}}$ and smaller $\overline{\text{FDR}}$ than ATUK. As we can note, this better performance is most evident for cases with a difference in variances, $\gamma = 2$ and $\gamma = 3$.

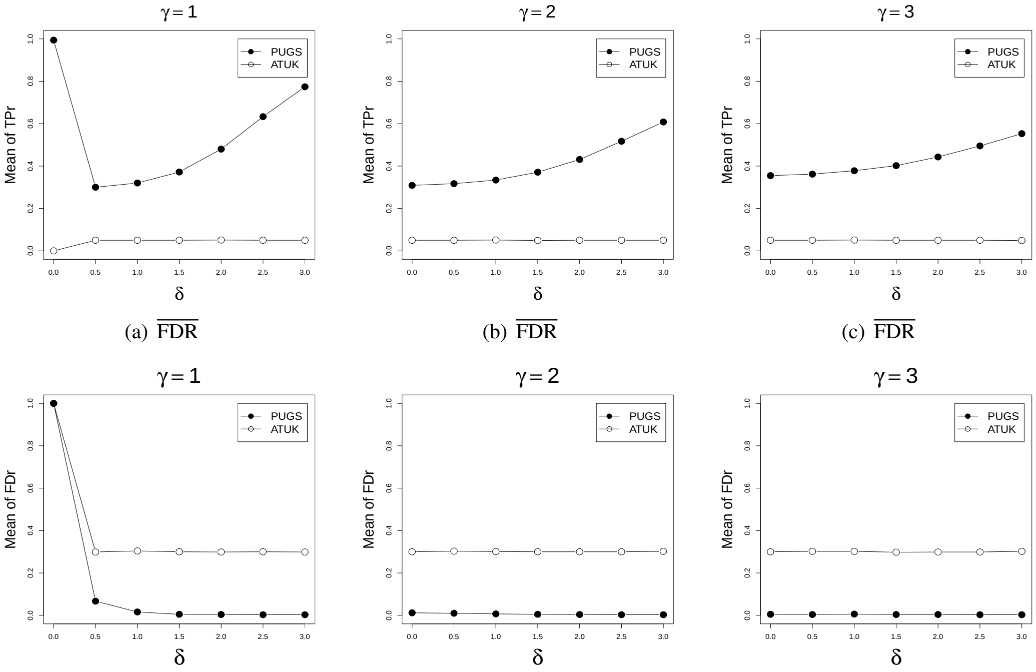
Appendix B: Simulated dataset for $T = 4$

Consider an experimental situation with a control and three treatments condition. For this case we have 15 possible models. Table B.1 describes these models written in terms of latent variables.

In order to generate the data sets we fix proportions generated from each configuration $\mathbf{c}_t$ as 0.30 from $\mathbf{c}_0$ and 0.05 from $\mathbf{c}_t$, $t = 1, \ldots, 14$. The data sets were generated in a similar way as for $M = 3$. Figure B.1 show the $\overline{\text{TPR}}$ and $\overline{\text{FDR}}$ for both methods. As we can note, PUGS also present better

Table B.1: Model configurations

$M_0 : \mathbf{c}_0 = (c_1 = c_2 = c_3 = c_3)$	$M_5 : \mathbf{c}_5 = (c_1 = c_2 \neq c_3 = c_4)$	$M_{10} : \mathbf{c}_{10} = (c_1 = c_4 \neq c_2 \neq c_3)$
$M_1 : \mathbf{c}_1 = (c_1 = c_2 = c_3 \neq c_4)$	$M_6 : \mathbf{c}_6 = (c_1 = c_3 \neq c_2 = c_4)$	$M_{11} : \mathbf{c}_{11} = (c_1 \neq c_2 = c_3 \neq c_4)$
$M_2 : \mathbf{c}_2 = (c_1 = c_2 = c_4 \neq c_3)$	$M_7 : \mathbf{c}_7 = (c_1 = c_4 \neq c_2 = c_3)$	$M_{12} : \mathbf{c}_{12} = (c_1 \neq c_2 = c_4 \neq c_3)$
$M_3 : \mathbf{c}_3 = (c_1 = c_3 = c_4 \neq c_2)$	$M_8 : \mathbf{c}_8 = (c_1 = c_2 \neq c_3 \neq c_4)$	$M_{13} : \mathbf{c}_{13} = (c_1 \neq c_2 \neq c_3 = c_4)$
$M_4 : \mathbf{c}_4 = (c_1 \neq c_2 = c_3 = c_4)$	$M_9 : \mathbf{c}_9 = (c_1 = c_3 \neq c_2 \neq c_4)$	$M_{14} : \mathbf{c}_{14} = (c_1 \neq c_2 \neq c_3 \neq c_4)$



(a) $\overline{\text{FDR}}$ (b) $\overline{\text{FDR}}$ (c) $\overline{\text{FDR}}$

(d) $\overline{\text{FDR}}$ (e) $\overline{\text{FDR}}$ (f) $\overline{\text{FDR}}$

Figure B.1: *Average of TPR and FDR for PUGS and ATUK. TPR = number of models correctly identified divided by N; FDR = number of models M_0 incorrectly selected divided by the number of rejected models M_0; ATUK = analysis of variance followed by Tukey test; PUGS = Polya urn within Gibbs sampling.*

performance than ATUK for this situation. This better performance happens for cases with a difference in variances, $\gamma = 2$ and $\gamma = 3$.

Acknowledgements

The author Erlandson F. Saraiva acknowledges the Brazilian institution CNPq.

References

Antoniak CE (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, **2**, 1152–1174.

Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, and Hatfield GW (2000). Global gene expression profiling in Escherichia coli K12: the effects of integration host factor, *Journal of Biological Chemistry*, **275**, 29672–29684.

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

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

DeRisi JL, Iyer VR, and Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680–686.

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

Ferguson TS (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **2**, 209–230.

Fox RJ and Dimmic MW (2006). A two-sample Bayesian *t*-test for microarray data, *BMC Bioinformatics*, **7**, 126.

Gelfand AE and Smith AFM (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

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

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

Guindani M, Müller P, and Zhang S (2009). A Bayesian discovery procedure, *Journal of the Royal Statistical Society Series B (Statistical Methodology),* , **71**, 905–925.

Hatfield GW, Hung SP, and Baldi P (2003). Differential analysis of DNA microarray gene expression data, *Molecular Microbiology*, **47**, 871–877.

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

Kim SG, Park JS, and Lee YS (2013). Identification of target clusters by using the restricted normal mixture model, *Journal of Applied Statistics*, **40**, 941–960.

Louzada F, Saraiva EF, Milan LA, 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.

MacEachern SN (2016). Nonparametric Bayesian methods: a gentle introduction and overview, *Communications for Statistical Applications and Methods*, **23**, 445–466.

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

Neal RM (1998). Markov chain sampling methods for Dirichlet process mixture models, Technical Report 4915, Retrieved September 1, 2017, from: http://cs.toronto.edu/ redford/mixmc.abstract.html

Oh HS and Yang WY (2006). A Bayesian multiple testing of detecting differentially expressed genes in two-sample comparison problem, *Communications for Statistical Applications and Methods*, **13**, 39–47.

Oh S (2015). How are Bayesian and non-parametric methods doing a great job in RNA-seq differential expression analysis?: a review, *Communications for Statistical Applications and Methods*, **22**, 181–199.

Parkitna JR, 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.

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

Saraiva EF and Milan LA (2012). Clustering gene expression data using a posterior split-merge-birth procedure, *Scandinavian Journal of Statistics*, **39**, 399–415.

Wu TD (2001). Analyzing gene expression data from DNA microarrays to identify candidate genes, *Journal of Pathology*, **195**, 53–65.

Zollanvari A, Cunningham MJ, Braga-Neto U, and Dougherty ER (2009). Analysis and modeling of time-course gene-expression profiles from nanomaterial-exposed primary human epidermal keratinocytes, *BMC Bioinformatics*, **10**, S10.

Zou F, Huang H, and Ibrahim JG (2010). A semiparametric Bayesian approach for estimating the gene expression distribution, *Journal of Biopharmaceutical Statistics*, **20**, 267–280.

Received June 30, 2017; Revised August 31, 2017; Accepted September 13, 2017