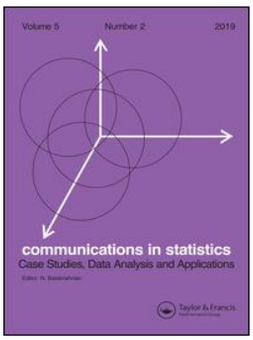


Communications in Statistics: Case Studies, Data Analysis and Applications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucas20

Bayesian criterion for identification of differentially expressed genes

Erlandson F. Saraiva , Luís A. Milan & Carlos Alberto B. Pereira

To cite this article: Erlandson F. Saraiva , Luís A. Milan & Carlos Alberto B. Pereira (2020): Bayesian criterion for identification of differentially expressed genes, Communications in Statistics: Case Studies, Data Analysis and Applications, DOI: 10.1080/23737484.2020.1800535

To link to this article: https://doi.org/10.1080/23737484.2020.1800535

COMMUNICATIONS IN STATISTICS: CASE STUDIES, DATA ANALYSIS AND APPLICATIONS
https://doi.org/10.1080/23737484.2020.1800535

Bayesian criterion for identification of differentially expressed genes

Erlandson F. Saraiva[a], Luís A. Milan[b], and Carlos Alberto B. Pereira[a,c]

[a]Institute of Mathematics, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil; [b]Departmento of Statistics, Federal University of São Carlos, São Carlos, SP, Brazil; [c]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil

ABSTRACT

In this article, we propose a Bayesian criterion for the identification of differentially expressed genes by using the Kullback–Leibler divergence. The advantage of using the Kullback–Leibler divergence is that it allows measuring the influence of the treatment average on the posterior distribution of the parameters of the control distribution. To verify the performance of the proposed method and compare it with the *t*-test and other two Bayesian methods, we developed a simulation study. The comparison is made in terms of the true positive rate and the false discovery rate. The results obtained show a better performance of the proposed method. We also apply the four methods to a real dataset publicly available on the internet.

ARTICLE HISTORY
Received 10 February 2020
Accepted 21 July 2020

KEYWORDS
Gene expression; Bayesian approach; Kullback–Leibler divergence; Cyber-T

1. Introduction

A common interest in the analysis of gene expression data is the identification of differentially expressed (DE) genes between a treatment condition and a control condition. The interest in the identification of these genes is that this allows to study and detect possible relationships between genes and between genes and proteins. In addition, it also allows identifying which genes may be involved in the origin and/or evolution of diseases with genetic origin or which genes react to a drug stimulus (Schena et al. 1995; DeRisi, Iyer and Brown 1997; Arfin et al. 2000; Lonnstedt and Speed 2001; Wu 2001).

A method commonly used to identify DE genes is the two-sample *t*-test (TT) for the log-transformed data (Baldi and Long 2001; Hatfield, Hung, and Baldi 2003). However, a problem often encountered in the application of the *t*-test for this kind of dataset is the small sample size, which may lead to low test power. As an alternative to the *t*-test, Baldi and Long (2001) consider a Bayesian approach and propose the Cyber-*t* test. Following the work of Baldi and Long (2001), Fox and Dimmic (2006) propose the Bayesian *t*-test (BT). In both methods, the

CONTACT Erlandson F. Saraiva erlandson.saraiva@ufms.br Institute of Mathematics, Federal University of Mato Grosso do Sul, Campo Grande, MS 79070-900, Brazil.

Supplemental data for this article can be accessed on the publisher's website.

authors consider modifications of the standard error estimate of the two sample differences present in the denominator of the standard t statistics.

In this article, we adopt a Bayesian approach and consider the identification of DE genes by using the influence of the treatment average on the posterior distribution of the parameters of the control distribution. To measure this influence, we consider the Kullback–Leibler divergence (Kullback and Leibler 1951; Cover and Tomas 1991), hereafter denoted by index KLD. This procedure was motivated by the work of Song (2002), who developed a goodness-of-fit test using the KLD. Also based on Song's (2002) work, Pérez-Rodrigues, Vaquera-Huerta, and Villaseñor-Alva (2009) propose a goodness-of-fit test for the Gumbel distribution. Girardin and Lequesne (2019) proposed the unification of the works of Song (2002) and Vasicek (1976) in a unique goodness-of-fit test and applied it to a DNA dataset.

To decide which are the DE genes, we consider a threshold κ in such a way that if the index KLD is larger than κ for a specific gene, it is considered DE, otherwise it is not. This threshold value was obtained according to the proposal of Peng and Dey (1995).

To verify the performance of the proposed method and compare it with the three methods cited earlier, we developed a simulation study. In this simulation study, we compare the performance of the methods in terms of the true positive rate and the false discovery rate. Results show a better performance for KLD, i.e., greater true positive rate and smaller false discovery rate, especially, for the case with the difference of means and variances. We also applied the four methods to a real dataset downloaded from the website http://cybert.ics.uci.edu/controlexp.

The remainder of the article is structured as follows. In Sec. 2, we describe the Bayesian approach and the criterion based on the KLD. In Sec. 3, the proposed method is applied to simulated datasets and to a real dataset. Sec. 4 concludes the article with the final remarks.

2. Bayesian model for gene expression data analysis

Following Louzada et al. (2014), consider a DNA array experiment with n genes and two experimental conditions which we name control (c) and treatment (t). Suppose that control and treatment are replicated n_c and n_t times, respectively. Denote by x_{ig_h} the ith observed expression level (or its logarithm) for gene g in experimental condition h, $h \in \{c,t\}$ and $g \in \{1,\ldots,n\}$. Let $\mathbf{x}_{g_h} = (x_{1g_h},\ldots,x_{n_hg_h})$ be realizations of the vector of independent random variables $\mathbf{X}_{g_h} = (X_{1g_h},\ldots,X_{n_hg_h})$, for $g = 1,\ldots,n$ and $h \in \{c,t\}$.

Assume that observed data (log-transformed) are generated from normal distributions with mean μ_{g_h} and variance $\sigma^2_{g_h}$, $X_{ig_h} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{g_h},\sigma^2_{g_h})$, for $i = 1,\ldots,n_h$, $h \in \{c,t\}$, and $g = 1,\ldots,n$. This normality assumption for the data is common in gene expression data analysis (see, e.g., Baldi and Long 2001;

Fox and Dimmic 2006; Hatfield, Hung, and Baldi 2003; Louzada et al. 2014; Saraiva and Milan 2012 and their references).

We specify the joint prior distributions for μ_{g_c} and $\sigma^2_{g_c}$ such that

$$\pi(\mu_{g_c}, \sigma^2_{g_c}) = \pi(\mu_{g_c}|\sigma^2_{g_c})\pi(\sigma^2_{g_c}),$$

for $g = 1, \ldots, n$. So, we consider the following prior distributions

$$\mu_{g_c}|\mu_{0g}, \lambda, \sigma^2_{g_c} \sim \mathcal{N}\left(\mu_{0g}, \frac{\sigma^2_{g_c}}{\lambda}\right),$$

$$\sigma^{-2}_{g_c}|\mathbf{x}_{g_c}, \alpha_g, \beta_g \sim \Gamma(\alpha_g, \beta_g),$$

where μ_{0g}, λ, α_g, and β_g are known hyperparameters, for $g = 1, \ldots, n$. The parameterization of the Gamma distribution is so that the mean is α/β and the variance is α/β^2. These prior distributions were also used by Casella, Robert, and Wells (2000), Nobile and Fearnside (2007), and Saraiva et al. (2016).

Since it may be unrealistic to assume the availability of strong prior information regarding parameters $(\mu_{g_c}, \sigma^2_{g_c})$ in practice, we specify the hyperparameters values as $\mu_{0g} = \varepsilon_g$ and $E(\sigma^{-2}_{g_c}) = \frac{1}{10R_g}$, where ε_g is the midpoint of the observed variation interval of the data $\mathbf{x}_{g_c}$ and R_g is the length of this interval. Thus, we obtain $\beta_g = \frac{\alpha_g}{10R_g}$ and we fix $\alpha_g = 1$, for $g = 1, \ldots, n$. We also fix the hyperparameter $\lambda = 10^{-1}$ to get a prior distribution for component means with large variance.

The joint posterior distribution upon which inference is based is given by

$$\pi(\mu_{g_c}, \sigma^2_{g_c}|\mathbf{x}_{g_c}) \propto L\left(\mu_{g_c}, \sigma^2_{g_c}|\mathbf{x}_{g_c}\right)\pi(\mu_{g_c}, \sigma^2_{g_c}),$$

where $L\left(\mu_{g_c}, \sigma^2_{g_c}|\mathbf{x}_{g_c}\right)$ is the likelihood function from the normal distribution.

The conditional posterior distributions for parameters are

$$\mu_{g_c}|\mathbf{x}_{g_c}, \mu_{0g}, \lambda_g, \sigma^2_{g_c} \sim \mathcal{N}\left(\mu_g^{\text{post}}, \frac{\sigma^2_{g_c}}{n_j + \lambda_g}\right) \quad \text{and}$$

$$\sigma^{-2}_{g_c}|\mathbf{x}_{g_c}, \alpha_g, \beta_g \sim \Gamma\left(\alpha_g^{\text{post}}, \beta_g^{\text{post}}\right), \tag{1}$$

where

$$\mu_g^{\text{post}} = \frac{\sum_{i=1}^{n_c} x_{ig_c} + \lambda\mu_0}{n_c + \lambda},$$

$$\alpha_g^{\text{post}} = \alpha_g + \frac{n_c + 1}{2},$$

$$\beta_g^{\text{post}} = \beta_g + \frac{\sum_{i=1}^{n_c} x^2_{ig_c} + \lambda\mu_0^2}{2} - \frac{\left(\sum_{i=1}^{n_c} x_{ig_c} + \lambda\mu_0\right)^2}{2(n_c + \lambda)},$$

for $g = 1, \ldots, n$.

Thus, we can estimate the parameters of control distribution using a Gibbs sampling algorithm (Geman and Geman 1984). In this algorithm, at each iteration, values for $(\mu_{g_c}, \sigma^2_{g_c})$ are generated from conditional posterior distributions in Eq. (1), for $g = 1, \ldots, n$. The estimates are given by the average of the generated values for each parameter.

2.1. Influence measure based on Kullback–Leibler divergence

We now consider the KLD to measure the influence of the observed treatment average $\bar{x}_{g_t}$ on the posterior distribution for $\theta_{g_c} = (\mu_{g_c}, \sigma^2_{g_c})$, for $g = 1, \ldots, n$. If $\bar{x}_{g_t}$ is influential then we consider that there exists evidence for the difference between expression levels of treatment and control.

Thus, consider $\mathbf{x}^*_g = \{\mathbf{x}_{g_c}\} \cup \{\bar{x}_{g_t}\}$ and $\pi_1 = \pi(\theta_{g_c}|\mathbf{x}^*_g)$ and $\pi_2 = \pi(\theta_{g_c}|\mathbf{x}_{g_c})$ be the posterior distributions for $\theta_{g_c} = (\mu_{g_c}, \sigma^2_{g_c})$ given the observed data $\mathbf{x}^*_g$ and $\mathbf{x}_{g_c}$, respectively, for $g = 1, \ldots, n$.

The KLD measure between π_1 and π_2 is given by

$$\mathrm{KLD}(\pi_1||\pi_2) = \int_0^{+\infty}\int_{-\infty}^{+\infty} \log\left(\frac{\pi_1(\theta_{g_c}|\mathbf{x}^*_g)}{\pi_2(\theta_{g_c}|\mathbf{x}_{gc})}\right)\pi_1(\theta_{g_c}|\mathbf{x}^*_g)d\mu_{g_c}d\sigma^2_{g_c} = \sum_{r=1}^{6} K_r, \quad (2)$$

where

$$K_1 = \frac{1}{2}\log\left(\frac{n_c+1+\lambda}{n_c+\lambda}\right) - \frac{1}{2} + \frac{n_c+\lambda}{2(n_c+1+\lambda)}, \quad K_4 = \log\left(\frac{\Gamma(\alpha_g^{\mathrm{post}*})}{\Gamma(\alpha_g^{\mathrm{post}})}\right),$$

$$K_2 = (n_c+\lambda)\frac{(\mu_g^{\mathrm{post}*} - \mu_g^{\mathrm{post}})^2}{2}\frac{\alpha_g^{\mathrm{post}*}}{\beta_g^{\mathrm{post}*}}, \qquad K_5 = \left(\alpha_g^{\mathrm{post}*} - \alpha_g^{\mathrm{post}}\right)\psi(\alpha_g^{\mathrm{post}*}),$$

$$K_3 = \alpha_g^{\mathrm{post}}\log\left(\frac{\beta_g^{\mathrm{post}*}}{\beta_g^{\mathrm{post}}}\right), \qquad K_6 = \left(\beta_g^{\mathrm{post}*} - \beta_g^{\mathrm{post}}\right)\frac{\alpha_g^{\mathrm{post}*}}{\beta_g^{\mathrm{post}*}},$$

in which, $\mu_g^{\mathrm{post}*}$, $\alpha_g^{\mathrm{post}*}$, and $\beta_g^{\mathrm{post}*}$ are calculated using $\mathbf{x}^*_g$ and $\psi(a)$ is the digamma function, for $g = 1, \ldots, n$.

However, the influence measure $\mathrm{KLD}(\pi_1||\pi_2)$ does not determine when an observation is influential. For this, we need to define a cut-off point κ to determine whether $\bar{x}_{g_t}$ is influential or not. Nonetheless, as discussed by Peng and Dey (1995) and Weiss (1996), it is a very difficult task to define a cut-off point for the divergence measure to determine whether an observation is influential or not. To overcome this difficulty, these authors propose to define a cut-off point using a procedure based on the divergence measure for a biased coin with success probability p, for $0 < p < 1$.

To define the cut-off point we consider the proposal of Peng and Dey (1995) and Weiss (1996) adapted for the gene expression data analysis context, explained next. Assume all genes have the same probability p of being considered

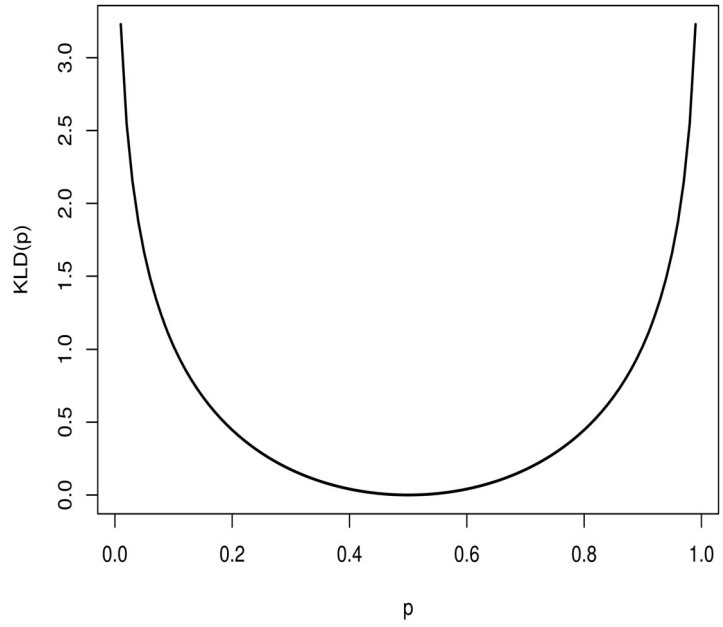


Figure 1. Kullback–Leibler divergence.

DE, for $0 < p < 1$. Let Z be an indicator variable that takes the value 1 whether a gene is DE and 0 otherwise. Thus, we have that Z follows a Bernoulli distribution wit success probability p, $Z \sim \text{Bernoulli}(p)$, and probability function $\pi_1(z|p) = p^z(1-p)^{1-z}$, for $z = 0, 1$. As a second scenery, consider that a gene is declared DE at random with $p = 0.5$. For this case, the probability function is given by $\pi_2(z|p = 0.5) = 0.5$. According to Peng and Dey (1995), the KLD between $\pi_1(z|p)$ and $\pi_2(z|p = 0.5)$ is given by

$$\text{KLD}(p) = \frac{-\log(2p) - \log(2(1-p))}{2}. \tag{3}$$

Figure 1 shows the graphic of KLD(p). As one can note, KLD(p) increases as p moves away from 0.5, is symmetric around $p = 0.5$ and achieves its minimum at $p = 0.5$. For $p = 0.5$, KLD$(0.5) = 0$ and $\pi_1(z|p) = \pi_2(z|p)$.

Thus, if we consider $p \geq 0.95$ (or $p \leq 0.05$) as an adequate probability to declare a gene as DE, then, since KLD$(0.95) = 0.8304$, we can indicate a gene as DE when KLD(p) > 0.8304. Thus, we set up the cut-off value $\kappa = 0.8304$ and consider that $\overline{x}_{g_t}$ is an influential observation if KLD$(\pi_1||\pi_2)$ given in Eq. (2) is greater than κ, for $g = 1, \ldots, n$. For these cases, there exist evidence for the difference between the observed expression levels in treatment in relation to the control condition.

At this point, it is important to note that as in the use of hypothesis tests where other values for significance level different from 5% can be chosen, in our case, the choice of $\kappa = 0.8304$ can also be questioned. However, a possible rationale may arise from looking at Fig. 1. If the interest is to use the

methodology described here to detect DE genes, then we can choose a value of κ that corresponds to a value of p far from $\frac{1}{2}$ and near to 1 or 0, in Eq. (3). Based on our experience with simulated data the value $\kappa = 0.8304$ obtained from the choice of $p = 0.95$ leads to satisfactory results.

3. Data analysis

In this section, we illustrate the performance of the proposed method by using simulated data sets and a real data set.

3.1. Simulated datasets

To generate the data sets, we follow the procedure proposed by Louzada et al. (2014). For this, we fix $\mu_{g_c} = -8.5$ and $\sigma^2_{g_c} = 0.40$. These values are the average of the observed means and variances from the control of the real dataset. We then set up

$$\mu_{g_t} = \mu_{g_c} \pm \delta\sigma_{g_c} \quad \text{and} \quad \sigma_{g_t} = \gamma\sigma_{g_c},$$

for $\delta = \{0, 0.50, 1, 1.50, 2, 2.50, 3, 3.50, 4\}$ and $\gamma = \{1, 2, 3, 4\}$, where the signal $\pm$ in expression for μ_{g_t} represent the situation over and under expressed, respectively. Besides, we fix $n = 1{,}000$ and $n_c = n_t = 4$.

The procedure to generate the artificial data sets is given by the following four steps:

(i) For $g = 1, \ldots, n$, generate $X_{ig_c} \sim \mathcal{N}(\mu_{g_c}, \sigma^2_{g_c})$, for $i = 1, \ldots, n_c$;
(ii) Choose randomly $w = 10\%$ of the indexes $\{1, \ldots, n\}$ to indicate the cases to be generated with a difference. We use $w = w_{\text{over}} + w_{\text{under}}$, for $w_{\text{over}} = w_{\text{under}} = 5\%$.
(iii) If the index $g \in \{1, \ldots, n\}$ was chosen, then consider an indicator variable $\mathbb{I}_g = 1$ and generate $X_{ig_t} \sim \mathcal{N}(\mu_{g_t}, \sigma^2_{g_t})$, for $i = 1, \ldots, n_t$;
(iv) If the index $g \in \{1, \ldots, n\}$ was not chosen, then set up $\mathbb{I}_g = 0$ and generate $X_{ig_t} \sim \mathcal{N}(\mu_{g_c}, \sigma^2_{g_c})$, for $i = 1, \ldots, n_t$.

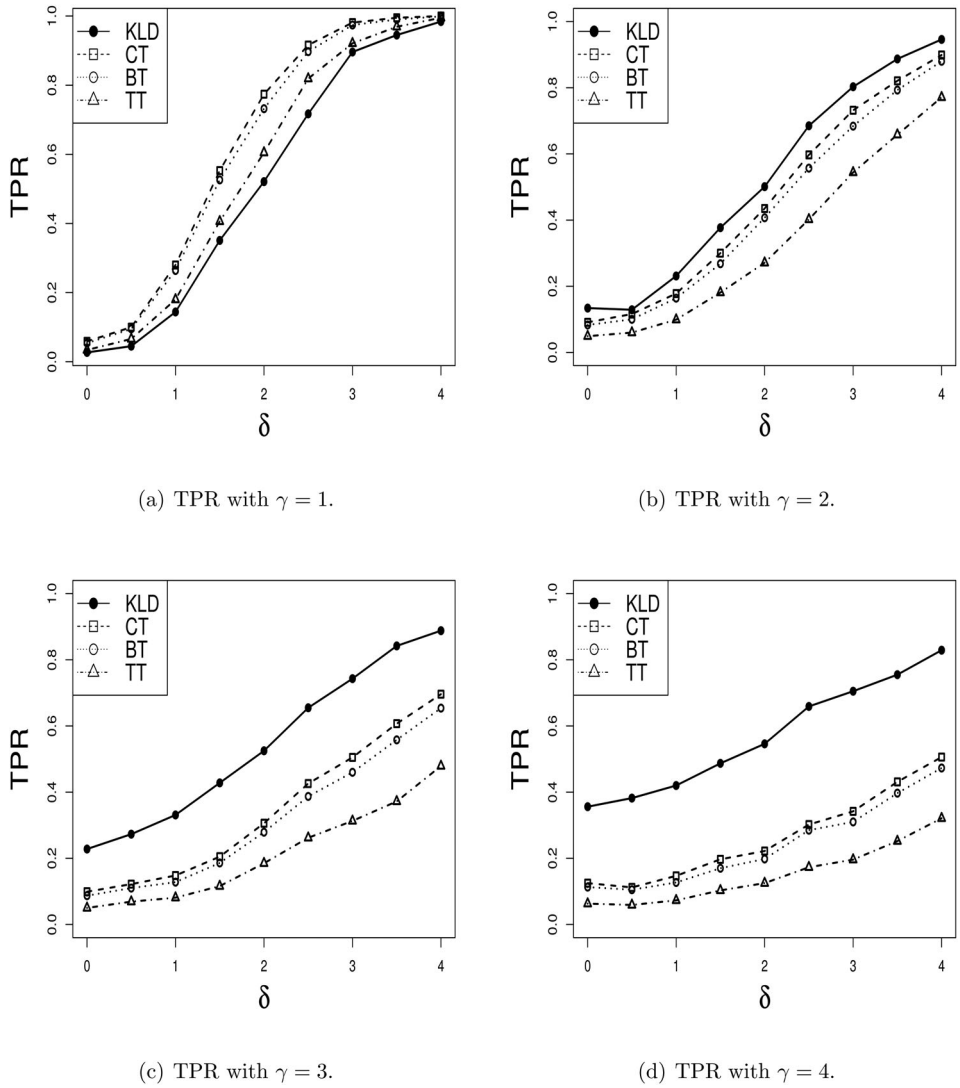
To record the cases identified with difference by KLD method, we consider an indicator variable $\mathbb{I}_g^{\text{KLD}}$ that assumes the value 1 if $\text{KLD}(\pi_1||\pi_2) > \kappa$ and $\mathbb{I}_g^{\text{KLD}} = 0$ otherwise. Analogously, for t-tests we consider $\mathbb{I}_g^{t\text{-tests}} = 1$ if $p\text{-value}_g < \frac{\alpha}{2}$ and $\mathbb{I}_g^{t\text{-tests}} = 0$ otherwise, for $\alpha = 0.05$. Then, we calculate the true positive rate and the false discovery rate,

$$\text{TPR}_M = \frac{\sum_{g=1}^{n} \mathbb{I}_g \cdot \mathbb{I}_g^M}{\sum_{g=1}^{n} \mathbb{I}_g} \quad \text{and} \quad \text{FDR}_M = \frac{\sum_{g=1}^{n} (1 - \mathbb{I}_g) \cdot \mathbb{I}_g^M}{\sum_{g=1}^{n} \mathbb{I}_g},$$

where $M \in \{\text{KLD}, \text{TT}, \text{CT}, \text{BT}\}$. The method with better performance should present a higher TPR values and smaller FDR values.

We also generate $L = 100$ different artificial data sets for each pair (δ, γ) according to steps (i)–(iv) described above. We present results according to $\overline{\text{TPR}}_M$, the average of TPR from $L = 100$ simulated data sets. Figures 2 and 3 show the plot of $\overline{\text{TPR}}_M$ and $\overline{\text{FDR}}_M$ for each pair (δ, γ) used, respectively, for each method $M \in \{\text{KLD}, \text{CT}, \text{BT}\}$.

The graphs in Figs. 2 and 3 show that the KLD method performs better than the t-tests for most of the simulated cases. An exception is the TPR cases with $\gamma = 1$ fixed, the condition of equal variances between case/control distributions.



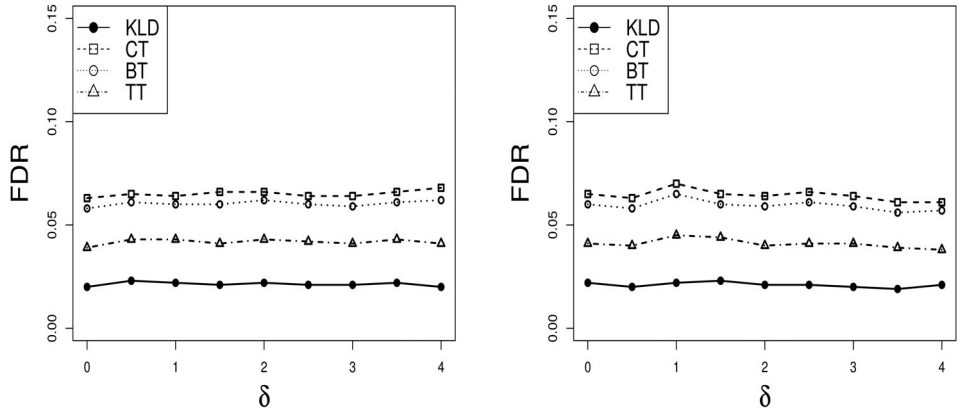
(a) TPR with $\gamma = 1$.

(b) TPR with $\gamma = 2$.

(c) TPR with $\gamma = 3$.

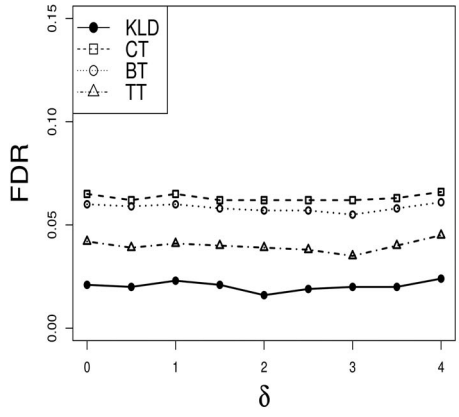
(d) TPR with $\gamma = 4$.

Figure 2. $\overline{\text{TPR}}_M$ for each pair (δ, γ) considered.

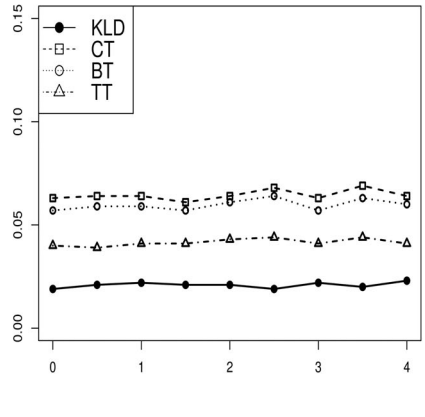


(a) FDR with $\gamma = 1$.

(b) FDR with $\gamma = 2$.



(c) FDR with $\gamma = 3$.



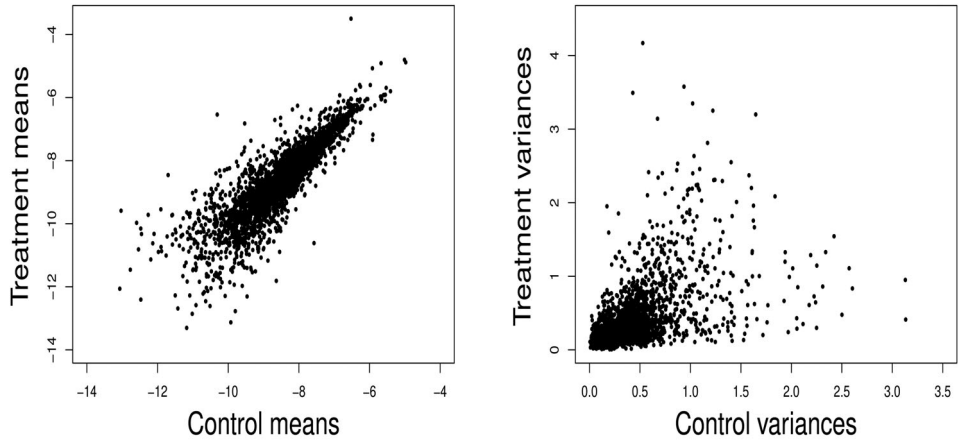
(d) FDR with $\gamma = 4$.

Figure 3. $\overline{\text{TPR}}_M$ for each pair (δ, γ) considered.

One possible explanation for the better performance of the KLD method in most cases is that it is a Bayesian method that uses parameter distributions with a focus on the parametric space, as a whole, and not only in the estimated central values as is the case of other methods.

3.2. Application

We now consider the gene expression dataset publicly available on the website http://cybert.ics.uci.edu/controlexp. This data set is composed of $n = 2{,}758$ genes and refers to a small amount of control and experimental data from a DNA microarray. Each gene g has four measures from control and four measures from the treatment. Figure 4 shows the observed averages and variances from



(a) Observed means. (b) Observed variances.

Figure 4. Treatment and control observed means and variances.

treatment versus the observed averages and variances from control for the n genes of this dataset.

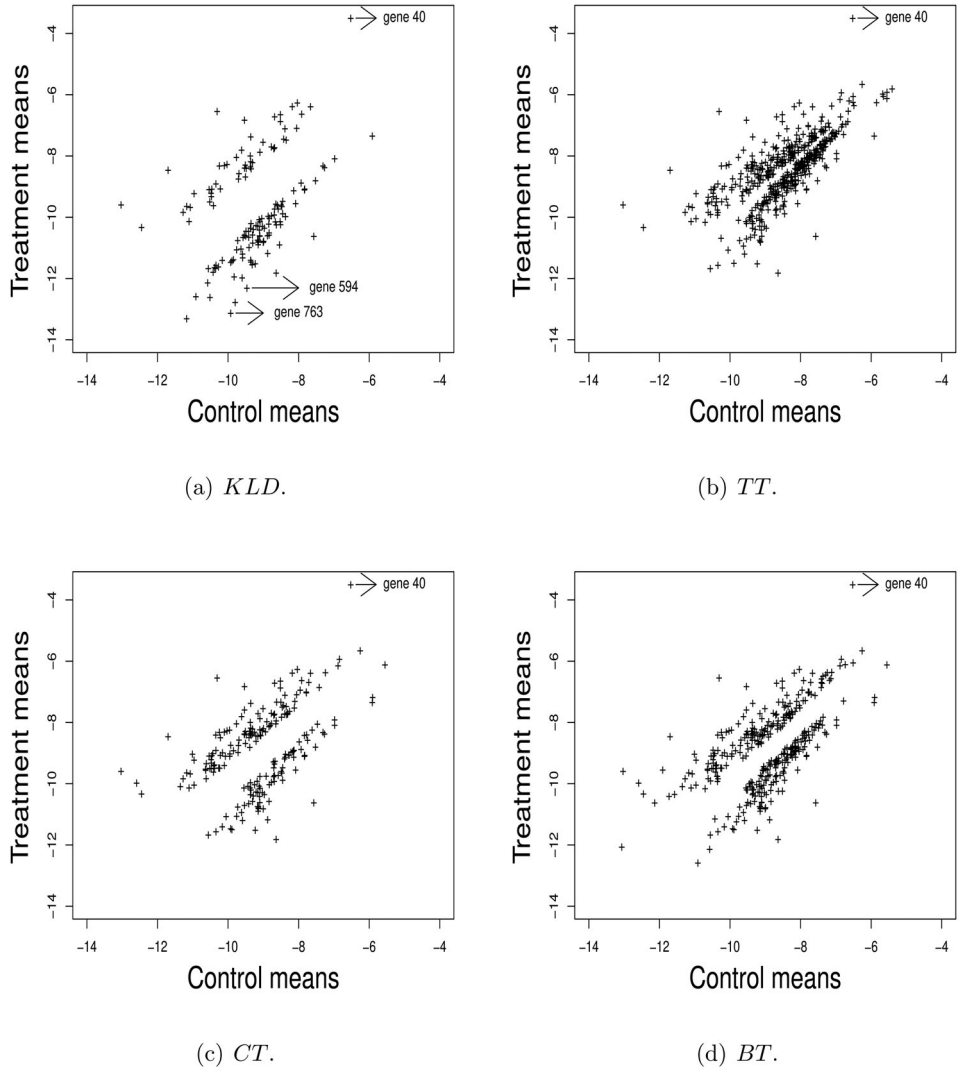
For the application of the KLD, we consider the same cut-off $\kappa = 0.8304$ value to define a gene as DE. The hyperparameters values were specified in the same way as in the simulation study. The t-tests were applied with a significance level $\alpha = 0.05$.

The KLD identifies 144 genes with evidence for difference, while TT identifies 482 genes, CT identifies 207, and BT identifies 347 genes. Out of case identified with difference, 90 were identified by the four methods and 120 cases were identified with difference by the three Bayesian methods (KLD, CT, and BT).

Figure 5 shows the observed treatment and control averages for the cases identified with evidence for the difference by four methods. Figure 6 shows the observed treatment and control variances of the cases identified with evidence for the difference.

As one can note in Fig. 5, the gene 40 (oppA) is identified as DE by the four methods. This gene has the highest absolute difference between averages of treatment and control. However, as can be viewed in Fig. 6, cases with the highest absolute differences of variances are not identified by the t-tests. Two examples are the genes 594 (menC) and 763 (yihT) that are highlighted in Figs. 5a and 6a. Table 1 shows the observed treatment and control observed means and variances for the cases cited above. This table also shows the KLD value and the p-value from t-tests.

Table 2 shows the observed treatment and control observed means and variances for the ten most significant case identified with difference but KLD. This table also shows the p-value from t-tests.



(a) KLD. (b) TT.

(c) CT. (d) BT.

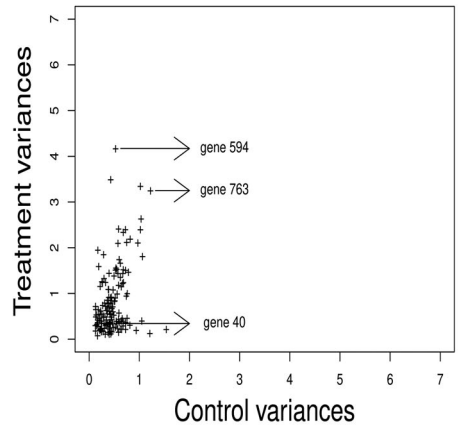
Figure 5. Treatment and control observed means for cases identified with evidence for difference.

4. Final remarks

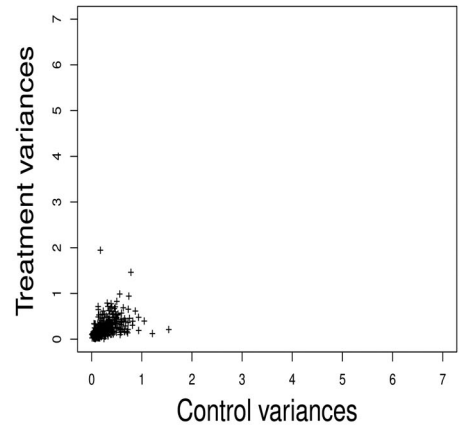
We proposed a Bayesian approach to identify DE genes considering the influence of the treatment average in the posterior distribution for the parameters of the control distribution. The influence was measured through the Kullback–Leibler divergence.

To identify the cases with evidence for the difference, we establish a criterion that considers the treatment average as being an influential observation if the KLD value is greater than the value $\kappa = 0.8304$.

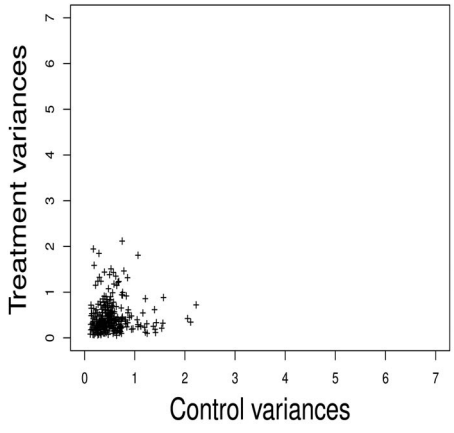
Although we have considered a cut-off value $\kappa = 0.8304$, another user has the option to consider another cut-off value which is calculated according to expression (3) for a fixed value of p. Thus, for example, if an user consider



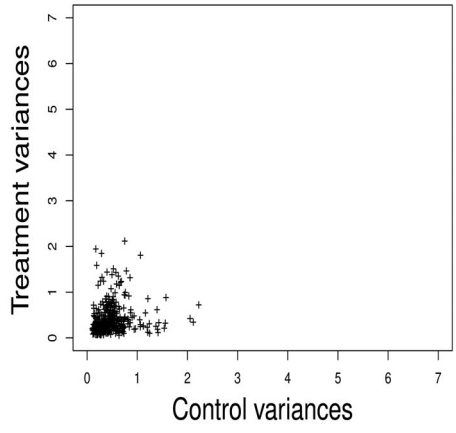
(a) KLD.



(b) TT.



(c) CT.



(d) BT.

Figure 6. Treatment and control observed variances for cases identified with evidence for difference.

Table 1. Treatment and control observed means and variances for genes 40, 594, and 763.

Gene	Control		Treatment		KLD value	*p*-value		
	Average	Variance	Average	Variance		TT	CT	BT
40	−6.5243	0.5491	−3.4981	0.3420	4.4751	0.0001	<0.0001	<0.0001
594	−9.4657	0.5280	−12.3068	4.1694	4.2603	0.1333	0.1096	0.0940
763	−9.9263	1.2219	−13.1280	3.2507	1.5890	0.0711	0.0540	0.0534

Table 2. The ten most significant genes identified with evidence by KLD.

Gene	Control		Treatment		KLD value	*p*-value		
	Average	Variance	Average	Variance		TT	CT	BT
15	−7.5710	0.2552	−10.6148	0.5542	6.5317	0.0002	<0.0001	<0.0001
11	−8.6385	0.1731	−11.8122	1.9511	6.3296	0.0234	0.0079	0.0053
51	−9.5336	0.3324	−6.8211	0.3296	5.4489	<0.0001	<0.0001	<0.0001
2752	−9.7992	0.4291	−12.7735	3.4927	5.2377	0.0934	0.0678	0.0555
52	−8.8761	0.1904	−11.1700	1.5937	4.5506	0.0313	0.0138	0.0094
40	−6.5243	0.5491	−3.4981	0.3420	4.4751	0.0001	<0.0001	<0.0001
594	−9.4657	0.5280	−12.3068	4.1694	4.2603	0.1333	0.1096	0.0940
230	−9.3605	0.2870	−11.3954	1.8535	4.1033	0.0572	0.0349	0.0264
55	−9.3584	0.3315	−7.3667	0.3052	3.8695	0.0001	0.0001	<0.0001
2180	−10.2446	0.3450	−8.3140	0.4271	3.5324	0.0002	0.0002	0.0001

$p = 0.90$ in Eq. (3), then a treatment average is an influential observation whether KLD value is greater than 0.5108. Reducing the value of p in Eq. (3) the κ value also reduces. As a consequence, the number of cases identified with evidence for the difference may increases. On the other hand, increasing the value of p the κ value also increases and the number of cases identified with evidence for the difference may reduce.

Results from simulated datasets show a better performance of the KLD in relation to the t-tests, i.e., greater true positive rate and smaller false discovery rate. An exception is the case with the difference of means ($\delta > 0$) and the same variance ($\gamma = 1$), in which the t-tests present a higher true positive rate than KLD. However, for this case, the false discovery rate of the KLD is smaller than t-tests. In the real dataset, the cases with the highest absolute difference between observed averages and variances are identified with evidence for the difference by KLD and are not identified by t-tests.

Although the article does not present a new theoretical result from the mathematical viewpoint, the simulation study and the application highlight the following three advantages of the proposed method: (1) it is easier to use like t-tests, (2) it performs well in situations with small sample sizes which are common in gene expression data analysis, and (3) present better performance than t-tests for cases with difference o means and variances. From the biological practical point of view, it indicates the KLD may identify case DE which are not identified by usual t-tests methods, TT, CT, and BT.

The computational codes used in the simulation study and in the application to the real dataset are in the R language. In Appendix S-3 of the Supplementary Material, we present the codes used in the application of the KLD method to the real dataset. The extension of the proposed method to the second level of analysis to identify clusters of genes can be viewed a future work.

Acknowledgments

The authors are grateful to the editor and referees for helpful comments and suggestions which have led to an improvement of this article.

Funding

This study was financed in part by the Fundação Universidade Federal de Mato Grosso do Sul—UFMS/MEC—Brasil and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

References

Arfin, S. M., A. D. Long, E. T. Ito, L. Tolleri, M. M. Riehle, E. S. Paegle, and G. W. Hatfield. 2000. "Global Gene Expression Profiling in *Escherichia coli* K12." *Journal of Biological Chemistry* 275:29672–29684.

Baldi, P., and D. A. Long. 2001. "Bayesian Framework for the Analysis of Microarray Expression Data: Regularized *t*-Test and Statistical Inferences of Gene Changes." *Bioinformatics* 17:509–519.

Casella, G., C. Robert, and M. Wells. 2000. "Mixture Models, Latent Variables and Partitioned Importance Sampling." Technical Report 2000-03, CREST, INSEE, Paris.

Cover, T. M., and J. A. Thomas. 1991. *Elements of Information Theory*. *Wiley Series in Telecommunications*. Hoboken, NJ: Wiley.

DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale." *Science* 278:680–686.

Fox, R. J., and M. W. Dimmic. 2006. "A Two-Sample Bayesian *t*-Test for Microarray Data." *BMC Bioinformatics* 7:126.

Geman, S., and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–741.

Girardin, V., and J. Lequesne. 2019. "Entropy-Based Goodness-of-Fit Tests—A Unifying Framework: Application to DNA Replication." *Communications in Statistics—Theory and Methods* 48:62–74.

Hatifield, G. W., S. Hung and P. Baldi. 2003. "Differential Analysis of DNA Microarray Gene Expression Data." *Molecular Microbiology* 47:871–877.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22:79–86.

Lonnstedt, I., and T. P. Speed. 2001. "Replicated Microarray Data." *Statistica Sinica* 12:31–46.

Louzada, F., E. F. Saraiva, L. A. Milan, and J. Cobre. 2014. "A Predictive Bayes Factor Approach to Identify Genes Differentially Expressed: An Application to *Escherichia coli* Bacterium Data." *Brazilian Journal of Probability and Statistics* 28:167–189.

Nobile, A., and A. T. Fearnside. 2007. "Bayesian Finite Mixtures With an Unknown Number of Components: The Allocation Sampler. *Statistics and Computing* 17:147–162.

Peng, F., and D. Dey. 1995. "Bayesian Analysis of Outlier Problems Using Divergence Measures." *The Canadian Journal of Statistics* 23:199–213.

Pérez-Rodríguez, P., H. Vaquera-Huerta, and J. A. Villaseñor-Alva. 2009. "A Goodness-of-Fit Test for the Gumbel Distribution Based on Kullback–Leibler Information." *Communications in Statistics—Theory and Methods* 38:842–855.

Saraiva, E. F., and L. A. Milan. 2012. "Clustering Gene Expression Data Using a Posterior Split-Merge-Birth Procedure." *Scandinavian Journal of Statistics* 39:399–415.

Saraiva, E. F., A. K. Suzuki, F. Louzada, and L. A. Milan. 2016. "Partitioning Gene Expression Data by Data-Driven Markov Chain Monte Carlo." *Journal of Applied Statistics* 43:1155–1173.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray." *Science* 270:467–470.

Song, K. S. 2002. "Goodness-of-Fit Tests Based on Kullback–Leibler Discrimination Information." *IEEE Transactions on Information Theory* 48:1103–1117.

Vasicek, O. 1976. "A Test for Normality Based on Sample Entropy." *Journal of the Royal Statistical Society: Series B (Methodological)* 38:54–59.
Weiss, R. 1996. "An Approach to Bayesian Sensitivity Analysis." *Journal of the Royal Statistical Society: Series B (Methodological)* 58:739–750.
Wu, T. D. 2001. "Analyzing Gene Expression Data From DNA Microarray to Identify Candidates Genes." *Journal of Pathology* 195:53–65.