Journal of Applied Statistics

Vol. 00, No. 00, August 2009, 1–27

RESEARCH ARTICLE

Evaluation of Laplace distribution-based ANOVA models applied to microarray data

Suzy Van Sanden* and Tomasz Burzykowski

Interuniversity Institute for Biostatistics and statistical Bioinformatics,

Universiteit Hasselt, B3590 Diepenbeek, Belgium,

and Katholieke Universiteit Leuven, Belgium

(*Received 27 August 2009*)

In a microarray experiment, intensity measurements tend to vary due to various systematic and random effects, which enter at the different stages of the measurement process. Common test statistics do not take these effects into account. An alternative is to use, e.g., ANOVA models. In many cases, we can, however, not make the assumption of normally distributed error terms. Purdom and Holmes [8] have concluded that the distribution of microarray intensity measurements can often be better approximated by a Laplace distribution. In this paper, we consider analysis of microarray data by using ANOVA models under the assumption of Laplace-distributed error terms. We explain the methodology and discuss problems related to fitting of this type of models. In addition to evaluating the models using several real-life microarray experiments, we conduct a simulation study to investigate different aspects of the models in detail. We find that, while the normal model is less sensitive to model misspecifications, the Laplace model has more power when the data are truly Laplace distributed. However, in the latter situation, neither of the models is able to control the FDR at the pre-specified significance level. This problem is most likely related to sample size issues.

Keywords: ANOVA models; gene expression; Laplace distribution; microarrays; simulation

*Corresponding author. Universiteit Hasselt; Campus Diepenbeek; Agoralaan - Gebouw D; 3590 Diepenbeek; Belgium; Email: suzy.vansanden@uhasselt.be; Tel: +32-11-26 82 81

ISSN: 0233-1934 print/ISSN 1360-0532 online

DOI: 10.1080/0266476YYxxxxxxxx

http://www.informaworld.com

study.

1. Introduction

Microarrays have been around since the early nineties of the XXth century. Though the technology has evolved quite a bit since then, some of the problems, inherent to a microarray experiment, are still present. Typically, such an experiment leads to vast amounts of data needing to be analyzed. Unfortunately, the data available per gene are generally limited, because, due to the high cost of a single slide, the number of arrays available to a researcher tends to be small. In addition, the technology itself is not without problems. Various stages of the experiment are vulnerable to inclusion of systematic variability, originating from sources other than the difference between the samples. Due to the complex design of a typical microarray experiment, more sophisticated statistical methods might be required for the normalization and analysis of the data.

Microarrays data are often analyzed by using non-parametric methods [13], empirical Bayes methods like, e.g., LIMMA [10], permutations-based methods like, e.g., SAM [12], or by ANOVA models [6]. In this paper, we focus on the latter. Parametric models are in general more powerful than non-parametric techniques. They offer more insight into the data and are less computationally intensive compared to resampling methods. An ANOVA model can take into account several of the systematic sources of variability present in microarray data. The error terms of the model are assumed to be independent and identically distributed. The assumption that they follow the normal distribution is, however, most of the time not fulfilled. Purdom and Holmes [8] indicate that the distribution of the residuals often tends to be heavy-tailed and/or to exhibit skewness of varying degrees.

Instead of relying on the normality assumption, a possible solution is to turn to

bootstrap techniques [4, 6]. This is practical when dealing with a moderate number of genes, but it becomes computationally intensive and time consuming for many current microarrays, which contain tens of thousands of genes.

Another possible approach is to find a distribution that approximates better that of microarray data. Purdom and Holmes [8] proposed to use the (asymmetric) Laplace distribution. They found it to fit microarray data often better than the normal distribution. The Laplace distribution is more peaked in the center and more heavy-tailed, as compared to the normal. However, this approach is also not without problems.

In this paper, we evaluate the use of ANOVA models with a Laplace distributed error term to analyze microarray data. The models are fitted by using procedure NLMIXED of SAS 9.1.3 with the quasi-Newton optimization algorithm. An example of the code that was used is given in Appendix 1. We discuss a number of issues related to the use of this type of models. The focus of the paper is concentrated on cDNA microarrays; however, the technique can be considered for other platforms as well, provided that the distributional assumption is fulfilled.

The paper is organized as follows. Section 2 contains details about the Laplace distribution. In Section 3, we shortly describe the methodology involved in fitting the Laplace model. Section 4 contains examples of the method applied to real-life microarray case studies, while Section 5 describes a simulation study. A short discussion is presented in Section 6.

2. Properties of the Laplace distribution

The symmetric Laplace distribution, also known as the double exponential distribution or the first law of Laplace, is commonly denoted by $L(\theta, \sigma)$, where $\theta \in (-\infty, \infty)$ is the location parameter and $\sigma > 0$ is the scale parameter. The density function

is given by

$$f(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}\,|x-\theta|}{\sigma}\right). \tag{1}$$

In the literature, several parameterizations are used to describe the Laplace distribution. Using the notation in (1), the mean of the Laplace distribution is equal to θ and the variance is equal to σ^2.

The distribution can be generalized to the asymmetric case, $AL(\theta, \mu, \sigma)$, by the inclusion of a skewness parameter μ [8]. Note that $AL(\theta, 0, \sigma)$ is equivalent to $L(\theta, \sigma)$.

Furthermore, the distribution can be standardized with respect to the location and scale parameters as follows:

$$X \sim AL\left(\theta, \mu, \sigma\right), \tag{2}$$

$$\frac{X-\theta}{\sigma} \sim AL\left(0, \frac{\mu}{\sigma}, 1\right). \tag{3}$$

However, standardization with respect to the skewness parameter μ is not possible.

The log-likelihood function $LL(\theta, \sigma)$ for a sample of n independent, identical random variables distributed according to the symmetric Laplace distribution, given in (1), is

$$\sum_{i=1}^{n} \left\{ -\log(\sqrt{2}\sigma^2) - \frac{\sqrt{2}\,|x_i-\theta|}{\sigma} \right\}. \tag{4}$$

The maximum likelihood estimators, obtained from the log-likelihood (4), are presented in Purdom and Holmes [8].

The estimation of the variance-covariance matrix of the maximum likelihood estimates is usually based on the inverse of the observed Hessian matrix. The latter contains the second-order partial derivatives of the log-likelihood function. However, the matrix is not positive definite when based on the log-likelihood defined

in (4), as its determinant is always negative. The inverse of the negative Hessian matrix is, therefore, also not positive definite. Thus, it does not produce valid estimates of the variance-covariance matrix.

3. Methodology

The process of removing systematic effects from microarray data consists of several steps. The first one usually involves applying a transformation $g(.)$ to the intensity measurements. Microarray data are commonly analyzed on a logarithmic scale. However, other transformations might be necessary to eliminate non-linear effects that can exist on the log-scale [2, 16]. Many of the remaining sources of variation can be taken into account by using an ANOVA model, as proposed by Kerr *et al.* [6]. An example of such a model is given as follows:

$$\begin{aligned} g(Y_{ijksgm}) = \mu + A_i + D_j + T_k + G_g + L + LD_j \\ + AG_{ig} + DG_{jg} + TG_{kg} + SG_{isg} + LGg + LDG_{jg} + \varepsilon_{ijksgm}, \end{aligned} \tag{5}$$

where Y_{ijksgm} are the signal intensity measures, μ is the overall mean, A_i, D_j, and G_g represent the effect of the i^{th} array ($i = 1, 2, ..., n_i$), the j^{th} dye ($j = 1, 2$), and gene g ($g = 1, 2, ... G$), respectively. T_k stands for the overall effect of treatment or group k ($k = 1, ..., K$), L represents the linear effect of the laser power used to scan the arrays [16], while LD_j stands for the interaction term between the laser power and dye. Terms AG_{jg}, DG_{jg}, TG_{kg}, LGg, and LDG_{jg} represent, respectively, the gene-specific array, dye, treatment, laser power, and laser power$*$dye interaction effects, while SG_{isg} stands for the gene-specific spot effect nested within an array. The error terms ε_{ijksgm} are assumed to be independent and identically distributed with mean zero.

Due to the large number of parameters, it is practically impossible to fit the model in one step. Instead, the model is split up into two stages [17]. The first

stage consists of a normalization model:

$$g(Y_{ijksgm}) = \mu + A_i + D_j + T_k + L + LD_j + \xi_{ijksgm}. \quad (6)$$

All the effects shared by the genes are thus removed from the data. The estimated residuals of this model, denoted by R_{ijksgm}, are used for the second stage, the gene-specific model. In this stage, the following model is fit for each gene separately:

$$R_{ijksgm} = \eta + A_i + D_j + T_k + S_{is} + L + LD_j + \varepsilon_{ijksgm}. \quad (7)$$

Whether the normalization and gene-specific model, presented in (6) and (7), respectively, represent the classic ANOVA model for normally distributed data (Normal model) or an ANOVA model for Laplace distributed data (Laplace model), depends on the assumption made concerning the distribution of the error terms. The latter is obtained when the model is fit under the condition that ξ_{ijksgm} and ε_{ijksgm} are independent and identically Laplace distributed. The corresponding log-likelihood function is then defined in (4), where $\mu = 0$ and θ is a linear function of the parameters involved in (6) and (7).

3.1 *Hypothesis testing*

As the estimation of variances of parameters is not straightforward for the ANOVA model with the Laplace distribution, hypothesis tests, based on model (7), are performed by means of likelihood ratio tests. The latter does not make use of the estimates of the variance-covariance matrix. The tests are conducted for every gene separately. To maintain an overall significance level of 5%, a multiplicity adjustment is required. We will consider the Benjamini-Hochberg procedure (BH, [1]) to control the false discovery rate (FDR).

4. Case studies

For illustration purposes, we use microarray data from case studies designed to investigate the effect of certain vegetable diets on gene-expression in colon and lung tissue of mice. Two different types of experiments were performed: a dose-response study and a treatment-control study. Both were conducted using colon tissue and repeated for lung tissue. This leads to four microarray datasets in total.

4.1 *Dose-response study*

The dose-response vegetable study was designed to investigate the dose effect of a mixture of four vegetables on the gene expression in colon and lung tissue of mice [14, 15]. Four groups of mice were formed, each receiving a different diet with either 0% (control group), 10%, 20%, or 40% of the vegetable mix. For each dose group, three samples consisting of equal amounts of total RNA pooled from two-three mice were available. The use of different pools assured the presence of biological variability in the data. To each of the three sets of pooled samples, a loop design with four cDNA microarrays was applied (0%→10%→20%→40%→0%, where each arrow indicates a microarray, with the head indicating the sample labeled by the red dye and the tail indicating the sample labeled with the green dye). In total, 12 arrays were used, each containing 602 genes that were spotted three times on every slide.

4.2 *Treatment-control study*

The vegetables, present in the mix used for the dose-response study (see Section 4.1), were investigated separately in a second vegetable study [14, 15]. Diets of cauliflower (T1), carrots (T2), peas (T3), and onions (T4) were compared with a control diet (R) that contained no vegetables. Three samples of genetic material,

pooled from two or three mice, were available for the control group and for the four treatment groups. Using each of the three sets of pooled samples, the four treatment groups were compared to the control group by applying a reference design. This led to 12 arrays. For each of them, dye-swap was performed, which doubles the number of arrays. Thus, in both studies, 24 arrays were used in total, each containing 602 genes, spotted three times on every slide.

4.3 *Analysis of the vegetable studies*

The ANOVA model, presented in equation (6), was applied to the intensity values, where $g(.)$ represents the background-dependent transformation [16]. For some of the case studies, the null-hypothesis that a particular effect of model (6) was equal to zero, could not be rejected. This is the case for $(LD)_j$ for the dose-response experiment in colon tissue, for T_k for the treatment-control experiment in colon tissue, and for L and $(LD)_j$ for the dose-response lung study. Those effects were removed from the model. Subsequently, for each gene we fit the gene-specific model, presented in equation (7), to the estimated residuals of normalization model (6).

In (6) and (7), ξ_{ijksgm} and ε_{ijksgm} are both assumed to be either normally- or Laplace distributed with mean zero and a gene-specific variance σ_g^2. The model thus intrinsically allows for between-gene heterogeneity. The symmetric Laplace is used in stead of the asymmetric Laplace distribution, as the distribution of the data appears to be symmetric in most cases (Figure 1). Under the assumption of normality, the model can be fit using two different methods: maximum likelihood estimation (ML) and restricted maximum likelihood estimation (REML). Both are applied to the datasets.

FIGURE 1

Figure 1 displays QQ-plots for the standardized residuals of the normal REML

and the Laplace model. The normal ML model leads to a very similar picture as the REML model (figures not shown). All QQ-plots contain a reference line obtained by a least squares regression. The distribution of the residuals deviates considerably from the normal distribution in the tails. For the dose-response studies in colon and lung, and for the treatment-control colon vegetable study, the Laplace distribution approximates the true distribution of the residuals more closely. This is consistent with the findings of Purdom and Holmes [8]. There are minor discrepancies in the center of the plots, as the data seem to be even more peaked than the Laplace distribution. However, the proportion of genes for which a deviation is seen in the tails, is much smaller as compared to the normal distribution. For the treatment-control lung study, deviations from the reference line are also visible for the Laplace distribution. The QQ-plots show that both the normality and Laplace assumption are not very appropriate for that dataset. We therefore delete it from the analysis.

The ANOVA models can be used for hypothesis testing. For the dose-response studies, the null hypothesis of interest states that there is no dose effect on gene expression. For the treatment-control studies, we test the hypothesis that none of the vegetables have an effect on the expression levels of the genes. These gene-specific hypothesis tests can be performed by the likelihood ratio or F-tests. Likelihood ratio tests are only valid under the maximum likelihood estimation, while the F-test can also be performed under the restricted maximum likelihood estimation. The tests are carried out at the 5% overall significance level; the BH method is used to correct for multiple testing. The results of the hypothesis tests based on the models fitted under the normality or Laplace assumption are presented in Table 1.

TABLE 1

In general, we would expect only a small number of genes from the entire genome to respond to the treatment. However, the experiments in the vegetable studies were

performed on a specific subset of approximately 600 genes. The genes represented a selection of biologically relevant gene sequences involved in inflammation, DNA damage and repair, oxidative stress, cell signaling, cell proliferation, metabolism, transcription, and apoptosis. It is therefore quite difficult to make any prediction about the number of genes from this subset that should be found to have significant results. We do see that more genes are found to be differentially expressed using the maximum likelihood under the normality assumption compared to the restricted maximum likelihood estimation. A similar trend can be seen for the LR-test as compared to the F-test. When evaluating the model fitted by ML under the normality and Laplace assumptions, the LR-test seems to lead to somewhat different results. For instance, for the dose-response colon study, 30 genes (413 minus 383 genes) are found to be differentially expressed when assuming normality in stead of a Laplace distribution, while the reverse is true for 70 genes (453 minus 383 genes). It is therefore important to base the conclusions on the proper model. For the case studies, the Laplace distribution fits better to the data. Thus, it seem that the list of genes declared significant by the model under the Laplace assumption, is more appropriate.

Examining the individual profiles of a number of the genes found to be significant by the Laplace model and not by the normal, and vice versa (data not shown), did not reveal a possible cause for the difference in conclusions, nor did it reveal these genes to be outliers in any way. **However, before drawing any conclusions, the Laplace model needs to be examined further to find out, for instance, why the LR-test based on the Laplace model always leads to more rejections of the null hypothesis as compared to the normal model. A similar trend is visible for the LR-test compared to the F-test under the normality assumption.**

FIGURE 2

Figure 2 shows the distribution of the different test statistics, for which results are presented in Table 1. The plots support the statements drawn based on Table 1. For instance, under the Laplace model, the distribution of the LR-test statistic is shifted slightly toward the larger values, as compared to the one obtained under the Normal model. This results in a larger number of positive findings for the former model. **We do not know, however, whether these positive findings are true positives or false positives.** To study these issues and other aspects of the model under controlled settings, we performed a simulation study.

5. Simulation study

In the previous section, we have applied an ANOVA model under the Laplace assumption to real-life microarray data. In this section, we describe a simulation study, conducted to investigate how the Laplace assumption affects the point estimated of the model parameters, and the distribution of the test statistics. We use a variation of the simulation model presented by Van Sanden *et al.* [16].

5.1 *Data simulation*

We assume the setting of an experiment corresponding to the real-life experiments presented in the previous section, using two-channel cDNA microarrays and a dye-swap design. Two classes of samples (a treatment and a control group, say) are compared. For every setting of interest, 100 simulation datasets are created, each containing 50 arrays. On each array we simulate 600 genes, spotted three times. The arrays contain 100 genes that are differentially expressed. A linear mixed effects model is utilized to simulate observations subject to various systematic and random effects usually present in real microarray experiments [6, 17].

An observation Y_{ijgk}, assumed to be the signal intensity for array i ($i = 1, \ldots, 50$), dye j ($j = 1, 2$) and gene g ($g = 1, \ldots, 600$), is generated by the following model:

$$log_2(Y_{ijgk}) = \mu + A_i + D_j + G_g + TG_g + \varepsilon_{ijgk}, \tag{8}$$

where μ is the overall mean, A_i stands for the array effect, D_j for the dye effect and G_g for the gene effect. Term TG_g represents the gene specific treatment effect, while ε_{ijgk} is a random error. Some of these effects are fixed: $\mu = 9$, $D_j = 1 \times I(j = 1)$, and $TG_g = T \times I(g \leq 100, i = \text{odd}, j = 1)$, where $I(X)$ is an indicator function, equal to one if X is true, and zero otherwise.

Other effects are random and drawn from a normal or symmetric Laplace distribution. The array and gene effects A_i and G_g are distributed according to distributions $F_1(0,\ 0.25)$ and $F_1(0,\ \sigma_G^2)$, respectively, while the error term ε_{ijgk} follows distribution $F_2(0,\ \sigma^2)$, where F_1 and F_2 are assumed to be either the normal or symmetric Laplace distribution. The chosen values for the parameters of the simulation model are based on the estimated parameters of an ANOVA-model fitted to the vegetable studies (see Section 4).

The five settings considered with regard to the size of the treatment effect and variance components (σ_G^2, T, σ^2) are defined as follows: S_1=(5, 0, 0.14), S_2=(0.14, 0, 5), S_3=(5, 0, 5), S_4=(5, 0.2, 0.14), S_5=(5, 0.1, 0.14). In what follows, the data will be referred to as N when they are normally distributed (F_1, F_2 are both normal distributions), L when they are Laplace distributed (F_1, F_2 are both Laplace), NL when F_1 is a normal distribution and F_2 is a Laplace distribution, and LN when F_1 is a Laplace distribution and F_2 is normal.

5.2 *Analysis of the simulated datasets*

The simulated data are analyzed by the two-stage fixed effects ANOVA model introduced in Section 3. For each of the 100 datasets of every simulation setting, the following normalization and gene-specific model are fit to the simulated intensity measures:

$$log_2(Y_{ijkm}) = \mu + A_i + D_j + \xi_{ijkm}, \tag{9}$$

$$R^g_{ijkm} = \eta^g + T^g_k + \varepsilon^g_{ijkm}, \tag{10}$$

where ξ_{ijkm} and ε_{ijkm} are assumed to be either normally or Laplace distributed with mean zero and variances σ^2_1 and σ^2_{2g}, respectively. Note that η^g, T^g_k, and σ^2_{2g} contain a gene-specific index g, as model (10) is fit for every gene separately. Only the results for the normal and Laplace distributed data (N and L) are included. The results of the NL and LN datasets are not shown, but are briefly discussed.

5.2.1 Estimation of the model parameters

Using the SAS NLMIXED procedure, we obtain the estimates for the model parameters. Each dataset consists of 50 arrays. We therefore have 49 estimated array effects (the estimate of the remaining array is always put equal to 0, and is thus contained in the intercept). For every dataset, we first take the average of the 50 array effects. Similarly, we also average over all 600 gene-specific model parameters. They are then summarized, together with the other parameter estimates, by the mean value and standard deviation over all simulations datasets. For illustration purposes, some of the parameter estimates for two particular settings, S_1 and S_2, are displayed in Table 2. In setting S_1, the variance of the gene-effect is large compared to that of the error term, while the reverse is true for setting S_2.

In Table 2, Var$[A_i]$ stands for the variance of the parameter estimates for A_i

over all arrays. Every array has 600 different genes spotted on it; η and σ_2^2 are the mean values of η^g and σ_{2g}^2, respectively, over the 600 genes. The estimate for the variance of η is denoted by Var[η].

TABLE 2

For most of the estimated parameters, the mean values correspond closely to the values used in the simulation model. Some deviations are noticeable between the normal and Laplace model. For instance, the normal model seems to underestimate the value of σ_1^2 when the distribution of the random effects, F_1, is Laplace. The Laplace model, on the other hand, overestimates the value when F_1 is normal. For setting S_2, where the residual variance is relatively large, the performance of the normal model is good, and thus less sensitive to the model misspecification. The Laplace model, on the other hand, remains sensitive to the model misspecification, but performs reasonably well when the error distribution, F_2, is Laplace.

5.2.2 Hypothesis tests

For the differentially expressed genes, a treatment effect ($T_k^g \neq 0$) is incorporated in the intensity measures of the treatment group. When we test the null hypotheses of no treatment effect ($H_0 : T_k^g = 0$) based on the gene-specific models, we should only reject the hypothesis for the models corresponding to the 100 differentially expressed genes. The tests are performed at the overall 5% significance level. As the hypothesis is tested for all 600 genes, we have to correct the obtained p-values for multiple testing. The BH procedures is considered to control the FDR, which can be estimated based on the 100 simulated datasets.

The results of these calculations for settings S_1-S_5 are displayed in Table 3. Even with the BH procedure, we do not always obtain an FDR smaller than or equal to 0.05. For the model fit under the Laplace assumption, the FDR is not

controlled. When the Laplace model is applied to normally distributed data, the FDR is generally over 50%. For Laplace distributed data, the FDR is still generally over 10%. On the other hand, when the normal model is applied to the Laplace distributed data, the FDR is also not controlled and, in most cases, even larger than that for the Laplace model. When the total variability of the data increases (setting S_3), so does the FDR, even for the normal model.

For settings S_4-S_5, for which a treatment effect is simulated, the deviations from 0.05 are less severe as compared to settings S_1-S_3. This is because the FDR measure takes the ratio of false discoveries into account, relative to the total number of discoveries. The FDR for the normal model applied to Laplace distributed data is again similar or higher as compared to the Laplace model.

The lack of control we observe for the FDR under several conditions can be a result of a deviation of the distribution of the LR-test statistic from the chi-squared distribution. As the test statistic is only asymptotically chi-squared distributed, the cause of the deviation could be related to the sample size of the experiment. We investigate this issue in Section 5.3.

TABLE 3

The power for a particular setting, for which differentially expressed genes exist, can be estimated by taking the average power over the 100 simulated datasets. The results for the power calculations are displayed in Table 3. In setting S_5, the treatment effect is relatively small compared to the variance in the data. Hence, it becomes more difficult to pick up the truly differentially-expressed genes, which results in a low power. While the power estimates for the normal model remain reasonably stable over the different data distributions, the Laplace model clearly has more power to detect the differentially expressed genes under the proper distributional assumption.

5.3 *Distribution of the LR-test*

To investigate the distribution of the LR-test and its sensitivity to sample size for the model fit under the Laplace distribution, we conducted a small and straightforward simulation for group comparison under different distributional assumptions. Measurements for various sample sizes were both simulated and analyzed by the following model:

$$Y_i = \mu + \varepsilon_i, \tag{11}$$

with $\varepsilon_i \sim N(0;1)$ or $L(0;1)$, and $\mu = 2$. Hence, there is no actual difference created between the measurements of the two groups. Per sample size setting, 10,000 datasets were simulated using model (11). The results for the hypothesis test of no difference between the two groups, performed for all the datasets, are displayed in Table 4. This table contains the fraction of hypotheses that were (falsely) rejected at the 5% significance level out of the 10,000 datasets. One would expect about 500 datasets, for which the null hypothesis should be rejected. When the sample size is small (10 per group), too many hypothesis are rejected.

TABLE 4

For both the normal and Laplace model, the Type I error converges approximately to the expected value when the sample size is large enough. However, for the Laplace distribution, the convergence rate is slower. Therefore, the increased FDR, seen in Table 3 for the model assuming Laplace-distributed error terms, seem to be sample size related as only 50 arrays were simulated.

A second simulation study was set up to study the distribution of the LR-test under circumstances closer to those encountered in the microarray experiments. We repeat the simulation of setting S_1 with a varying number of arrays and replicates. However, to keep the computation time of the study feasible, we only simulate data

for two genes. A separate model has to be fit for every gene. Therefore, the number of genes has a significant influence on the computation time of the simulation study. Moreover, we do not expect that the number of tests to be performed and taken into account using the Benjamini-Hochberg procedure has a considerable influence on the results and conclusions of the study.

For every sample size setting, we calculate the FDR based on the 100 simulated datasets. The whole experiment is repeated 50 times and the average FDR and standard deviation over the 50 repetitions are reported in Table 5.

TABLE 5

The normal model leads again to very stable results, as the FDR barely changes with the increasing sample size. As the number of replicates on an array increase, the FDR for the Laplace model, applied to Laplace distributed data, decreases to even below 0.05. Using more arrays does, however, not have the same effect. The FDR seems to remain almost constant as the number of arrays varies. This might be explained by the following argument. When a gene is spotted several times on an array, we have uncontaminated replicates of the measurements. If, on the other hand, arrays are added to the experiment, the sample size becomes larger, but we also increase the variability in the data and the number of model parameters that have to be estimated. We suspect this to be the reason why we do not see any benefit, with respect to controlling the FDR, of increasing the number of arrays.

FIGURE 3

Figure 3 displays the distribution of the LR-test statistic, corresponding to the first part of Table 5. We do not present all the figures, as the results for the settings with 30, 40, or 50 arrays are similar to the setting with 20 arrays.

The distributions in Figure 3 illustrate the findings in Table 5. When both the model and the distribution of the data are normal, the empirical and theoretical **(i.e., under the null hypothesis)** distributions coincide. The FDR is therefore controlled at 0.05%. If the normal model is applied to the Laplace distributed data, the tail of the empirical distribution starts to fall below the tail of the theoretical one when the number of replicates increases. Thus, the FDR decreases with the increasing number of replicates. In the case of a Laplace model and Laplace distributed data, the tails of the empirical and theoretical distribution become close to each other for 6 and 9 replicates, so only then the FDR is controlled. When applying the Laplace model to normally distributed data, however, we always observe a thicker tail for the empirical distribution, so the FDR is not controlled.

Based on the results of the simulation studies, we can reevaluate the findings for the case studies. As the sample sizes are small, 12 or 24 arrays with three replicates per gene, we can expect an increase of the probability of the Type I error for the hypothesis tests based on the Laplace model. Due to model misspecification, such an increase is also expected for the normal model. When the distribution of the data closely approximates the Laplace, as it is the case for the dose-response studies, the Laplace model has more power compared to the normal. The increase of both the probability of the Type I error and the power could be the cause of the high rate of positive findings for the Laplace model. Based on the simulations, we cannot unambiguously explain the difference seen between the F-test and the LR-test, as such difference did not occur in the simulated data.

6. Discussion

It has been shown by Purdom and Holmes [8] that the Laplace distribution can provide a better fit to microarray data as compared to the normal distribution.

This appears to be also the case for the vegetable studies presented in this paper. However, when trying to incorporate the Laplace distribution in an ANOVA modeling approach, we encountered several problems.

Caution is needed when interpreting the hypothesis testing results of the Laplace model. Using a simulation study, we showed that the distribution of the LR-test statistic deviates from the chi-square for small sample sizes. The convergence rate to this asymptotic distribution is slower compared to the test-statistics based on the normal distribution. As a result, the Benjamini-Hochberg multiple testing procedure does not control the FDR of the Laplace model at the pre-specified significance level.

Small sample sizes are very common for microarray experiments. The problem is therefore particularly relevant for this type of experiments. We can only conclude that when the data are clearly Laplace distributed, the Laplace model is preferred over the normal. In this situation, both models suffer from an increased FDR, but the Laplace model has more power.

For the case studies at hand, the distribution of the error terms was fairly symmetric. Therefore, the symmetric Laplace distribution was favored over the asymmetric distribution. The ANOVA model considered in this paper can also be applied using the asymmetric Laplace distribution. There are, however, several issues related to the use of the latter distribution. For instance, it requires the estimation of one extra parameter, the skewness. Moreover, assessing the fit of such a model is not straightforward, as standardization of the Laplace distribution with respect to the skewness parameter is not possible. Therefore, QQ-plots for the standardized residuals, as those presented in Section 4, cannot be easily constructed.

In this paper, we focus on the use of the Laplace distribution in a gene-by-gene analysis. In recent years, more emphasis has been put on methods that borrow

information across genes [3, 9]. Such methods should also be incorporated into the Laplace modeling approach.

The effects of model misspecification on the parameter estimates, the Type I error, and the power were also examined by simulation data using model (8), where the random effects were normally distributed and the error terms were Laplace distributed, and vice versa (results not shown). The normal model seemed to be less affected by to this type of model misspecification, as compared to the Laplace model.

Microarray experiments lead to datasets, which can differ a lot in properties, often depending on the technology used. The latter can be very diverse. Therefore, we do not expect that the Laplace distribution will always be a perfect match to data. The properties of every dataset have to be examined independently.

Acknowledgment

We gratefully acknowledge support of the IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

References

[1] Y. Benjamini, and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B 57 (1995), pp. 289-300.

[2] X. Cui, M.K. Kerr, and G.A. Churchill, *Transformations for cDNA microarray data*, Statistical Applications in Genetics and Molecular Biology 2 (2003), Art. 4.

[3] X. Cui, J.T.G. Hwang, and J. Qiu, *Improved statistical tests for differential gene expression by shrinking variance components estimates* , Biostatistics 6(1) (2005), pp. 59–75.

[4] B. Efron, and R. Tibshirani, *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy* , Statistical Science 1 (1986), pp. 54–77.

[5] Y. Hochberg, and Y.C. Tamhane, *Multiple comparison procedures*, Wiley, New York, 1987.

[6] M.K. Kerr, M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data*, Journal of Computational Biology 7 (2000), pp. 819–838.

[8] E. Purdom, and S.P. Holmes, *Error distribution for gene expression data*, Statistical Applications in Genetics and Molecular Biology 4(1) (2005), Art. 16.

[9] G.K. Smyth, *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments*, Statistical Applications in Genetics and Molecular Biology 3(1) (2004), pp. 1–29.

[10] G.K. Smyth, *Limma: linear models for microarray data*, In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, 2005, pp. 397–420.

[12] V.Tusher, R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*, Proceedings of the National Academy of Sciences of the United States of America 98 (2001), pp. 5116-5121.

[13] O.G. Troyanskaya, M.E. Garber, P.O. Brown, and D. Botstein, *Nonparametric methods for identifying differentially expressed genes in microarray data*, Bioinformatics 18 (2002), pp. 1454–1461.

[14] S. Van Breda, E. van Agen, and S. Van Sanden, *Vegetables affect the expression of genes involved in anticarcinogenic processes in the colonic mucosa of C57BL/6 female mice*, Journal of Nutrition 135(8) (2005a), pp. 1879–1888.

[15] S. Van Breda, E. van Agen, and S. Van Sanden, *Vegetables affect the expression of genes involved in carcinogenic and anticarcinogenic processes in the lungs of female C57BL/6 mice*, Journal of Nutrition 135(11) (2005b), pp. 2546–252.

[16] S. Van Sanden, and T. Burzykowski, *The use of background signal in the transformation of cDNA-microarray measurements*, Applied Bioinformatics 5(3) (2006), pp. 161–172.

[17] R.D. Wolfinger, G. Gibson, and E.D. Wolfinger, *Assessing gene significance from cDNA microarray expression data via mixed models*, Journal of Computational Biology 8 (2001), pp. 625–637.

Appendix 1: SAS code for the Laplace model

In the first step, the normalization model is fit to the data. Starting values can, for instance, be obtained from a standard ANOVA model.

```
title 'Normalization model';
proc nlmixed data=dataset tech=QUANEW;
parms /data=startval;
mu=b0+b1*treat1+b2*treat2+ ...;
ll=-log(sqrt(2)*sig)-(sqrt(2)/sig)*abs(logexpvalue-mu);
model logexpvalue~general(ll);
predict mu out=norm;
run;
```

From the results, the residuals of the normalization model can be calculated as follows:

```
data norm;
set norm;
residual=logexpvalue-pred;
run;
```

In the next step, a gene-specific model is fit to every gene separately, once with (full model) and onces without the treatment effects (reduced model). Using the estimated log likelihood from the two models, a likelihood ratio test can then be performed to test for a significant treatment effect.

```
title 'Full model';
proc nlmixed data=norm_gene1 tech=QUANEW;
parms /data=startval;
```

```
mu=b0+b1*treat1+b2*treat2+ ...;

ll=-log(sqrt(2)*sig)-(sqrt(2)/sig)*abs(residual-mu);

model residual~general(ll);

ods output FitStatistics=testsfull;

run;


title 'Reduced model';

proc nlmixed data=norm_gene1 tech=QUANEW;

parms /data=startval;

mu=b0+ ...;

ll=-log(sqrt(2)*sig)-(sqrt(2)/sig)*abs(residual-mu);

model residual~general(ll);

ods output FitStatistics=testsreduced;

run;
```

Figure legends

Figure 1: Standardized residuals of the two-stage ANOVA model, fitted under the assumption of normally (resp. Laplace) distributed data, plotted versus the quantiles of the normal (resp. Laplace) distribution, with a reference line obtained by least squares regression.

Figure 2: Empirical and theoretical (i.e., under the null hypothesis) distribution of the test statistics (first row: F-test Normal model under REML and under ML; second row: LR-test Normal model and Laplace model) for the D-R, T-C colon, and D-R lung study.

Figure 3: Empirical and theoretical (i.e., under the null hypothesis) distribution of the LR-test statistic from the second simulation study described in Section 5.3 (setting S_1; 20 arrays; 3, 6 or 9 replicates). Each figure is followed by a close-up of the tails of the distributions.

Tables legends

Table 1: Number of genes, for which the null hypothesis is rejected (in parenthesis the number of genes in common between the model fit under the normality and Laplace assumption). To correct for multiple testing, the Benjamini-Hochberg (BH) procedure is applied.

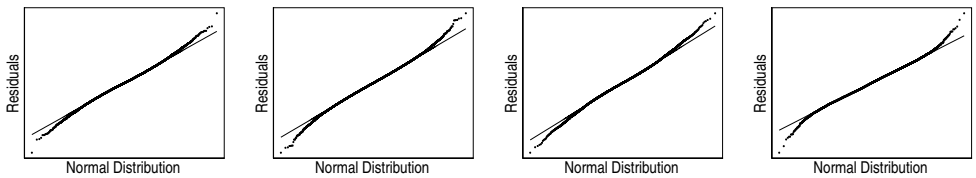
Table 2: The mean value (standard deviation $\times$10E03) of some of the estimated model parameters and related values for the 100 simulated datasets from settings S_1 and S_2. The models are fit under the assumption of normally distributed (Normal) and Laplace distributed (Laplace) errors, using the ML estimation method.

Table 3: FDR and average power (standard deviation) for the different settings of the simulation study. The models are fit under the assumption of normally

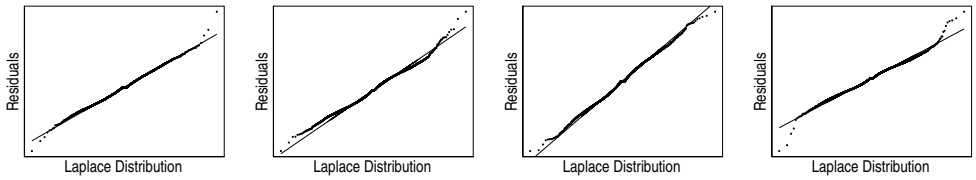
distributed (Normal) and Laplace distributed (Laplace) errors, using the ML estimation method. To correct for multiple testing (600 genes), the Benjamini-Hochberg (BH) procedure is applied.

Table 4: Rate of rejected null hypothesis at the 5% s.l. for the 10000 datasets, and a 95% confidence interval.

Table 5: The average FDR (standard deviation) for the different sample size settings of the simulation study for setting S_1. The models are fit using the ML estimation method. The LR-test is used for hypothesis testing at the 5% significance level. To correct for multiple testing, the Benjamini-Hochberg (BH) procedure is applied. (A: number of arrays, R: number of replicates per gene on one array).



(a) D-R colon study (b) T-C colon study (c) D-R lung study (d) T-C lung study



(e) D-R colon study (f) T-C colon study (g) D-R lung study (h) T-C lung study

Figure 1. Standardized residuals of the two-stage ANOVA model, fitted under the assumption of normally (resp. Laplace) distributed data, plotted versus the quantiles of the normal (resp. Laplace) distribution, with a reference line obtained by least squares regression.

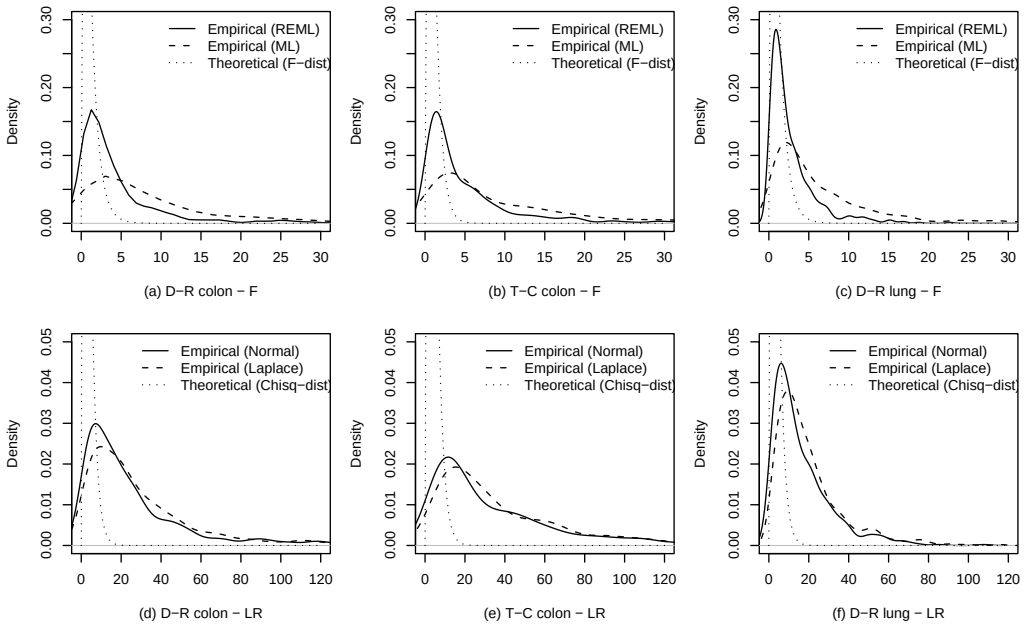


Figure 2. Empirical and theoretical **(i.e., under the null hypothesis)** distribution of the test statistics (first row: F-test Normal model under REML and under ML; second row: LR-test Normal model and Laplace model) for the D-R, T-C colon, and D-R lung study.

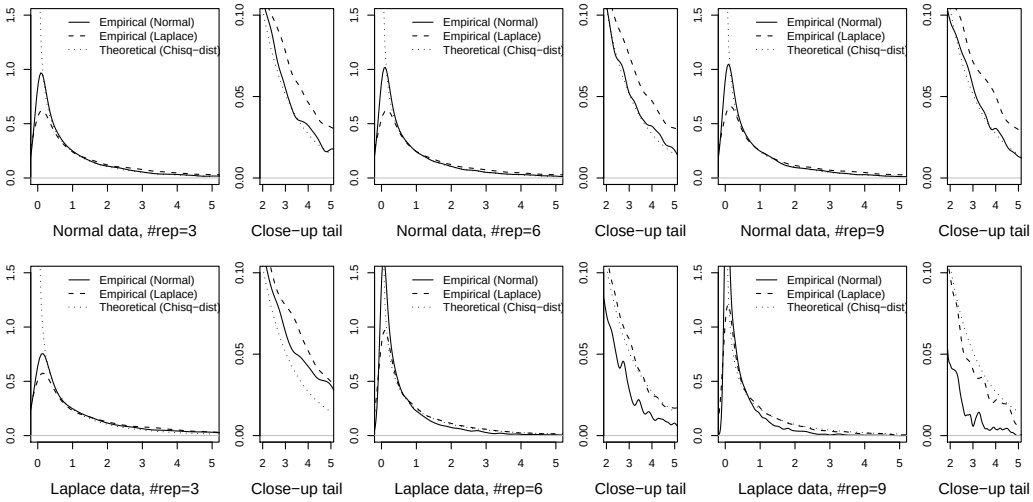


Figure 3. Empirical and theoretical **(i.e., under the null hypothesis)** distribution of the LR-test statistic from the second simulation study described in Section 5.3 (setting S_1; 20 arrays; 3, 6 or 9 replicates). Each figure is followed by a close-up of the tails of the distributions.

Table 1. Number of genes, for which the null hypothesis is rejected (in parenthesis the number of genes in common between the model fit under the normality and Laplace assumption). To correct for multiple testing, the Benjamini-Hochberg (BH) procedure is applied.

$\varepsilon_{ijksm} \sim$	$N(0, \sigma_g^2)$						$L(0, \sigma_g)$
Study	F-test (REML)		F-test (ML)		LR-test (ML)		LR-test (ML)
D-R, colon	208	(206)	398	(372)	413	(383)	453
T-C, colon	273	(273)	436	(421)	440	(424)	496
D-R, lung	102	(102)	327	(315)	342	(327)	435

Table 2. The mean value (standard deviation ×10E03) of some of the estimated model parameters and related values for the 100 simulated datasets from settings S_1 and S_2. The models are fit under the assumption of normally distributed (Normal) and Laplace distributed (Laplace) errors, using the ML estimation method.

Par.	Setting	Data	Normal	Laplace
σ_1^2	S_1	N	5.09 (3.79)	6.37 (5.07)
		L	4.31 (1.38)	4.67 (3.04)
	S_2	N	5.14 (19.1)	6.55 (26.6)
		L	5.14 (47.6)	5.24 (43.4)
σ_2^2	S_1	N	0.14 (0.51)	0.18 (0.68)
		L	0.14 (1.29)	0.14 (1.25)
	S_2	N	4.96 (18.7)	6.31 (25.7)
		L	4.97 (47.3)	5.02 (42.6)
$\mathrm{Var}[A_i]$	S_1	N	0.24 (47.0)	0.24 (49.4)
		L	0.26 (78.4)	0.26 (78.7)
	S_2	N	0.26 (55.4)	0.26 (57.5)
		L	0.25 (92.5)	0.25 (92.5)
$\mathrm{Var}[\eta]$	S_1	N	4.96 (5.49)	4.96 (6.81)
		L	4.18 (4.48)	4.18 (3.74)
	S_2	N	0.18 (6.21)	0.20 (7.15)
		L	0.16 (4.93)	0.16 (4.46)

Table 3. FDR and average power (standard deviation) for the different settings of the simulation study. The models are fit under the assumption of normally distributed (Normal) and Laplace distributed (Laplace) errors, using the ML estimation method. To correct for multiple testing (600 genes), the Benjamini-Hochberg (BH) procedure is applied.

	Setting	Data	Normal F-test	Normal LR-test	Laplace LR-test
FDR	S_1	N	0.05	0.05	0.62
		L	0.29	0.29	0.24
	S_2	N	0.03	0.03	0.55
		L	0.27	0.27	0.26
	S_3	N	0.08	0.08	0.64
		L	0.35	0.36	0.32
	S_4	N	0.04	0.04	0.16
		L	0.10	0.10	0.10
	S_5	N	0.04	0.04	0.22
		L	0.13	0.13	0.10
Power	S_4	N	0.976 (0.020)	0.976 (0.020)	0.950 (0.023)
		L	0.956 (0.021)	0.957 (0.020)	0.997 (0.005)
	S_5	N	0.216 (0.070)	0.218 (0.071)	0.320 (0.072)
		L	0.288 (0.055)	0.289 (0.054)	0.632 (0.069)

Table 4. Rate of rejected null hypothesis at the 5% s.l. for the 10000 datasets, and a 95% confidence interval.

	Sample size per group			
ε_i	10	100	1000	10,000
$N(0;1)$	0.0684	0.0544	0.0511	0.0485
	[0.0635; 0.0733]	[0.0500; 0.0588]	[0.0468; 0.0554]	[0.0443; 0.0527]
$L(0;1)$	0.0784	0.0579	0.0511	0.0514
	[0.0731; 0.0837]	[0.0533; 0.0625]	[0.0468; 0.0554]	[0.0471; 0.0557]

Table 5. The average FDR (standard deviation) for the different sample size settings of the simulation study for setting S_1. The models are fit using the ML estimation method. The LR-test is used for hypothesis testing at the 5% significance level. To correct for multiple testing, the Benjamini-Hochberg (BH) procedure is applied. (A: number of arrays, R: number of replicates per gene on one array).

		Normal model		Laplace model	
A	R	Normal data	Laplace data	Normal data	Laplace data
20	3	0.066 (0.025)	0.118 (0.037)	0.149 (0.039)	0.168 (0.045)
20	6	0.059 (0.023)	0.014 (0.014)	0.154 (0.039)	0.057 (0.026)
20	9	0.054 (0.028)	0.001 (0.004)	0.139 (0.043)	0.031 (0.019)
30	3	0.061 (0.021)	0.119 (0.034)	0.144 (0.044)	0.161 (0.044)
30	6	0.057 (0.022)	0.016 (0.012)	0.148 (0.033)	0.064 (0.027)
30	9	0.054 (0.023)	0.002 (0.005)	0.143 (0.035)	0.031 (0.021)
40	3	0.064 (0.027)	0.114 (0.034)	0.159 (0.031)	0.156 (0.040)
40	6	0.056 (0.024)	0.014 (0.013)	0.146 (0.041)	0.054 (0.027)
40	9	0.056 (0.020)	0.002 (0.005)	0.143 (0.035)	0.036 (0.019)
50	3	0.064 (0.025)	0.109 (0.031)	0.145 (0.041)	0.152 (0.041)
50	6	0.060 (0.021)	0.012 (0.011)	0.145 (0.034)	0.056 (0.025)
50	9	0.055 (0.023)	0.001 (0.004)	0.140 (0.038)	0.031 (0.018)