Biological assessment of robust noise models in microarray data analysis

A. Posekany¹, K. Felsenstein² and P. Sykacek¹,*

¹Chair of Bioinformatics, Department of Biotechnology, University of Natural Resources and Life Sciences, Gregor Mendel Straße 33, 1180, Vienna and ²Department of Statistics, Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria

ABSTRACT

Motivation: Although several recently proposed analysis packages for microarray data can cope with heavy-tailed noise, many applications rely on Gaussian assumptions. Gaussian noise models foster computational efficiency. This comes, however, at the expense of increased sensitivity to outlying observations. Assessing potential insufficiencies of Gaussian noise in microarray data analysis is thus important and of general interest.

Results: We propose to this end assessing different noise models on a large number of microarray experiments. The goodness of fit of noise models is quantified by a hierarchical Bayesian analysis of variance model, which predicts normalized expression values as a mixture of a Gaussian density and $t$-distributions with adjustable degrees of freedom. Inference of differentially expressed genes is taken into consideration at a second mixing level. For attaining far reaching validity, our investigations cover a wide range of analysis platforms and experimental settings. As the most striking result, we find irrespective of the chosen preprocessing and normalization method in all experiments that a heavy-tailed noise model is a better fit than a simple Gaussian. Further investigations revealed that an appropriate choice of noise model has a considerable influence on biological interpretations drawn at the level of inferred genes and gene ontology terms. We conclude from our investigation that neglecting the over dispersed noise in microarray data can mislead scientific discovery and suggest that the convenience of Gaussian-based modelling should be replaced by non-parametric approaches or other methods that account for heavy-tailed noise.

Contact: peter.sykacek@boku.ac.at
Availability: http://bioinf.boku.ac.at/alexp/robmca.html.

Received on July 15, 2010; revised on January 5, 2011; accepted on January 6, 2011

1 INTRODUCTION

The importance of microarray data for the biological sciences has generated a large number of sophisticated analysis methods. Approaches like $t$-tests (Baldi and Long, 2001; Tusher et al., 2001), linear models (Smyth, 2005) and many Bayesian methods (Bae and Mallick, 2004; Ibrahim et al., 2002; Ishwaran and Rao, 2003; Lewin et al., 2007; Zhao et al., 2008) consider data to be approximately Gaussian distributed. Recent investigations have, however, cast doubt on the correctness of the Gaussian assumption. By testing for Gaussianity, Hardin and Wilson (2009) find that microarray data does not follow a Gaussian distribution. The observed overdispersion leads to a large number of outlying values which can have a considerable influence on the inference results. The cost of measurements and the possibility that outlying data points are caused by biological processes rule out that such samples get removed. All samples must thus be taken into account carefully, as excluding outlying values or including them based on incorrect distribution assumptions would falsify the biological findings. The adverse effects of outliers in microarray data can be overcome with non-parametric approaches (cf. de Haan et al., 2009; Gao and Song, 2005; Lee et al., 2005; Troyanskaya et al., 2002; Tusher et al., 2001; Zhao and Pan, 2003). Non-parametric methods replace the restrictive assumptions linked with the Gaussian distribution with very general ones, however, at the expense of losing some power of tests (cf. Whitley and Ball, 2002). Alternatively, we can analyze overspersed data with robust parametric noise models like Student’s-$t$ distributions (cf. Gottardo et al., 2006).

The issue of appropriate noise models led to an ongoing discussion, with Giles and Kipling (2003) arguing that microarray data are Gaussian distributed. Similar methods let Hardin and Wilson (2009) conclude that microarray data require heavy-tailed noise models. The conclusion of Novak et al. (2006) was that 5–15% of genes are non-Gaussian distributed, with the majority following Gaussian distributions. Finding such diverse conclusions about noise in microarray data suggest an in-depth investigation of this issue. We propose to this end inferring the appropriate degree of overdispersion in microarray data with a hierarchical Bayesian model, which is inspired by the proposal of Gottardo et al. (2006). Built in means for ranking genes according to differential expression enable investigations of the biological implications of deviating from the optimal noise model. The essential components of the proposed model are thus two indicator variables, one decoding whether a gene is differentially expressed, the other decoding the most appropriate noise model. These variables are built into a hierarchical Bayesian analysis of variance (ANOVA) model which can be used for analyzing a variety of experimental designs.

Inferring the proposed model with uninformative prior settings provides reliable probability measures, which quantify the suitability of competing noise models. This mode of operation compares the goodness of fit of a Gaussian noise model with $t$-distributions of different degrees of freedom and infers the appropriate robustness level required for analyzing a microarray dataset. The ultimate
goal of microarray data analysis is, however, obtaining sound biological conclusions about which transcripts are involved in a particular process. Judgements about different noise models should therefore be linked with their implications on biological findings. The proposed model provides for this purpose a second mode of operation, in which we fix the noise model either to a Gaussian density or to a t-distribution with optimal degrees of freedom as found in the adaptive mode of operation. The biological implications of deviating from the optimal noise model can then be assessed from the noise model-dependent gene rankings.

To warrant reliable conclusions, we calibrated the model on synthetic data and the golden-spoke experiment from Choe et al. (2005), before analysing 14 microarray datasets. Independent of normalization and preprocessing, we found in every case that a t-distribution with small degrees of freedom provides a much better fit of the noise characteristics than a Gaussian density. The importance of robust inference is apparent from our observation that exchanging the optimal Student’s-t density with a Gaussian leads to between 119 and 3561 differences in gene lists and to between 14 and 316 differences in Gene Ontology (GO) (cf. Ashburner et al., 2000) term lists. We have thus strong evidence that opting for Gaussian noise models in microarray data analysis may result in seriously misleading biological leads. Microarray data analysis should thus preferably use non-parametric approaches (cf. de Haan et al., 2002; Tusher et al., 2001; Zhao and Pan, 2003) or approaches that allow for heavy-tailed noise models (cf. Gottardo et al., 2006).

2 METHODS
The methods in this article provide a framework for thoroughly investigating whether microarray data analysis requires robust approaches, or whether we may safely rely on Gaussian assumptions. The Bayesian ANOVA model shown in Figure 1 as directed acyclic graph (DAG) (inferred) in this optimal robustness levels and a measure whether genes are differentially expressed. The proposed approach achieves robustness by using a parametric heavy-tailed noise model, with non-parametric, methods (cf. de Haan et al., 2009; Gao and Song, 2005; Lee et al., 2005; Troyanskaya et al., 2002; Tusher et al., 2001; Zhao and Pan, 2003) or approaches that allow for heavy-tailed noise models (cf. Gottardo et al., 2006).

2.1 Bayesian ANOVA with flexible noise model
The Bayesian one-way ANOVA model shown in Figure 1 as DAG constitutes the core of our evaluation. ANOVA models are commonly used for analysing multi-level microarray experiments like time course data. The model is based on a linear relation between the gene expression $y_{ng}$, measured for sample $n$ and gene $g$ and the mean expression $\mu_g$. The $S$-dimensional vector $x_n$ is an indicator for the biological state. If sample $n$ belongs to state $s$, $x_n$ has a 1 at the $s$-th position and zeros everywhere else. Depending on whether gene $g$ is differentially expressed or not, the latent indicator variable $I_g$ switches between two different dimensional representations of $\beta_g$ (cf. Holmes and Held, 2006; Sykacek et al., 2007). The case that gene $g$ is not differentially expressed is coded by $I_g=0$ and corresponds to the null hypothesis of the classical ANOVA that all groups have the same mean. Vector $\beta_g$ contains in this case $S$ identical entries of mean expression $\beta_{g0}$, the latter being equipped with a Gaussian prior with mean $\mu$ and prior precision $\lambda$. The alternative hypothesis that gene $g$ is differentially expressed is coded by $I_g=1$ with $\beta_{g1}$ being multivariate with a Gaussian prior with mean $\mu$ and diagonal precision matrix $\lambda$. The indicator $I_g$ is a priori binomially distributed with probability $p$ of differential expression. To reduce the sensitivity of the approach, the model is extended hierarchically by allowing for a beta prior over $p$ with hyperparameters $a$ and $b$. The observations $y_{ng}$ follow a symmetric distribution centred around $x_n \beta_g$ with precision $\nu_g \tau$. Different robustness levels are achieved by selecting the noise model for the observations $y_{ng}$ from a set containing $K=1$ Student’s-$t$-distributions of different degrees of freedom, $v$, and a Gaussian distribution ($v=\infty$). For obtaining computationally tractable representations of Student’s-$t$-distributions with arbitrary degrees of freedom, we introduce the auxiliary variables $y_{ng}$, represent $p(x_{ng}y_{ng}|\beta_g,\tau,v)$ as a certain Gaussian-Gamma density and integrate over $y_{ng}$ (cf. Bernardo and Smith, 1994).

An essential aspect of robustness is adjusting the degrees of freedom $v$ to the level required by the data. We propose to this end selecting the best fitting degrees of freedom from a finite set of possible choices (cf. Berger, 1994), which includes the Gaussian ($v=\infty$) as the non-robust special case. The proposed model implements this selection via the multinomial-one distributed indicator variable $J$, which chooses a particular $v$ from the set $v = \{1\ldots K\} \cup \{\infty\}$, with $p(J|\nu) \propto \Gamma(\nu_J+1)^{-1} \prod_{j=1}^K \Gamma(\nu_j+1)$, with $j \in \{1\ldots K\}$ and $\nu_\infty = 1$. This
Algorithm 1 illustrates MCMC sampling as pseudo-code. Inference requires a Metropolis Hastings step is used for updating \( p \) random draws from a given distribution. The random samples are realizations (MCMC) in a bioinformatics context. MCMC is an application of the Law from MCMC as (2004). Denoting the sampling density as \( f \) \( Ig \) and \( \beta \) respectively, and \( p(\mu, \lambda) \) denotes a Beta density, \( \beta(\alpha, \beta) \) denotes a Gamma density, \( p(\beta) \) denotes a Multinomial-one density, \( p(\tau) \) a Binomial-one density and \( \beta(\beta_i, \beta_j, \mu, \lambda) \) \( p(\nu_0, \nu_1, \nu_2, \nu_3) \) denote Gaussian densities.

### 2.2 Algorithm

The complexity of the model requires approximate inference. Although closed form solutions (Liu et al., 2006, Sykaec et al., 2007) have computational advantages, we prefer here an unbiased approximation and follow (Bae and Mallick, 2004; Gottardo et al., 2006; Huang et al., 2002; Lewin et al., 2007; Shahbaba and Neal, 2006; Tadesse et al., 2003): we, among many others, have previously used Markov chain Monte Carlo (MCMC) in a bioinformatics context. MCMC is an application of the Law of Large Numbers and allows approximating expectations by averages of random draws from a given distribution. The random samples are realizations of a Markov chain that behave under certain conditions like draws from a single stationary distribution (cf. Gilks et al., 1996; Robert and Casella, 2004). Denoting the sampling density as \( f \) and the random samples obtained from MCMC as \( \beta_i^{(t)} \). MCMC allows us for example to approximate the expectation of the group-specific mean expression \( \beta \) as

\[
\mathbb{E}_\text{MCMC}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \beta_i^{(t)}
\]

Algorithm 1 illustrates MCMC sampling as pseudo-code. Inference requires a combination of Gibbs, Metropolis Hastings and Reversible Jump steps. Gibbs steps are used for updating the prior probability of differential expression, \( p \), the prior precision \( \lambda \), the error precision, \( \tau \), and, and keeping the expression dependent indicator \( I_j \) fixed, for updating the group means \( \beta_i \). A Metropolis Hastings step is used for updating \( J \) as long as we keep the Student’s \( t \) noise model. Updates of \( J \) that propose changing from a Student’s \( t \) to a Gaussian density and vice versa and updates of \( I_j \) rely on the reversible jump introduction introduced in Green (1995). Further details about the model, the algorithm and a Matlab implementation are provided in http://bioinfo.boku.ac.at/alex/rebromca.html.

### 2.3 Data collection

For reliably inferring the optimal noise characteristics and evaluating the implications of potentially oversimplified Gaussian assumptions, we have to consider two aspects. A reliable assessment of different noise models requires calibrating the proposed inference scheme. Calibration makes sure that MCMC converges rapidly and that inference result is insensitive to the chosen hyperparameters. These aspects are best assessed when knowing the expected outcome by using synthetically generated data and dedicated spike-in experiments. Warranting that our findings are generally applicable the chosen normalization method (cf. Bolstad et al., 2003). It is well known that results from microarray data analysis may depend on the chosen normalization method (cf. Bolstad et al., 2003). To ensure that our conclusions hold in general, we repeated the analysis for a subset of the data in Table 2 with additional normalization methods. Guided by their popularity in applied microarray papers, we chose loess (cf. Yang et al., 2002) and quantile (cf. Bolstad et al., 2003) normalization.

In the light of recent findings that intensities of highly expressed targets cross-talk in neighbouring probes due to scanner inadequacy (cf. Upton and Harrison, 2010), we may expect that Affymetrix probe sets contain outlier measurements. Being designed to alleviate the effect of artefacts contaminating individual probes, the mmgMOS approach (cf. Liu et al., 2005) and the PPLR method (cf. Liu et al., 2006) could help improving the expression. The mean structure and fraction of occurrence of each group are reported in Table 1. Variances have been chosen in the range of 0.1–10, without altering the reported results. To mimic a realistic microarray scenario, we generated five replicates per group, resulting in 10 data points per gene. Some aspects of computer-generated data might deviate from real microarray measurements. We endorse therefore our respective conclusions by including the spike-in experiment of Choe et al. (2005) in our analysis.

For warranting far reaching validity of our results, we analysed 14 microarray experiments covering various organisms and measurement platforms. The data include investigations of plant soil responses, drosophila sleep deprivation, primates dietary comparisons and animal liver metabolism. The experiments, which are summarized in Table 2, are identified by the Gene Expression Omnibus (GEO) reference number (cf. Edgar et al., 2002). Further details about each dataset can be found in the corresponding reference. The selection provided in Table 2 covers several different platforms and quantification algorithms (cf. column ‘Prep.’). We used all data as provided by the owner and applied the conservative normalization method vsn (cf. Huber et al., 2002).

### 2.4 Alternative normalization and analysis methods

The proposed model can be summarized by the joint density

\[
\begin{align*}
&\text{Random initialization of parameters} \\
&\text{for } n = 1 \text{ to burnin do} \\
&\text{update parameters}\{ \\
&\quad \text{update } \nu, J \text{ and } \nu_{g, h} \text{ jointly} \\
&\quad \text{update } \beta \\
&\quad \text{update } \lambda \\
&\quad \text{update } \beta_g \text{ and } I_g \text{ jointly} \\
&\quad \text{update } \tau \\
&\text{end for} \\
&\text{for } n = 1 \text{ to burnin+simulationlength do} \\
&\text{update parameters (see first burn-in)} \\
&\text{end for}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Subset i</th>
<th>( \mu_{1,1} )</th>
<th>( \mu_{1,2} )</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-12.0</td>
<td>12.0</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>-5.0</td>
<td>5.0</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>-1.0</td>
<td>1.0</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>-0.5</td>
<td>0.5</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>20</td>
</tr>
</tbody>
</table>

The proportion of genes in subset \( i \) is shown in column %.

Table 1. Depending on sample type which is either 1 or 2, genes from subset \( i \) are drawn from distributions with means equal to \( \mu_{1,1} \) and \( \mu_{1,2} \), respectively.

...
An important aspect in our assessment of different noise models for Arabidopsis thaliana GDS3216 Dinneny/GDS3225 Van Hoewyk et al. (2008). The GEO entry GDS1686 (I) refers to the behavioural subset of the data (only the sleep-deprived flies). In column Prep., we use MAS5.0 to refer to the Affymetrix MAS 5.0 (cf. de Haan et al., 2004; Dennis et al., 2003) -value threshold of 0.01 for assigning differential expression.

2.5 Biological implications

An important aspect in our assessment of different noise models for microarray data analysis is evaluating the biological implications of deviations from the appropriate noise model. The implication of choosing Gaussian noise instead of the optimal noise model can be quantified by deviations from the appropriate noise model. The implication of choosing the Gaussianity of residuals. For testing whether such sophisticated methods (cf. de Haan et al., 2009; Gao and Song, 2005; Lee et al., 2007) on the gene lists obtained with different noise models represented by Ig models, we obtain the following GO terms:

a. Using a distributional assumptions and using non-parametric methods (cf. de Haan et al., 2009; Gao and Song, 2005; Lee et al., 2007; Tusher et al., 2001), we investigate whether non-parametric approaches are a viable alternative for robust analyses of microarray data, we compare gene rankings obtained with such approaches with gene rankings we obtain (i) with the Bayesian ANOVA when using the optimal (possibly heavily tailed) noise distribution and (ii) with the proposed model when assuming Gaussian distributed noise. Compatibility with our ANOVA model suggests applying a Kruskal–Wallis permutation test (cf. Lee et al., 2005) and a robust ANOVA (cf. de Haan et al., 2009). Unknown differences in scale which we have to expect when comparing $P$-values and Bayesian probabilities are overcome by using a $P$-value threshold of 0.01 for assigning differential expression in the statistical test and adjusting the probability threshold such that the number of differentially expressed genes matches.

The GEO entry GDS1686 (I) refers to the behaviouralsubset of the data (only the sleep-deprived flies). In column Prep., we use MAS5.0 to refer to the Affymetrix MAS 5.0 quantization method. $RMA$ to refer to the Robust Multi-array Average method by Irizarry et al. (2003) (both used for Affymetrix arrays), PathStat for referring to the package described in Mulholland et al. (2004), CLSS4.1 to refer to the Codelink Software Suite 4.1 and RGP3.0 to refer to Research Genetics’ Pathway software v. 3.0.

<table>
<thead>
<tr>
<th>Org.</th>
<th>GEO ID</th>
<th>Reference</th>
<th>Prep.</th>
<th>$N$</th>
<th>Comm. genes</th>
<th>Diff. genes</th>
<th>Comm. GO terms</th>
<th>Diff. GO terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>GDS3216</td>
<td>Dinneny et al. (2008)</td>
<td>MAS5.0</td>
<td>12</td>
<td>4.71</td>
<td>1176</td>
<td>150</td>
<td>111</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>GDS3225</td>
<td>Van Hoewyk et al. (2008)</td>
<td>MAS5.0</td>
<td>4</td>
<td>5.50</td>
<td>832</td>
<td>290</td>
<td>41</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>GDS1404</td>
<td>Cameron et al. (2005)</td>
<td>PathStat</td>
<td>10</td>
<td>13.58</td>
<td>1776</td>
<td>136</td>
<td>11</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>CAMDA 08</td>
<td>Alifara et al. (2007)</td>
<td>CLSS4.1</td>
<td>24</td>
<td>4.04</td>
<td>400</td>
<td>304</td>
<td>26</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>GDS1375</td>
<td>Talantsev et al. (2005)</td>
<td>MAS5.0</td>
<td>70</td>
<td>3.25</td>
<td>6861</td>
<td>3561</td>
<td>160</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>GDS810</td>
<td>Blalock et al. (2004)</td>
<td>MAS5.0</td>
<td>31</td>
<td>4.37</td>
<td>72</td>
<td>135</td>
<td>9</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>GDS2960</td>
<td>Yao et al. (2007)</td>
<td>RQPS</td>
<td>101</td>
<td>4.33</td>
<td>318</td>
<td>106</td>
<td>17</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>GDS660</td>
<td>MacLennan et al. (2004)</td>
<td>MA</td>
<td>50</td>
<td>22.10</td>
<td>584</td>
<td>126</td>
<td>20</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>GDS3221</td>
<td>Somel et al. (2008)</td>
<td>$RMA$</td>
<td>24</td>
<td>4.21</td>
<td>180</td>
<td>119</td>
<td>108</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>GDS3162</td>
<td>Someya et al. (2008)</td>
<td>MAS5.0</td>
<td>10</td>
<td>4.38</td>
<td>797</td>
<td>446</td>
<td>112</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>GDS1555</td>
<td>MacLennan et al. (2006)</td>
<td>MAS5.0</td>
<td>8</td>
<td>3.90</td>
<td>131</td>
<td>183</td>
<td>24</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>GDS2946</td>
<td>Li et al. (2008)</td>
<td>MAS5.0</td>
<td>15</td>
<td>4.57</td>
<td>146</td>
<td>157</td>
<td>14</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>GDS5972</td>
<td>Jin et al. (2003)</td>
<td>MAS5.0</td>
<td>44</td>
<td>4.98</td>
<td>369</td>
<td>163</td>
<td>94</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>golden-spike</td>
<td>Choe et al. (2005)</td>
<td>MAS5.0</td>
<td>6</td>
<td>3.74</td>
<td>401</td>
<td>1748</td>
<td>–</td>
</tr>
</tbody>
</table>
chosen carefully to avoid side effects. The graphs show the ordered posterior probabilities of differential expression $P(D|X,a,b,c,d,e,h,K)$ with the legend denoting the corresponding $c,d$ pair for a Gaussian and Student's-$t$ noise model. Choices larger than around 100 increase the influence of these hyperparameters on the posterior of interest. This motivates our choice of an improper prior $c=0,d=0$.

Figure 2 illustrate this sensitivity analysis for synthetic data that was generated according to the description in Section 2.3. Choosing the Jeffreys prior $c=0,d=0$ is justified by observing that up to a prior variance of less than 1/100, the hyper-parameters $c$ and $d$ have, independently of the noise model, little influence on the posterior probabilities of differential expression.

Another important aspect of our inference scheme is providing accurate assessments of noise characteristics. This requires a clear distinction between Gaussian and Student's-$t$ noise models and thus an appropriate choice for the upper limit for the degrees of freedom parameter $\nu$, which marks the bound between Student's-$t$ and Gaussian distributions. Simulations on synthetic data showed that taking $\nu_{\text{max}}=45$ as upper limit is a good choice because larger degrees of freedom parameters render Student's-$t$ densities as indistinguishable from Gaussians and smaller values would unnecessarily misjudge Student's-$t$ densities as Gaussians.

Further calibration efforts were concerned with assuring fast convergence of the sampling algorithm to the stationary distribution. Our simulations showed that convergence speed can be dramatically improved by adjusting the grid size $c_{\text{grid}}$ between two burn-in phases. After starting with an initial value in the range of 1–5, we switch to a smaller value of about 0.05 which is then also used for sampling. A large initial grid size allows the algorithm to quickly determine the approximately correct error model with the smaller grid size improving the convergence properties of the Markov Chain and leading to better approximations of the true continuous degrees of freedom. Convergence towards the stationary distribution was assessed with the R package coda (cf. Plummer et al. (2005)). Resulting from a wet lab experiment, these data are both a paired with a by far larger evidence in favour of this noise model. This observation allows the conclusion that already the technical noise component in microarray data, which is the only remaining source of variation in the golden-spike data, requires considering robust models.

### Table 3. An assessment of robustness levels in dependence of normalization and preprocessing showing the expected degrees of freedom parameters

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>Loess</th>
<th>Quantile</th>
<th>mmgMOS</th>
<th>PPLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDS5216</td>
<td>2.02</td>
<td>1.13</td>
<td>2.23</td>
<td>1.17</td>
</tr>
<tr>
<td>GDS810</td>
<td>1.13</td>
<td>1.18</td>
<td>3.23</td>
<td>1.14</td>
</tr>
<tr>
<td>GDS5225</td>
<td>1.24</td>
<td>1.29</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CAMDA08</td>
<td>1.06</td>
<td>1.11</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GDS1375</td>
<td>1.14</td>
<td>1.15</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GDS2960</td>
<td>2.94</td>
<td>2.85</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GDS1555</td>
<td>1.15</td>
<td>1.17</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GDS972</td>
<td>1.38</td>
<td>1.4</td>
<td>3.67</td>
<td>1.15</td>
</tr>
</tbody>
</table>

For various datasets. Dashes indicate unavailable results, which require for mmgMOS and PPLR to have Affymetrix CEL files available. The results confirm that neither normalization nor sophisticated preprocessing compensate for the need of heavy-tailed noise models.

### 3 RESULTS

To highlight the importance of choosing valid noise models for microarray analysis, we applied the proposed inference scheme to 14 microarray datasets, which are summarized in Table 2. The arresting result of our evaluation is that a heavy-tailed Student's-$t$ noise model is a better fit than a Gaussian noise model for every dataset we looked at (cf. column ‘P’). For most datasets, a $t$-distribution with degrees of freedom between 3 and 5 got the highest posterior probability. This indicates the need of robust noise models, which can handle outlying data points well and suggests that Gaussian noise models are unsuitable for microarray data analysis, even if according to Novak et al. (2006) only about 5–15% of samples are non-Gaussian distributed.

Our assessments also revealed that biological inference depends considerably on the chosen noise model. For obtaining a quantitative statement, we inferred the differentially expressed genes set twice: once with a Gaussian noise model and once with the optimally inferred $t$-distribution. This approach provided for every dataset two gene lists with the intersect representing agreement and the symmetric difference representing different biological interpretations, which are solely caused by the different noise models (cf. Table 2, columns ‘Comm. genes’ and ‘Diff. genes’).

Microarray data analysis depends often critically on chosen preprocessing and normalization (cf. Bolstad et al., 2003). To rule out being misled by a particular choice, we repeated the assessment of optimal noise models using loess and quantile normalization and mmgMOS and PPLR preprocessing. The expected degrees of freedom, $\nu$ we report in Table 3 for these data allow concluding that our observations are independent of normalization and even sophisticated analysis methods do not compensate for the need of robust noise models. The robust model is in general less sensitive to outlying values. Models with $t$-distributed noise will therefore assign lower posterior probabilities of differential expression.
This is about one tenth to two times the number of genes, which Table 2 shows that the number of genes with noise model-dependent probabilities of differential expression obtained by a Gaussian-based analysis (probabilities shown as black line). The probabilities obtained for the same gene from an optimally adjusted robust analysis are shown as dots.

Subplot (A) in Figure 3 is ranked by the posterior probabilities obtained with the optimal \( t \)-distributed noise model (probabilities shown as black line). A subset of posteriors obtained with Gaussian noise is shown as dots. Subplot (B) is ranked by the posterior probabilities obtained with a Gaussian noise model (probabilities shown as black line). A subset of posteriors obtained with an optimally adjusted robust analysis are shown as dots. Subgraph (A) illustrates the Human melanoma data (GDS1375) ranked by the posterior probabilities of differential expression obtained by a Gaussian-based analysis (probabilities shown as black line). The probabilities obtained for the same genes from an optimally adjusted robust analysis are shown as dots.

When differential expression is caused by one or a few outlying measurements. In situations where outliers lead to an increase of variance or a decrease in average differential expression, the Gaussian noise model will overlook differentially expressed genes, which would be captured by the more appropriate \( t \)-distributed noise model. A wrongly chosen noise model will therefore lead to false positives and false negatives. Both error types are confirmed by the illustrations in Figure 3, which show many genes with noise model-dependent probabilities of differential expression.

The columns under KW perm. illustrate the agreements of the Kruskal–Wallis permutation test with the robust parametric method (columns ‘\( T \)\%)’ and with a Gaussian-based analysis (column ‘\( N \)\%). The two columns under RANOVA show the same information for the robust ANOVA method. Dashes indicate that the non-parametric method did not find differentially expressed genes. These results allow the conclusion that non-parametric methods are viable for analysing microarray data robustly, as long as we have sufficiently many samples.

Having gathered substantial evidence that microarray data should be analysed by considering heavy-tailed noise, the question arises whether non-parametric approaches can help solving this issue. We compare to this end the agreement in gene lists obtained with non-parametric tests with our robust Bayesian ANOVA and compare this with the agreement we observe between the same tests and the Gaussian version of our Bayesian ANOVA. The results in Table 4 show a better agreement of rankings between the robust methods, which suggests that non-parametric methods should be considered for analysing microarray data. Our results do, however, in agreement
with Whitley and Ball (2002) also reveal the loss in power inherent to non-parametric methods. In our analysis this manifests in finding no significant $P$-values with the robust ANOVA method for GEO ID GDS3216. For GEO ID GDS3225, GEO ID GDS1555 and the CAMDA 08 data, both non-parametric methods fail in finding significant $P$-values (data omitted from table). From Table 4, it is also obvious that small sample sizes (cf. Table 2, column ‘$N$’) lead in general to poor agreement. If sample sizes permit application, we can however recommend non-parametric methods for microarray data analysis.

4 DISCUSSION

This article provides an in-depth assessment of two competing assumptions about the noise characteristics in microarray data. Assuming Gaussian noise has the benefit of leading to highly efficient analysis methods. A considerable sensitivity to outlying observations is, however, an unfortunate weakness of Gaussian noise-based data analysis. This weakness may be overcome with non-parametric methods or by methods which assume heavy-tailed noise distributions. Applying robust analysis methods to microarray data has the disadvantage of introducing more involved computations. The application of non-parametric methods is in addition limited to problems with sufficiently many samples.

Comparing robust analysis methods with Gaussian-based microarray data analysis has to provide conclusions, which are relevant for biological practise. Certain technical aspects can be tested by gold standards like the spike-in data from Choe et al. (2005). Other aspects like, for example, biological variation are only captured by data analysing real-world biology. Although certain facts about individual experiments are well known, complete knowledge of ground truth is not available for any biological microarray experiment. An assessment of biological implications has thus to resort to indirect strategies. The route chosen in this article first compares the technical suitability of Gaussian noise and heavy tailed r-distributions. This requires a mode of operation in data analysis, which allows comparing different noise models. Once we established which noise model is preferred for technical reasons, we can turn to investigating the biological implications caused by changing the noise model. This mode of operation relies in our analysis on counting the number of genes which show a noise model-dependent difference in differential expression. These gene counts are complemented by investigating which GO terms are significantly affected from the noise model-dependent gene lists. For providing conclusions of far-reaching validity, we analysed I4 carefully chosen microarray experiments, covering a wide range of model organisms and measurement platforms. To avoid reporting spurious results, our simulations included careful tuning of hyperparameters to minimize model sensitivity, steps for assessing convergence of the algorithm and applied different normalization and preprocessing methods.

The arresting result of our assessment is that we find highly decisive evidence in favour of r-distributions with high kurtosis for every experiment we looked at. The significance of this finding is backed up by the observation that the choice of error model considerably influences the biological conclusions drawn from the analyses. Gene lists differ in dependence of the noise model by between 119 and 3561 genes. These differences have a substantial influence on the conclusions we draw on a higher level of biological abstraction. The number of differences in the GO term lists we find in dependence of the chosen noise model ranges from 14 to 316. For many datasets, the number of GO terms with noise model-dependent equivocal assessment is larger than the number of GO terms we can unambiguously assign to these experiments irrespective of the chosen noise model. We may thus conclude that a substantial number of outlying measurements is present in many microarray studies. Relying on implicit Gaussian assumptions means ignoring the heavy tails of the residuals and that can have adverse effects on biological conclusions drawn from microarray data. Practitioners should thus apply robust approaches for microarray data analysis, which work reliably irrespective of whether noise is Gaussian or heavy tailed. We suggest for this purpose considering non-parametric approaches (cf. de Haan et al., 2009; Gao and Song, 2005; Lee et al., 2005; Troyanskaya et al., 2002; Tuscher et al., 2001; Zhao and Fan, 2003), or, for small sample sizes, apply Bayesian approaches like Gottardo et al. (2006) or the MatLab implementation which accompanies this paper at http://bioinf.boku.ac.at/alex/p/robmca.html.

ACKNOWLEDGEMENTS

The authors would like to thank D. Kreil for his input on microarray normalization. We are also grateful to our anonymous reviewers who helped improving our article with their valuable comments.

Funding: A. Posekany and P. Sykacek are grateful to the Vienna Science and Technology Fund for funding this work within the project WWTF-LSC-2005-335.

Conflict of Interest: none declared.

REFERENCES

Cameron, D. et al. (2005) Gene expression profiles of intact and regenerating retinal. Mol. Vis., 11, 775–791.


