BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 2 2006, pages 215–219
doi:10.1093/bioinformatics/bti790

Gene expression

Statistical estimation of gene expression using multiple laser scans of microarrays

Mizanur R. Khondoker[1,2,*], Chris A. Glasbey[1] and Bruce J. Worton[2]

[1]Biomathematics & Statistics Scotland and [2]School of Mathematics, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JZ, Scotland, UK

Received on August 9, 2005; revised on November 10, 2005; accepted on November 15, 2005
Advance Access publication November 22, 2005
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We propose a statistical model for estimating gene expression using data from multiple laser scans at different settings of hybridized microarrays. A functional regression model is used, based on a non-linear relationship with both additive and multiplicative error terms. The function is derived as the expected value of a pixel, given that values are censored at 65 535, the maximum detectable intensity for double precision scanning software. Maximum likelihood estimation based on a Cauchy distribution is used to fit the model, which is able to estimate gene expressions taking account of outliers and the systematic bias caused by signal censoring of highly expressed genes. We have applied the method to experimental data. Simulation studies suggest that the model can estimate the true gene expression with negligible bias.

Availability: FORTRAN 90 code for implementing the method can be obtained from the authors.

Contact: mizanur@bioss.ac.uk

1 INTRODUCTION

DNA microarrays are proving immensely valuable to cell biologists, scientists and drug researchers, by being able to track tens of thousands of molecular reactions in parallel. Microarray technology aims at assessing the transcript abundances (measured in terms of fluorescence intensity) of thousands of genes in response to different experimental conditions or in different tissue samples. One of the major problems of microarray analysis is that the quantification of fluorescence intensity does not give direct measurement of messenger RNA (mRNA) abundance of the gene of interest. In addition to the random noise, measured expression levels are disturbed by a number of systematic factors. One of the sources of systematic bias in the intensity measurements is laser scanner setting. The sensitivity level of microarray scanners is adjustable and plays a crucial role in getting reliable measurement of the fluorescence intensity. A change in scanner setting transforms the intensity measurements by a multiplicative constant. A scanner's sensitivity has to be raised to a certain level to ensure that the intensity levels of weakly expressed genes exceed the intrinsic noise level of the scanner and so become measurable. This may, however, cause another problem: signal censoring for highly expressed genes. Scanners cannot record pixel intensities above some software-dependent threshold ($2^{16}-1 = 65\,535$, for a 16 bit computer storage system), so highly expressed genes can have pixel values which are right censored at the largest possible value that the scanner software allows. It is not usually possible to find a scanner setting which is optimal for both weakly and highly expressed genes. So, it seems reasonable to consider multiple scanning of the same microarray at different scanner settings and to estimate spot intensities from these combined data. To illustrate, Figure 1 shows data from four scans of a single channel of a microarray. The experiment, conducted at the Scottish Centre for Genomic Technology and Informatics, University of Edinburgh, was designed to examine the effects of ingestion of apoptotic cells on macrophage gene expression 24 h after administration and to compare this expression profile against a control of untreated cells. Each of two arrays containing 9248 spots (representing 4624 genes each replicated twice) was scanned with an Affymetrix 428 scanner at four different sensitivity levels and analysed using Quantarray. Here, the estimated expression level from scans 1 to 4 for each of 9248 spots has been plotted against that for scan 1. We see the multiplicative change due to scanner setting and the effect of pixel censoring at $T = 65\,535$. The challenge is to estimate the expression level of each gene from data such as these.

Little work has been done so far on adjustment of pixel censoring. Depending on the type of data used two types of methods have been found in the literature: methods using pixel level data and methods using spot summary data. Spatial statistical models on the pixel level, termed spot shape models, were considered by Ekstrøm *et al*. (2004) to predict signal intensities of the censored pixels. Glasbey C. A., T. Forster and P. Ghazal (submitted for publication) (2006) proposed a linear model to impute censored pixels based on the principal components of the uncensored spots on the same array. The idea of using multiple scan data is also fairly new. Dudley *et al*. (2002) used summary data from multiple scans to correct pixel censoring by combining the linear ranges of each scan onto a common linear scale. Romualdi *et al*. (2003) used multiple scan data to get improved spot summaries through image integration. The problem of addressing downward bias in the spot summary measures of highly expressed genes arising due to pixel censoring was considered, on the basis of summary data from a single scan, by Wit and McClure (2003). The authors suggested statistical adjustment for pixel censoring based on typically available spot summaries. For every spot, the method uses the observed values of mean, median and variance statistics to fit a two-parameter probability model. The median or mean of the fitted distribution, according to the paper, is a good alternative to the observed median or mean intensity of that spot. However, the result is dependent on the choice of distribution for the pixel values and the method is likely to

*To whom correspondence should be addressed.

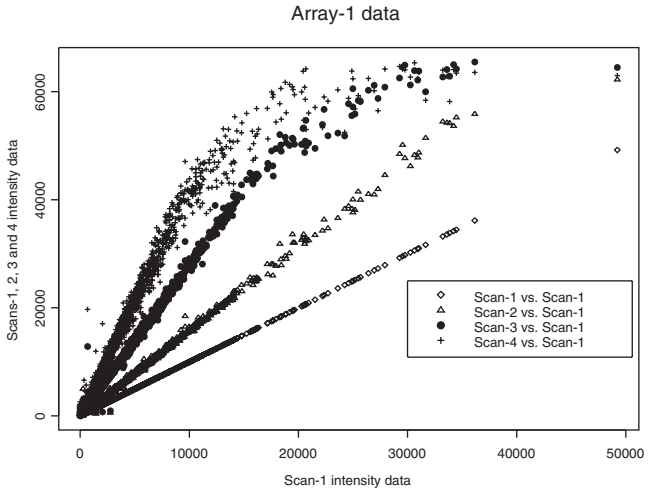

Fig. 1. Scatterplot of scans 1, 2, 3 and 4 versus scan-1 intensity data from a single channel of a microarray.

produce unstable estimates as it uses only three observations to estimate two parameters. Although better adjustment may be possible using pixel level data, such data are generally unavailable as this would involve handling vast datasets.

In Section 2 we propose a statistical model for estimating gene expressions using spot summary data from multiple scans, and in Section 3 we fit the model to experimental data. We validate our approach by simulation in Section 4 and finally in Section 5 we review the method.

2 MODEL AND ESTIMATION

Suppose that the same microarray has been scanned several (say, m) times at different sensitivity levels of the scanner. Let y_{ij} denote the observed intensity of the i-th of n spots in the j-th scan. In the absence of censoring, we assume that the expectation of y_{ij} would be $\mu_i\beta_j$, where μ_i is the expression level of gene i and β_j is the multiplicative scaling effect due to scanner setting j. The observed intensity is the average of pixel values. For example, the data plotted in Figure 1 were produced by Quantarray, using the average of pixels between the 80th and 95th percentiles contained in a 25 × 25 square centred on each spot. If some of these pixels are censored at T then the expectation of y_{ij} will be less than $\mu_i\beta_j$. If pixel values associated with a spot are normally distributed with mean $\mu_i\beta_j$ and variance $\mu_i^2\beta_j^2\nu^2$, where ν is a variance scaling term, then

$$\begin{aligned} E(y_{ij}) &= T + (\mu_i\beta_j - T)\Phi\left(\frac{T-\mu_i\beta_j}{\mu_i\beta_j\nu}\right) \\ &\quad - \mu_i\beta_j\nu\phi\left(\frac{T-\mu_i\beta_j}{\mu_i\beta_j\nu}\right) \\ &= g(\mu_i\beta_j,\nu),(\text{say}), \end{aligned} \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution functions of the standard Gaussian random variable, respectively, using expressions for truncated normal distributions (Johnson *et al*., 1994, p. 156). We do not believe the normal distribution to be entirely appropriate, but it yields a mathematically tractable expression for g, whose precise functional form is probably not important, beyond it being hyperbolic in shape. Typical curves are shown in Figure 2.

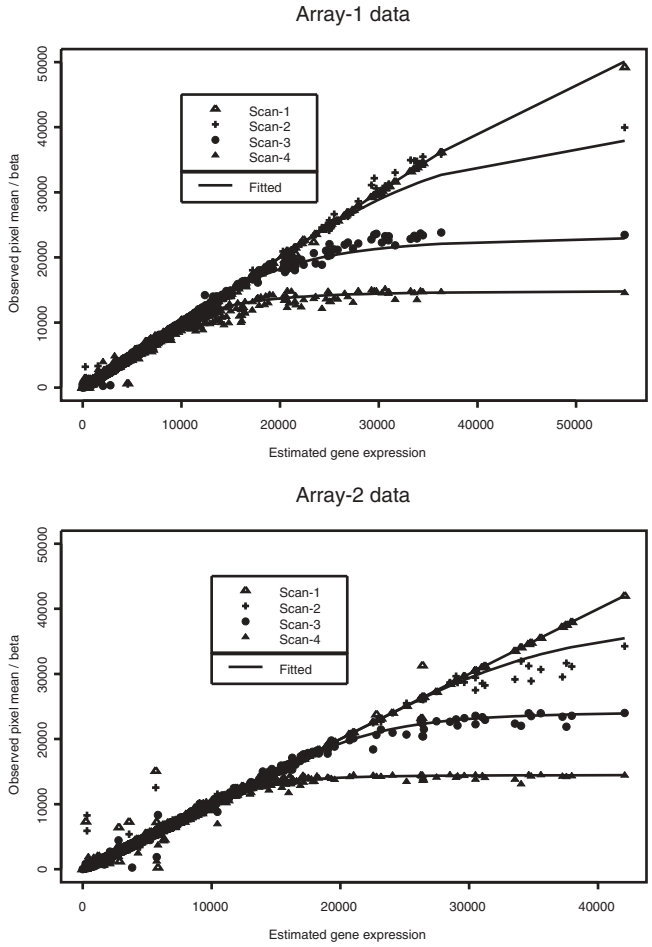

Fig. 2. Rescaled intensities $(y_{ij}/\hat{\boldsymbol{\beta}}_j)$ plotted against estimated gene expressions $(\hat{\boldsymbol{\mu}}_i)$. The solid lines indicate the corresponding fitted model.

We assume that y_{ij} is distributed with location $g(\mu_i\beta_j,\nu)$. However, rather than assuming a normal distribution, we choose to use a heavy-tailed distribution to account for the outliers, which are a feature of the data as illustrated in Figure 1. Specifically, we assume a Cauchy distribution with scale $\sigma_{ij} = \sqrt{(\sigma_1^2 + \sigma_2^2\mu_i^2)\beta_j^2}$. In passing, we note that Cauchy distributions do not have expectations, and so g could not be derived from it. The combined additive and multiplicative nature of error variability has been proposed previously by Ideker *et al*. (2000), Rocke and Durbin (2001), Huber *et al*. (2002, 2003) and Durbin and Rocke (2003), and is consistent with the data in Figure 1. Purdom and Holmes (2005) used a heavy-tailed distribution, though in their case they used a Laplace distribution. The proposed model therefore is

$$y_{ij} \sim C(g(\mu_i\beta_j,\nu),\sigma_{ij}^2), \tag{2}$$

where $\beta_1 \equiv 1$ for identifiability. The notation $C(a,b^2)$ represents a Cauchy distribution with location and scale parameters a and b, respectively.

Model (2) belongs to the class of functional regression model, a form of Measurement Error model (Cheng and Van Ness, 1999).

Additive and multiplicative dispersion parameters σ_1 and σ_2 are scaled by the corresponding scanning effects (β_j) to allow for increasing variability, as evident in Figure 1, across scans of increasing sensitivity. For functional regression models it is problematic to estimate separate scaling terms for each variable. It is shown in the literature of functional relationships (Cheng and Van Ness, 1999) that for a simpler model, such as $y_{ij} \sim N(\mu_i\beta_j, \delta_j^2)$, the log-likelihood function $L \to \infty$ as any one of the variance parameters $\delta_j^2 \to 0$. Another option is to consider a structural relationship, by treating μ_i as a latent random variable. In the Gaussian case, as pointed out by Mardia *et al*. (1979, Exercise 9.2.7, p. 277), this problem can be approached using a factor analysis model. However, we prefer not to make assumptions about the distribution of the μs. Therefore, to circumvent problems of estimation, we make the simplifying model assumption that the scale parameters increase in proportion to β across scans.

The log-likelihood function for estimating the parameters of model (2) can be expressed as follows:

$$L(\mu, \beta, \sigma_1, \sigma_2, \nu) = \sum_{i=1}^{n} L_i(\mu_i, \beta, \sigma_1, \sigma_2, \nu), \tag{3}$$

where

$$L_i(\mu_i, \beta, \sigma_1, \alpha_2, \nu) = -\sum_{j=1}^{m}\left[\log \sigma_{ij} + \log\left\{1 + \left(\frac{y_{ij} - g(\mu_i\beta_j, \nu)}{\sigma_{ij}}\right)^2\right\}\right]. \tag{4}$$

A challenge of working with this model is the estimation of the large number ($n + m + 2$) of parameters. We propose an alternating algorithm for simultaneous estimation of all the parameters of model (2) as follows:

(1) Set $\mu = y_{\cdot 1}$ (intensity data of scan-1) as the starting values and maximize L with respect to all other parameters $(\beta, \sigma_1, \sigma_2, \nu)$, where μ is a vector of dimension n, β is an $(m-1)$ vector and σ_1, σ_2 and ν are scalars. Denote the updated values of other parameters by $(\beta^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \nu^{(1)})$.

(2) Update each μ_i, $(i = 1, \ldots, n)$ individually according to the following substeps:

  (a) For each j, set $\mu_i = g^{-1}(y_{ij}, \nu^{(1)})/\beta_j^{(1)}$.

  (b) Maximize L_i with respect to μ_i alone.

  (c) Repeat (a) and (b) for $j = 1, \ldots, m$.

  (d) From among the m updated values of μ_i, choose the one with maximum L_i value. Denote the updated vector by $\mu^{(1)}$.

(3) Update the $(m+2)$ parameters in $(\beta, \sigma_1, \sigma_2, \nu)$ by maximizing $L(\beta, \sigma_1, \sigma_2, \nu, \mu^{(1)})$ for given values of the gene expression parameters in $\mu^{(1)}$.

Continue repeating steps (2) and (3), replacing the previous estimates by the updated ones, until gain in the log-likelihood function is negligible. The substeps under step (2), that update each μ_i starting from m different initial values, are essential. Otherwise, the algorithm may be trapped in a local optimum. The simplex method of Nelder and Mead (1965) using FORTRAN 90 and IMSL Library was used as an optimization tool. The IMSL routine `DUMPOL` implements the simplex method of function minimization.

At each iteration L increases. Therefore, because L is bounded above with probability 1, the alternating algorithm is guaranteed to terminate at a local stationary point. For the gene expression parameters (μ) the likelihood naturally decomposes into n components, and μ_i can be estimated by maximizing the i-th component (L_i), which generally has m peaks, one near to the intensity value for each scan. Multiple starts for each μ_i therefore improves the chance of finding the highest peak. However, as is usually the case with optimization algorithms, there is no guarantee that the global maximum will be found.

Table 1. Estimates of the scanning effects and scale parameters

Dataset	Scanning effects			Scale		
	β_2	β_3	β_4	σ_1	σ_2	ν
Array-1	1.56	2.75	4.32	5.36	0.0068	0.42
Array-2	1.71	2.71	4.53	5.36	0.0051	0.27

3 EXAMPLE

We apply the method to data from a single channel of two microarrays, one of which is plotted in Figure 1, for the experiment described in Section 1. CPU time (with a single processor Ultra-1 Sun machine) for executing the program to apply the method of Section 2 to each microarray took 11 minutes. Estimates of the parameters (other than μ) for both sets of data are tabulated in Table 1.

Observed intensity data divided by the corresponding scanning effects (β) for both sets of data are plotted against the corresponding estimated gene expressions (μ) in Figure 2. It is seen that the estimated gene expressions, particularly for the highly expressed genes, are more consistent with scan-1 intensity data. This is the desired case because the data of scan-1, scanned at the lowest level of scanner's sensitivity, are likely to be least affected by the pixel censoring. For weakly expressed genes the model has sufficient information, from all scans of data combined, for reliable estimation of the expression values.

Figure 3 shows a plot of standardized residuals against the rank of estimated gene expressions from one microarray and does not indicate any obvious model violations. Assessment of model fit is also possible via likelihood-based criteria such as AIC and GAIC. However, more pertinent is whether the use of multiple scans can reduce the signal-to-noise ratio in the estimates of gene expression.

On each array each gene has been replicated twice in such a way that spot i and $i + n/2$ represent the same gene where $i = 1, \ldots, n/2$. To compare the between-replicate variations in the data and fit, we compute

$$S(\tilde{\boldsymbol{\mu}}) = \sum_{i=1}^{n/2} \frac{(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i+n/2})^2}{[(\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\mu}}_{i+n/2})/2]^2}, \tag{5}$$

where $\tilde{\boldsymbol{\mu}}$ is replaced by $\hat{\boldsymbol{\mu}}$ to assess the multi-scan estimate, and by $y_{\cdot j}/\beta_j$ to assess the use of scan j alone. Because variability increases approximately as the square of the expression level, we give equal weight in S to genes at low and high levels by dividing by the square of the estimated expression level for each gene. However, rather than computing this using $\tilde{\boldsymbol{\mu}}$, which is downward biased for censored spots, we use $\hat{\boldsymbol{\mu}}$ in all cases.

The results are summarized in Table 2. It is seen that between-replicate variation in the estimated gene expressions is less than that

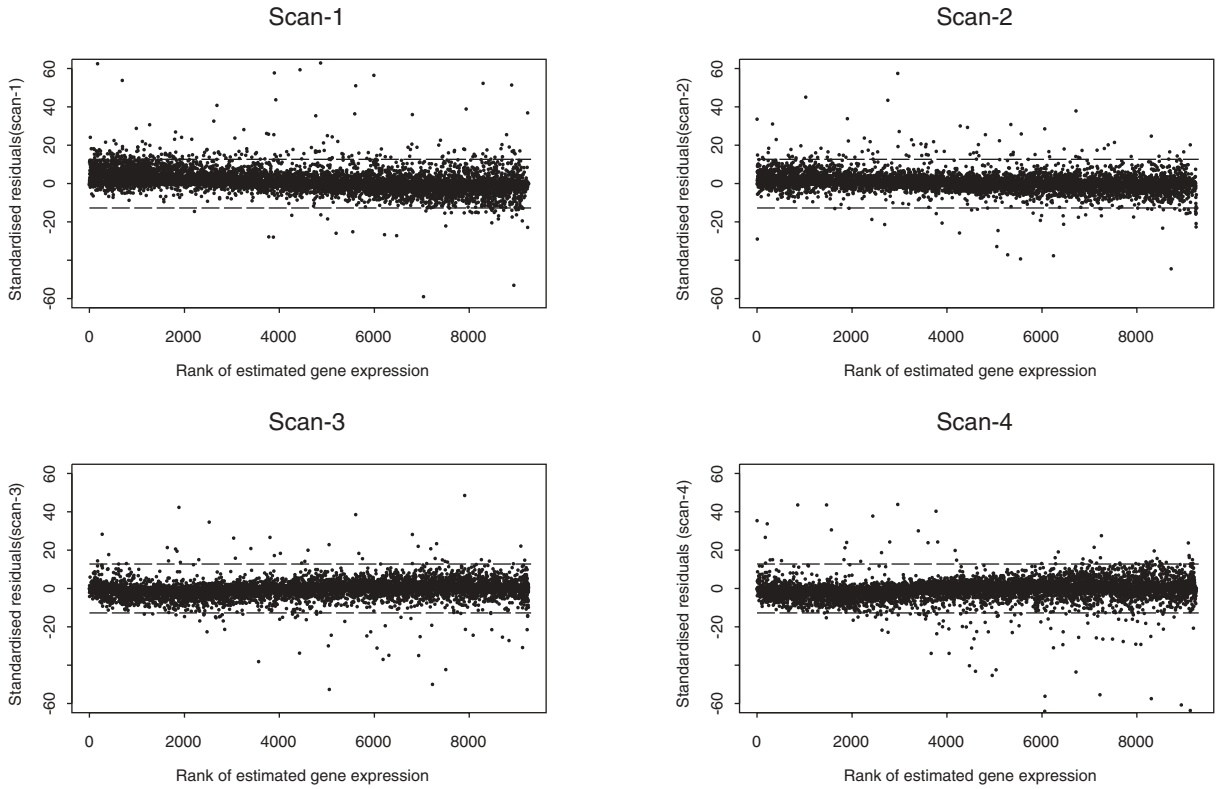

Fig. 3. Standardized residuals against the ranks of estimated gene expressions. The dashed lines show 95% probability limits (±12.71).

Table 2. Comparison of between-replicate variation in data and fit

Dataset	Between-replicate variation				
	$S(\hat{\boldsymbol{\mu}})$	$S(y_{\cdot 1}/\beta_1)$	$S(y_{\cdot 2}/\beta_2)$	$S(y_{\cdot 3}/\beta_3)$	$S(y_{\cdot 4}/\beta_4)$
Array-1	812	958	913	823	927
Array-2	858	1683	1768	882	863

Table 3. Estimated biases and standard errors

	Parameters					
	β_2	β_3	β_4	σ_1	σ_2	ν
True	1.71	2.71	4.53	5.36	0.0051	0.27
Bias	−0.00005	−0.00007	−0.00015	−2.036	−0.00187	0.00077
SE	0.00038	0.00069	0.00111	0.053	0.00008	0.00235

The results are based on 100 simulated datasets.

in any individual scan of data. This suggests that it is possible to reduce the between-replicate variation of the gene expression measurements by combining the data according to the proposed model from several scans. Results of Table 2 indicate that by combining scans we improve the signal-to-noise ratio in the data, particularly relative to scan 1, which would be the scientists' preferred single scan, as the other ones are affected by censoring bias.

4 SIMULATION STUDY

We performed some simulation experiments to check the validity of the estimation algorithm. We simulated 100 datasets from model (2) using the parameter values as estimated for Array-2 data (Table 1). For the gene expression parameters we used the same set of values for both replicates, obtained as the average of the estimated gene expressions of the two replicates for Array-2 data. Empirical biases and standard errors of the parameter estimates (other than μ) are summarized in Table 3. It is seen that the parameters (except for σ_1 and σ_2) are estimated with high precision and negligible bias. There is substantial downward bias in the maximum likelihood estimates of σ_1 and σ_2. This bias, however, does not affect the estimation of the other parameters and in particular the gene expression parameters (μ_i). We have conducted some additional investigation of the bias in scale parameter estimation. It is seen that when we estimate scale (σ_i) from the model $y_{ij} \sim C(\mu_i, \sigma_i^2)$ it appears unbiased. However if we combine the observations over i to estimate a common σ from the model $y_{ij} \sim C(\mu_i, \sigma^2)$ there is downward bias similar to that shown in the above simulations. The amount of bias depends on the value of n (number of spots) and m (number of scans) but the changes are negligible when n exceeds some large (say, 100) value. From the simulation results we found that $E(\hat{\sigma}^2) \approx 0.4\sigma^2$ for $n \geq 100$, $m = 4$ but each μ_i is approximately unbiased. We think, however, that there is little concern as this bias does not affect the estimation of gene expression parameters. Simulation results suggest no systematic bias in the gene expressions. We plot empirical biases (as percentage of true values) against the rank of true values in Figure 4. The bias in estimating gene expression parameters is seen to be in an acceptable range, in most cases <0.5%.

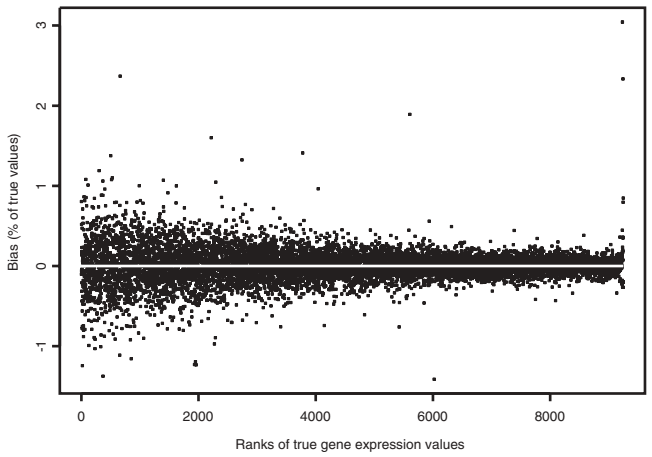

Fig. 4. Percentage of bias against the ranks of true gene expression values.

5 DISCUSSION AND CONCLUSIONS

Microarray gene expression data obtained as the output of typical image analysis steps are contaminated, in addition to other factors, by the scanner's intrinsic noise level (at the lower end) and by the pixel censoring (at the upper end). As the problems at the two ends are in conflict, no unique scanner setting is optimal. Moreover, there is no objective guideline to date for choosing optimum scanner setting to address these issues. It therefore seems reasonable to consider multiple scanning, some at relatively lower sensitivity levels (ensuring that there is no censoring at the upper end) and the others at higher sensitivity levels (ensuring the visibility of the weakly expressed genes over the scanner's intrinsic noise level) and combine the information together to get final gene expression measures. The simplest approach of combining the data through simple or weighted average over the scans will give a biased result as some individual scans of data are likely to be affected by pixel censoring. The proposed model can successfully combine the data of multiple scanning to get improved gene expression measures throughout the entire range of intensity data. As the simulation results suggest, the model is capable of estimating gene expressions adjusting for outliers and pixel censoring with reasonable precision and negligible bias. One strength of the model is that the location function specified in (1) explicitly captures the trend of the possibly censored spot summary data. Also, the derivation of the function has a natural correspondence with the data generation mechanism of microarray scanners. The choice of the Cauchy distribution for handling outliers proved to be better than the robust methods with which we have experimented. For example, methods of fitting based on using *M*-estimation or Least Trimmed Squares (Rousseeuw and Leroy, 1987) require subjectivity about the amount of robustness needed, e.g. the proportion of observations to be considered as outlying. The Cauchy distribution is however a reasonable choice on the grounds of simplicity and objectiveness. Among the few available methods of its kind in the literature, Dudley *et al*.'s (2002) method also considers multiple scan data but loses information discarding data outside the linear range. The method of Wit and McClure (2003) considers single scan data and does not suggest a general pixel distribution. The authors note that their method may produce unstable estimates as it estimates two parameters using only three summary statistics, mean, median and variance.

Finally, we consider how the model may be extended. A natural extension would be to replace the Cauchy distribution by a *t*-distribution. This would introduce an additional degrees of freedom parameter which would ideally be estimated from the data, and depend on the tail behaviour. We have conducted some simulation experiments with such a model. The bias in the estimation of the scale parameter noted in Section 4 for the Cauchy model is also present in the estimation of the scale parameter for the *t*-distribution model but additionally there is a corresponding bias in the estimation of the degrees of freedom parameter. However, we found that we get very similar maximum likelihood estimates of the μ_i as with the Cauchy model and therefore there was little advantage in using the slightly more complex model. The Cauchy distribution has a very slightly heavier tail than the data required, but this did not cause any problems with the estimation as would have been the case if the error distribution for the model had been taken to be too light-tailed, e.g. a normal or a *t*-distribution with large degrees of freedom. The use of the Cauchy distribution is convenient and although it is perhaps slightly too heavy tailed it provides some extra robustness in the estimation procedure.

ACKNOWLEDGEMENTS

We thank Scottish Centre for Genomic Technology and Informatics for providing the data used in this study. M.R.K. and C.A.G. were supported by the Scottish Executive Environment and Rural Affairs Department (SEERAD).

Conflict of Interest: none declared.

REFERENCES

Cheng,C.-L. and Van Ness,J.W. (1999) *Statistical Regression with Measurement Error*. Arnold, London.

Dudley,A.M. *et al*. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl Acad. Sci. USA*, **99**, 7554–7559.

Durbin,B. and Rocke,D.M. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 1360–1367.

Ekstrøm,C.T. *et al*. (2004) Spot shape modelling and data transformations for microarrays. *Bioinformatics*, **20**, 2270–2278.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Huber,W. *et al*. (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol*., **2**:1, article 3.

Ideker,T. *et al*. (2000) Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *J. Comp. Biol*., **7**, 805–818.

Johnson,N.L., Kotz,S. and Balakrishnan,N. (1994) *Continuous Univariate Distributions*, Vol. 1. John Wiley and Sons, NY.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press Inc., London.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Comp. J*., **7**, 308–313.

Purdom,E. and Holmes,S.P. (2005) Error distribution for gene expression data. *Stat. Appl. Genet. Mol. Biol*., **4**:1, article 16.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol*., **8**, 557–569.

Romualdi,C. *et al*. (2003) Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Res*., **31**, e149.

Rousseeuw,P.J. and Leroy,A.M. (1987) *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Wit,E. and McClure,J. (2003) Statistical adjustment of signal censoring in gene expression experiments. *Bioinformatics*, **19**, 1055–1060.