Statistical adjustment of signal censoring in gene expression experiments

Wit, Ernst; McClure, John

BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1055–1060
DOI: 10.1093/bioinformatics/btg003

Statistical adjustment of signal censoring in gene expression experiments

Ernst Wit* and John McClure

Department of Statistics , University of Glasgow, Glasgow G12 8QW, UK

Received on April 30, 2002; revised and accepted on September 17, 2002

ABSTRACT

Motivation: Numerical output of spotted microarrays displays censoring of pixel intensities at some software dependent threshold. This reduces the quality of gene expression data, because it seriously violates the linearity of expression with respect to signal intensity. Statistical methods based on typically available spot summaries together with some parametric assumptions can suggest ways to correct for this defect.

Results: A maximum likelihood approach is suggested together with a sensible approximation to the joint density of the mean, median and variance—which are typically available to the biological end-user. The method 'corrects' the gene expression values for pixel censoring. A by-product of our approach is a comparison between several two-parameter models for pixel intensity values. It suggests that pixels separated by one or two other pixels can be considered independent draws from a Lognormal or a Gamma distribution.

Availability: The R/S-Plus code is available at http://www.stats.gla.ac.uk/∼microarray/software.

Contact: ernst@stats.gla.ac.uk

1 INTRODUCTION

Typically the gain of microarray scanners is adjustable. This transforms the scanned intensities approximately by a multiplicative constant. The main reason for such adjustment is generally to increase the intensity level of lowly expressed genes to a level that exceeds the intrinsic noise level of the scanner. Low frequency noise could potentially swamp all low level gene expression as well as subtle differences between such lowly expressed genes. After adjustment unexpressed genes and differences between expressions are more clearly visible.

Unfortunately this advantage doesn't extend indefinitely because a similar problem exists at the very upper range of the scanner's sensitivity. Increasing the gain of the scanner too much has as a result that highly expressed genes get spot values close to or at the largest possible value that the scanner software allows. In a 16-bit (double precision) computer storage system, this is equal to $2^{16} - 1 = 65\,535$.

These two complementary problems require real skill on the part of the experimenters to make sure that the main bulk of the expression levels lie in the central region of the scanner sensitivity range. However, careful execution of the experimental methods is not sufficient. Figure 1 shows that the problems with so-called intensity censoring already begin in the middle region of the normal scanner sensitivity range. At an intensity of approximately 45 000 the mean becomes smaller than the median. Also at that level the variance starts to decrease. In other words, already when the mean and median spot intensity are only two thirds of the maximum intensity certain pixels are expressed higher than the highest possible value and are therefore censored. For this reason, it is paramount to consider ways to postprocess the intensity data in order to correct for such spurious effects.

Early papers (Chen *et al.*, 1997) make assumptions of normality for the raw pixel values. (Yang *et al.*, 2000) and many other commentators suggest that the logarithm of the pixel values can be assumed normal, although others (Kerr *et al.*, 2000) notice large tails even in the log-transformed data and suggest to use bootstrap. However, none mention the curious cut-off effect at the higher expression levels. Amaratunga and Cabrera (2001) suggest the use of bilinear or spline transformations to adjust for the apparent clustering of expression values at the upper threshold, although also they do not seem to be aware of the underlying reason. Similar alternatives involve modelling the expression data by a distribution that is defined on a finite interval, such as the Beta distribution. However, all these methods fail to address the issue of pixel censoring specifically and as a result may lead to bias.

Our method does not suffer from these aforementioned defects. It uses several pixel summaries such as the mean, median and variance to estimate the fraction of pixels that are censored and on the basis of a pixel distribution model it proposes adjustments. A fast maximum likelihood algorithm is deduced. Complications due to lack of independence of the pixel values and due to approximations are also addressed.

*To whom correspondence should be addressed.

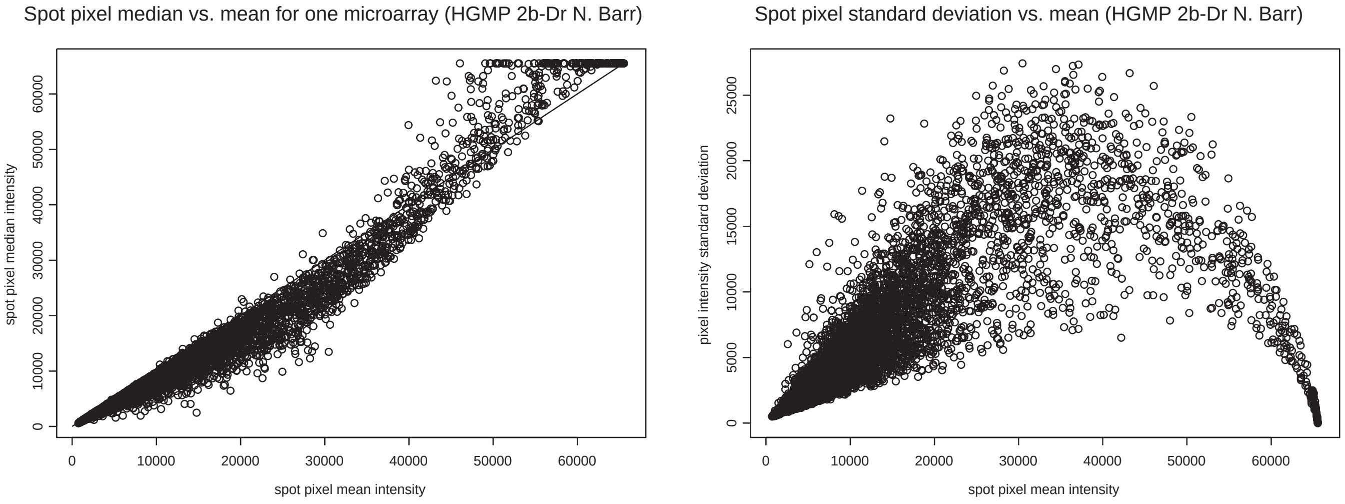

Fig. 1. Most scanner software give several spot pixel summaries. Plotting the median and the standard deviation versus the mean reveals that pixel values are censored at 65 535. (Data are provided by Dr N.Barr from the Beatson Institute, The University of Glasgow.)

2 METHODS

2.1 Maximum likelihood

Random variables are quantities subject to noise or uncertainty. Pixel values are the random variables considered in this paper. The expressions 'censored at c' and 'restricted on interval I' are used in the following ways. A random variable X is censored at c, if we observe $X^* = \min\{X, c\}$. If X is distributed according to distribution $\mathcal{D}(\alpha, \beta)$, then we indicate the censored distribution by $\mathcal{D}^*(\alpha, \beta)$. A 'random variable X is restricted on I' stands for the conditional distribution of X given $X \in I$. This conditional distribution is written as $\mathcal{D}(\alpha, \beta)|_I$. The distribution function F_X is defined as the non-exceedence probability of random variable X, i.e. $F_X(x) = P(X \leq x)$. The notation $p_X(x)$ is used for the probability density function, which loosely stands for the probability that random variable X takes on value x.

We assume that a parametric family with two parameters $\mathcal{D}(\alpha, \beta)$ is an appropriate description of the uncensored pixel intensity distribution. The original expression values for each spot are assumed independent draws from this distribution, i.e.

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{D}(\alpha, \beta),$$

where n is taken to be large, typically several hundred pixels. For convenience of notation we assume that the number of pixels is odd, $n = 2h + 1$. This assumption is not essential because n is large and can easily be replaced by $n - 1$ in case it is even. The observed pixel intensity is censored at a certain *maximum intensity*, MaxInt. The basis of the summary statistics are the observed pixel values,

$$X_i^* = \min\{X_i, \text{MaxInt}\}, \quad i = 1, \ldots, n.$$

We denote by m, $\bar{x}$ and s^2 the median, mean and variance respectively of the censored observations, i.e.

$$m = \text{Median}\{X_1^*, \ldots, X_n^*\}, \quad \bar{x} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i^*,$$
$$s^2 = \textstyle\sum_{i=1}^n (X_i^* - \bar{x})^2/(n-1).$$

Using the standard multiplication formula from probability theory (Grimmett and Welsh, 1988, p. 12), the likelihood of α, β given the observed spot mean, median and variance can be factorized as:

$$\begin{aligned} L(\alpha, \beta|\bar{x}, m, s^2) &= p(\bar{x}, m, s^2|\alpha, \beta) \\ &= p(m|\alpha, \beta)\, p(\bar{x}|m, \alpha, \beta)\, p(s^2|\bar{x}, m, \alpha, \beta) \qquad (1) \end{aligned}$$

2.1.1 Contribution of median to likelihood. The probability density of the median is a continuous distribution with an atom at MaxInt. A standard result (Grimmett and Welsh, 1988, p. 138) yields an expression for the continuous part, when $m <$ MaxInt,

$$p(m|\alpha, \beta) \propto F_X(m)^h\, p_X(m)\, (1 - F_X(m))^h\, .$$

At m =MaxInt the likelihood is a point mass,

$$p(m = \text{MaxInt}|\alpha, \beta) = \binom{n}{h+1} F_X(\text{MaxInt})^h\, (1 - F_X(\text{MaxInt}))^{h+1}\, .$$

2.1.2 Contribution of the mean to likelihood. For the conditional distribution of the mean given the median, we use several approximations. Notice that given $m <$

MaxInt, the mean $\bar{x}$ can approximately be written as the convolution of the mean of h draws from $\mathcal{D}(\alpha, \beta)$ restricted to $(0, m)$ and the mean of h draws from $\mathcal{D}(\alpha, \beta)$ restricted to (m, ∞) and censored at MaxInt. On the assumption that n is large (typically over 200 pixels), these two means can both be adequately approximated by a normal distribution. This assumption may be in some doubt for the second mean when m is close to MaxInt and a lot of the pixels are censored.

Given the median $m <$ MaxInt the density of the mean can approximately be written as convolution of the mean $\bar{X}_1$ of h pixels below m and the mean $\bar{X}_2$ of h pixels above m. We approximate $\bar{X}_1$ and $\bar{X}_2$ by $N(\mu_1, \sigma_1^2/h)$ and $N(\mu_2, \sigma_2^2/h)$ respectively, where the parameters are the means and variances of $\mathcal{D}^*(\alpha, \beta)$ restricted to $(0, m)$ and $(m, \text{MaxInt}]$ respectively. Although no closed form expressions are available for the coefficients, accurate numeric approximations are readily found due to the parametric structure of $\mathcal{D}(\alpha, \beta)$. Given these definitions we write for $\bar{x} <$ MaxInt,

$$\begin{aligned} p(\bar{x}|m, \alpha, \beta) &= n\ p\left(\sum_{i=1}^{n} X_i^* = n\bar{x}|m, \alpha, \beta\right) \\ &\approx n\ p(h\bar{X}_1 + m + h\bar{X}_2 = n\bar{x}|\alpha, \beta) \\ &\approx n\ \varphi\left(\frac{(n\bar{x} - m) - (h\mu_1 + h\mu_2)}{\sqrt{h\sigma_1^2 + h\sigma_2^2}}\right), \end{aligned}$$

where φ is the density of a standard normal.

A case that deserves some attention is when the median is equal to MaxInt, but the mean is not. In this case, at least half of the observations are known to be equal to MaxInt. To find the conditional density for the mean $\bar{x}$, we consider the mean of the censored distribution over the remaining h random variables,

$$\bar{X}_1 = \frac{1}{h}\sum_{i=1}^{h} Y_i^*, \quad Y_1, \ldots, Y_h \overset{\text{i.i.d.}}{\sim} \mathcal{D}^*(\alpha, \beta).$$

Again, using a normal $N(\mu_1, \sigma_1^2/h)$ approximation for $\bar{X}_1$, we can write for $\bar{x} <$ MaxInt

$$\begin{aligned} &p(\bar{x}|m = \text{MaxInt}, \alpha, \beta) \\ &= n\ p\left(\sum_{i=1}^{n} X_i^* = n\bar{x}|m = \text{MaxInt}, \alpha, \beta\right) \\ &\approx n\ p(h\bar{X}_1 + (h+1)m = n\bar{x}|\alpha, \beta) \\ &\approx n\ \varphi\left(\frac{(n\bar{x} - (h+1)m) - (h\mu_1)}{\sqrt{h\sigma_1^2}}\right), \end{aligned}$$

where φ is the density of a standard normal.

For the atom $\bar{x} =$ MaxInt, we don't have to resort to approximations. Notice that this can only occur in the case that the median is also equal to MaxInt and that therefore already half of the observations are known to be equal to MaxInt.

$$p_{\bar{x}}(\text{MaxInt}|m = \text{MaxInt}, \alpha, \beta) = (1 - F_X(\text{MaxInt}))^h$$

2.1.3 Contribution of variance to likelihood. Similarly for the conditional density of the variance given the mean and median, we shall employ a series of normal approximations. First we consider the case that both the mean and the median are less than the maximum intensity, i.e. $m, \bar{x} <$ MaxInt. The density of the variance can approximately be written as convolution of S_1 and S_2, where

$$S_1 = \sum_{i=1}^{h} Y_i^2,\ Y_1, \ldots, Y_h \overset{\text{i.i.d.}}{\sim} \mathcal{D}(\alpha, \beta)|_{(0,m)}$$

$$S_2 = \sum_{i=1}^{h} Z_i^{*2},\ Z_1^*, \ldots, Z_h^* \overset{\text{i.i.d.}}{\sim} \mathcal{D}^*(\alpha, \beta)|_{(m,\text{MaxInt}]}$$

We approximate S_1 and S_2 by $N(\mu_3, h\sigma_3^2)$ and $N(\mu_4, h\sigma_4^2)$ respectively, where the parameters are the means and variances of the distribution of the square of $\mathcal{D}^*(\alpha, \beta)$ restricted to $(0, m)$ and $(m, \text{MaxInt}]$ respectively. Just as before, accurate numeric approximations are readily found due to the parametric structure of $\mathcal{D}(\alpha, \beta)$. Given these definitions we write for $m, \bar{x} <$ MaxInt,

$$\begin{aligned} &p(s^2|\bar{x}, m, \alpha, \beta) \\ &\approx n\ p\left(\sum_{i=1}^{n} X_i^{*2} - n\bar{x}^2 = (n-1)s^2|m, \alpha, \beta\right) \\ &\approx n\ p(S_1 + m^2 + S_2 = (n-1)s^2 + n\bar{x}^2|\alpha, \beta) \\ &\approx n\ \varphi\left(\frac{((n-1)s^2 + n\bar{x}^2 - m^2) - (h\mu_3 + h\mu_4)}{\sqrt{h\sigma_3^2 + h\sigma_4^2}}\right) \end{aligned}$$

where φ is the density of a standard normal.

If the median is equal to MaxInt and the mean is not, then at least half of the observations are known to be equal to MaxInt. To find the conditional density for the variance s^2, the sum of the squares of the remaining h random variables is considered,

$$S_1 = \sum_{i=1}^{h} Y_i^{*2}, \quad Y_1, \ldots, Y_h \overset{\text{i.i.d.}}{\sim} \mathcal{D}^*(\alpha, \beta).$$

Using a normal approximation for S_1, we can write for $\bar{x} < \text{MaxInt}$

$$\begin{aligned}
&p(s^2|\bar{x}, m = \text{MaxInt}, \alpha, \beta)\\
&\approx np\left(\sum_{i=1}^{n} X_i^{*2} = (n-1)s^2 + n\bar{x}^2 | m{=}\text{MaxInt}, \alpha, \beta\right)\\
&\approx np(S_1 + (h+1)m^2 = (n-1)s^2 + n\bar{x}^2 | \alpha, \beta)\\
&\approx n\varphi\left(\frac{((n-1)s^2 + n\bar{x}^2 - (h+1)m^2) - (h\mu_3)}{\sqrt{h\sigma_3^2}}\right).
\end{aligned}$$

If both the mean and median are equal to MaxInt, then all pixel intensities are equal to MaxInt and therefore the variance is equal to zero with probability one.

2.1.4 Maximum likelihood procedure. Given the approximation of the likelihood $L(\alpha, \beta)$ in Equation (1), the function can now be maximized over the parameters α and β. The values

$$\hat{\alpha}, \hat{\beta} = \arg\max L(\alpha, \beta \mid m, \bar{x}, s^2)$$

are called the maximum likelihood estimates. They represent the 'most likely' values of α and β for the pixel distribution $\mathcal{D}(\alpha, \beta)$ for some spot, given the mean, median and variance of that spot. For each of the spots on the array a different set of parameters $(\hat{\alpha}_1, \hat{\beta}_1), \ldots, (\hat{\alpha}_{n_s}, \hat{\beta}_{n_s})$ is estimated from $(m_1, \bar{x}_1, s_1^2), \ldots, (m_{n_s}, \bar{x}_{n_s}, s_{n_s}^2)$. The median or mean of the distribution $\mathcal{D}(\hat{\alpha}_i, \hat{\beta}_i)$ is a better estimate of the true gene expression than the observed median or mean intensity of spot i.

2.2 Complications: approximations

The calculation of the likelihood involves a series of approximations. Each of these approximations affects the best possible estimate of the parameters of the distribution. Particularly when the median is close to the maximum intensity the approximation using the normal is highly suspect for reasons explained in the previous paragraph.

Similarly, we have assumed that the number of pixels per spot is large. Whereas this is true for typical spotted cDNA microarrays, it is not the case for Affymetrix chips. The U95 chips, for instance, have only some 25 pixels per probe. The question is whether this small amount of pixels justifies the normal approximation.

To evaluate the practical impact of these approximations, we simulate several 'spots' with different number of pixels (25, 201) and for different levels of censoring. Each of the pixels are sampled from a lognormal distribution. The results of this simulation study in Table 1 suggest that the approximations do not harm the estimates even when 50% of the data are censored and only 25 pixels are used. At higher levels of censoring, it seems that the approximation procedure slightly underestimates the mean and overestimates the variance of the pixel values.

Table 1. Simulation results for the mean of 10 replicates. The pixel data are generated according to a censored lognormal. The estimates are based only on the mean, median and variance of the pixel data

Lognormal		%	25 pixels		201 pixels	
μ	σ	Censored	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
7	2	2.0	7.1	1.9	7.0	2.0
9	2	14.8	8.8	1.7	9.1	1.9
10	2	29.3	9.6	2.2	10.0	1.8
11	0.2	27.3	11.0	0.4	11.0	0.2
11.1	0.2	52.6	11.1	0.4	11.1	0.4
11.3	0.2	91.9	11.2	0.4	11.2	0.3

2.3 Complications: lack of independence

In the calculation of the likelihood we assumed that the pixel values constituted independent draws from some intensity distribution. This assumption is generally made off-hand without much justification (Brown *et al.*, 2001, p. 8945). However, spatial effects such as print-pin effects invalidate independence and hybridization patterns even undermine conditional independence.

Nevertheless, independence is a very powerful assumption that makes calculations computationally tractable and fast. In the interest of calculability we propose to limit the independence assumption to non-neighbouring pixels of a certain separation. For this assumption to hold, the spatial hybridization effect is thought to be constant over the spot and the print-pin dependence is thought to act within the limits of the proposed pixel separation.

Given a k-pixel separation, each pixel value is replaced by the mean of the $(k+1)^2$ pixels neighbouring that pixel. We make the explicit assumption that the means over $(k+1)^2$ pixel values are 'approximately' independent. Theoretically, this approach is supported by the idea that for sufficient separation points in a two-dimensional first order Markov process are uncorrelated. Moreover, ergodicity results suggest that using the mean over more but correlated observations—rather than the independent ones—improves efficiency (Gamerman, 1997, Section 5.3.3.).

As a result, all the calculations in Section 2.1 are still valid with the only exception that the effective number of spot pixels n in the calculation of the likelihood is reduced by a factor $(k+1)^2$. If one pixel separation is sufficient for independence, then effectively the number of pixels in the normal calculations is divided by four; for two pixels of separation this number is nine.

2.4 Goodness-of-fit statistics

The methodology described allows us to fit any two parameter pixel intensity probability model to the data for each of the spots. It is of considerable interest to find out which of the models proposed seems the most adequate for pixel intensities of spotted microarrays. Goodness-of-fit statistics comparing the pixel probability models are therefore interesting. Because none of the pixel probability models are nested, typical likelihood based goodness-of-fit statistics are irrelevant. Instead we consider a variant of the chi-squared statistic.

This goodness-of-fit statistic is calculated for only the observed means $\bar{x}_i$ of the pixel values.

$$X^2 = \sum_{i=1}^{n_s} \frac{(\bar{x}_i - E_{\hat{\alpha}_i,\hat{\beta}_i}\bar{X})^2}{\mathrm{Var}_{\hat{\alpha}_i,\hat{\beta}_i}\bar{X}} \qquad (2)$$

Under the null-hypothesis that the uncensored pixel distribution is indeed $\mathcal{D}(\alpha, \beta)$, the mean of $\bar{X}$ is identical to the mean of a single pixel value generated by $\mathcal{D}^*(\alpha, \beta)$, whereas the variance is approximately equal to the variance of a single pixel value divided by the number of independent pixels in the spot. Therefore, under the null-hypothesis the statistic X^2 is approximately distributed like a chi-squared $\chi^2_{n_s}$ with approximately $3n_s - 2n_s$ degrees of freedom.

3 IMPLEMENTATION

In order to maximize the likelihood, several methods are available. Usual Newton–Raphson, gradient or other maximization procedures are infeasible due to the non-explicit nature of the parameters μ_i, σ_i^2 $(i = 1, \ldots, 4)$. We opt instead for maximizing the likelihood by an iterative grid-search procedure. Depending on the distribution $\mathcal{D}$, a grid of values for α and β is chosen over which the likelihood is evaluated. Around the maximum likelihood value in this grid a new, narrower grid is selected over which the likelihood is calculated. This procedure is repeated several times, until a numerically stable value of $(\hat{\alpha}, \hat{\beta})$ that maximize the likelihood is attained.

4 RESULTS

4.1 Corrected expressions

For the purposes of testing our methodology on a real set of microarray data, we use unpublished data from Dr N. Barr and her co-workers at the Beatson Institute at the University of Glasgow. The data consist of 4 cDNA spotted arrays, each with 9216 targets. Among the targets were 4224 known genes and ESTs and 384 control spots, each replicated twice on a single array. The probes were Cy3 and Cy5 labelled mRNA from a skin cancer cell-line (Bicr6) and a normal cell-line (Hec94). Dye swapping took place for 2 of the 4 arrays. We selected at random 1000 spots and calculated the maximum likelihood estimates of the parameters $(\alpha_1, \beta_1), \ldots, (\alpha_{1000}, \beta_{1000})$ for three different parametric distributions (Gamma, Lognormal, Weibull) for three different independence scenarios (full independence, independence with 1 pixel separation, independence with 2 pixel separation). Figure 2 shows that for both the Gamma and the Lognormal pixel model the adjusted pixel mean is larger than the unadjusted pixel mean for the higher pixel spot intensities.

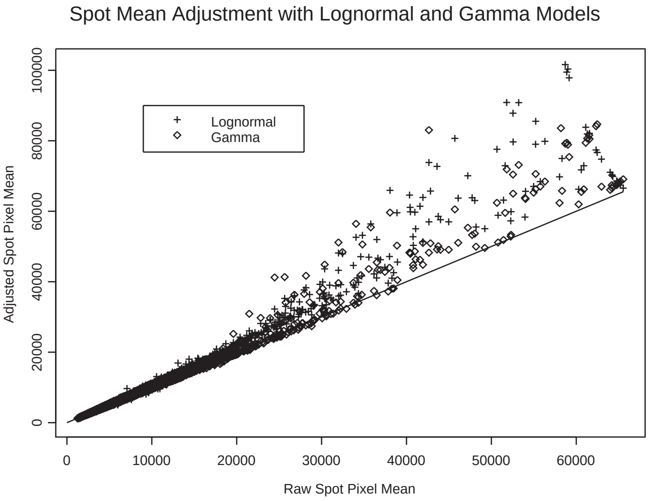

Fig. 2. Gamma and Lognormal full independent pixel model used to adjust the mean spot intensities of a single slide (Hgmp2b, Cy3, Dr N.Barr).

4.2 Impact on inference

The effect of signal censoring on inference can be addressed by considering the issue of differential expression, in particular by looking at its effect on the common expression ratio. We simulate pixel values from a lognormal distribution for the treatment and the reference channel so that before censoring the treatment is twice upregulated compared to the reference. This is done for several median pixel levels between 44 000 and 120 000 for the treatment sample and between 22 000 and 60 000 for the reference sample. Then the pixels are censored at 65 535 and median pixel levels are calculated for both reference and treatment. Figure 3 plots the reference sample on the x-axis and the expression ratios on the y-axis. If there was no censoring, the expression ratio should stay at 2. However, the raw expression ratio drops dramatically, whereas the corrected expression ratio using the method outlined in this paper performs much better.

4.3 Independence and distribution

In Table 2 the chi-square test shows that the hypotheses that the pixels are independent draws can be rejected. It suggests however, that the lognormal and Gamma model with one pixel separation is consistent with the data.

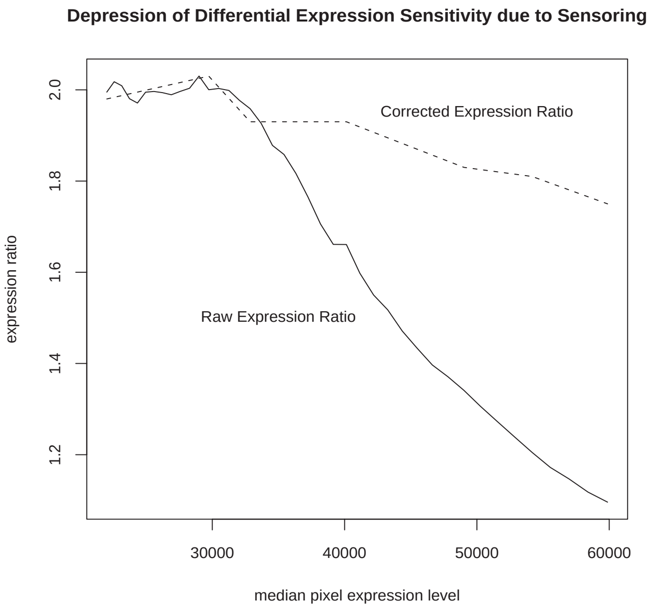

Fig. 3. Censoring affects corrected signal less.

Table 2. Chi-squared statistic for three parametric models and three different independence scenarios

$\mathcal{D}(\alpha, \beta)$	Pixel separation	Chi-squared Mean	p-value
Gamma	0	1856	0.000
	1	866	0.993
	2	881	0.978
Lognormal	0	1772	0.000
	1	875	0.985
	2	875	0.985
Weibull	0	3403	0.000
	1	1562	0.000
	2	1434	0.000

5 DISCUSSION

The maximum likelihood methodology described in this paper gives a quick way to adjust actual mean or median spot values for artificial pixel value censoring observed in microarray gene expression studies with spotted arrays. This method only uses the observed mean, median and standard deviation which are typically provided by the scanner output and to which the biologist has access. On top of that the paper provides a method to compare different two-parameter models for the pixel intensities.

However, three objections may be raised against this method. First of all, if the bioinformatician has access to the raw pixel values, much more accurate and analytically simple adjustments to the observed mean or median spot intensity can be proposed. Secondly, a much wider class of two-parameter distributions could have been considered. Finally, the estimate of the two parameters is based on only three observations per spot, i.e. the mean, the median and the variance. This may lead to unstable estimates.

We would like to counter each of these three arguments. Although better adjustments based on all the pixel values are possible, these data are generally unavailable to the general scientific community—typically because it challenges the scientist's disk space. Nevertheless, we do urge developers of scanner software to build-in options that perform adjustments based on all pixel values, analogous to the way we described, but possibly with more complicated pixel distributions. Secondly, although more complicated models *are* possible, it is encouraging that some of the two-parameter models that we considered, namely the Gamma and the Lognormal, are consistent with the data. Finally, although the two parameters of each spot pixel distribution were estimated based only on three observations, i.e. the mean, median and variance, these observations are highly reliable, because they themselves were calculated on the basis of hundreds of pixel values. Moreover, bioinformaticians normally use only one of the values (mean, median), which is subject to more variation than an estimate based on all three, as we propose.

ACKNOWLEDGEMENTS

We would like to thank Dr N.Barr for providing the data and the two anonymous reviewers for their constructive comments.

REFERENCES

Amaratunga,D. and Cabrera,J. (2001) Analysis of data from viral DNA microchips. *JASA*, **96**, 1161–1170.

Brown,C,S., Goodwin,P.C. and Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *PNAS*, **98**, 8944–8949.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.

Gamerman,D. (1997) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London.

Grimmett,G. and Welsh,D. (1988) *Probability, an Introduction*. Clarendon Press, Oxford.

Kerr,M.K., Kartin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Yang,Y.H., Buckley,M.J., Dudoit,S. and Speed,T.P. (2000) Comparison of methods for image analysis on cDNA microarrays. *Technical Report 584*. Department of Statistics, University of California at Berkeley.