

Microarray Background Correction: Maximum Likelihood Estimation for the Normal-Exponential Convolution

Jeremy D. Silver^{1, 2}, Matthew E. Ritchie³ & Gordon K. Smyth^{1, 4}

July 28, 2008

Abstract

Background correction is an important pre-processing step for microarray data which attempts to adjust the data for the ambient intensity surrounding each feature. The *normexp* method models the observed pixel intensities as the sum of two random variables, one normally distributed and the other exponentially distributed, representing background noise and signal respectively. Using a saddlepoint approximation, Ritchie *et al.* (2007) found *normexp* to be the best background correction method for two-color microarray data. This article develops the *normexp* method further by improving the estimation of the parameters. A complete mathematical development is given of the *normexp* model and the associated saddlepoint approximation. Some subtle numerical programming issues are solved which caused the original *normexp* method to fail occasionally when applied to unusual data sets. A practical and reliable algorithm is developed for exact maximum likelihood estimation (*MLE*) using high quality optimisation software and using the saddlepoint estimates as starting values. *MLE* is shown to outperform heuristic estimators pro-

¹Bioinformatics Division, Walter and Eliza Hall Institute, Parkville 3050, Australia

²Department of Biostatistics, University of Copenhagen, Copenhagen DK-1014, Denmark

³Department of Oncology, University of Cambridge, Cambridge CB2 0RE, United Kingdom

⁴Corresponding author. Phone: +613 9345 2326, fax: +613 9347 0852, email: smyth@wehi.edu.au

posed by other authors, both in terms of estimation accuracy and in terms of performance on real data. The saddlepoint approximation is an adequate replacement in most practical situations. The performance of *normexp* for assessing differential expression is improved by adding a small offset to the corrected intensities.

1 Introduction

Fluorescence intensities measured by microarrays are subject to a range of different sources of noise, both between and within arrays. Sources leading to within-array variation include non-specific hybridisation and debris left after the wash stage, while between-array variation includes differing amounts of dye, scanner settings and RNA quality. Background correction aims to adjust for these effects by taking account of ambient fluorescence in the neighbourhood of each microarray feature.

Ritchie *et al.* (2007) compared a range of background correction methods for two-colour microarrays. A method *normexp* was introduced which models the observed intensities as the sum of exponentially distributed signals and normally distributed background values. The corrected intensities are obtained as the conditional expectations of the signals given the observations. The *normexp* method is an adaptation of the background correction method proposed by Irizarry *et al.* (2003) for Affymetrix single-channel arrays, as the first step of the popular *RMA* algorithm for pre-processing Affymetrix expression data. Ritchie *et al.* (2007) showed that *normexp*, followed by a started-log transformation, gave the lowest false discovery rate of any commonly available background correction method for two-color microarrays.

The convolution model underlying the *normexp* method involves three unknown parameters, all of which must be estimated before the method can be applied. In the two-color context, the parameters must be estimated for each channel on each array, by fitting the convolution model to the observed intensities for that channel. Ritchie *et al.* (2007) suggested an approximate likelihood method for estimating the parameters, based on a saddlepoint approximation, but did not give mathematical details.

This article develops the *normexp* method further by improving the estimation of the parameters. Firstly, a complete mathematical development is given of the *normexp* model and the associated saddlepoint approximation. Secondly, some subtle numerical programming issues are solved which caused the original *normexp* method to fail oc-

casionally when applied to unusual data sets. Thirdly, we show that exact maximum likelihood estimation (*MLE*) of the parameters can be made practical and reliable using high quality optimisation software and using the saddlepoint estimates as starting values. Fourthly, we compare exact and approximate *MLE* with estimators proposed by other authors.

Maximum likelihood estimation has previously proved difficult because of numerical sensitivity of the likelihood function (Irizarry *et al.* 2003, Bolstad 2004, McGee & Chen 2006). Instead of *MLE*, the *RMA* algorithm, implemented in the **affy** software package for R (Gautier *et al.* 2004), uses simple heuristic estimators obtained by smoothing the histogram of observed intensities and partitioning the distribution about its mode (Bolstad 2004, Irizarry *et al.* 2003). McGee & Chen (2006) observed that the *RMA* estimators are highly biased, and proposed two new estimators. These methods are based on the *RMA* kernel smoothing approach, but partition the distribution about its mean (the *RMA-mean* method) or about 75th percentile (the *RMA-75* method) and then apply a one-step correction. The *RMA-mean* and *RMA-75* estimators are far less biased than those of *RMA* but apparently do not improve the performance of the *RMA* algorithm on real data (McGee & Chen 2006).

The saddlepoint approximation avoids the sensitivity of the likelihood function by providing a closed-form expression for the probability density on the log-scale, ensuring good relative accuracy. However the saddlepoint itself must first be found for each data value. This article provides a globally convergent iterative scheme which locates the saddlepoint to full accuracy in floating point arithmetic in all cases.

The accuracy of the different estimators are compared in a simulation study. The estimators are also compared using the extensive battery of calibration data sets assembled by Ritchie *et al.* (2007). This allows the estimators to be compared according to their ability to estimate fold changes and to detect differential expression on real data. As in Ritchie *et al.* (2007), the assumed context is that of a small microarray experiment in which popular differential expression methods are to be applied. *MLE* is shown to

have markedly better performance than the heuristic estimators.

Section 2 describes the *normexp* convolution model, presents the *MLE* and *saddle* procedures, and addresses some challenges in their implementation. Section 3 briefly describes the three test data sets with known levels of differential expression. Section 4 compares the four estimation schemes both by simulation and by performance on the test datasets.

2 Correction methods

2.1 The normal-exponential convolution model

Image analysis software for two-colour microarrays produces red foreground and background intensities R_f and R_b and green foreground and background intensities G_f and G_b for each spot on each array. Our aim is to appropriately adjust the foreground intensities R_f and G_f for the ambient intensities represented by R_b and G_b .

The *normexp* model for the red channel assumes $R_f = R_b + B + S$ where S is the true expression intensity signal and B is the residual background not captured by R_b . The model for the green channel is similar. The signal S is assumed exponentially distributed with mean α while B is normally distributed with mean μ and variance σ^2 . The parameters μ , σ^2 and α are assumed different for each channel on each array. All variables are assumed independent.

Write $X = R_f - R_b$ for the background-subtracted observed intensity. The *normexp* model becomes

$$X = B + S \tag{2.1}$$

The joint density of B and S is just the product of densities

$$f_{B,S}(b, s; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp(-s/\alpha) \phi(b; \mu, \sigma^2) \tag{2.2}$$

where $s > 0$ and $\phi()$ is the Gaussian density function. A simple transformation gives

the joint density of X and S as

$$f_{X,S}(x, s; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x - \mu}{\alpha}\right) \phi(s; \mu_{S.X}, \sigma^2)$$

where $\mu_{S.X} = x - \mu - \sigma^2/\alpha$. Integrating over s gives the marginal density of X

$$f_X(x; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x - \mu}{\alpha}\right) [1 - \Phi(0; \mu_{S.X}, \sigma^2)] \quad (2.3)$$

where $\Phi()$ is the Gaussian distribution function. Dividing the joint by the marginal gives the conditional density of S given X as

$$f_{S|X}(s|x; \mu, \sigma, \alpha) = \frac{\phi(s; \mu_{S.X}, \sigma^2)}{1 - \Phi(0; \mu_{S.X}, \sigma^2)}$$

for $s > 0$, which is a truncated Gaussian distribution. Our estimate of the signal given the observed intensities is the conditional expectation

$$\mathbb{E}(S|X = x) = \mu_{S.X} + \frac{\sigma^2 \phi(0; \mu_{S.X}, \sigma^2)}{1 - \Phi(0; \mu_{S.X}, \sigma^2)} \quad (2.4)$$

2.2 Saddle-point approximation

Maximum likelihood estimation requires the marginal density (2.3) which, although seemingly simple, turns out to be difficult to compute with full relative accuracy in floating point arithmetic, due to subtractive cancellation affecting both factors in the expression. As an alternative, the saddlepoint approximation, or tilted Edgeworth expansion, provides a means of approximating the density of any random variable from its cumulant generating function (Barndorff-Nielsen & Cox 1981, pp. 104). The approximation is attractive because it typically remains accurate far into the tails of the distribution.

The cumulant generating function of X is immediately available as the sum of those

for B and S ,

$$K_X(\theta) = K_B(\theta) + K_S(\theta) = \mu\theta + \sigma^2\theta^2/2 - \log(1 - \alpha\theta)$$

where $\theta < 1/\alpha$. The definition of the cumulant generating function implies that $g(x; \theta) = f_X(x) \exp[y\theta - K_X(\theta)]$ integrates to unity for all θ . Here we suppress the dependence of f_X on μ , σ and α for notational simplicity. The density $g(x; \theta)$ defines a linear exponential family with canonical parameter θ and with r th cumulant $\kappa_r = K_X^{(r)}(\theta)$.

The second-order Edgeworth expansion for g (Barndorff-Nielsen & Cox 1981, pp. 106) is

$$\log \tilde{g}(x; \theta) = -\frac{1}{2} \log(2\pi\kappa_2) + \frac{\kappa_4}{8\kappa_2^2} - \frac{5\kappa_3^2}{24\kappa_2^3}$$

yielding the approximation for f_X

$$\log \tilde{f}_X(x; \theta, \mu, \sigma, \alpha) = \log \tilde{g}(x; \theta) - y\theta + K_X(\theta).$$

The key feature which makes the saddlepoint approximation so effective is the ability to choose θ to make the Edgeworth expansion as accurate as possible for each x . This is done by choosing θ so that x is the mean of the distribution, i.e., θ is chosen to solve the saddlepoint equation

$$K'_X(\theta) = \mu + \sigma^2\theta + \frac{\alpha}{1 - \alpha\theta} = x \tag{2.5}$$

for $\theta < 1/\alpha$. Although this equation has a simple analytic solution, computing the solution is subject to catastrophic subtractive cancellation for certain values of σ and α . Details of how we avert this numerical issue are provided in Section A of the supplementary material (<http://www.biostatistics.oxfordjournals.org>).

2.3 Optimisation

Given a set of observed intensities x_i , $i = 1, \dots, n$, the unknown parameters μ , σ and α must be estimated before the correction formula (2.4) can be applied. Starting values are obtained as follows. The initial estimate $\hat{\mu}_0$ of μ is the 5% quantile of the x_i . The initial variance $\hat{\sigma}_0^2$ is the mean of $(x_i - \hat{\mu}_0)^2$ for $x_i < \hat{\mu}_0$. The initial $\hat{\alpha}_0 = \bar{x} - \hat{\mu}_0$.

Next the saddlepoint approximation to the likelihood is maximised using the Nelder-Mead simplex algorithm (Nelder & Mead 1965). Nelder-Mead is slower than gradient-based optimisation algorithms, but has the advantage of robust convergence from a very wide range of starting values.

Finally, using the saddlepoint estimates as starting values, the exact likelihood is maximised using the `nlminb` function of R, which performs unconstrained minimisation using PORT routines (Gay 1981, Gay 1983, Gay 1990). First and second derivatives of f_X with respect to μ , $\log \alpha$ and $\log \sigma^2$ are supplied. Optimising the likelihood with respect to $\log \alpha$ and $\log \sigma^2$, rather than α and σ^2 , avoids parameter constraints and improves convergence.

The algorithm is implemented in the `limma` software package for R (Smyth 2005). Saddlepoint parameter estimation takes about one second per channel with 20,000 probe arrays on a 2GHz Windows PC. Exact *MLE* takes about 50% longer. Time taken is roughly linear with the number of probes.

2.4 Transformation and offset

The *normexp* background correction (2.4) is performed for each channel on each array, yielding adjusted strictly positive red and green intensities R and G for each spot on each array. These are then converted to log-ratios, $M = \log_2(R/G)$, and log-averages, $A = \frac{1}{2} \log_2(RG)$ (Yang *et al.* 2001).

It also proves useful to offset the intensities by a small positive value k , giving offset log-ratios $M = \log_2[(R + k)/(G + k)]$. This simple transformation shifts the intensities

away from zero and serves to stabilise the variance of the log-ratios at low intensities (Rocke & Durbin 2003, Ritchie *et al.* 2007). The value $k = 50$ was chosen for this study on the basis of our previous experience with cDNA microarray experiments (Ritchie *et al.* 2007).

3 Test Data

3.1 Spike-in experiment

This article uses the same three calibration data sets as Ritchie *et al.* (2007). The first uses Lucidea Universal ScoreCard controls (Amersham Biosciences) to assess bias. Twelve copies of the LUS control probe set were printed in duplicate on 9 cDNA microarrays, along with a 13K clone library. Only the control probes are analyzed here. Prior to labelling, test and reference control RNA were spiked into RNA samples to produce known fold changes (Supplementary Table 1). All 8 background correction methods (*RMA*, *RMA-75*, *saddle* and *MLE*, with and without the offset) were applied. The resulting log-ratios were normalised and duplicate spots were combined to give an estimate of the \log_2 -fold change as described by Ritchie *et al.* (2007).

3.2 Mixture experiment

The second data set is from Holloway *et al.* (2006). Six RNA mixtures consisting of mRNA from MCF7 and Jurkat cell-lines in known relative concentrations (100%:0%, 94%:6%, 88%:12%, 76%:24%, 50%:50% and 0%:100%) were compared to pure Jurkat reference mRNA on 12 cDNA microarrays printed with a Human 10.5K clone set. Dye-swap pairs were performed for each of the 6 mixtures. All 8 background correction methods were applied and the data were normalised using print-tip loess (Yang *et al.* 2001). Probe-wise non-linear regression equations were fitted to the normalised log-ratios (Holloway *et al.* 2006). This produced for each probe a reliable consensus estimate of the MCF7 to Jurkat fold change and a standard deviation which estimates

the between-array measurement error.

3.3 Quality control study

The final data set is from Ritchie *et al.* (2006), and comprises 111 replicate arrays printed with the same 10.5k clone set as in the mixture study and hybridised with MCF7 (Cy3) and Jurkat mRNA (Cy5). Spot image data was morph background corrected and print-tip loess normalised. A large proportion of the genes are expected to be differentially expressed (DE) between the two samples, and this provided an independent source of truth for a false discovery rate comparison based on the Mixture data.

4 Results

4.1 Reliability

The estimation scheme outlined in Section 2.3 has proved to be extremely reliable. It has converged successfully for all datasets the authors have encountered so far, including thousands of simulated and real microarrays. This contrasts with earlier experiences reported by McGee & Chen (2006), who used Newton’s method to find maximum likelihood estimates numerically. They attempted to maximise the likelihood of samples with $\mu = 30$, $\sigma = 10$ and $\alpha = 100$. Their optimisation algorithm converged in only 15% of cases, even when initial estimates were equal to the true parameter values.

RMA estimation also returned useable values for all datasets. The *RMA-mean* and *RMA-75* methods each failed for some simulated data sets, the former slightly more often than the latter. Since the two are otherwise similar in performance, results will be presented here only for *RMA-75*. *RMA-75* returned NaNs for 32% of simulated data sets with $\sigma = 5$ and $\alpha = 10^4$ and for 0.3% of data sets with $\sigma = 20$ and $\alpha = 10^4$.

4.2 Estimation accuracy

Data was simulated for all combinations of $\mu \in \{30, 100, 500\}$, $\sigma \in \{5, 20, 100\}$ and $\alpha \in \{10^2, 10^3, 10^4\}$. These values represent a very wide range of scenarios in terms of the distribution of foreground values typically observed in microarray data. For each combination of parameter values, 1000 replicate samples of 20,000 observed intensities X were generated. Results are presented only for $\mu = 100$ as the other results are almost identical.

The *MLE* bias and standard deviation were the smallest followed closely by those of *saddle* (Tables 1, 2 and 3). *RMA-75* is much more biased and *RMA* is by far the worst. Parameter estimates for individual datasets for the representative parameter values $\sigma = 20$ and $\alpha = 1000$ are plotted in Figure 1. It is telling that the estimates from *RMA* were so biased that they fall outside the range of this plot. *MLE* is the most precise with almost no bias. *Saddle* is equally precise but with some bias, tending to underestimate σ . *RMA-75* and *RMA* on the other hand overestimate σ .

Another way to view accuracy is in terms of ability to return the correct signal values. Figure 2 shows the bias with which $E(S|X)$ estimates S on the \log_2 scale, for $\mu = 0$ $\sigma = 20$ and $\alpha = 1000$. Here *RMA-75* and especially *RMA* yield far more biased estimates of the signal than *MLE* or *saddle*, which are relatively accurate. Although *MLE* and *saddle* do tend to overestimate the true signal at lower intensities, this is indistinguishable from the bias that arises from inserting the true parameter values into $E(S|X)$.

4.3 Implicit offsets

The normalised M and A -values for one array from the Mixture experiment are shown in figure 3. This array has 100% Jurkat on both channels, so there is no true differential expression.

Some fanning of M -values is apparent at low A -values in the *MLE* and *saddle* panels.

This fanning is essentially eliminated in the corresponding offset panels, at the cost of compressing the range of A -values. Compared with MLE , $RMA-75$ and especially RMA show a somewhat compressed range of A and M -values even before the offset is applied. Our interpretation is that these estimation schemes incorporate offsets implicitly, which arise from the fact that they tend to overestimate the quantity $\mu_{S,X}$. Adding an offset to RMA is therefore in effect a double offset.

For this array, high offset and low M -value variability is desirable because the true M -values are zero. For arrays with genuine differential expression, compression of the M -values might appear as bias, and this is examined in Section 4.5.

4.4 Precision of expression values

We now examine the precision of the background corrected intensities, using results from the Mixture experiment. The residual standard deviation for each probe, $\hat{\sigma}_i$, is a measure of the precision with which the M -values returned by the microarrays follow the pattern of the mixing proportions. Figure 4 shows the trend in variability for each background method as a function of intensity. The vertical scale is \log_2 -variance, so each unit on the vertical axis corresponds to a 2-fold change in variance or a halving of statistical information.

As expected, precision improves with intensity for all the background correction methods prior to applying an offset. MLE and *saddle* have the best precision of the four methods for most of the intensity range. $RMA-75$ is relatively poor at higher intensities. After adding an offset, MLE and *saddle* have roughly constant variance across the intensity range, whereas the offset seems overdone for RMA and $RMA-75$, which now show a reversed trend in precision.

4.5 Bias of expression values

It is to be expected that higher precision, purchased by compressing the intensity range, will also result in attenuated signal. This is confirmed by examining the MCF7-Jurkat

log-fold changes estimated from the Mixture experiment. Supplementary figure 1 shows boxplots of the log-fold changes arising from each method. The spread of fold changes narrows when offsets are added, although the largest fold changes remain nearly as great.

To confirm whether attenuated fold changes can be interpreted as bias, we turn to the Spike experiment data. Supplementary figure 2 shows the M -values for a typical slide for the non-DE calibration controls and for the DE D03Med ratio controls, theoretically having 3-fold change down ($-\log_2 3 = -1.58$). All methods give log-ratios which are slightly biased towards zero, and the bias increases when offsets are added. There is surprisingly little difference between the four estimation algorithms, all leading to broadly similar bias.

4.6 Assessing differential expression

We now assess the ability of the background corrected expression values to correctly identify DE genes. Apart from the self-self hybridisations, the Mixture experiment consists of 5 dye-swap pairs of arrays. We assessed differential expression between MCF7 and Jurkat using each pair of arrays separately. The RNA mixtures vary from 100% to 50% MCF7, so the magnitude of the fold changes will vary from one pair of the arrays to another, but the set of DE genes should be the same in each case.

Using only two arrays to find DE genes presents a challenging problem, because there is only one degree of freedom available to estimate gene-wise standard deviations. The level of difficulty further increases with the concentration of Jurkat in the MCF7:Jurkat RNA mixture. The use of ordinary t -tests or other traditional univariate statistics to assess differential expression would be disastrous (Smyth 2004). Instead we use two of the most popular algorithms for microarray differential expression which have the characteristic of ‘borrowing’ information between genes. These algorithms have the ability to make statistical inferences with some confidence even for small numbers of replicate arrays. Genes were ranked in terms of evidence for differential

expression using SAM regularised t -statistics (Tusher *et al.* 2001) and using empirical Bayes moderated t -statistics (Smyth 2004). The statistics were calculated using the `samr` (<http://www-stat.stanford.edu/~tibs/SAM/>) and `limma` (Smyth 2005) software packages respectively.

To assess the success of the differential expression analyses, an independent determination of which genes are truly DE is required. The top 30% of genes, as ranked by moderated t -statistic (Smyth 2004), from the Quality Control study were selected as unambiguously DE and the bottom 40% as unambiguously non-DE. This gave 3098 DE and 4130 non-DE genes. The remaining 30% of genes were treated as indeterminate and are not used in the analysis.

Figure 5 shows the number of false discoveries for each method versus the number of genes selected by ranking the genes using absolute t -statistics, from largest to smallest for (a) `limma` and (b) SAM. The curves have been averaged over the 5 dye-swap pairs. The `limma` curves show that adding an offset reduces the false discovery rate, with the best performance achieved by *MLE* and *saddle*, followed by *RMA-75* and then *RMA*. For SAM, the advantage of *MLE+offset* and *saddle+offset* over the methods is even more marked. SAM appears to penalise methods which do not stabilise the variance.

5 Discussion

In this article, we have shown that exact *MLE* gives the most accurate estimation of the *normexp* parameters, and this accuracy translates into higher precision for the computed log-ratios of expression. The saddlepoint approximation is a very close competitor. The heuristic *normexp* estimators are markedly poorer in estimation accuracy. Furthermore, *RMA-mean* and *RMA-75* fails occasionally, and even frequently for some simulated scenarios. However *MLE* and *saddle* converged successfully in all of our tests.

The performance of *normexp* for assessing differential expression on real data is improved when combined with an offset, as a result of stabilising the variance as a

function of intensity. *MLE+offset* and *saddle+offset* gave the lowest false discovery rates. Although exact *MLE* does slightly better, the saddlepoint approximation could be considered an adequate replacement in most practical situations.

Estimation accuracy did not however directly translate to practical performance in all cases. *RMA* gives easily the most biased parameter estimates. Yet when we turned to the real-data examples, *RMA* yielded higher precision and fewer false-positives than *RMA-75*. Prior to offset, *RMA* is the best of all the methods when used with SAM significance analysis. This can be understood in terms of noise-bias trade-off. It appears that the biased *RMA* estimators have the fortuitous effect of introducing an implicit offset into the corrected intensities, and this has a variance stabilising effect. This partly explains why the *RMA* algorithm has been so successful for Affymetrix data. *RMA* also tends to return roughly similar parameter estimates regardless of the data, producing more consistent parameter estimates between arrays than the other methods. We speculate that this consistency may also help its performance on real data.

Since our study was completed, Ding *et al.* (2008) developed a *normexp*-type background correction method for Illumina microarray data. They proposed a Markov chain Monte Carlo (MCMC) simulation method to approximate the maximum likelihood parameter estimates. Their method is not directly applicable to non-Illumina data, because it requires Illumina negative controls. MCMC is far more computationally intensive than our Newton-Raphson *MLE*, and returns approximate estimates which vary stochastically from run to run.

Our algorithm is the first to reliably return exact maximum likelihood estimates for the *normexp* model. Achieving this required careful attention to a number of numerical analysis issues. In initial attempts, numerical issues including subtractive cancellation prevented us from computing the likelihood for some data sets. Several ingredients were required before reliable success was achieved including: 1) good initial estimates provided by the *saddle* procedure, 2) optimising with respect to $\log \alpha$ and $\log \sigma$ instead of α and σ (to enforce $\alpha > 0$ and $\sigma > 0$), and 3) optimising using both first and second

derivatives. Note that the Nelder-Mead algorithm was used first with the saddlepoint likelihood, then a pseudo-Newton-Raphson algorithm was used on the exact likelihood once a focused parameter range was established. The Nelder-Mead algorithm could not have been used directly on the exact likelihood because of the much wider range of parameter values under which the likelihood would need to be evaluated. Nor could Newton-Raphson have been applied to the saddlepoint approximation, because of the lack of good starting values.

Although we have focused exclusively here on two-color microarrays, our algorithmic development has obvious applications to other microarray platforms as well.

Funding

JDS was supported by an Allan Harris memorial scholarship. MER was supported by a grant from the Isaac Newton Trust. GKS was funded by National Health and Medical Research Council Program Grant 406657.

Acknowledgments

We are grateful to Terry Speed and Ben Bolstad for valuable discussions, and to an Associate Editor and an anonymous reviewer for useful suggestions.

References

- Barndorff-Nielsen, O. E. & Cox, D. R. (1981), *Asymptotic techniques for use in statistics*, Chapman & Hall, London, UK.
- Bolstad, B. M. (2004), Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization, PhD thesis, University of California, Berkeley.

- Cleveland, W. S. (1979), ‘Robust locally weighted regression and smoothing scatter-plots’, *Journal of the American Statistical Association* **74**, 829–836.
- Ding, L.-H., Xie, Y., Park, S., Xiao, G. & Story, M. D. (2008), ‘Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology’, *Nucleic Acids Research* **36**, e58.
- Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. (2004), ‘**affy** – analysis of Affymetrix GeneChip data at the probe level’, *Bioinformatics* **20**, 307–315.
- Gay, D. M. (1981), ‘Computing optimal locally constrained steps’, *SIAM Journal on Scientific and Statistical Computing* **2**, 186–197.
- Gay, D. M. (1983), ‘Algorithm 611 – subroutines for unconstrained minimization using a model/trust-region approach’, *ACM Trans. Math. Software* **9**, 503–524.
- Gay, D. M. (1990), Computing science technical report no. 153: Usage summary for selected optimization routines, Technical report, AT & T Bell Laboratories, Murray Hill, USA.
- Holloway, A. J., Oshlack, A., Diyagama, D. S., Bowtell, D. D. & Smyth, G. K. (2006), ‘Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis’, *BMC Bioinformatics* **7**, 511.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. (2003), ‘Exploration, normalization and summaries of high density oligonucleotide array probe level data’, *Biostatistics* **4**, 249–264.
- McGee, M. & Chen, Z. (2006), ‘Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data’, *Statistical applications in genetics and molecular biology*. Article 24.

- Nelder, J. A. & Mead, R. (1965), ‘A simplex algorithm for function minimization’, *Computer Journal* **7**, 308–313.
- Ritchie, M. E., Diyagama, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A. & Smyth, G. K. (2006), ‘Empirical array quality weights for microarray data’, *BMC Bioinformatics*. Article 261.
- Ritchie, M. E., Silver, J. D., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. & Smyth, G. K. (2007), ‘A comparison of background correction methods for two-colour microarrays’, *Bioinformatics* **23**, 2700–2707.
- Rocke, D. M. & Durbin, B. (2003), ‘Approximate variance-stabilizing transformations for gene-expression microarray data’, *Bioinformatics* **19**, 966–972.
- Smyth, G. K. (2004), ‘Linear models and empirical Bayes methods for assessing differential expression in microarray experiments’, *Statistical Applications in Genetetics and Molecular Biology*. Article 3.
- Smyth, G. K. (2005), Limma: linear models for microarray data, in R. Gentleman, V. Carey, S. Dudoit, R. Irizarry & W. Huber, eds, ‘Bioinformatics and Computational Biology Solutions using R and Bioconductor’, Springer, New York, pp. 397–420.
- Tusher, V. G., Tibshirani, R. & Chu, G. (2001), ‘Significance analysis of microarrays applied to the ionizing radiation response’, *PNAS* **98**, 5116–5121.
- Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. (2001), Normalization for cDNA microarray data, in M. L. Bittner, Y. Chen, A. N. Dorsel & E. R. Dougherty, eds, ‘Microarrays: Optical Technologies and Informatics’, Vol. 4266, Proceedings of SPIE.

List of Figures

1	Box-plots of parameter estimates for the three best-performing methods. The true values of the parameters are indicated by a dashed horizontal line. Estimates of <i>RMA</i> were so far from those of other methods that they do not appear when plotted on this scale (see tables 1, 2 and 3).	21
2	Smoothed \log_2 -ratio of the true to the estimated signal versus the true signal. The black line show this relationship if the true parameter values are used instead of estimates. The data used for this figure includes 100,000 observations simulated with $\mu = 0$, $\sigma = 20$ and $\alpha = 1000$. Quantiles for the signal distribution are marked. The curves were smoothed using LOWESS (Cleveland, 1979).	22
3	MA-plots obtained using different background correction methods for a self-self hybridisation from the Mixture experiment.	23
4	Smoothed $\hat{\sigma}^2$ from the nonlinear fits versus intensity for the Mixture experiment. The <i>A</i> -values have been standardised between methods, and plotted from the 5th to the 95th percentiles. The quantiles of the <i>A</i> -values are marked.	24
5	Number of false discoveries from the Mixture data set using moderated <i>t</i> -statistics from (a) limma and (b) SAM. Each curve is an average over the 5 mixtures.	25

List of Tables

- 1 Bias and standard deviation (shown in brackets) in estimating μ for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures. 26
- 2 Bias and standard deviation (shown in brackets) in estimating σ for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures. ∞^a and ∞^b indicate, respectively, where 32.4% and 0.3% of replicates yielded infinite estimates. 26
- 3 Bias and standard deviation (shown in brackets) in estimating α for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures. 27

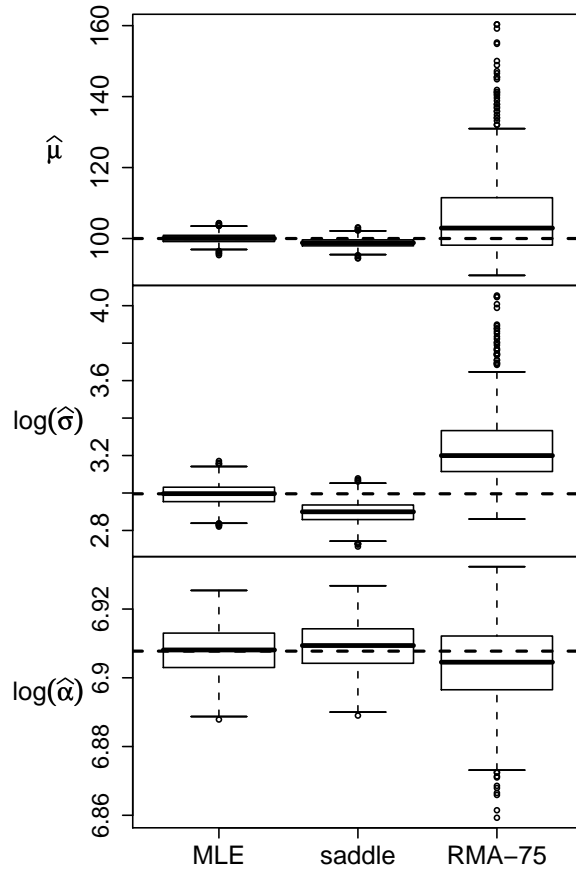


Figure 1: Box-plots of parameter estimates for the three best-performing methods. The true values of the parameters are indicated by a dashed horizontal line. Estimates of *RMA* were so far from those of other methods that they do not appear when plotted on this scale (see tables 1, 2 and 3).

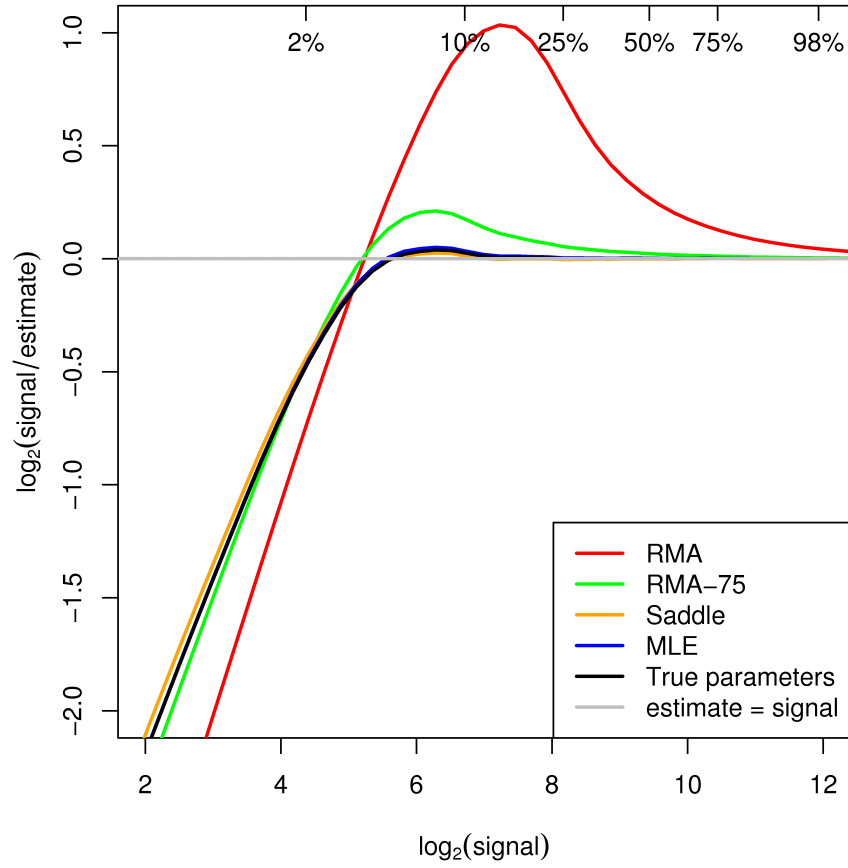


Figure 2: Smoothed \log_2 -ratio of the true to the estimated signal versus the true signal. The black line show this relationship if the true parameter values are used instead of estimates. The data used for this figure includes 100,000 observations simulated with $\mu = 0$, $\sigma = 20$ and $\alpha = 1000$. Quantiles for the signal distribution are marked. The curves were smoothed using LOWESS (Cleveland, 1979).

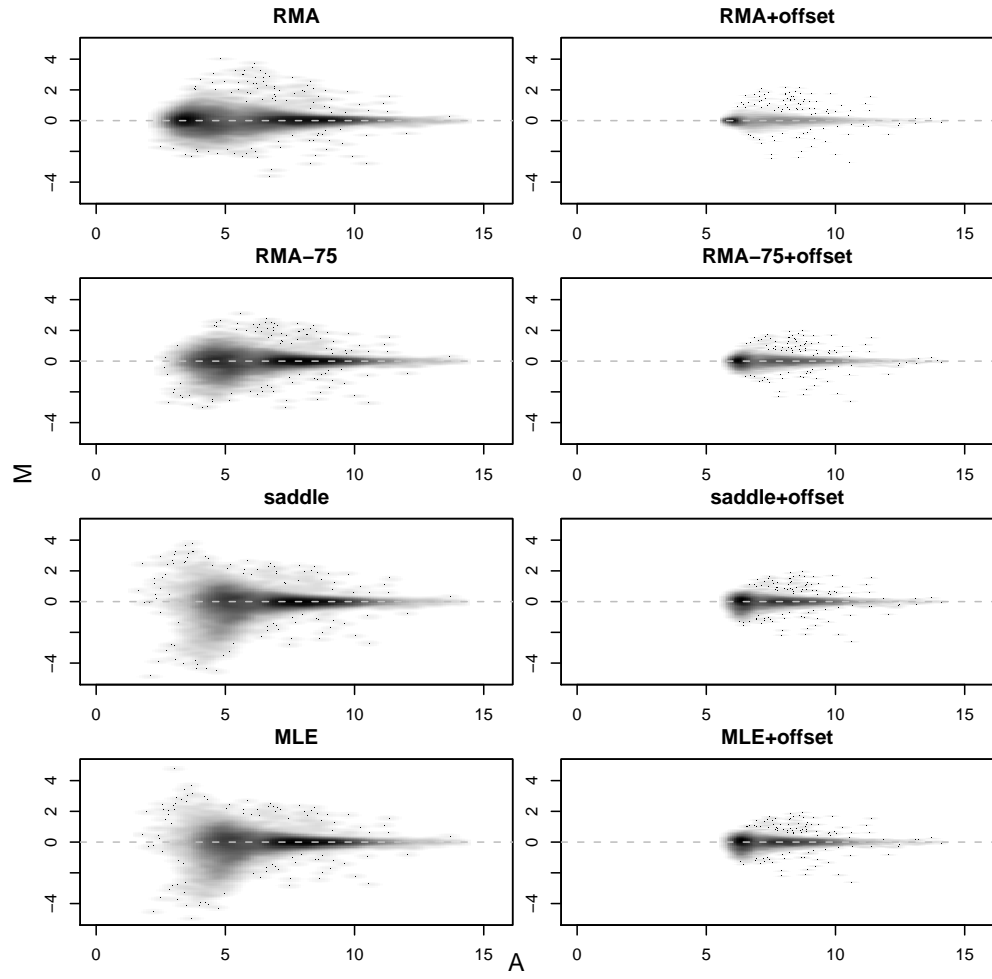


Figure 3: MA-plots obtained using different background correction methods for a self-self hybridisation from the Mixture experiment.

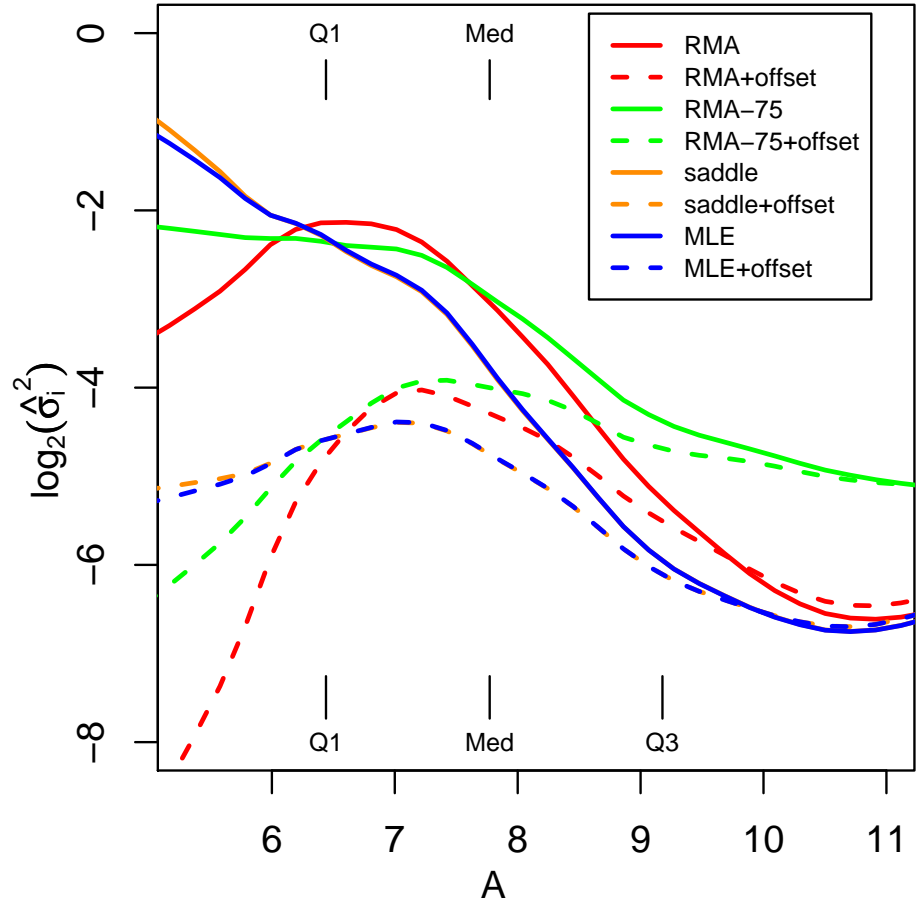


Figure 4: Smoothed $\hat{\sigma}^2$ from the nonlinear fits versus intensity for the Mixture experiment. The A -values have been standardised between methods, and plotted from the 5th to the 95th percentiles. The quantiles of the A -values are marked.

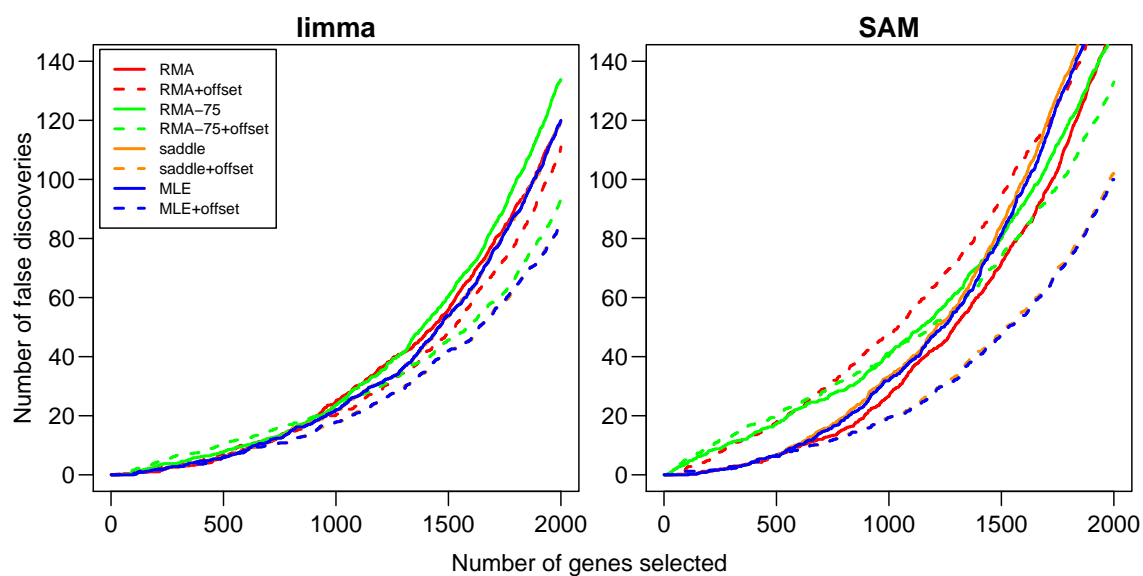


Figure 5: Number of false discoveries from the Mixture data set using moderated t -statistics from (a) limma and (b) SAM. Each curve is an average over the 5 mixtures.

σ	α	<i>MLE</i>		<i>saddle</i>		<i>RMA-75</i>		<i>RMA</i>	
5	10^2	0.0079	(0.22)	-0.25	(0.22)	1.7	(1.6)	12	(2.7)
20	10^2	0.0024	(0.47)	0.013	(0.50)	5.4	(2.3)	25	(2.6)
100	10^2	0.013	(1.6)	11.0	(1.5)	4.8	(11)	47	(9.0)
5	10^3	-0.023	(0.67)	-0.37	(0.65)	4.2	(7.5)	44	(23)
20	10^3	-0.025	(1.4)	-1.3	(1.4)	6.6	(11)	69	(24)
100	10^3	-0.098	(3.1)	-3.4	(3.1)	26.0	(18)	170	(24)
5	10^4	0.022	(2.3)	-0.36	(2.2)	32.0	(64)	380	(230)
20	10^4	0.20	(4.2)	-1.3	(4.0)	32.0	(66)	390	(220)
100	10^4	0.069	(9.2)	-6.5	(9.0)	41.0	(85)	520	(240)

Table 1: Bias and standard deviation (shown in brackets) in estimating μ for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures.

σ	α	<i>MLE</i>		<i>saddle</i>		<i>RMA-75</i>		<i>RMA</i>	
5	10^2	0.00059	(0.20)	-0.40	(0.19)	1.5	(0.71)	7.0	(1.9)
20	10^2	-0.0069	(0.40)	-0.46	(0.43)	5.6	(1.0)	15.0	(1.6)
100	10^2	0.003	(1.0)	7.3	(0.99)	25.0	(5.0)	45.0	(5.1)
5	10^3	-0.067	(0.62)	-0.56	(0.56)	3.2	(4.7)	32.0	(19)
20	10^3	-0.11	(1.2)	-1.9	(1.1)	6.0	(5.3)	44.0	(19)
100	10^3	-0.00048	(2.8)	-5.9	(2.7)	27.0	(7.8)	100.0	(16)
5	10^4	-0.72	(2.4)	-1.2	(2.1)	∞^a	∞^a	310.0	(190)
20	10^4	-0.40	(4.0)	-2.5	(3.6)	∞^b	∞^b	300.0	(180)
100	10^4	-0.52	(8.5)	-10.0	(7.8)	36.0	(46)	360.0	(190)

Table 2: Bias and standard deviation (shown in brackets) in estimating σ for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures. ∞^a and ∞^b indicate, respectively, where 32.4% and 0.3% of replicates yielded infinite estimates.

σ	α	<i>MLE</i>		<i>saddle</i>		<i>RMA-75</i>		<i>RMA</i>	
5	10^2	-0.00013	(0.75)	0.25	(0.75)	-1.1	(1.5)	-80	(0.31)
20	10^2	-0.013	(0.82)	-0.023	(0.84)	-2.5	(1.9)	-79	(0.40)
100	10^2	-0.046	(1.6)	-11.0	(1.5)	27.0	(8.1)	-69	(4.4)
5	10^3	0.021	(6.8)	0.37	(6.8)	-2.8	(10)	-800	(2.9)
20	10^3	0.11	(6.8)	1.4	(6.8)	-4.6	(12)	-800	(2.9)
100	10^3	-0.16	(7.5)	3.2	(7.5)	-15.0	(16)	-790	(3.2)
5	10^4	0.50	(72)	1.0	(72)	-28.0	(100)	-8000	(28)
20	10^4	-3.2	(69)	-1.6	(69)	-29.0	(100)	-8000	(28)
100	10^4	3.1	(71)	9.5	(71)	-23.0	(110)	-8000	(30)

Table 3: Bias and standard deviation (shown in brackets) in estimating α for the four estimation methods in 9 different scenarios. The true values of α and σ in each scenario are shown in the first two columns, and $\mu = 100$ for all scenarios. All values are given to two significant figures.