Variance models for microarray data

January 27, 2002

Mats Rudemo,[1,⋆] Tatsiana Lobovkina,[1] Petter Mostad,[1] Stefan J. Scheidl,[2] Sven Nilsson[2] and Per Lindahl[2]

[1]Department of Mathematical Statistics, Chalmers University of Technology and Göteborg University, SE – 412 96 Göteborg, Sweden

[2]Department of Medical Biochemistry, Göteborg University, Box 440, SE – 405 30 Göteborg, Sweden

⋆*email:* rudemo@math.chalmers.se

Summary

Variance estimation for estimated gene effects is crucial for identifying differentially expressed genes in micro-array experiments. Current methods are briefly reviewed and a new variance component model is suggested. One major component in this model corresponds closely to a Poisson model. A possible source of Poisson variation is spatial variability of fluorescently tagged nucleotides in spots on the microarray slide. For optimal weighting of global variance estimates and individual gene-based variance estimates an empirical Bayes procedure based on cross-validation is suggested. The suggested models and methods are found to perform well in the analysis of a dye-swap experiment.

Keywords: Microarrays, expression profiling, variance components, Poisson variation, fluorescently tagged nucleotides, empirical Bayes estimation.

1 Introduction

With microarrays the expression levels of thousands of genes are estimated simultaneously. In cDNA microarray analysis DNA fragments corresponding to different genes are typically arrayed on glass slides in spots with a diameter of the order 100 μm. Two pools of RNA of different origin are used to synthesize cDNA (complementary DNA) labeled with two specific fluorescent dyes. These are mixed together and allowed to hybridize with the DNA fragments on the glass slide. The intensity of fluorescent light is then measured across the slide surface at two wavelengths corresponding to the two dyes.

Regard an experiment with two treatments, typically a control and a specific treatment under study, and a corresponding cDNA experiment repeated for a number of slides. A crucial quantity for judging whether a gene is significantly differentially expressed or not is the associated variance estimate. Microarray experiments are subject to comparatively large experimental variation, and one critical source of variation may be the number of fluorescently tagged nucleotides in a spot on the microarray slide.

In the present paper we study and compare models for the error variance of the logarithm of estimated intensity ratios: (i) constant variance, (ii) one variance for each gene, (iii) a variance that decays exponentially with the expression level. Our conclusion is to recommend a weighted combination of these models. To find optimal weights we suggest an empirical Bayes procedure. The main results of the paper are illustrated in Figure 3 which gives the results of the empirical Bayes procedure in the example studied and in Figure 4 which shows the estimated variances after optimal weighting to

be compared with the individual gene variance estimates in Figure 2.

A number of papers have been devoted to the problem of choosing a suitable variance model for microarray data. The straightforward individual t-test for each gene gives problems, particularly for experiments with a small number of repetitions for which small variance estimates may lead to a number of corresponding spurious significant differential-expression levels.

Tusher, Tibshirani and Chu (2001) propose a model where a constant is added to the standard error in the denominator of the t-statistics. The method is further studied in Efron et al. (2001) where a Bayesian model with affected and unaffected genes is formulated for a specific $2\times2\times2$-experiment. The data allow computation of an empirical scores distribution and by re-arrangement also computation of an empirical null scores distribution. The resulting empirical Bayes procedure is used to find an optimal constant to be added to the estimated standard deviations.

Kerr et al. (2001) consider a model where the variance of the expression level of a gene depends on the intensity of the gene and the function describing this dependence is estimated nonparametrically. This intensity-depending model is compared with two other models where the variance is either gene-specific or homoscedastic. The conclusion is that the intensity-dependent variance model gives a better performance than the other two models.

Baldi and Long (2001) suggest a Bayesian model with conjugate priors both for the mean and the variance in a normal model for log-intensity expression ratios. The posterior distribution for the variance of a specific gene has then a mode which is a weighted average of a global experiment-specific variance

and the gene-specific variance. It is shown, see particularly Fig. 2 in Baldi and Long (2001), that the resulting regularized standard deviations decrease substantially with the expression level.

A Bayesian model with conjugate priors is also suggested in Lönnstedt and Speed (2001). The model gives an explicit formula for the log-posterior odds for a gene to be differentially expressed. In this formula a hyperparameter is added to the gene-specific variance. In a simulated data set the method is compared with three other methods: (i) the t-test with gene-specific variance estimates, (ii) the constant variance method and (iii) the method described above where a constant is added to the standard error in the t-statistic. The Bayesian method is found to be best in the comparison, but method (iii) turns out to be almost as good.

It may be noted that the two Bayesian methods mentioned above require some hyperparameter value to be supplied. Both Baldi and Long (2001) and Lönnstedt and Speed (2001) give recommendations for these choices.

Let us also mention two papers, where intensities are modeled directly without taking logarithms. A Bayesian model with gamma distributions for the intensities in cDNA microarray experiments is suggested in Newton et al. (2001). Conjugate distributions are assumed for parameters, and an empirical Bayes method is used to estimate parameters by maximizing a marginal log-likelihood. A frequentist approach is suggested by Ideker et al. (2000) in a model for the two intensities in cDNA microarrays. Maximum likelihood is used to compute estimates of variance and correlation parameters corresponding to two two-dimensional normal distributions for multiplicative and additive error components in the two channels.

In Section 2 we specify the models to be discussed in the present paper. We regard experiments with cDNA microarrays and two treatments. One model corresponds to exponential decay of the variance as a function of the logarithm of the harmonic mean intensity, which may be interpreted as Poisson variation at some stage in the generation of the intensity signals. The Poisson model, which is specified in Section 3, can be motivated by spatial variation of the fluorescently tagged nucleotides in spots, cf. Section 6.

Parameter estimation is discussed in Section 4. Our suggested method proceeds in five steps. First slide effects are estimated by least squares in a normalization procedure similar to the nonparametric normalization method with the lowess function suggested by Yang et al. (2001). As an alternative we use a parametric model with slide effects described by a spline function. Another modification we suggest is to measure the intensity level by the logarithm of the harmonic mean intensity rather than the arithmetic mean log-intensity. The main reason for using the harmonic mean intensity is the above mentioned Poisson variance model, but it also turns out to give some improvement in the normalization in the examples we have studied. In the second step gene effects are estimated from the normalized data. The estimation of slide and gene effects is in the third step followed by an estimation of the variance parameters using residuals from the second step. In the fourth step we use an empirical Bayes method to find optimal weighting of the gene-specific variance and a global variance (modified by expression level effects). In the final fifth step we provide variance estimates and confidence limits for the gene effects using the weighted variance and a t-distribution with degrees of freedom obtained from the preceding empirical Bayes method.

In Section 5 our methods are applied to a data set with mRNA obtained from two different cell types. Parameter estimates and results are given in Tables 1–2 and Figures 1–4. We also briefly describe corresponding results for a dataset available on the Internet and analysed in Kerr et al. (2001).

In the final Section 6 we draw some general conclusions from our analyses and compare our suggested methods with methods from the literature. We describe briefly how our models and methods can be extended to other types of microarrays experiments.

2 Models for microarray data with two treatments

Suppose that we have microarray data from S slides, denoted $s = 1, \ldots, S$. For each slide there are two treatments $t = 1, 2$. Let

$$Z_{gts}, \quad g = 1, \ldots, G, \; t = 1, 2, \; s = 1, \ldots, S, \tag{1}$$

denote the observed intensity value for the spot corresponding to gene g and treatment t in slide s, where G is the number of genes. We assume here that each spot corresponds to one gene.

Let Y_{gs} denote the observed relative effect of treatment 1 compared to treatment 2 on a log-scale

$$Y_{gs} = \log \frac{Z_{g1s}}{Z_{g2s}}, \tag{2}$$

where log denotes natural logarithms, and let

$$x_{gs} = -\log(\frac{1}{2}(\frac{1}{Z_{g1s}} + \frac{1}{Z_{g2s}})) \tag{3}$$

be the log-harmonic mean intensity. We regard the model

$$Y_{gs} = \mu_g + v_s(x_{gs}) + \epsilon_{gs}, \tag{4}$$

where μ_g is the effect of gene g, $v_s(x_{gs})$ is the effect of slide s for gene g and ϵ_{gs} are independent with $\mathbf{E}(\epsilon_{gs}) = 0$ and $\text{var}(\epsilon_{gs}) = \sigma^2_{gs}$. Here μ_g measures on a log-scale the relative effect of treatment $t = 1$ compared to treatment $t = 2$ for gene g. The function v_s gives a slide effect that for slide s and gene g depends on the mean intensity x_{gs}. We assume that

$$v_s(x) = \sum_{j=0}^{J} \beta_{js} f_{js}(x) \tag{5}$$

with known functions f_{js}, $j = 0, \ldots, J$, $s = 1, \ldots, S$.

A flexible class of functions is the set of cubic spline functions,

$$v_s(x) = \sum_{j=0}^{3} \beta_{js} x^j + \sum_{j=4}^{J} \beta_{js} (x - k_{js})^3_+ \tag{6}$$

where $(x)_+$ denotes the positive part of x, and k_{js}, $j = 4, \ldots, J$, denote inner knots for the spline function v_s, cf. Eubank (1984). Here $J - 3 \geq 0$ is the number of inner knots.

Instead of the log-harmonic mean intensity (3) the traditional choice, cf. Dudoit et al. (2000), is to use arithmetic mean log-intensity

$$\tilde{x}_{gs} = \frac{1}{2}(\log Z_{g1s} + \log Z_{g2s}) \tag{7}$$

as covariate in normalization. In the data analysis we will compare the use of the covariates x_{gs} and $\tilde{x}_{gs}$.

The variance structure could be modeled in several ways. Simple variance models are one variance for each gene, $\sigma^2_{gs} = \sigma^2_g$, and one global variance,

$$\sigma^2_{gs} = \sigma^2. \tag{8}$$

We also consider models with $\sigma_{gs}^2 = h(z_{gs}, \alpha)$, where h is a known function, z_{gs} is a covariate (vector) and α is a parameter vector to be estimated from our data. In particular, we will regard the models

$$\sigma_{gs}^2 = \sigma^2 \exp(-\alpha_1 x_{gs}), \tag{9}$$

$$\sigma_{gs}^2 = \sigma^2 \exp(-\alpha_1 x_{gs}) + \alpha_2, \tag{10}$$

and the model where x_{gs} in (10) is replaced by $\tilde{x}_{gs}$ as covariate,

$$\sigma_{gs}^2 = \sigma^2 \exp(-\alpha_1 \tilde{x}_{gs}) + \alpha_2. \tag{11}$$

3 A Poisson model for the log-harmonic mean

The model (9) for the variance contains as a special case a simple Poisson model. In this model we assume that $Z_{gts} = cN_{gts}$, where c is a constant and N_{gts} is supposed to be Poisson distributed with expectation λ_{gts}. For a Poisson distributed variable N with a large expectation λ we have $\mathrm{var}(\log N) \approx \lambda^{-1} \approx N^{-1}$, which implies $\mathrm{var}(\log Z_{gts}) \approx cZ_{gts}^{-1}$. Assuming that Z_{g1s} and Z_{g2s} are independent we find

$$\mathrm{var}(Y_{gs}) \approx c\left(\frac{1}{Z_{g1s}} + \frac{1}{Z_{g2s}}\right) = 2c\exp(-x_{gs}), \tag{12}$$

which corresponds to (9) with $\alpha_1 = 1$ and $\sigma^2 = 2c$. In the examples studied in Section 5 it will turn out that a variance component model which includes (10) gives a good description of the observed variance estimates. One possible source of Poisson variability is the spatial distribution of the number of fluorescently tagged nucleotides, cf. Section 6.1 below.

4 Parameter estimation

To estimate parameters we will use a five-step procedure. First we will use least squares to normalize our data by subtracting slide effects. Then we estimate the gene effects μ_g. From the residuals of this analysis we estimate in the third step parameters in the variance models. The fourth step is an empirical Bayes method for weighting global variance with gene-specific variances. In the fifth step we obtain confidence intervals for the gene effects.

4.1 Step 1: Normalization

The parameter vector β of the cubic spline functions according to (5)–(6) is estimated by least squares by minimizing

$$Q(\beta) = \sum_{gs} \left(Y_{gs} - \sum_{j=0}^{J} \beta_{js} f_{js}(x_{gs}) \right)^2 . \tag{13}$$

To choose the number $J-3$ of (inner) spline knots for the spline function (13) we may compare the corresponding sum of squares. For nested spline functions we may also use standard analysis-of-variance F-tests. To choose spline knots for a given number of knot points we suggest to use knot points k_{js} chosen individually for each slide such that each resulting subinterval of $(\min_g x_{gs}, \max_g x_{gs})$ contain the same number of x_{gs}-points.

4.2 Step 2: Estimation of gene effects

Let $\hat{\beta}$ denote the estimated parameter vector from Step 1 and let $\hat{v}_s$ denote the corresponding estimate in (5). To estimate gene effects we put

$$\hat{\mu}_g = (1/S) \sum_{s=1}^{S} (Y_{gs} - \hat{v}_s(x_{gs})). \tag{14}$$

A corresponding estimate of the global constant variance σ^2 is given by

$$\hat{\sigma}^2 = (1/\mathrm{df}) \sum_{gs} (Y_{gs} - \hat{\mu}_g - \hat{v}_{gs}(x_{gs}))^2 \tag{15}$$

with $\mathrm{df} = GS - G - S - JS + 1$.

4.3 Step 3: Estimation of variance parameters

Let $\hat{\epsilon}_{gs} = Y_{gs} - \hat{\mu}_g - \hat{v}_{gs}(x_{gs})$ denote the residuals from Steps 1 and 2 above. For different genes these residuals are approximately independent, but within genes they are strongly dependent as $\sum_s \hat{\epsilon}_{gs} = 0$ by (14). In the third step we use these residuals to estimate a gene-specific variance and the parameters σ^2 and α in (9), (10) and (11). The gene-specific variance estimate is

$$\hat{\sigma}_g^2 = \frac{1}{S-1} \sum_s (\hat{\epsilon}_{gs})^2, \tag{16}$$

which apart from a multiplicative constant is essentially chi-square-distributed with $S - 1$ degrees of freedom.

The most satisfactory method to estimate the parameters in (9), (10) and (11) is presumably the nonlinear restricted maximum likelihood method (REML), cf. Pinheiro and Bates (2000). Here we will use a closely related but simpler method based on a chi-square distribution for the sum of squares in (16). Regard estimation of parameters in (9). Put

$$\bar{x}_g = (1/S) \sum_s x_{gs}, \tag{17}$$

$\chi_g^2 = \sigma^{-2} \exp(\alpha_1 \bar{x}_g) \sum_s (\hat{\epsilon})_{gs}^2$, and let $f_{\chi^2}(\cdot, r)$ denote the probability density of a chi-squared variable with r degrees of freedom. To estimate σ^2 and

α_1 we maximize the log-likelihood corresponding to $\sum_s(\hat{\epsilon}_{gs})^2$, $g = 1, \ldots, G$, regarded as observations, that is

$$L_{\chi^2}(\sigma^2, \alpha) = \sum_g \log\{\sigma^{-2}\exp(\alpha_1\bar{x}_g) f_{\chi^2}(\chi^2_g, S-1)\}. \tag{18}$$

With small modifications the same method may used to estimate the parameters in (10) and (11). In fact, it may also be used for model (8) producing an estimate close to (15) but then also providing a corresponding maximum of the log-likelihood function, which may be useful for testing if we get a significant improvement by adding more parameters.

4.4 Step 4: Empirical Bayes estimation of weights

Let us now find a suitable weighting of the gene-specific variance estimate (16) and the expression-level modified global variance estimate from (9) or (10) but with x_{gs} replaced by $\bar{x}_g$, that is,

$$\hat{\sigma}^2_{\text{glob}}(\bar{x}_g) = \hat{\sigma}^2\exp(-\hat{\alpha}_1\bar{x}_g), \tag{19}$$

and

$$\hat{\sigma}^2_{\text{glob}}(\bar{x}_g) = \hat{\sigma}^2\exp(-\hat{\alpha}_1\bar{x}_g) + \hat{\alpha}_2. \tag{20}$$

Recall that (16) is associated with $S-1$ degrees of freedom. Regard the variance estimate (19) or (20) as an additional independent estimate of the variance for gene g with r degrees of freedom. The weighted variance estimate

$$\tilde{\sigma}^2_g = \frac{(S-1)\hat{\sigma}^2_g + r\hat{\sigma}^2_{\text{glob}}(\bar{x}_g)}{S-1+r} \tag{21}$$

is an improved variance estimate associated with $S-1+r$ degrees of freedom.

In a Bayesian model with a random sample from a normal distribution and appropriate conjugate distributions, cf. DeGroot (1970), p 169-170, and Lönnstedt and Speed (2001), the posterior marginal distribution of the mean is a t-distribution. Assuming an uninformative prior for the mean we get a weighted variance as in (21) with $S-1+r$ degrees of freedom in the t-distribution. Minor modifications of the number of degrees of freedom and coefficients in the weighting can be made, cf. Baldi and Long (2001), and our choice here is motivated by the formulas for weighting variances and computing degrees of freedom in ordinary analysis of variance.

To estimate r we will use a cross-validation technique with a predictive t-distribution with $S-2+r$ degrees of freedom. Put $\tilde{Y}_{gs} = Y_{gs} - v_{gs}(\hat{\beta})$, $\tilde{Y}_{g,-s} = (S-1)^{-1}\sum_{s' \neq s} \tilde{Y}_{gs'}$,

$$\hat{\sigma}^2_{g,-s} = \frac{1}{S-2}\sum_{s' \neq s}(\tilde{Y}_{gs'} - \tilde{Y}_{g,-s})^2,$$

and

$$\tilde{\sigma}^2_{g,-s} = \frac{(S-2)\hat{\sigma}^2_{g,-s} + r\hat{\sigma}^2_{\text{glob}}(\bar{x}_g)}{S-2+r}. \tag{22}$$

Note that $\tilde{Y}_{gs}$, $s = 1, \ldots, S$, are approximately independent with variance σ_g^2 and thus $\tilde{Y}_{gs} - \tilde{Y}_{g,-s}$ has approximately variance $(S/(S-1))\sigma_g^2$ which we estimate by $(S/(S-1))\tilde{\sigma}^2_{g,-s}$. Put

$$t_{gs} = (\tilde{Y}_{gs} - \tilde{Y}_{g,-s})((S/(S-1))\tilde{\sigma}^2_{g,-s})^{-1/2} \tag{23}$$

and let $f_t(\cdot, r)$ denote the probability density of a t-distribution with r degrees of freedom. To estimate r we regard the sum of log-likelihoods obtained by considering $\tilde{Y}_{gs} - \tilde{Y}_{g,-s}$, $g = 1, \ldots, G$, $s = 1, \ldots, S$, as observations

$$L_{\text{student}}(r) = \sum_{gs} \log\{((S/(S-1))\tilde{\sigma}^2_{g,-s})^{-1/2}\ f(t_{gs}, S-2+r)\} \tag{24}$$

and maximize it as a function of r. Note that the variables t_{gs} in (23) are dependent. Thus (24) is not a proper log-likelihood function, but rather the logarithm of a pseudolikelihood function well-known from spatial statistics, cf. Hjort and Omre (1994), and there known to have good properties. Note also that when we use the constant global variance model (8) we can estimate the corresponding optimal r by replacing $\hat{\sigma}^2_{\text{glob}}(\bar{x}_g)$ in (21) and (22) with the global variance estimate that now does not depend on $\bar{x}_g$.

4.5 Step 5: Weighted variances and confidence intervals for gene effects

As a fifth and final step in the parameter estimation we compute the weighted variances and the corresponding confidence intervals for the gene effects. Let $\tilde{\sigma}^2_g$ be the estimated variance given by (21) with r estimated by maximization of (24). Then we get confidence intervals for the gene effects by regarding

$$t_g = (\hat{\mu}_g - \mu_g)/\tilde{\sigma}_g \tag{25}$$

as t-distributed with $S - 1 + r$ degrees of freedom. For possibilities of improving the $\hat{\mu}_g$-estimates, see Section 6.5.

5 Data analysis

5.1 Analysis of data from a dye-swap experiment with four slides

The methods described in the previous section were applied to data from an experiment with two treatments corresponding to mRNA obtained from

different types of cells, treatment 1 called "heart" with cells from dissected mouse hearts and treatment 2 from a reference of cultured cells, cf. Fig. 3 in Scheidl et al. (2001) where cells from mouse heart are compared with cells from mouse kidney. There were four slides in the experiment: 249, 286, 287 and 346, and 2208 genes on all slides.

Let (1) denote the observed intensities, more precisely, the observed mean signals for spots corrected for background by subtracting local mean background levels in the program ImaGene® (BioDiscovery Inc, CA). Model (4) was used to analyse the data with cubic spline functions (6).

A preliminary analysis with two inner knots for the cubic spline functions and with individual estimates of standard deviations for the four slides showed that five genes had one residual (out of four) that numerically exceeded four times the maximal of these four estimated standard deviations. For all these five genes there was one numerically large residual, while the other three residuals had the opposite sign and were quite close to each other, indicating that the numerically large residual corresponds to an outlier. For the main part of the present study it was found suitable to exclude these five genes and also one additional gene with a very small observed intensity for one of the slides. The reduced data set thus contained $G = 2208 - 6 = 2202$ genes. The analysis was also carried through for all the genes with not too different results, which we comment briefly upon in Section 6.4.

Figure 1 shows the data for the four slides together with estimated spline functions with two inner knots chosen as described in Section 4.1. An F-test of the hypothesis of $J-3 = 2$ inner knots with the nested model with $J-3 = 5$ inner knots as alternative, showed a barely significant improvement for five

inner knots ($p \approx 3\%$). The decrease in the residual variance estimate (15) for five inner knots was small, and we preferred to keep the model with two inner knots. On the other hand, the improvement with two inner knots compared to no inner knot was highly significant ($p \approx 0.0001\%$). The estimated spline functions are shown in Figure 1.

The variance parameter estimates computed according to Section 4.2 for the models (8), (9), (10) and (11) are given in Table 1 together with the corresponding maximum log-likelihoods. The first three models are nested and standard asymptotic theory implies that twice the differences in log-likelihood should be approximately chi-squared distributed with one degree of freedom if one additional parameter did not have any effect. Thus we see from the log-likelihood values that the improvement in going from model (8) to model (9) is overwhelmingly significant as is the improvement in going from model (9) to model (10). For the models (8), (9) and (10) the corresponding curves are shown in Figure 2 together with the gene-specific variance estimates (16). Although the parameter estimates for the models look very different in Table 1 the curves in Figure 2 are rather close.

For the choice of the optimal degrees of freedom r in the estimate (21) we used the empirical Bayes method as described in Section 4.3 with the log-pseudolikelihood (24). The results are given in Table 2 and the log-pseudolikelihood functions are shown in Figure 3. For the optimal weighting according to model (10) the corresponding weighted variances are shown in Figure 4, which should be compared with Figure 2.

Standard asymptotic theory is not directly applicable to (24), which as mentioned above is strictly the logarithm of a pseudolikelihood. However,

comparing the log-pseudolikelihoods for weights $w = 0$ and $w = 1$, corresponding to $r = 0$ and $r = \infty$, with the values for $r = \hat{r}$, the large differences strongly indicate that the value of the optimally weighted variance gives an overwhelmingly significant improvement.

5.2 Analysis of an experiment with data available from the Internet

The methods above have also been applied to a data set, called NIEHS, with 1907 genes and 6 slides described in Kerr et al. (2001). The results were broadly similar to those described in Tables 1–2 and Figures 1–4. We found for the NIEHS dataset the same ranking of the four models as given in both Table 1 and Table 2 and in Figure 3. However, while Figure 3 shows optimal weights close to one, the optimal weights were for the NIEHS dataset close to one half or in one case close to zero. For the models (8), (9), (10) and (11) the estimated optimal weights, cf. Table 2, were 0.14, 0.45, 0.50, 0.49, respectively, with corresponding estimated degrees of freedom 0.7, 3.3, 4.0 and 3.8, respectively.

6 Discussion

6.1 The Poisson model

The Poisson model in Section 3 can be motivated by sampling variation either in time or in space. Let us first consider spatial variation of fluorescently tagged nucleotides. The rough estimate in Duggan et al. (1999) gives 12 fluor tags in a 100 μm^2 scanned pixel from a probe spot. In addition, as shown in

Brown et al. (2001), there is granularity within spots. Thus spatial sampling variation should be considerable and a Poisson variability as discussed in Section 3 above gives one model catching this type of variation.

Consider now variation in time. In the experiment described in Section 5.1 the scanner spends 15 minutes for each 25×75 mm^2 slide which gives roughly $8 \cdot 10^4$ ns per pixel of size 160μm^2. The fluorescence lifetime of the cyanine fluorophores Cy3 and Cy5 used in the experiment is of the order 1 ns, cf. Widengren and Schwille (2000). Even if only a fraction of the emitted photons enter the microscope, and the fluor tags after emission spend some time before becoming re-excited, the scanning time seems large enough to exclude substantial sampling variation for the logarithms of the spot intensities (1). Thus our conclusion is that one simple possible explanation of the observed, roughly Poisson-type, variability is spatial sampling variation.

Let us also give an argument for the approximate variance in (12) by conditioning in a binomial distribution. Suppose that there in a spot are N tags that will give a signal contribution and that N_1 of them give a contribution to the treatment signal $Z_1 \approx cN_1$ and $N_2 = N - N_1$ of them give a contribution to $Z_2 \approx cN_2$. We condition upon $N = n$ and assume that N_1 is Binomial(n, p) given $N = n$ with a proportion p corresponding to the gene tested at the regarded spot. Then

$$\begin{aligned} \mathrm{var}(Y_{gs}) &\approx \mathrm{var}(\log \frac{N_1}{n - N_1}) \approx \mathrm{var}(N_1) \left(\frac{\partial}{\partial x} \log \frac{x}{n-x} \Big|_{x=np} \right)^2 \\ &= \frac{1}{np} + \frac{1}{n(1-p)} \approx \frac{1}{N_1} + \frac{1}{N_2} \end{aligned}$$

The assumption that N_1 is Binomial(n, p) given that $N = n$ seems relevant also without the assumption of a Poisson distribution, and we should then

get approximately the same variance as in (12). Thus we would expect that this approximate formula is robust with respect to the assumption of Poisson variability and also with respect to the assumption in Section 3 of independence of N_1 and N_2.

It is conspicuous that the estimates of the parameter α_1 in Table 1 for models (10) and (11) are rather close to the theoretical value $\alpha_1 = 1$ from Section 3. For the NIEHS data set we similarly got the estimates 0.93 and 0.95 for α_1 with the models (10) and (11).

The estimate $\hat{\sigma}^2 = 7.0$ in Table 1 may be regarded as a rough estimate of $2c$ in (12). Combining with the median value value 5.3 of x_{gs} for all genes we get a corresponding rough estimate of the median of the Poisson parameters for all genes, $\hat{\lambda}_{\text{median}} = (2/7.0)\exp(5.3) = 60$.

6.2 The log-harmonic mean intensity

The use of the log-harmonic mean intensity (3) in the normalization instead of (7) gave about 4 % lower estimated variance (15) for the data set analysed in Section 5.1 and about 7% lower estimated variance for the NIEHS data set. Similarly, the results in Table 1, Table 2 and Figure 3 indicate that the variance model (10) is superior to model (11). In addition, we have for the log-harmonic mean model the Poisson model interpretation as described in Section 3. For most genes the intensity means (3) and (7) are close. They are identical if Z_{g1s} and Z_{g2s} are equal, but if the two Z-variables differ considerably x_{gs} is essentially determined by the smaller of the two Z-values, while $\tilde{x}_{gs}$ is essentially determined by the larger.

6.3 Bayes, empirical Bayes and frequentist approaches to finding differentially expressed genes

Figure 3 shows that weighting of gene-specific variances and the global intensity-modified variance gives considerable improvement, particularly compared to the use of the individual gene-specific variance estimates (the left endpoint of the curves), when we, as in this dataset, have a small number of slides. But we see from Table 2 that also the differences between the maxima and the right endpoint values in Figure 3 are substantial. The analyses of the two datasets in Section 5 indicates that weights can vary considerably from one experiment to another. Thus a method such as the empirical Bayes method in Section 4.3 should be quite useful in finding the appropriate weights for a particular experiment.

The previously mentioned papers Baldi and Long (2001) and Lönnstedt and Speed (2001) give Bayesian and empirical Bayes methods related to the approach described in the present paper. Baldi and Long (2001) suggest that a total number K of degrees of freedom could be specified and give a default value $K = 10$. This is about the same as we get in Section 5.2, where we find $\hat{r} + S - 1 \approx 9.0$, but somewhat too low for the data analysed in Section 5.1, where we find $\hat{r} + S - 1 \approx 16.4$ for our preferred model (10).

For the experiment analysed in Efron et al. (2001) a constant a_0 is added to the estimated gene-specific standard deviations. Different percentiles P of standard deviations for all genes are used as a_0. The optimal value corresponds to $P = 90\%$ while $a_0 = 0$ is worst (indicating a situation similar to the one in Figure 3 above).

The three methods compared in Table 5 in Kerr et al. (2000) can be related to Figure 3. Their method with gene-specific variance corresponds to the left end-points of the curves in the figures, their homoscedastic method corresponds to the right end-point in the curve with circles, while their method with intensity-dependent variance roughly corresponds to the right end-point of the upper curve — although their intensity-dependent curve is estimated nonparametrically instead of via a model such as (10).

6.4 Computations

All computations in Section 5 were performed with the Matlab® software with the optimization algorithm `fminunc` for log-likelihood maximization. The computation time for the results shown in Tables 1–2 and Figures 1–4 was about two minutes on a SunSparc® Ultra 10 workstation.

Use of data for all 2208 genes in the original dataset for Section 5.1 instead of the reduced dataset with 2202 genes gave broadly similar results. The estimated optimal degrees of freedom for the global variance were somewhat reduced, which could be expected as the added 6 genes had large estimated gene-specific variances. Thus we got, for instance, $\hat{r} = 11.4$ for model (10) with the full data set compared to $\hat{r} = 13.4$ in Table 2.

6.5 Extensions

In this paper we have concentrated upon variance estimation with the immediate goal to compute improved confidence intervals for gene effects. It is possible to use the use the variance estimates also to improve estimation of gene effects by appropriate weighting.

Estimation of gene effects with corresponding standard errors is often only a first step. It may be followed by refined multiple comparisons as Dudoit et al. (2000) or by cluster analysis as in Eisen et al. (1998).

We have here considered cDNA microarrays with two treatments, but the methods can be extended to analyse experiments with several treatments and to oligonucleotide microarrays.

Acknowledgements

We thank John Gustafsson for suggestions which drastically reduced the computation time and Olle Nerman for valuable comments on the interpretation of the Poisson-type variability.

References

[1] Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, *Bioinformatics*, **17**, 509–519.

[2] Brown, C.S., Goodwin, P.C. and Sorger, P.K. (2001). Image metrics in the statistical analysis of DNA microarray data, *Proceedings National Academy of Sciences USA* **98**, 8944-8949.

[3] DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

[4] Dudoit, S., Yang, Y.H., Speed, T.P. and Callow, M.J. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA

microarray experiments. Tech report #578, Statistics Dep., Univ. Calif., Berkeley. To appear in *Statistica Sinica*.

[5] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, **21**, 10–14.

[6] Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. Tech report #216 , Stanford University. To appear in *Journal of American Statistical Association*.

[7] Eisen, M.B, Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis of genome-wide expression patterns. *Proceedings National Academy of Sciences USA* **95**, 14863–14868.

[8] Eubank, R.L. (1984). Approximate regression models and splines. *Communications Statistical Theory and Methods B*, **13** 433-484.

[9] Hjort, N.L. and Omre, H. (1994). Topics in spatial statistics. *Scandinavian Journal of Statistics*, **21**, 289 – 358.

[10] Ideker, T., Thorsson, V., Siegel, A.F. and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805–817.

[11] Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N.J. and Churchill, G.A. (2001). Statistical analysis of a gene expression microarray experiment with replication. To appear in *Statistica Sinica*.

[12] Lönnstedt, I. and Speed, T. (2001). Replicated microarray data. To appear in *Statistica Sinica*.

[13] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37–52.

[14] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-effect Models in S and S-plus*. New York: Springer.

[15] Scheidl, S.J., Nilsson, S., Kalén, M., Hellström, M., Takemoto, M., Håkansson, J. and Lindahl, P. (2001). mRNA expression profiling of laser microbeam microdissected cells from slender embryonic structures. To appear in *American Journal of Pathology*.

[16] Tusher, V.G., Tibshirani, R. and Chu,G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings National Academy of Sciences USA*, **98**, 5116–5121.

[17] Widengren, J. and Schwille, P. (2000). Characterization of photoinduced isomerization and back-isomerization of the cyanine dye Cy5 by fluorescence correlation spectroscopy. *Journal of Physical Chemistry A*, **104**, 6416–6428.

[18] Yang, Y.H., Dudoit, S., Luu, P., and Speed, T.P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Microarrays: Optical Technologies and Informatics, Vol. 4266 of Proceedings of SPIE.

7 Tables and figures

Table 1: Variance parameter estimates (with standard error *s.e.* below each estimate) and maximum log-likelihoods for the variance models (8), (9), (10) and (11) computed as described in Section 4.2, for instance by maximizing (18) for model (9).

Model	$\hat{\sigma}^2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$L_{\chi^2}(\hat{\sigma}^2, \hat{\alpha})$
(8)	0.0747			3400.69
s.e.	0.0013			
(9)	0.232	0.203		3536.88
s.e.	0.016	0.012		
(10)	7.0	1.05	0.0441	3588.96
s.e.	3.3	0.10	0.0019	
(11)	13.9	1.20	0.0472	3571.28
s.e.	6.7	0.10	0.0016	

Table 2: Optimal weights $\hat{w}$ of the global variance and the corresponding optimal degrees of freedom $\hat{r}$ and maximum log-pseudolikelihoods $L_{\text{student}}(\hat{r})$ from (24) for the variance models (8), (9), (10) and (11) computed as described in Section 4.3, see also Fig. 3. The log-pseudolikelihoods for weights $w = 0$ and $w = 1$, that is for $r = 0$ and $r = \infty$, are also given.

Model	$\hat{w}$	$\hat{r}$	$L_{\text{student}}(\hat{r})$	$L_{\text{student}}(0)$	$L_{\text{student}}(\infty)$
(8)	0.840	10.50	-2174.87	-3831.97	-2342.39
(9)	0.862	12.47	-2003.20	-3831.97	-2160.81
(10)	0.870	13.43	-1938.05	-3831.97	-2091.36
(11)	0.867	13.01	-1955.44	-3831.97	-2114.94

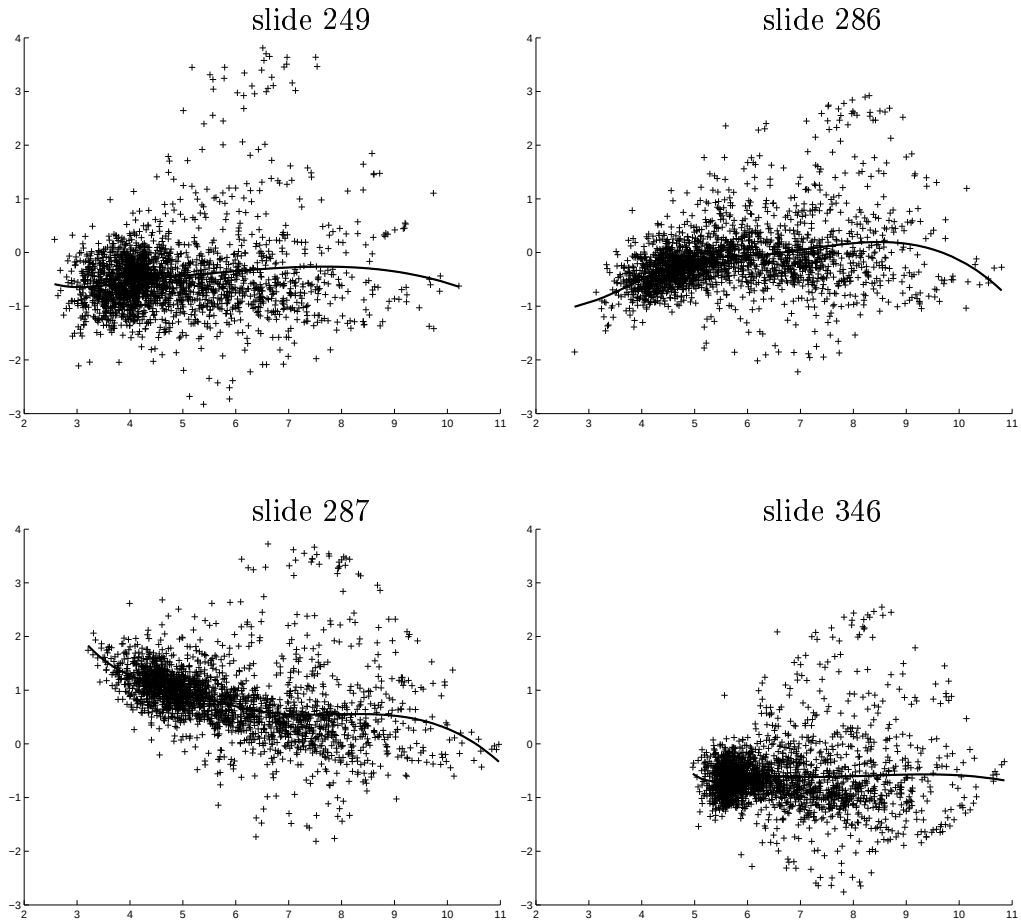

Figure 1: Data from an experiment with 2208 genes and 4 slides. The plots show for each of the four slides the values Y_{gs} in (2) plotted against x_{gs} in (3) for the slightly reduced data set with $G = 2202$ genes, cf. Section 5. The heart treatment (t=1) was assigned to the green channel (Cy3) in slides 286 and 346 and to the red channel (Cy5) in slides 249 and 287. The curves show the estimated slide effects as cubic spline functions with 2 inner knots.

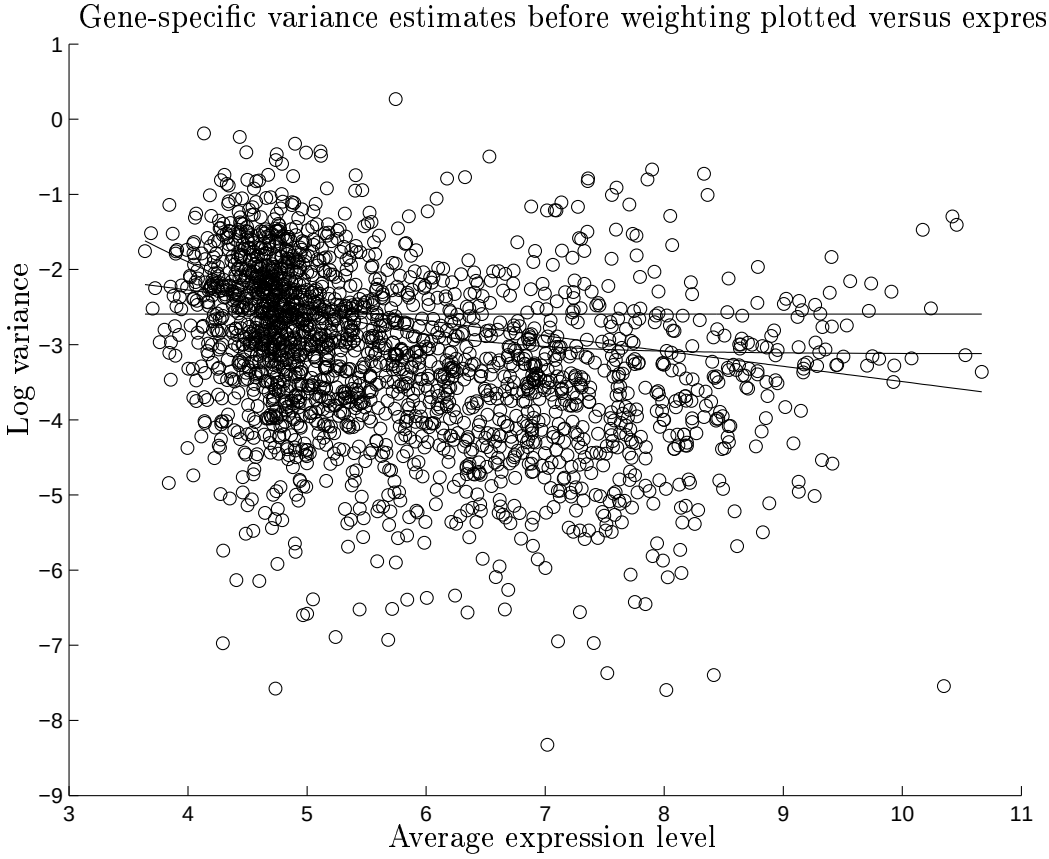

Figure 2: Natural logarithms of the estimated variances (16) for the individual genes plotted against the average log-harmonic mean intensity $\bar{x}_g$ in (17). The figure also shows the estimated curves corresponding to the variance models (8), (9) and (10) with parameters estimated as described in Section 4.2.

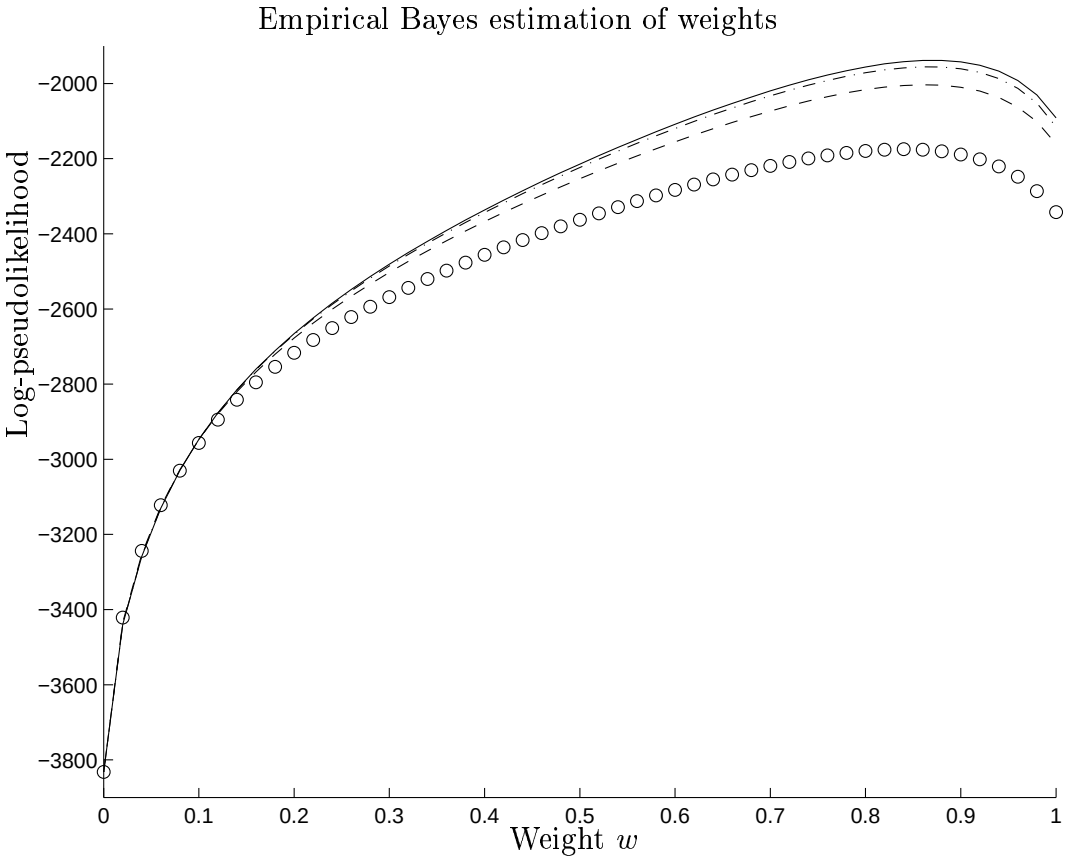

Figure 3: The log-pseudolikelihod (24) here shown as a function of weight $w = r/(S-2+r)$ of the global variance for the variance models (8) (circles), (9) (dashed curve), (11) (dash-dotted curve) and (10) (solid curve).

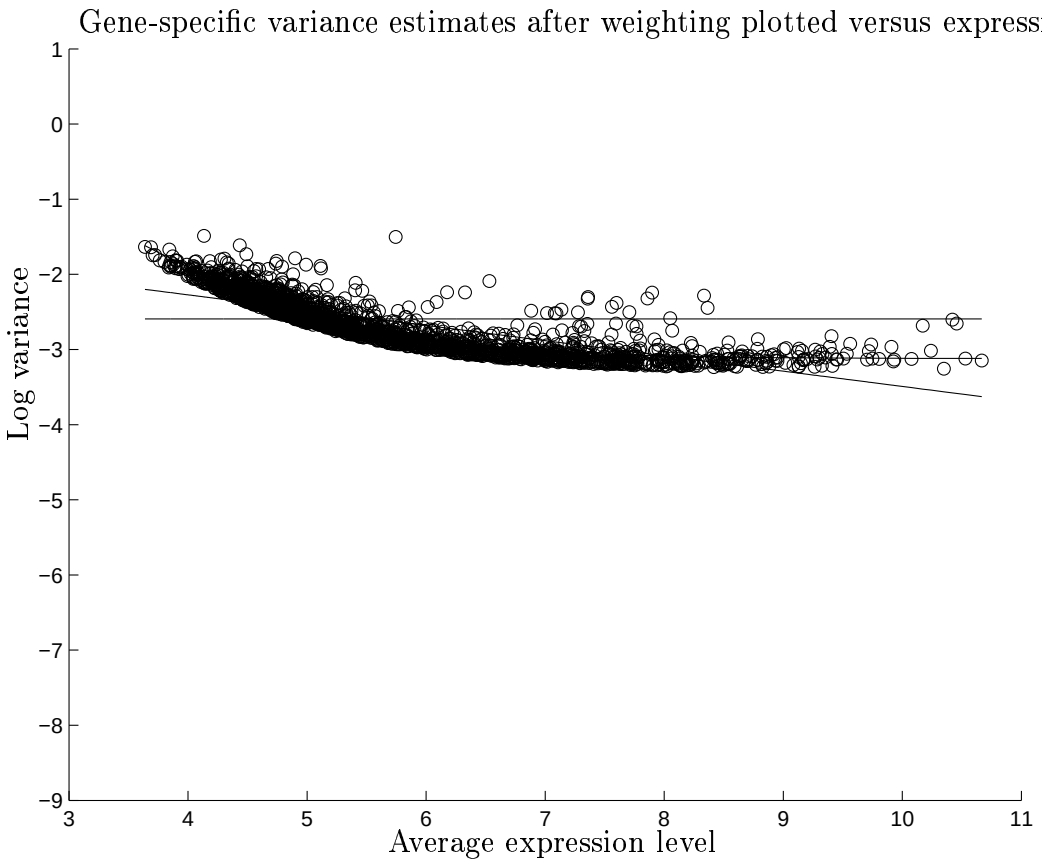

Figure 4: Natural logarithms of the estimated variances (16) for the individual genes after optimal weighting according to (21) for model (10) plotted against the the average log-harmonic mean intensity $\bar{x}_g$ in (17). The figure also shows the estimated curves corresponding to the variance models (8), (9) and (10) with parameters estimated as described in Section 4.2.