Incorporating Multiple cDNA Microarray Slide Scans - Application to Somatic Embryogenesis in Maize

Tanzy Love  
*Iowa State University*

Alicia L. Carriquiry  
*Iowa State University*, alicia@iastate.edu

Follow this and additional works at: [http://lib.dr.iastate.edu/stat_las_preprints](http://lib.dr.iastate.edu/stat_las_preprints)

Part of the [Statistics and Probability Commons](http://lib.dr.iastate.edu/stat_las_preprints)

Recommended Citation

[http://lib.dr.iastate.edu/stat_las_preprints/37](http://lib.dr.iastate.edu/stat_las_preprints/37)

This Article is brought to you for free and open access by the Statistics at Digital Repository @ Iowa State University. It has been accepted for inclusion in Statistics Preprints by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Incorporating Multiple cDNA Microarray Slide Scans
- Application to Somatic Embryogenesis in Maize
Abstract

Microarray data are subject to multiple sources of measurement error. One source of potentially significant error is the settings of the instruments (laser and sensor) that are used to obtain the measurements of gene expression. Because ‘optimal’ settings may vary from slide to slide, operators typically scan each slide multiple times and then choose the reading with the fewest over-exposed and under-exposed spots. We propose a hierarchical modeling approach to estimating gene expression that combines all available readings on each spot. The basic premise is that all readings contribute some information about gene expression and that after appropriate re-scaling, it would be possible to combine all readings into a single estimate. We illustrate the use of the model using expression data from a maize embryogenesis experiment and assess the statistical properties of the proposed expression estimates using a simulation experiment. As expected, combining all available scans using a reasonable approach to do so results in expression estimates with noticeably lower bias and root mean squared error relative to other approaches that have been proposed.

1 Introduction

Large cDNA microarray studies are carried out to investigate complex processes in cyclic and developmental behavior. Typically, genes from biological materials that have been subjected to different treatments or that arise from different tissues or from the same tissue at different stages of development are spotted on different arrays (or slides). The objective is to draw inferences about differential gene expression levels across treatments, tissues or developmental stages. Gene expression levels can only be compared across different slides after the appropriate background cleaning and normalization procedures have scaled the data to the same range. There are several important sources of variation in gene expression measurement that must be accounted for in statistical analyses, and much of this variation is slide specific.

The data generated from cDNA microarray experiments are obtained by combining two types of images of the microarray slide. The two images are obtained while the slide is excited with a laser tuned to Cy5 and to Cy3 fluorescent dyes, respectively. Different laser strengths and the sensitivities of the camera result in different images. While a setting for laser and camera may produce a large number of saturated spots on the slide image, other settings may result in too many spots with measured expression below the minimum that can be captured by the instruments. The laser strength and the sensitivity of the camera to light can be adjusted by the operator to find a ‘best’ picture, one where most of the spots show some measurable expression and where very few of the spots reach saturation. Once the best image is obtained, the spots of cDNA must be found, a process called segmentation, and the background of each
spot must be calculated. Finally, before statistical analysis of the data can be underway, each slide must be normalized within itself to recover from any systematic dye bias (usually Cy3 is stronger) and all the slides must be normalized together to make them comparable (Smyth, Yang, and Speed 2002).

Methods for estimating gene expression that use the multiple slide scans that are typically produced in microarray experiments have been discussed in the literature in recent months. Lyng et al. (2004) and Skibbe, Nettleton, and Schnable (2004) investigated the effects of scanner settings on expression ratios and significant differential expression, respectively. Both found that scanner settings have an important impact on the quality of and conclusions from microarray data. Lyng et al. (2004) suggests using two scans at different settings to increase the usable range of expression values. Rombualdi et al. (2003) ingrated multiple slide readings before segmentation to create more uniform spots. However, the same scanner settings were used, so that there was no improvement of the dynamic range of expression estimates. Dudley et al. (2002) and Garcia de la Nava, van Hijum, and Trelles (2004) used multiple slide readings at varying settings to extend the dynamic range and address the truncation error. However, the former uses only the estimate from one reading for each gene (possibly linearly transformed) while the latter accomodates only two scans.

In this work, we focus on the measurement error that is introduced when scientists vary the strength of the laser and the sensitivity of the camera used to record expression levels and propose a statistical approach that allows incorporating multiple readings of each slide into the analysis. We show that under relatively lax model assumptions, expression levels can be estimated with significantly lower bias and higher precision when combining multiple readings for each gene into the statistical analysis than when choosing only the ‘best’ reading.

This paper is organized as follows. In Section 2 we describe in more detail the process by which multiple slide readings are obtained and hypothesize about the association between various slide readings. Section 3 develops the Bayesian hierarchical model for jointly analyzing multiple cDNA slide scans. Sections 4 and 5 apply the proposed approach in a simulation experiment and in a real maize embryogenesis experimental dataset. The simulation study was carried out to compare the proposed approach to other approaches published in the literature and that address this type of measurement error. Section 6 dicusses the implications of incorporating multiple slide scans into the statistical analyses of microarrays.
2 Multiple Laser and Sensor Settings

Different laser and sensor settings can be used to read a cDNA microarray slide. Stronger laser settings create more fluorescence and stronger sensor settings pick up more signal. There is a balance to be struck between picking up signal from the lowly fluorescing spots and over-exposing the highly expressing genes. There is an upper limit of 65535 to the measurement of fluorescence; readings of spots which are brighter are truncated. Over-exposing the high intensity spots will cause them to be artificially near other high expression values. Correspondingly, low signals will be artificially assigned to 0 if the laser and sensor settings are too low.

Figure 1 illustrates the two scenarios. The expression estimates used are background corrected average pixel intensities. An overexposed spot will have variation in its pixels due to inconsistencies in spot printing and irregular spot shape. Further, background correction will reduce the measured expression value so that 65535 is not the true point of truncation. See Figure 1 (b) where there is truncation at around 55000. An underexposed spot will also have variation in its pixels for the same reasons. Background correction, however, will create expression estimates that are negative or near zero. The truncated values in the figure correspond to those genes that are not expressed in that particular slide, but also to those genes which may have exhibited measurable expression levels had the spot been more exposed. Negative expression measurements are routinely set to zero, however, often the true point of truncation is not zero. See Figure 1 (a) where there is truncation at around ten.

Multiple readings of the microarray slides can be taken for both fluorescence channels. Since all of the readings at different settings attempt to capture true expression levels for the genes on the slide, it is reasonable to assume that all readings contribute useful information about true expression levels and to think of combining the multiple readings into one estimate of gene expression for each spot. If the readings at different settings contain information about the true expression of the gene, then the variance in estimated gene expression that is due to the measurement process should be reduced in the estimate that is based on all available readings.

Several aspects of the measurement process of gene expression create challenges for statistical modeling. As discussed earlier, many microarray experiments include pseudo-replicates, which we define as multiple readings of the slide under different laser and sensor settings. Generally, settings for different slides are very different because of the large experimental variation between slides. That is, one slide may result in a good reading at low laser and sensor settings while another may require higher settings to reduce the number of expression levels below threshold while keeping the number of overexposed spots to a minimum. Because of this practice, we are typically unable to assume that the settings act as blocks in a traditional experimental design.
Figure 1: (a) Many spots are truncated above in the higher reading. (b) Many spots are truncated below in the lower reading.
However, since the settings to read the two channels are almost always picked independently across slides, we can model each slide/dye combination separately. In what follows, we consider an arbitrary slide and dye channel in the experiment and propose a hierarchical model for estimating gene expression levels that permits incorporating multiple measurements for each gene into a single analysis.

3 Bayesian Hierarchical Gamma Model

In order to estimate gene expression, we propose a Bayesian hierarchical model. This model incorporates all slide scans into one estimate of expression per spot. To formulate the model, we rely on the natural ordering of slide readings. For instance, if we have two readings with the same sensor setting and different laser settings, the measurements on the reading with the higher laser setting will tend to be larger. Dudley et al. (2002) discuss gene expression and its dependence on changes in one of the experimental settings. Here we consider multiple settings simultaneously and to do so order the slides from smallest to largest median reading. Clearly, the median-based ordering is subject to some uncertainty because of the measurement error in observed gene expressions.

3.1 Likelihood Function

Suppose that there are \( m + 1 \) readings taken on \( n \) spots for each combination of a slide and a dye. For a given gene \( i \), we use \( S_{i1}, \ldots, S_{i(m+1)} \) to denote the \( m + 1 \) ordered signal measurements after background correction. Here \( S_{i1} \) is the gene expression measurement from the smallest (in median expression units) scan and \( S_{i(m+1)} \) denotes the reading for gene \( i \) on the scan with the highest median reading. Truncated spots are recorded as missing data, however, a spot truncated below in the highest scan or above in the lowest scan is not recorded as missing. In practice a spot can be designated as truncated below if any of its pixels are less than the background median. A spot can be designated as truncated above if any of its pixels are saturated. Alternatively, exploratory data analysis can be used to decide appropriate cut-off values for a particular slide, such as 20 and 50,000.

We assume that all readings measure the same quantity – actual gene expression – with error. Therefore, under suitable scaling the readings would be identically distributed. We assume that the strictly positive scaled readings can be represented by a Gamma distribution. The Gamma has support on the positive real line and depending on parameter values, exhibits noticeable skewness. Therefore, we can model the background corrected signals for each gene \( i \).
across the $m + 1$ readings in the following way:

\[
S_{i1} * \chi_1 = S_{i1}' \sim \Gamma(a, \psi_i) \\
S_{i2} * \chi_2 = S_{i2}' \sim \Gamma(a, \psi_i) \\
\vdots \\
S_{i(m+1)} * \chi_{m+1} = S_{i(m+1)}' \sim \Gamma(a, \psi_i)
\]

where $\chi_1, \ldots, \chi_{m+1}$ are constant for all genes in a given slide and dye combination. This amounts to assuming that the changes in laser and sensor settings affect the amount of each spot’s fluorescence equivalently.

As formulated, the model is not identified in that there is no way to estimate the parameters $\psi_i$ directly. Thus, we do not attempt to estimate $\chi_1, \ldots, \chi_{m+1}$ and instead focus on estimating $\theta_i = \chi_{m+1}\psi_i$. We choose one of the $m + 1$ readings as a reference reading and scale all other readings to that level. By scaling all readings upwards to the highest one we are increasing the effective range of gene expression measurement. This does not limit the usefulness of the model in any way, because all measures of gene expression are relative and normalization is performed on the expression estimates.

We now have the following model:

\[
S_{i1} * \delta_1 = S_{i1}' \sim \Gamma(a, \theta_i) \\
S_{i2} * \delta_2 = S_{i2}' \sim \Gamma(a, \theta_i) \\
\vdots \\
S_{im} * \delta_m = S_{im}' \sim \Gamma(a, \theta_i) \\
S_{i(m+1)} \sim \Gamma(a, \theta_i)
\]

where $\delta_1, \ldots, \delta_m$ are constant for all genes in a given slide and dye combination. For notational convenience, let $\delta_{m+1} = 1$. The unknown parameters in this model are $a, \theta_1, \ldots, \theta_n$, and $\delta_1, \ldots, \delta_m$.

The mean of each of the Gamma distributions for the $i$th spot is $a/\theta_i$. Within a classical framework, an estimate of expression level for the $i$th gene would be based on the corresponding mean or perhaps on a suitable function of the $m + 1$ means. These estimates still require normalization so that expressions observed for different slide/dye combinations can be compared.
3.2 Prior Distributions

We adopt a Bayesian approach to estimating the parameters in the model. In order to do so, we must complete the specification of the model by assigning prior distributions to each parameter. We restrict our attention to proper prior distributions to guarantee integrability of the posterior, and within the family of proper distributions we focus on the conjugate or semi-conjugate families to attempt to simplify computations wherever possible. If the prior distribution for the parameters is conjugate, then the posterior will have the same form as the likelihood function.

We assume that the scale parameters, $\theta_1, \ldots, \theta_n$, arise from a common population distribution independently (a priori) of the scaling parameters $\delta_1, \ldots, \delta_m$. Let
\[
p(\theta_1, \ldots, \theta_n, \delta_1, \ldots, \delta_m) = p(\theta_1, \ldots, \theta_n) \times p(\delta_1, \ldots, \delta_m)
\]
represent a joint prior distribution that for now will remain unspecifed. We derive a posterior distribution for the vectors $\theta = (\theta_1, \ldots, \theta_n)$ and $\delta = (\delta_1, \ldots, \delta_m)$ and then determine the form of the prior distribution $p(\theta, \delta)$ that would be conjugate for the likelihood.

Conditional on $a$ and on the data $S$, the joint posterior distribution of $(\theta, \delta)$ is given by
\[
p(\theta, \delta|S, a) \propto p(\theta, \delta) \times p(S|\theta, \delta, a)
\]
\[
= p(\theta, \delta) \prod_{i=1}^{n} \prod_{j=1}^{m+1} \Gamma(S_{ij}\delta_j|a_i, \theta_i)
\]
\[
= p(\theta, \delta) \prod_{i=1}^{n} \prod_{j=1}^{m+1} \frac{\theta_i^{e_i}}{\Gamma(a_i)} (S_{ij}\delta_j)^{a_i-1} e^{-\theta_i(S_{ij}\delta_j)}
\]
\[
\propto p(\theta, \delta) \prod_{i=1}^{n} \theta_i^{e_i(a+1)} \prod_{j=1}^{m} \delta_j^{n(a)-1} e^{-\sum_{i=1}^{n} \theta_i \sum_{j=1}^{m+1} \delta_j S_{ij}}.
\]

A true conjugate prior for $\theta$ and $\delta$ would have the form
\[
p(\theta, \delta) \propto \prod_{i=1}^{n} \theta_i^{\psi_i} \prod_{i=1}^{m} \delta_j^{\psi_j} e^{-\psi_3 \sum_{i=1}^{n} \theta_i \sum_{j=1}^{m+1} \delta_j}.
\]

This distribution is difficult to interpret from a biological viewpoint and further, implies a prior dependency between $\theta$ and $\delta$ which we cannot justify. Thus, the conjugate prior option, while convenient from a mathematical viewpoint appears to be unsuitable from a biological
viewpoint. We consider instead independent Gamma prior distributions for each of the \( n + m \) parameters. Gamma distributions can be justified from a biological point of view because typically genes spotted on a slide exhibit low expression levels and only some of them exhibit high levels of expression. The Gamma distribution would appear to be an appropriate model for the population distribution because the expression values of the genes, estimated by \( a/\theta_i \), will be skewed. Thus

\[
\theta_i \sim \Gamma(a_0, \nu)
\]

for \( i = 1, \ldots, n \). The Gamma model may also be reasonable for the strictly positive scaling parameters, so that

\[
\delta_j \sim \Gamma(\alpha_1, \alpha_2)
\]

for \( j = 1, \ldots, m \). The joint Gamma prior has the form

\[
p(\theta, \delta) \propto \prod_{i=1}^{n} \theta_i^{a_0} \prod_{i=1}^{m} \delta_j^{\alpha_1} e^{-\nu \sum_{i=1}^{n} \theta_i - \alpha_2 \sum_{j=1}^{m+1} \delta_j}.
\]

(3)

The conditional posterior distributions of \( \theta|\delta \) and \( \delta|\theta \) are Gamma distributions under this prior, but the joint posterior of \( (\theta, \delta) \) is not. Therefore, the prior in (3) is a semi-conjugate prior distribution.

The hyperparameters in the model are \( \eta = (a, a_0, \nu, \alpha_1, \alpha_2) \). We must either specify prior distributions for these hyperparameters or fix the parameters at some appropriate value. The hyperparameters \( \alpha_1 \) and \( \alpha_2 \) are both chosen to be 10 to create a relatively noninformative prior on the \( \delta \)'s. Specifying a value for the other hyperparameters \( a, a_0 \) and \( \nu \), however, requires some thought since these parameters can have a significant effect on the estimates of expression levels.

One approach to obtaining values for hyperparameters is to find the values \( (\hat{a}, \hat{a}_0, \hat{\nu}) \) that maximize the marginal likelihood of the parameters (MMLEs, e.g., Carlin and Louis, 2003). The marginal likelihood \( p(a, a_0, \nu|S) \) is obtained by integrating \( (\delta, \theta) \) out of the joint likelihood function as follows (the complete derivation of \( p(a, a_0, \nu|S) \) is presented in the Appendix):

\[
p(a, a_0, \nu|S) = \int \int p(a, a_0, \nu, \theta, \delta|S) d\delta d\theta
\]

\[
= \Gamma(a)^{-n(m+1)} \Gamma(n(a - 1) + \alpha_1)^m \prod_{i=1}^{n} \prod_{j=1}^{m+1} \theta_i^{a_0 - 1} e^{-\nu \sum_{i=1}^{n} \theta_i} \prod_{j=1}^{m} (\sum_{i=1}^{n} \theta_i S_{ij} - \alpha_2) d\theta.
\]

(4)
This marginal distribution is not analytically tractable. However, we could integrate $\delta$ out analytically if instead of conditioning on $S$ we were to derive the marginal likelihood given only the largest reading, $S_{(m+1)}$. Here, we have chosen to use the largest reading arbitrarily. Any reading could be used as the standard and the model would be adjusted to estimate $\theta = \chi h \psi$. The subsequent normalization that expression estimates must undergo makes any choice of standard reading equivalent. In this case,

$$p(a, a_0, \nu|S_{(m+1)}) = \int \int p(a, a_0, \nu, \delta|S_{(m+1)})d\delta d\theta \tag{5}$$

$$\propto \int \int p(S_{(m+1)}|a, a_0, \nu, \delta)p(\theta|a_0, \nu)p(\delta)d\delta d\theta$$

$$= \int p(S_{(m+1)}|a, a_0, \nu, \theta)p(\theta|a_0, \nu)d\theta$$

$$= \int \Gamma(a)^{-n} \prod_{i=1}^{n} \theta_i^{a_i S_{i(m+1)}^{a_i - 1}} \exp(-\theta_i S_{i(m+1)})$$

$$\Gamma(a_0)^{-n} \nu^{na_0} \prod_{i=1}^{n} \theta_i^{a_0 - 1} \exp(-\nu \theta_i) d\theta$$

$$\propto (\Gamma(a) \Gamma(a_0))^{-n} \prod_{i=1}^{n} S_{i(m+1)}^{a_i - 1} \nu^{na_0}$$

$$\int \prod_{i=1}^{n} \theta_i^{a_i + a_0 - 1} \exp \left( - \sum_{i=1}^{n} \theta_i (S_{i(m+1)} + \nu) \right) d\theta$$

$$= (\Gamma(a) \Gamma(a_0))^{-n} \prod_{i=1}^{n} S_{i(m+1)}^{a_i - 1} \nu^{na_0} \prod_{i=1}^{n} (a + a_0)(S_{i(m+1)} + \nu)^{-a - a_0}$$

$$= \left( \frac{\Gamma(a + a_0)}{\Gamma(a) \Gamma(a_0)} \right)^{n} \prod_{i=1}^{n} S_{i(m+1)}^{a_i - 1} \nu^{na_0} \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-a - a_0}.$$

The resulting expression can now be maximized with respect to $a, a_0$ and $\nu$ using standard nonlinear optimization techniques.

One disadvantage of proceeding within an empirical Bayes framework and fixing hyperparameters to point values is that we ignore the uncertainty about the hyperparameters when estimating other parameters in the model. To ameliorate this problem we consider placing a prior distribution on $\nu$ that again has the form of a Gamma($\beta_1, \beta_2$). Here, we specify $\beta_1 = 100$ and $\beta_2 = 1$ to create a relatively noninformative prior for $\nu$. Prior sensitivity has been checked for many values of $\beta_1$. If we now attempt to obtain the MMLE of $(a, a_0)$ as before using data from all the readings, we again obtain an expression for the marginal likelihood that is computationally intractable. However, approximating the marginal likelihood by conditioning only
on the highest reading $S_{(m+1)}$ again results in an approximate marginal distribution for $a$ and $a_0$ that can be maximized. The derivation is presented in the Appendix. The expression to be maximized with respect to $(a, a_0)$ is the following:

$$p(a, a_0 | S_{(m+1)}) = \int \int \int p(a, a_0, \theta, \delta, \nu | S_{(m+1)}) d\delta d\theta d\nu$$

$$= \left( \frac{\Gamma(a + a_0)}{\Gamma(a) \Gamma(a_0)} \right)^n \prod_{i=1}^{n} S_{i(m+1)}^{a_0-1}$$

$$\int \nu^{na_0 + \beta_1 - 1} \exp(-\beta_2 \nu) \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a_0+a_0)} d\nu.$$  (6)

An efficient approach to finding the values $(\hat{a}, \hat{a}_0)$ that maximize (6) is the EM algorithm (Dempster, Laird, and Rubin 1977). The algorithm cycles between two steps: an expectation step (E-step) and a maximization step (M-step) until reaching convergence. In the E-step, we find the conditional posterior expectation of $\nu$ given $a$ and $a_0$:

$$E(\nu | a, a_0, S_{(m+1)}) = \int \nu p(\nu | a, a_0, S_{(m+1)}) d\nu$$

$$= \int \frac{\nu^{na_0 + \beta_1 - 1} \exp(-\beta_2 \nu) \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a_0+a_0)}}{\int \nu^{na_0 + \beta_1 - 1} \exp(-\beta_2 \nu) \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a_0+a_0)} d\nu} \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a_0+a_0)} d\nu.$$  (7)

In the M-step, we compute the MLE for $a$ and $a_0$ given $\nu$. The conditional likelihood to be maximized is

$$L(a, a_0 | \nu, S_{(m+1)}) = \int \int \int p(a, a_0, \theta, \delta | \nu, S_{(m+1)}) d\delta d\theta$$

$$= \left( \frac{\Gamma(a + a_0)}{\Gamma(a) \Gamma(a_0)} \right)^n \nu^{na_0} \prod_{i=1}^{n} S_{i(m+1)}^{a_0-1} \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a_0+a_0)}.$$  (8)

The algorithm iterates between the E-step and the M-step until reaching convergence. The derivation of expressions (7) and (8) is given in the Appendix.

### 3.3 Posterior distributions

The joint posterior distribution of $(\delta, \theta, \nu | S, a, a_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ is given by:
We use Markov chain Monte Carlo (MCMC) methods to approximate the marginal posterior distributions of each of the $m + n + 1$ parameters in the model. To do so, we first derive the full conditional distributions for each of the parameters:

\[
p(\delta, \theta, \nu | S, a, a_0, \alpha_1, \alpha_2, \beta_1, \beta_2) \propto \prod_{j=1}^{m} \delta_j^{a_j-1} e^{-\sum_{j=1}^{m+1} \delta_j (\sum_{i=1}^{n} \delta_i S_{ij} - \alpha_2)} \times \prod_{i=1}^{n} \theta_i^{a_i+1} e^{-\nu (\sum_{i=1}^{n} \delta_i + \alpha_2)} \beta_1^{-1} \tag{9}
\]

Notice that all full conditional distributions have standard form, and thus the Gibbs sampler can be used to sequentially draw parameter values from the conditionals. A point estimate for the expression of the $i$th gene is the posterior mean of $\theta_i$. These estimates may be subsequently used as the expression values for further normalization.

### 4 Comparisons via a simulation experiment

We ran a simulation study to examine the differences in bias and root mean squared error (RMSE) of gene expression estimates between different approaches to estimate gene expression. A cDNA microarray dataset read at $m + 1$ reading levels was simulated from the hierarchical model. Gene expression was then estimated using the Bayesian hierarchical model we propose.
Table 1: Values of the hyperparameters used in simulation experiment.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>1,000</td>
</tr>
<tr>
<td>$m + 1$</td>
<td>3</td>
</tr>
<tr>
<td>$a$</td>
<td>20</td>
</tr>
<tr>
<td>$a_0$</td>
<td>1</td>
</tr>
<tr>
<td>$\nu$</td>
<td>12</td>
</tr>
<tr>
<td>$\alpha_1 = \alpha_2$</td>
<td>10</td>
</tr>
</tbody>
</table>

here, the average gene expression over the $m + 1$ readings, the geometric mean of the gene expression over the $m+1$ readings, and a linear extrapolation method that was recently proposed in the literature (Dudley et al. 2002). We replicated the experiment 100 times, and computed average bias and RMSE over the 100 replicates for each gene.

The values of the hyperparameters that we chose for the simulation are given in Table 1. These values were used in every replicate. We then generated values for the $\theta$’s and $\delta$’s from the population distributions given the chosen hyperparameters. These $\theta$’s and $\delta$’s were in turn used to generate a set of observed expression values for $m + 1 = 3$ readings on $n = 1,000$ spots.

4.1 Estimation of scale parameters

The model is very good at estimating the scale parameters $\delta_1, \ldots, \delta_m$. The average bias over 100 replications was -0.00047 and the average root mean squared error (RMSE) was 0.00568. The linear extrapolation method proposed by Dudley et al. (2002) relies on the same assumption that the scaling between scans is constant across the slide. However, only one scan value (possibly scaled) is used to estimate gene expression. The average bias in estimating the scaling constants under the linear model was 0.07597 and the average RMSE was 0.10260. Both of the estimation methods use the same assumption that there is a linear relationship between the readings, so we expect their estimates to be close.

4.2 Estimation of gene expression

We fitted the hierarchical model we propose to each of the 100 replicated datasets using the $m + 1$ readings available for each spot in each replicate. The posterior means of the $a/\theta$’s (mean expression values) were used as estimates of the true expression values. We also calculated the average observed expression value from $m + 1$ readings, the geometric average observed expression value, the estimates obtained by linearly extrapolation between readings (Dudley
Table 2: Simulation Results for 1000 genes after 100 simulations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Absolute Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical model using $m + 1$ readings</td>
<td>257.45</td>
<td>499.71</td>
</tr>
<tr>
<td>Average observed gene expression</td>
<td>1065.94</td>
<td>1117.12</td>
</tr>
<tr>
<td>Geometric average observed expression</td>
<td>1217.33</td>
<td>1286.10</td>
</tr>
<tr>
<td>Linear extrapolation</td>
<td>441.89</td>
<td>1565.42</td>
</tr>
<tr>
<td>Hierarchical model using highest reading</td>
<td>622.10</td>
<td>864.14</td>
</tr>
<tr>
<td>Naive (highest reading)</td>
<td>619.11</td>
<td>863.15</td>
</tr>
</tbody>
</table>

e et al. 2002) and the posterior mean expression values under a Bayesian hierarchical model similar to ours but that relies on only one reading per slide (Newton et al. 2001). All of the estimates were compared to a naive gene expression estimate obtained by simply using the value from the highest scan (by median ranking) as the estimate. The biases for each of the 1,000 expression values were calculated for each of the 100 replicates. The average absolute bias and RMSE for each expression value were calculated over the replicates. The range of expression values for these simulations was 424891; of the 1000 expression values, three were saturated in all three readings and around three more were saturated in one or two scans. Simulations were run with many more or no genes being always saturated and the results are the same; bias increases for all methods with more always-saturated genes and our method remains better in both the bias and RMSE sense. Results are presented in Table 2.

Results suggest that gene expression estimates obtained by implementing the hierarchical model that we propose are better (in the minimum bias and RMSE sense) than estimates obtained as arithmetic or geometric averages and from the hierarchical model that relies on a single slide reading. Even though gene expression is estimated with similar bias by linearly interpolating between the multiple readings, the hierarchical model proposed here for all scans has a significantly lower RMSE. This means that the sampling variance of the estimates from the model is much smaller. For expression values below the truncation limit (65535), the biases are evenly spread above and below zero, see Figure 2. For values above the the truncation limit, biases are uniformly negative.

### 4.3 Improvements over estimates that rely on a single reading

The model we propose here is an extension of the Gamma hierarchical model proposed by Newton et al. (2001) that relies on a single reading per slide and can be used to obtain the posterior distribution of differential expression.
Figure 2: Log true expression by average bias over 100 simulations for expression values less than 65535.
<table>
<thead>
<tr>
<th>Method</th>
<th>Average Range</th>
<th>Average Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical model using $m + 1$ readings</td>
<td>240732</td>
<td>-184159</td>
<td>225372</td>
</tr>
<tr>
<td>Hierarchical model using one reading</td>
<td>65993</td>
<td>-358898</td>
<td>358898</td>
</tr>
<tr>
<td>Naive (highest reading)</td>
<td>65507</td>
<td>-359385</td>
<td>359385</td>
</tr>
</tbody>
</table>

Table 3: Average range in simulation.

Results presented in Table 2 suggest that the sampling variance of gene expression estimates is reduced when all available readings are used for estimation. We now argue that by incorporating all available information about gene expression, it is also possible to increase the dynamic range of gene expression estimates in the microarray. Table 3 shows the average ranges that result from the application of three different methods. The average range (over the $m + 1$ readings) of the actual gene expressions in these simulations was 424891, so Table 3 shows that by appropriately combining the three available scans we manage to recover more of the range of the actual expressions than using one scan. Here, the bias is defined as the difference between the simulated range (maximum true expression minus minimum true expression) and the estimated range (maximum estimate minus minimum estimate), and is averaged over the 100 replicates. The RMSE is defined in a similar manner.

5 Maize Embryogenesis Experiment

The program of gene expression associated with somatic embryo maturation and germination was examined in callus cultures from a regeneration-proficient hybrid line of *Zea mays*, Hi II (Che et al. 2004). 12,060 element maize cDNA microarrays were used to generate gene expression profiles from embryogenic calli induced to undergo embryo maturation and germination.

5.1 Improvements over estimates that rely on a single reading

Two hierarchical models were fit to one dye channel of one slide: the hierarchical model proposed here, that incorporates all readings and the hierarchical model proposed by Newton et al. (2001) that incorporates information from only one scan. In this example, the Newton model was applied to the highest available reading for the gene. We present the results obtained for two arbitrarily chosen genes. Examination of those results suggests that by combining all readings for a spot we realize several improvements.

For the slide we used for illustration, the point estimates (posterior means) for the scaling
parameters $\delta_1, \delta_2$ were $\hat{\delta}_1 = 2.181$ and $\hat{\delta}_2 = 1.142$ so that scaling of the three gene expression measurements results in $2.2S_{i1} \approx 1.1S_{i2} \approx S_{i3}$. The first gene on which we focus is gene labeled #1 for which we obtained a posterior expression estimate of 276.2 based on its three measurements of (51.3, 211.0, 227.3). When using only its highest measurement (227.3), the estimate obtained was 593.5 with a 95% posterior credible set of (170, 1735). Notice that the gene expression estimate based on the three readings is within the 95% posterior probability interval. Consider now gene #1735, for which the highest reading was truncated at 0 due to very large within-spot measurement error. The posterior point estimate of expression for this gene was 153.9, based on three measurements of (59.9, 77.2, 0). When using only the highest measurement, the estimate of gene expression was 368.7 with a 95% posterior credible set of (112, 1080). In this case, the estimate obtained by combining the three readings for the gene was also contained in the 95% posterior probability interval.

We argue that by combining multiple measurements into the estimate of gene expression for a single gene the resulting estimator has lower standard error. In fact, the posterior standard deviations of expression of genes #1 and #1735 in our example were 124 and 70, respectively when using the three measurements but increased to 919 and 307 when only one scan was used for estimation. Therefore, the posterior distributions of gene expression are much less concentrated around the mean when only one reading is used in estimation.

6 Discussion

Data collected in the course of microarray experimentation is subject to multiple sources of measurement error. Some of the measurement error may actually introduce biases and analysts typically attempt to reduce those biases by re-scaling and normalizing the data prior to analyses. One source of potentially significant measurement error is the settings of the instruments (laser and sensor) that are used to obtain the data. Because the 'optimal' settings may vary from slide to slide, operators often obtain multiple readings of each slide and then choose the 'best', meaning the reading that includes the fewest saturated spots and the fewest under-exposed spots.

The use of multiple scans obtained under the same laser and sensor settings have been proposed as a means to reduce the variability of gene expression estimates (Romualdi et al. 2003). Yet improving homogeneity of spots and accounting for the purely random measurement error should be possible using effective segmentation and background cleaning methods. It has been only recently that some attention has been focused on analytical methods that might permit incorporating multiple slide scans obtained under different measurement conditions into statis-
tical analyses. Several approaches have been proposed in the literature for doing so (Dudley et al., 2002; Lyng et al., 2004; Garcia de la Nava et al., 2004). In this manuscript, we propose a general hierarchical modeling approach that allows incorporation of as many readings as may be available for each slide into the model, even if the number of readings per slide vary across slides. The basic premise is that each reading of a spot contains some information about the true expression of the gene and that if an appropriate scaling factor for each spot can be estimated, then all readings for a spot estimate the same quantity and can be combined. If so, then it is to be expected that the estimate of gene expression will have smaller variance than it would have if based on a single spot measurement.

We make several modeling assumptions in our work. For example, we assume that a single multiplicative factor is applicable to expression levels of all spots on a slide. That is, if a specific laser and sensor settings tends to increase expression levels, we assume that the multiplicative factor is uniform across all spots on a slide. This assumption may not hold in all situations, but modeling each spot within a slide individually makes the problem intractable from an analytical point of view. Simulation results show that the bias with which we can estimate gene expression is associated to expression levels, indicating that different spots on the slide might require different scalings to correct for the effect of the same laser and sensor settings.

To determine whether the modeling approach we propose results in estimators of gene expression with good statistical properties, we ran a simulation study and assessed bias and root mean squared error of the estimators over repeated sampling. The simulation experiment is described in some detail in Section 4. Using simulated gene expression data, we applied several of the approaches (including the approach proposed here) to estimate gene expression for 1,000 genes and compared the various methods on the basis of bias, RMSE and dynamic range of the estimates that were obtained. The hierarchical modeling approach we propose had smaller bias and smaller RMSE than all other estimators, suggesting that basing estimation on as many readings for each spot as might be available is probably a reasonable idea. As mentioned earlier, genes with very high and with very low true expression levels were subject to the larger biases; this is to be expected since these are the genes that are likely to have truncated expression measurements under very high or very low laser and sensor settings. Thus, changes in instrument settings cause not only a shift but also a truncation of the expression measurements in those genes.

Our method has an obvious non-model-based corollary. The scaling factors, \( \delta_1, \ldots, \delta_m \), can be estimated by the ratios of the medians of each scan in the following way:

\[
\hat{\delta}_1 = \frac{M_{m+1}}{M_{1}}
\]
\[ \hat{\delta}_m = \frac{M_{(m+1)}}{M_{(m)}} \]

where \( M_{(1)}, \ldots, M_{(m+1)} \) are the ordered scan median values. Then, the expression of gene \( i \) can be estimated as the scaled mean of the \( m + 1 \) scans, \( \frac{1}{m+1} \sum_{j=1}^{m+1} S_i \hat{\delta}_j \) where \( \hat{\delta}_{m+1} = 1 \) for notational convenience. Once again, truncated values are recorded as missing data and not included in the estimates. This method performs almost identically to point estimates from the hierarchical model in simulation. The great loss here are the uncertainty around the \( \hat{\delta} \)'s and the posterior distributions for expression estimates.

While promising, conclusions drawn from the simulation experiment may be overly optimistic as is often the case in simulation. Since the model used to generate the data is similar to that used for analyzing the data, biases and uncertainties in the estimates that may result from actually fitting the wrong model cannot be assessed. Thus, data obtained through simulation, while often quite informative, must be cautiously interpreted.

We illustrate the use of the hierarchical model we propose on a set of slides obtained in a maize experiment. While we present very limited results here, they serve to highlight some of the improvements that appear to be associated to the use of the three scans available for each slide. When comparing our results to those obtained from fitting the Newton (2001) hierarchical model using only one reading per slide, we note that the variance of expression estimates is lower when based on three readings, as would be expected. We also notice that expression levels are not as shrunken towards the mean expression (2594).
References


7 Appendix

7.1 Marginal likelihood of \((a, a_0, \nu)\)

Consider the joint likelihood function \(p(a, a_0, \nu, \theta, \delta|S)\). The marginal likelihood function of \((a, a_0, \nu)\) \(p(a, a_0, \nu|S)\) is obtained by integrating the likelihood function with respect to \(\delta\) and \(\theta\) as follows:

\[
p(a, a_0, \nu|S) = \int \int p(a, a_0, \nu, \delta|S) d\delta d\theta
\]

\[
\propto \int \int \Gamma(a)^{-n(m+1)} \prod_{i=1}^{n} \theta_i^{a(m+1)+a_0-1} \prod_{i=1}^{m} \delta_i^{n(a-1)+\alpha_1-1} \prod_{i=1}^{n} \prod_{j=1}^{m+1} S_{ij}^{a-1} \times e^{-\sum_{i=1}^{n} \theta_i \sum_{j=1}^{m+1} \delta_j S_{ij} - \nu \sum_{i=1}^{n} \theta_i - \alpha_2 \sum_{j=1}^{m+1} \delta_j} d\delta d\theta
\]

\[
= \Gamma(a)^{-n(m+1)} \prod_{i=1}^{n} \prod_{j=1}^{m+1} S_{ij}^{a-1} \int \prod_{i=1}^{n} \theta_i^{a(m+1)+a_0-1} \times e^{-\sum_{i=1}^{n} \theta_i} d\theta
\]

\[
= \Gamma(a)^{-n(m+1)} \Gamma(n(a-1) + \alpha_1) \prod_{i=1}^{n} \prod_{j=1}^{m+1} S_{ij}^{a-1} \times \int \prod_{i=1}^{n} \theta_i^{a(m+1)+a_0-1} \times e^{-\sum_{i=1}^{n} \theta_i} d\theta
\]

\[
= \Gamma(a)^{-n(m+1)} \Gamma(n(a-1) + \alpha_1) \prod_{i=1}^{n} \prod_{j=1}^{m+1} S_{ij}^{a-1} \times \int \prod_{i=1}^{n} \delta_i^{a(m+1)+a_0-1} \times e^{-\sum_{i=1}^{n} \theta_i} \prod_{j=1}^{m+1} \theta_i S_{ij} - \alpha_2 d\theta.
\]

7.2 Marginal likelihood of \((a, a_0)\) conditional only on largest reading

We can derive the marginal distribution of the hyperparameters \((a, a_0)\) conditional on the highest gene expression reading \(S_{(m+1)}\) as follows:

\[
p(a, a_0|S_{(m+1)}) = \int \int \int p(a, a_0, \theta, \delta, \nu|S_{(m+1)}) d\delta d\theta d\nu
\]

\[
\propto \int \int \int p(S_{(m+1)}|a, a_0, \theta, \nu) p(\theta|a_0, \nu) p(\delta)p(\nu) d\delta d\theta d\nu
\]
7.3 Expectation-Maximization to estimate \((a, a_0)\)

We wish to maximize expression (6) with respect to \((a, a_0)\) using the E-M algorithm. Below we derive expressions to be used in the E- and the M-steps in the algorithm.

E-step:

\[
\begin{align*}
\mathbb{E}(\nu | a, a_0, S_{(m+1)}) &= \int \nu p(\nu | a, a_0, S_{(m+1)}) d\nu \\
&= \int \nu \int p(\theta, \nu | a, a_0, S_{(m+1)}) d\theta d\nu \\
&= C \int \nu \int p(S_{(m+1)} | \theta, \nu | a, a_0) d\theta d\nu \\
&= C \frac{\prod_{i=1}^{n} S_{(m+1)}^{\alpha-1} \beta_2^{\alpha} \Gamma(a + a_0)^n}{\Gamma(\beta_1) \Gamma(a)^n \Gamma(a_0)^n}
\end{align*}
\]
\[ L(a, a_0) = \int \int p(a, a_0, \theta, \delta | \nu, S_{(m+1)}) d \delta d \theta \]

\[ = \int \int p(S_{(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) p(\delta) d \delta d \theta \]

\[ = \int p(S_{(m+1)} | a, \theta) p(\theta | a_0, \nu) d \theta \]

\[ = \int \Gamma(a)^{-n} \prod_{i=1}^{n} \theta_i^{a-1} S_{(m+1)}^{a-1} \exp(-\theta_i S_{i(m+1)}) \]

\[ = \Gamma(a_0)^{-n} \nu^{n a_0} \prod_{i=1}^{n} \theta_i^{a_0-1} \exp(-\nu \theta_i) d \theta \]

\[ = (\Gamma(a) \Gamma(a_0))^{-n} \prod_{i=1}^{n} S_{i(m+1)}^{a-1} \nu^{n a_0} \]

\[ \int \prod_{i=1}^{n} \theta_i^{a+a_0-1} \exp \left( -\sum_{i=1}^{n} \theta_i (S_{i(m+1)} + \nu) \right) d \theta \]

\[ = (\Gamma(a) \Gamma(a_0))^{-n} \prod_{i=1}^{n} S_{i(m+1)}^{a-1} \nu^{n a_0} \prod_{i=1}^{n} \Gamma(a + a_0) (S_{i(m+1)} + \nu)^{-(a+a_0)} \]

\[ = \left( \frac{\Gamma(a + a_0)}{\Gamma(a) \Gamma(a_0)} \right)^n \nu^{n a_0} \prod_{i=1}^{n} S_{i(m+1)}^{a-1} \prod_{i=1}^{n} (S_{i(m+1)} + \nu)^{-(a+a_0)} \]