Comparison of Li–Wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays

Tzu-Ming Chu1,2,∗, B.S. Weir1,2 and Russell D. Wolfinger1,2

1Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA and 2SAS Institute Inc., Cary, NC 27513, USA

Received on April 22, 2002; revised on June 25, 2003; accepted on July 1, 2003
Advance Access publication January 22, 2004

ABSTRACT
Motivation: Li and Wong have described some useful statistical models for probe-level, oligonucleotide array data based on a multiplicative parametrization. In earlier work, we proposed similar analysis-of-variance-style mixed models fit on a log scale. With only subtle differences in the specification of their mean and stochastic error components, a question arises as to whether these models could lead to varying conclusions in practical application.

Results: In this paper, we provide an empirical comparison of the two models using a real data set, and find the models perform quite similarly across most genes, but with some interesting and important distinctions. We also present results from a simulation study designed to assess inferential properties of the models, and propose a modified test statistic for the Li–Wong model that provides an improvement in Type 1 error control. Advantages of both methods include the ability to directly assess and account for key sources of variability in the chip data and a means to automate statistical quality control.

Availability: The Li–Wong models are available in dChip: http://www.biostat.harvard.edu/complab/dchip/, and both methods will be commercially available in the forthcoming SAS Microarray Solution.

Contact: tzu-ming.chu@sas.com

Supplementary information: Supplementary material is available at http://statgen.ncsu.edu/ggibson/Pubs.htm

INTRODUCTION
The Affymetrix GeneChip™ is currently the most widely used commercial expression array technology. A typical chip contains 11–20 oligonucleotide 25mer probes for each of thousands of genes. Although each oligonucleotide within a probe set is designed to interrogate the same gene, there are well-known and very strong differences between the performance of individual probes, as well as the potential for probes to interact in different ways with other sources of variation. Biologically sound statistical analysis of the probe-level data is therefore a critical challenge that must be addressed in order to effectively assess results from the chips.

Affymetrix itself has been responsive by providing a new summary measure based on Tukey’s biweight function in the software accompanying the chips (www.affymetrix.com/product/). Although this certainly represents an improvement over previous methods, by its nature the Affymetrix summary prevents analysts from making their own adjustments for individual probe effects. Research addressing this concern has arisen from numerous sources, including Teng et al. (1999), Schadt et al. (2000), Efron et al. (2001), Lemon et al. (2002); Irizarry et al. (2003) etc. and we forgo an attempt to provide a comprehensive review here.

What we do attempt to accomplish is to provide an empirical comparison between one of the first and most compelling approaches, that of Li and Wong (2001a,b), and the more recent recommendations of Chu et al. (2002). Li and Wong’s models for the probe-level measurements are a combination of multiplicative and additive terms, and have been applied successfully in a number of contexts (see references on the dChip web site). A key advantage of such an approach is a direct, parametric specification of important sources of variability in the chip data. This enables rigorous statistical quality control by automatically flagging observations that significantly deviate from the model, as well as quantified statistical inference about experimental effects. With similar advantages, Chu et al. (2002) describe a systematic statistical linear modeling approach based on the mixed linear model of the logarithms of the probe-level data.

So how do the Li–Wong and loglinear mixed models compare in practice? To investigate this question, we first consider a data set from the literature (Tusher et al., 2001) and compare two specific forms of the models applicable for these data in terms of their theoretical structure, their goodness-of-fit, and their standardized residuals. We then conduct a simulation study to investigate statistical inference properties of

∗To whom correspondence should be addressed.
the models, and propose a new test statistic for the Li–Wong model that provides improved control of Type I error.

**SYSTEM AND METHODS**

**The ionizing radiation data and associated Li–Wong and mixed models**

We consider the models in the context of the ionizing radiation data of Tusher et al. (2001). These data are from eight HUGeneFL GeneChips™, arising from two replicates of a $2 \times 2$ factorial design (Table 1). Measurements for 6810 genes are available from each chip, and for each gene, 40 values corresponding to 20 perfect match and mismatch pairs arise from quantification of a fluorescently derived image. The perfect match and mismatch values are usually subtracted to form a paired difference, although there is considerable debate about how the mismatch data should be incorporated. For simplicity of exposition and comparison, we consider only the 160 perfect-match data points for each gene, and do not perform any normalization on the data prior to analysis.

The Li–Wong model we consider is as follows:

$$[\text{LW}] \quad PM_{ijkl} = v_k + \theta_{ij} \phi_k + \gamma_{ijkl}, \quad \sum \phi_k^2 = K.$$  

Here, $PM_{ijkl}$ denotes the perfect match expression measurement of the $i$th cell line receiving the $j$th treatment at the $k$th probe in the $l$th replicate. The parameter $v_k$ is the baseline response of the $k$th probe. $\theta_{ij}$ is an expression index for the different samples, $\phi_k$ is the multiplicative effect of the $k$th probe, and $\gamma_{ijkl}$ is the stochastic error term. The $\gamma_{ijkl}$ are assumed to be independent and identically distributed normal random variables with mean 0 and variance $\omega^2$. The summation constraint is imposed for identifiability, with $K$ equal to the total number of probes within a probe set.

The mixed analysis-of-variance model we consider here, based on Chu et al. (2002), is as follows:

$$[\text{MM}] \quad \log_2(PM_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{ij} + \epsilon_{ijkl}.$$  

The indices are the same as before, and the symbols $L$, $T$, $LT$, $P$, $LP$, $TP$ and $A$ represent cell line, treatment, cell-line-by-treatment interaction, probe, cell-line-by-probe interaction, treatment-by-probe interaction and chip effects, respectively. They are the analogs of the $v$, $\theta$, $\phi$ parameters in [LW]. The $A_{ij}$ are assumed to be independent and identically distributed normal random effects with mean 0 and variance $\sigma^2_A$, and induce a common correlation across all observations on the same chip. The $\epsilon_{ijkl}$ are assumed to be independent and identically distributed normal random variables with mean 0 and variance $\sigma^2$, and are independent of the $A_{ij}$.

Note that [MM] has 65 degrees of freedom in its mean model, whereas [LW] has 47. It is possible to adjust either model to make these numbers closer to each other, but, as specified, they represent what we reckon analysts would most likely use in practice. A principal structural difference is that the stochastic errors are additive in [LW] and multiplicative in [MM] (in the original scale). Since the data are counts of pixel intensities, one might expect their variance to be proportional to their mean, providing some motivation for the latter approach; however, since we are fitting the models to data for only one gene at a time, it is not obvious which form might better track the noise in the data. The cell-line-by-probe interaction effect in [MM] is interesting because it can detect potential divergences between the RNA of the cell lines as compared to the oligos on the chip. The treatment-by-probe interaction is more difficult to justify a priori and could potentially be deleted, but is included to check whether or not the probes are telling a consistent story about the change induced by the ionizing radiation.

**ALGORITHM**


To conduct statistical inference with [LW], we use the confidence interval approach as described in Li and Wong (2001b). As noted in Li and Wong (2001a), this conditional method should be valid for a large number of arrays since the probe-specific parameters, $v_j$s and $\phi_j$s, can be estimated accurately. However, for small experiments like the ionization radiation data, an unconditional approach may be more appropriate. To this end, we also investigate a Wald Z statistic (Wald, 1943) for [LW], as derived in the Supplementary information. Statistical inferences for [MM] are conducted via approximate $t$-tests derived from standard mixed-model theory (Searle et al., 1993; Littell et al., 1996; SAS Institute Inc., 1999).

---

**Table 1. Experimental design for the ionizing radiation data**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell line</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Replicate</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

*a 1' and '0' indicate irradiated and unirradiated treatments, respectively.*
**IMPLEMENTATION**

**Goodness-of-fit**

Figure 1A shows the regression of the model-based fitted values from [LW] on those from [MM] for data from all of the 6810 genes. (The fitted values from [MM] are exponentiated by 2 for comparability.) The fitted values from both models are consistently close except for a few points in the lower off-diagonal.

Figure 1B plots the $R^2$ values from the model fits for each gene, defined as

$$R^2 = 1 - \frac{\sum (P_{ijkl} - \hat{P}_{ijkl})^2}{\sum (PM_{ijkl} - PM)^2},$$

where $PM$ is the average of all $PM_{ijkl}$s and $\hat{P}_{ijkl}$ is the model-based fitted value of $P_{ijkl}$. The lower 5th percentile of $R^2$ equals 0.96 for both models, indicating excellent fit in most of the cases. For a few genes the $R^2$ values of [MM] are much less than the corresponding $R^2$ values of [LW]. We inspected the expression profiles of those 66 genes that have $R^2$ values for [MM] less than 0.95 and at least 0.1 less than the respective $R^2$ values for [LW], and all of the 66 cases have outlying observations.

Figure 2A plots the raw data and model fits for gene number 3096 (U35451), which has [LW] $R^2 = 0.989$ and [MM] $R^2 = 0.171$. An apparent PM outlier is for probe 18 in array ‘112’ (note the change in scale in the y-axis for this array’s plot). [LW] fits this outlier very well whereas [MM] does not, leading to the discrepancies in $R^2$. This is a direct result of the model specifications, since [LW] has distinct parameters for individual replicates whereas [MM] averages across replicates. Averaging may be more appropriate in this case, as it allows one to automatically detect this outlier and remove it from the analysis. Figure 2B for gene number 1860 (M25753) represents a converse situation, for which [LW] $R^2 = 0.987$ and [MM] $R^2 = 0.991$. Note the PM profiles of the probes are very consistent throughout all eight arrays except for probe 1, which is relatively low in untreated arrays but relatively high in treated arrays. This is a treatment-by-probe interaction, and is captured by [MM] but not by [LW]. Under a conservative Bonferroni correction for multiple testing to ensure a false positive rate less than 0.05, only 0.59% of the probe sets have a significant treatment-by-probe interaction effect, according to [MM], whereas 9.34% have a significant cell-line-by-probe interaction effect.

**Outlier detection and normality diagnostics**

Figure 3A and B show standardized residual plots for all of the model fits of [LW] and [MM], respectively (the [LW] fitted values are rescaled for comparison purposes). The standardized residuals are the differences of observed and fitted values, divided by the square root of estimated variance. The diagonal truncation on the right of both graphs results from data censored at the upper detection limit, and decile density contours are superimposed to indicate the distribution of the points. In Figure 3A, the [LW] standardized residuals fall roughly in an oval shape, having the largest spread near the middle of the fitted values. The [MM] residuals in Figure 3B, in contrast, decrease roughly linearly in variance throughout the primary range of the data and have contours that are much more triangular. These displays provide some visual clues as to how the two models differ across a wide range of potential probe intensities.

A common rule of thumb for outlier detection from the quality control literature is based upon ‘six sigma’; i.e. delete all data points with standardized residuals greater than 3 in magnitude. With this criterion, the proportion of outliers for all genes is 0.567 and 0.077% for [LW] and [MM], respectively. The order of magnitude difference in these percentages can largely be attributed to the different scales on which the models operate.

As a check for normality, we applied the Kolmogorov–Smirnov test with Bonferroni adjustment to the standardized
Comparison of Li–Wong and loglinear mixed models

Fig. 2. (A) Expression profiles from gene 3096 (U35451). The dots, circles and squares indicate raw PM data, fitted values from [LW], and exponentiated fitted values from [MM], respectively. Those circles and squares within single probe set are jointed by black dotted lines and gray solid lines, respectively. The x-axis represents the probe number. The three digits on the top right corner of each plot indicate the experimental condition applied to GeneChip. The first digit indicates the two different cell lines (1 or 2). The second digit indicates the treatment condition (1: treated, 0: untreated). The third digit indicates the replicate number (1 or 2). Note the change in scale in the y-axis for array 112; others are consistent. (B) Profiles from gene 1860 (M25753).

residuals of each gene. Almost all genes (97.47%) pass the test in [MM], but only three-fourths of genes (74.24%) pass the test in [LW].

Simulations
To investigate the inferential operating characteristics of [LW] and [MM], we conducted a simulation study using parameter values estimated from the data for gene number 2863 (U18300) from the ionizing radiation data. This gene has \( R^2 \) values of 0.949 and 0.986 from [LW] and [MM], respectively. The corresponding log-transformed estimates of the variance components are \(-1.12, -1.10 \) and \( 5.17 \) for \( \hat{\sigma}^2, \hat{\alpha}_k^2 \) and \( \omega^2 \), respectively. Histograms across all genes of these parameter estimates (data not shown) reveal that these estimated values for gene 2863 fall near the middle of their respective distributions, and so gene 2863 is ‘typical’ in this respect.

Scenario 1—[LW] is the true model
Ten simulation cases were created as follows:

1. Obtain the estimates of \( \nu_k, \phi_k, \theta_i \) and \( \omega^2 \) by fitting [LW] to the data from gene 2863, and set \( \theta_{\mu} \) equal to the average of all the \( \theta_i \) estimates.
2. Set the scale parameter \( \delta \), which represents the fold change, to one of 10 different values ranging from 1 to 2.
3. Set the parameters \( \theta_1, \theta_2, \theta_5 \) and \( \theta_6 \) equal to \( \theta_{\mu} \) and parameters \( \theta_3, \theta_4, \theta_7 \) and \( \theta_8 \) equal to \( \delta \times \theta_{\mu} \).
4. Use all of the estimates of \( \nu_k, \phi_k, \omega^2 \) and the new parameters \( \theta_i \) in step 3 as the true parameters for [LW].
5. For each value of the scale parameter \( \delta \), generate 2000 simulated data sets.

These settings create 10 simulation cases of different fold changes, ranging from 1 to 2, between treatment groups and no changes between cell line groups. Therefore, testing the
Table 2. Simulation results from using [LW] as the true model—the rejection rates of 2000 simulations are displayed for each case, using a nominal 0.05 cutoff.

<table>
<thead>
<tr>
<th>Fold</th>
<th>[MM]</th>
<th>[LW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>M_l</td>
<td>M_t</td>
<td>M_lt</td>
</tr>
<tr>
<td>1</td>
<td>0.049</td>
<td>0.041</td>
</tr>
<tr>
<td>1.05</td>
<td>0.049</td>
<td>0.156</td>
</tr>
<tr>
<td>1.1</td>
<td>0.050</td>
<td>0.468</td>
</tr>
<tr>
<td>1.15</td>
<td>0.048</td>
<td>0.779</td>
</tr>
<tr>
<td>1.2</td>
<td>0.050</td>
<td>0.930</td>
</tr>
<tr>
<td>1.3</td>
<td>0.051</td>
<td>0.997</td>
</tr>
<tr>
<td>1.4</td>
<td>0.051</td>
<td>1</td>
</tr>
<tr>
<td>1.5</td>
<td>0.051</td>
<td>1</td>
</tr>
<tr>
<td>1.75</td>
<td>0.053</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.053</td>
<td>1</td>
</tr>
</tbody>
</table>

'M', 'CI', and 'W' indicate the test is performed by [MM], [LW] confidence interval, and [LW] Wald Z approaches, respectively. The characters 'l', 't', 'lt', 'p', 'lp' and 'tp' indicate the main or interaction effects of line, treatment, line-treatment, probe, line-probe and treatment-probe, respectively. R^2_MM and R^2_LW indicate the average R^2 values for models [MM] and [LW], respectively.

Scenario 2—[MM] is the true model

In order to have ten cases comparable with Scenario 1, we set the [MM] parameters as follows:

1. Obtain the estimates for \( P_k \), \( \sigma^2 \) and \( \sigma_a^2 \) by applying the Mixed procedure to the logarithms (base 2) of the data from gene 2863.

2. Set the scale parameter \( \delta \) to 10 different values from 1 to 2.

3. Set the parameters \( L_i \), \( LT_{ij} \), \( LP_{ik} \) and \( TP_{jk} \) to 0.

4. Set parameter \( T_1 \) to 0 and parameter \( T_2 \) to \( \log_2(\delta) \).

5. Put all estimates and parameters in steps 1, 3 and 4 into [MM].

6. Change the scale parameter and generate 2000 simulated data sets in each case.

Again, the line effects are tested to examine the false positive rate, and treatment effects are tested to examine power.

Simulation results

Table 2 and the left three plots in Figure 4 show the results from Scenario 1. Table 3 and the right three plots in Figure 4 show the results from Scenario 2. The curves in Figure 4 indicate the results from the confidence interval tests in [LW] (long black dashed curves) the Wald Z tests in [LW] (black solid curves) and tests in [MM] (gray curves). From the top two plots in Figure 4, the tests performed by [MM] control the nominal significance level (0.05) regardless of which model is true. The [LW] confidence interval approach is significantly too liberal; its 95% confidence band lies above 0.05 in all 10 cases where cell line effects allow us examine how well the tests control Type I error (false positive rate). Testing the treatment effects allows us to simultaneously examine power. We set the nominal significance level for the tests to 0.05.

Fig. 3. (A) Standardized residuals from [LW] plotted against fitted values of the ionizing radiation data; with decile contours superimposed. (B) Same plot for [MM].
Comparison of Li–Wong and loglinear mixed models

Fig. 4. Comparison of simulation results for Scenarios 1 and 2. The top two plots compare the Type 1 Error (false positive) rate for testing line effect. The short dashed curves are 95% confidence bands. The middle two plots compare power for testing the treatment effect. The bottom two plots compare average $R^2$ values. The three plots on the left are results from the simulation using [LW] as the true model, and the three plots on the right are those using [MM] as the true model. Gray curves represent the results from [MM], long black dashed curves from [LW] and black solid curves from the proposed [LW] Wald Z. The $X$-axes in all six plots represent the $\delta$-value (treatment fold change). The $Y$-axes in the top four plots represent the testing rejection rate, and those in the bottom two plots represent $R^2$ values.

[LW] is true. When [MM] is true, the simulated false positive rate is even worse and exceeds 0.2 for larger fold changes. The Wald Z in [LW] appropriately controls the significance level under [LW], but rises above 0.15 when [MM] is true.

The middle two plots of Figure 4 show power curves for testing treatment fold change. The plots of the [LW] confidence interval tests are not strictly comparable here because these tests are too liberal in their Type 1 error rates. The [MM] and [LW] Wald Z have similar power under the Scenario 1, with power near 100% for a small 1.3-fold change. This implies that [MM] can perform well as [LW] even when [LW] is the true model. In the middle right plot of Figure 4, the [MM] tests have the best power, as expected, since [MM] is the true model. The bottom two plots in Figure 4 show the average $R^2$ values for the two models in the 10 simulation cases. Under the first scenario, the average $R^2$ value for [MM] is about 0.014 lower than the average $R^2$ value for [LW]. Under the second scenario, the average $R^2$ value for [MM] is about 0.005 higher than the average $R^2$ value for [LW]. All values are about 0.92, indicating excellent goodness-of-fit.
Table 3. Simulation results from using [MM] as the true model—the rejection rates of 2000 simulations are displayed for each case, using a nominal 0.05 cutoff (notation is the same as in Table 2)

| Fold | [MM] M_l | M_t | M_lt | M_p | M_lp | | [LW] CI_l | CI_t | W_l | W_t | R^2_MM | R^2_LW |
|------|---------|-----|------|-----|------| |        |      |     |    |     |        |        |
| 1    | 0.055   | 0.056 | 0.058 | 1 | 0.049 | 0.051 | 0.189 | 0.208 | 0.093 | 0.108 | 0.977  | 0.971 |
| 1.05 | 0.049   | 0.855 | 0.055 | 1 | 0.049 | 0.055 | 0.203 | 0.954 | 0.085 | 0.598 | 0.977  | 0.971 |
| 1.1  | 0.056   | 1    | 0.051 | 1 | 0.050 | 0.046 | 0.207 | 1     | 0.092 | 0.775 | 0.978  | 0.971 |
| 1.15 | 0.065   | 1    | 0.048 | 1 | 0.048 | 0.042 | 0.189 | 1     | 0.103 | 0.916 | 0.978  | 0.972 |
| 1.2  | 0.043   | 1    | 0.048 | 1 | 0.054 | 0.052 | 0.205 | 1     | 0.12  | 0.965 | 0.979  | 0.972 |
| 1.3  | 0.063   | 1    | 0.046 | 1 | 0.053 | 0.058 | 0.207 | 1     | 0.14  | 0.993 | 0.978  | 0.974 |
| 1.4  | 0.058   | 1    | 0.051 | 1 | 0.048 | 0.044 | 0.211 | 1     | 0.144 | 1     | 0.981  | 0.976 |
| 1.5  | 0.046   | 1    | 0.057 | 1 | 0.051 | 0.045 | 0.247 | 1     | 0.173 | 1     | 0.982  | 0.978 |
| 1.75 | 0.043   | 1    | 0.043 | 1 | 0.052 | 0.05  | 0.224 | 1     | 0.154 | 1     | 0.985  | 0.981 |
| 2    | 0.055   | 1    | 0.06  | 1 | 0.054 | 0.043 | 0.241 | 1     | 0.166 | 1     | 0.987  | 0.984 |

DISCUSSION
We have empirically compared instances of two statistical modeling approaches, Li–Wong and loglinear mixed models, for oligonucleotide expression array data at the probe level. Both models fit both real and simulated data very well, and have great ability to capture the key measurable sources of variability of oligonucleotide arrays. In our real-data example, we did find data from genes that differentiate the two methods, and were able to provide reasonable explanations for the differences based on the particular models we selected. Of course both models can be adjusted to accommodate different kinds of data patterns, but the mixed model is operationally more convenient because of its linearity and rich tradition in handling sources of variation expressed as effects and their interactions. The mixed model is also easily applicable to experimental designs more complex than simple factorials, whereas the basic Li–Wong framework is nonlinear and operates within a one-way analysis-of-variance treatment framework. The two models naturally differ in their outlier selection criteria because they are on different scales. According to our simulations, the conditional Li–Wong confidence interval method has excessive Type I error rates, but these can be corrected by using larger, unconditional standard errors. The mixed model performed well in all cases, even when the simulated data arose from a Li–Wong model.

ACKNOWLEDGEMENTS
This work was supported in part by NIH grant GM 45344.

REFERENCES


