On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles

C.M. Kendziorski, Ph.D.
M.A. Newton, Ph.D.
H. Lan, Ph.D.
M.N. Gould, Ph.D.
1 Introduction

Enabled by resources created from genome sequencing projects, DNA microarray technology has emerged as a fundamental measurement tool in the study of diverse biological systems. Microarrays offer an unprecedented ability to perform large-scale studies of gene expression. As a result, the focus of many research efforts has shifted from individual genes to multiple genes and the complicated and orchestrated ways in which they interact to maintain life.

With the shift from individual to integrated analysis in molecular biology comes a shift in the related statistical questions posed and methods required. The number of measurements of distinct genes across an array greatly exceeds that for any individual gene. Thus, we as statisticians are faced with the “large p, small n” paradigm (West et al. 2000a, 2000b). Empirical Bayes methods provide a natural approach to microarray data analysis because they can reduce significantly the dimensionality of an inference problem involving many unknown parameters (e.g., Efron and Morris 1973, 1977). Our earlier work described a version of parametric empirical Bayes analysis for spotted microarrays and was restricted to so-called “single-slide” data in which each gene produces two measurements, one from each cell condition (Newton et al. 2001). The methodology provides improved estimation of expression fold-change and allows for the assessment of differential expression by the calculation of a posterior odds. In spite of there being very little data per gene, the methodology works because inference about a given gene uses information on the fluctuations of expression measurements from all genes. One goal of the present paper is to extend the parametric empirical Bayes calculations beyond the single-slide case, and thus to allow replicate expression profiles in multiple cell conditions.

The methodological work presented here is motivated in part by an experiment to study gene expression in a rat model of breast cancer. Microarray data were obtained from mammary epithelial cells harvested from 12 week old females representing four distinct inbred lines (two parentals and two congenic lines; see Section 5). The parental strains
method proposed by Newton et al. (2001) amounts to a simple two group mixture-model calculation. Stochastically, each gene is either differentially expressed or not. Those genes which are not present data according to some background distribution, and those which are present data according to a different distribution. The specific forms of these distributions arise by another layer of mixing over the latent mean expression level for each gene. In that work, the expression measurements are independent and follow a Gamma distribution conditional upon the latent mean expression level. The Gamma model is convenient numerically and analytically, but also has some justification in the modeling of abundances in a large population. The latent mean values are treated not as fixed effects (as they would be in the standard analyses outlined above) but follow the conjugate, inverse-Gamma distribution. Two measurements that happen to have the same latent mean value represent equivalent expression; otherwise there is differential expression. With these components in place, inference about differential expression amounts to computing the odds of that event, conditional on the measurements. The analysis is empirical Bayes because the small number of unknown parameters which index the component distributions are estimated from the data. In Section 2 we describe an extension of this approach to replicate profiles in multiple conditions.

There are other mixture-modeling approaches to expression data analysis. Working with a specific experimental design, Efron et al. (2000, 2001) describe empirical Bayesian calculations which relax the parametric assumptions. After a long series of pre-processing steps, each gene yields a one-dimensional test statistic whose marginal distribution turns out to be known and whose null distribution (i.e., on equivalent expression) can be nonparametrically estimated. Lee al. 2000 also use the idea of a two group mixture model for expression analysis; their calculations were in a slightly different context and were applied to parameter estimates from a first-stage analysis. Here we do not endeavor to extend either of these approaches to the case of multiple conditions, but in Section 4 we do offer some
in the sampled cells. Let us initially consider comparing two conditions, with data $x_g = (x_{g,1}, x_{g,2}, \ldots, x_{g,n_1})$ from the $n_1$ replicate measurements in the first condition and $y_g = (y_{g,1}, y_{g,2}, \ldots, y_{g,n_2})$ from the second condition. Two levels of mixing characterize the distribution of these data.

One stage of mixing is discrete and captures the different patterns of differential expression. On a null hypothesis there is equivalent expression (EE) between the two conditions, and the data arise from a joint probability density (pdf) $f_0(x_g, y_g)$. Alternatively there is differential expression (DE), and the joint pdf is $f_1(x_g, y_g)$. A priori we do not know which situation manifests itself for gene $g$, and so we introduce the discrete mixing parameter $p$ to denote the unknown probability of differential expression. Thus, the marginal distribution of the data is

$$p f_1(x_g, y_g) + (1 - p) f_0(x_g, y_g).$$

(1)

If we know the parameter $p$ and the form of $f_0$ and $f_1$ then by Bayes rule, the posterior probability of differential expression may be calculated:

$$p \frac{f_1(x_g, y_g)}{p f_1(x_g, y_g) + (1 - p) f_0(x_g, y_g)}.$$

Equivalently, one could consider the posterior odds of DE:

$$\text{odds}_g = \frac{P(\text{DE}|x_g, y_g)}{P(\text{EE}|x_g, y_g)} = \frac{p}{1 - p} \frac{f_1(x_g, y_g)}{f_0(x_g, y_g)}.$$

We use these summary measures to enable gene-specific inferences concerning differential expression. (Though it is masked in the notation, events DE and EE are specific to the gene $g$.)

The calculations are further specified by a second stage of mixing so that we obtain particular forms for the pdfs $f_0$ and $f_1$. Intuitively, the measurements $x_g$ and $y_g$ should tend to be closer together in $f_0$ since there the variations do not include any systematic shifts between the cell types. The situation is naturally characterized by mixing over
conditions, \( \binom{3}{0} + \binom{3}{1} + \binom{3}{2} = 5 \) expression patterns are possible. These include equivalent expression across the three conditions, altered expression in just one condition, and distinct expression in each condition. With microarrays from four cell conditions there are 15 different patterns. (The total number of patterns is equal to the Bell exponential number of possible set partitions, in fact.) As we see in the rat mammary study which has four cell conditions, we can sometimes reduce the total number of patterns to a more manageable level, and in that case we reduce to four interesting patterns (Section 5).

Suppose that \( m + 1 \) distinct patterns of expression are possible for a data vector \( d_g = (d_{g,1}, \ldots, d_{g,N}) \) on some gene \( g \). Then, generalizing (1), \( d_g \) is governed by a mixture of the form

\begin{equation}
\sum_{k=0}^{m} p_k f_k(d_g),
\end{equation}

where \( \{p_k\} \) are mixing proportions and component densities \( \{f_k\} \) give the predictive distribution of measurements for each pattern of expression. Consequently, the posterior probability of expression pattern \( k \) is

\begin{equation}
P(k|d_g) \propto p_k f_k(d_g).
\end{equation}

Furthermore, generalizing (2), the pattern-specific predictive density \( f_k(d_g) \), for \( k > 0 \), will be a product across subsets of the data vector, say,

\begin{equation}
f_k(d_g) = \prod_{s \in S} f_0(d_{g,s})
\end{equation}

where \( S \) is a set partition of \( \{1, 2, \ldots, N\} \) constructed in such a way that any measurements contained in a component subset \( s \) in \( S \) share a common mean value, \( \mu_s \).

The posterior probabilities summarize our inference about expression patterns at each gene. They can be used to identify genes with altered expression in at least one group, to classify genes into distinct expression groups, or to order genes within groups. Before posterior summaries can be evaluated, however, we must first specify distributional forms for the components of the hierarchical mixture model.
where
\[
K' = \frac{\nu_0^2 \Gamma(n_1\alpha + \alpha_0) \Gamma(n_2\alpha + \alpha_0)}{\Gamma(\alpha_0) \Gamma(N\alpha + \alpha_0)}.
\]
and recall that \(N = n_1 + n_2\) is the total number of observations on gene \(g\). The odds may be computed as soon as we have estimates in hand for \(\theta = (\alpha, \alpha_0, \nu)\). In Section 6 we point out an interesting connection between these posterior odds and the arithmetic-geometric mean inequality.

The GG calculations derived above extend those presented in Newton et al. 2001 to replicates and multiple conditions. Many investigators would consider as reasonable a different model for the array measurements— one in which the log-transformed measurements have a Gaussian observation component. We may use this in our hierarchical mixture model as follows. Let us say the natural logarithms of the measurements are denoted \(\bar{x}_g\) and \(\bar{y}_g\). The latent gene-specific mean \(\mu_g\) is now a mean for the log-transformed measurements, and these measurements have a sampling variance \(\sigma^2\) which we treat as common to all genes. Note that the coefficient of variation for the original measurements becomes \(\sqrt{\exp(\sigma^2) - 1}\) in this model. A conjugate prior for the \(\mu_g\) is normal with some underlying mean \(\mu_0\) and variance \(\tau_0^2\). Integrating as in (3), the joint predictive density \(f_0\) for an \(n\)-dimensional input becomes Gaussian with mean vector \(\mu_0 = (\mu_0, \mu_0, \ldots, \mu_0)^t\) and exchangeable covariance matrix
\[
\Sigma_n = \left(\sigma^2\right) I_n + \left(\tau_0^2\right) M_n
\]
where \(I_n\) is an \(n \times n\) identity matrix and \(M_n\) is an \(n \times n\) matrix of ones. This basic formulation has been well studied (e.g., Carlin and Louis, 1996). In our context there is an additional layer of discrete mixing, and we may derive the posterior probability of different expression patterns following (6). For the special case of two conditions, the odds of differential expression (4) may be written in terms of quadratic forms. Let \(\delta_g = (\bar{x}_g, \bar{y}_g)^t - \mu_0\) denote the centered, transformed full data vector for gene \(g\),
\[
\text{odds}_g = \sqrt{\frac{|\Sigma_N|}{|\Sigma_*|}} \exp \left\{ -\frac{1}{2} \delta_g^t (\Sigma_*^{-1} - \Sigma_N^{-1}) \delta_g \right\}
\]
different from the one which generated the data, and perhaps most importantly, they provide
information on error rates in the inference of differential expression.

First, we simulated the GG model with 10,000 genes in two conditions, having three
replicates in each condition. We took model parameters similar to those obtained in Newton
et al. 2001 (\(\alpha = 10, \alpha_0 = 0.9,\) and \(\nu = 0.5\)). The prior probability that a gene is differentially
expressed was set to \(p = 0.2\). The GG and LNN mixture models described in Section 3 were
each fit to these simulated data. Histograms of the simulated data along with the fitted
marginal densities are shown in the left panel of Figure 1. As expected, the fitted GG
marginal density more closely describes the simulated data.

Next, we simulated a similar data set under the LNN model (\(\mu_0 = 2.3, \sigma = 0.3,\) and \(\tau_0 =
1.39\)); each mixture model was again fit to the simulated data. As shown in the right panel
of Figure 1, the simulated data is better described by the LNN density. Although expected,
this result illustrates that comparing the marginal densities to the empirical distribution can
give insight into which model assumptions are more appropriate.

We did a more formal comparison of GG and LNN by calculating a log Bayes factor to
measure the relative fit of these models (Kass and Raftery 1995). We take this simply as the
difference of the log predictive densities (given by equation 9) calculated under GG or LNN
assumptions. For each simulated data set, the Bayes factor correctly identified the model
generating the simulated data. Considering the success of either approach in identifying the
underlying model, one might think that the parametric assumptions have a substantial effect
on which genes are identified as differentially expressed. We find that this is not the case.

The differences in the simulated data which allow for model identification do not seem to
impact the mixture model's ability to identify differentially expressed genes. Figure 2 shows
the average intensities (across replicates) for spots identified as changed (odds > 1) using the
GG or LNN model applied to GG or LNN data. The odds plots look similar within simulated
data set, independent of model assumption. The numerical results are in fact similar (see
The FDR estimates suggest that using an odds value greater than one as a rejection rule results in a type I error rate near 0.05. Interestingly, the estimates of the FDR are similar to those reported by Efron et al., 2000, in assessment of their empirical Bayes approach. A lower bound on $p$ for the data set considered there is estimated to be 0.189. The authors consider the FDR rates using the posterior probability of differential expression at values greater than and equal to 0.9. This corresponds to an odds $> 9$. They report an FDR of 0.0048 at this level. Our results are similar. For $p = 0.2$ and odds $> 9.1$, the FDR averaged over 100 simulations was 0.0054 (GG on GG), 0.0052 (GG on LNN), 0.0057 (LNN on GG), and 0.0049 (LNN on LNN); standard errors were all less than 0.0003.

5 Data Analysis

Rat strains vary greatly in their resistance to carcinogen-induced mammary cancer. The inbred Copenhagen (COP) strain is almost completely resistant to carcinogenesis induced by DMBA, while the inbred Wistar Furth (WF) strain is highly susceptible (Gould et al. 1989). By careful breeding, intermediate inbred lines can be produced which carry the homozygous WF/WF genotype throughout the genome except on a relatively small and interesting region where the animals are homozygous COP/COP. Such animal populations are referred to as congenic lines (Figure 4). The size of the homozygous COP/COP region is approximately 30 cM in congenic line CI and 1.5 cM in congenic line CII.

In this experiment, we are interested in the identification of genes that are differentially regulated among the parental strains (COP and WF) and the derived congenic lines (CI and CII) in mammary epithelial cells. By a standard protocol, mammary epithelial cells were harvested from untreated 12 week old females. Messenger RNAs were extracted, prepared, and then probed using a set of three Affymetrix Rat Genome U34 chips. In most cases, these mRNAs were pooled from samples of four genetically identical animals to reduce animal to animal variation. Intensity measurements were obtained for 26,379 genes recorded on 10
were 24,119 (null), 217 (pattern 1), 51 (pattern 2), and 78 (pattern 3) genes.

Three genes identified as pattern 3 by the GG model are shown in Table 7. Two of these genes (J00801 and L08100) are known markers of mammary gland differentiation, and a common belief is that differentiation protects against tumor development. For each of these genes, the average intensity in the WF condition is higher than that observed in the COP or congenic lines. This indicates increased expression (and increased differentiation) in the WF, which is unexpected since the WF strain is tumor susceptible. It may be the case that not all forms of differentiation are associated with resistance. Preliminary data in other rat strains and other experiments are supporting this hypothesis (Gould, unpublished data). The third gene (J00772) is rat prostatein. Recent work suggests that this gene, normally associated with the ventral prostate, is strongly expressed in the stromal cells of the rat mammary gland (Watson and Gould, unpublished data). The GG calculations classify this gene as having elevated expression in WF, but the LNN calculations are equivocal, and consider it to be unchanged. Further study of this gene is warranted.

As a separate calculation, we analyzed the data from the WF and COP parentals only, omitting the congenics. Table 6 contains parameter estimates. The odds calculation under GG assumptions estimates 58 genes to be differentially expressed. Of these, 57 are also identified by the LNN model. Figure 5 gives a graph of the average intensities (across replicates) for the spots identified as changed in the GG model. These results are consistent with the multiple group analysis. Each of the 58 genes identified as differentially expressed in the two group analysis is also identified when comparing multiple groups. 48 of the 58 genes have posterior probability larger than 0.5 of being in pattern 1, 4 of the genes are in pattern 2, and 5 of the genes are in pattern 3. One gene was not classified. For both the four and two group analysis, Bayes factors indicated that the GG model fits better than the LNN model.
error rate is inherent to this empirical Bayes approach. Virtually identical numerical results were obtained under LNN assumptions. In addition, there was much overlap between specific genes identified using either approach.

These results suggest that error rates are reasonably low and that particular modeling assumptions might have only a minimal impact on the accurate identification of differentially expressed genes. We note that such results are preliminary, and further work is required before any such conclusions can be made in general. Only two groups having a fixed number of replicates in each group were considered in our simulation study. The study could be extended to evaluate error rates in the case of multiple conditions for a varying number of replicates. Additional model forms should also be considered, both for data simulation and odds calculations. We are currently investigating the effects of nonparametric assumptions on the latent mean distribution $\pi(\mu_g)$.

The method assumes that intensity measurements approximate some true underlying expression level. Thus, expression profiles must be normalized in such a way so that any systematic sources of variation have been removed. DNA Chip Analyzer (Li and Wong, 2001) was used here, but many other methods are available. We also note that mRNA samples were pooled across subjects. Of course, under some conditions, this can decrease measurement variability, thereby reducing the number of replicates required. However, owing to array specific effects, pooling does not eliminate the need for replication. Both Kerr et al. 2000 and Lee et al. 2000 stress the importance of replication in microarray studies. In addition to array effects, if outliers (e.g. contaminated samples) are present, pooling can lead to biased estimates of underlying expression. Optimal experimental designs which provide for maximum measurement accuracy using a minimum number of arrays have yet to be developed. This is an area that requires further investigation.

Finally, we note an interesting statistic which emerges from the odds of differential expression in the GG model (8) when the number of replicates $n_1 = n_2 = n$ per group
Acknowledgements: Code and data used in this article are available at the first author’s web site www.biostat.wisc.edu/~kendzior/. This research was funded in part by training grant TA-CA 09565 for CMK, the NIH R01 grants CA64364 to MAN, and CA28954 and CA77494 to MNG.

References


Table 2: Summary of parameter estimates for GG model applied to GG simulated data. Parameter estimates are averaged over 100 simulations; standard error is shown in parentheses. For each simulation, 
\((\alpha, \alpha_0, \nu) = (10, 0.9, 0.5)\).

<table>
<thead>
<tr>
<th>(p)</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\alpha})</td>
<td>10.001 (0.0098)</td>
<td>9.997 (0.0087)</td>
<td>9.995 (0.0104)</td>
<td>9.993 (0.0099)</td>
<td>10.009 (0.01)</td>
</tr>
<tr>
<td>(\hat{\alpha}_0)</td>
<td>0.900 (0.0016)</td>
<td>0.900 (0.0015)</td>
<td>0.897 (0.0015)</td>
<td>0.900 (0.0014)</td>
<td>0.901 (0.0013)</td>
</tr>
<tr>
<td>(\hat{\nu})</td>
<td>0.499 (0.0012)</td>
<td>0.500 (0.0011)</td>
<td>0.500 (0.0011)</td>
<td>0.500 (0.0012)</td>
<td>0.500 (0.0011)</td>
</tr>
<tr>
<td>(\hat{\rho})</td>
<td>0.101 (0.0005)</td>
<td>0.201 (0.0007)</td>
<td>0.298 (0.0008)</td>
<td>0.401 (0.0008)</td>
<td>0.501 (0.0009)</td>
</tr>
</tbody>
</table>
Table 4: Operating characteristics of GG model evaluated on GG (upper) and LNN (lower) simulated data. Estimates are averaged over 100 simulations; standard error is shown in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sens</td>
<td>0.670(0.002)</td>
<td>0.703(0.002)</td>
<td>0.728(0.001)</td>
<td>0.753(0.001)</td>
<td>0.778(0.001)</td>
</tr>
<tr>
<td>Spec</td>
<td>0.996(0.000)</td>
<td>0.990(0.000)</td>
<td>0.982(0.000)</td>
<td>0.968(0.000)</td>
<td>0.948(0.001)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.950(0.001)</td>
<td>0.947(0.001)</td>
<td>0.945(0.001)</td>
<td>0.940(0.001)</td>
<td>0.938(0.001)</td>
</tr>
<tr>
<td>NPV</td>
<td>0.964(0.000)</td>
<td>0.930(0.000)</td>
<td>0.894(0.001)</td>
<td>0.855(0.001)</td>
<td>0.810(0.001)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.050(0.001)</td>
<td>0.053(0.001)</td>
<td>0.055(0.001)</td>
<td>0.060(0.001)</td>
<td>0.062(0.001)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sens</td>
<td>0.688(0.002)</td>
<td>0.718(0.002)</td>
<td>0.742(0.001)</td>
<td>0.764(0.001)</td>
<td>0.785(0.001)</td>
</tr>
<tr>
<td>Spec</td>
<td>0.996(0.000)</td>
<td>0.991(0.000)</td>
<td>0.983(0.000)</td>
<td>0.972(0.000)</td>
<td>0.955(0.000)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.953(0.001)</td>
<td>0.950(0.001)</td>
<td>0.950(0.001)</td>
<td>0.948(0.001)</td>
<td>0.946(0.000)</td>
</tr>
<tr>
<td>NPV</td>
<td>0.966(0.000)</td>
<td>0.934(0.000)</td>
<td>0.899(0.000)</td>
<td>0.861(0.001)</td>
<td>0.816(0.001)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.047(0.001)</td>
<td>0.050(0.001)</td>
<td>0.050(0.001)</td>
<td>0.052(0.001)</td>
<td>0.054(0.000)</td>
</tr>
</tbody>
</table>
Table 6: Parameter estimates for GG \((\theta = (\alpha, \alpha_0, \nu))\) and LNN \((\theta = (\mu_0, \sigma, \tau_0))\) models used in two group comparisons between parentals and in four group comparisons among the parentals and derived inbred lines.

<table>
<thead>
<tr>
<th>Model</th>
<th>(\hat{\theta}_1)</th>
<th>(\hat{\theta}_2)</th>
<th>(\hat{\theta}_3)</th>
<th>(\hat{p}_0)</th>
<th>(\hat{p}_1)</th>
<th>(\hat{p}_2)</th>
<th>(\hat{p}_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GG (2 groups)</td>
<td>12.490</td>
<td>0.919</td>
<td>35.842</td>
<td>0.998</td>
<td>0.002</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>LNN (2 groups)</td>
<td>6.775</td>
<td>0.292</td>
<td>1.193</td>
<td>0.993</td>
<td>0.007</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>GG (4 groups)</td>
<td>16.738</td>
<td>0.883</td>
<td>24.398</td>
<td>0.985</td>
<td>0.012</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>LNN (4 groups)</td>
<td>6.741</td>
<td>0.257</td>
<td>1.221</td>
<td>0.975</td>
<td>0.017</td>
<td>0.004</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Figure 1: Histograms are of intensities (log scale) simulated under the GG (left) or LNN (right) model. Solid line is fitted marginal density from the GG model and dashed line is fitted marginal density from the LNN model.
Figure 3: Operating characteristics: Results are shown for the GG model applied to simulated data. To minimize overlap, we jittered the horizontal component. Closed characters imply identical model and simulation assumptions (GG model applied to GG data); open characters imply the opposite (GG model applied to LNN data).
Figure 5: Average intensities across replicates for the WF and COP data. Only spots which exhibit significant differential expression (as determined by the GG model) are shown.