Parametric Empirical Bayes Methods for Microarrays

M.A. Newton
C.M. Kendziorski

ABSTRACT We have developed an empirical Bayes methodology for gene expression data to account for replicate arrays, multiple conditions, and a range of modeling assumptions. The methodology is implemented in an R library called EBARRAYS. Functions in the library calculate posterior probabilities of patterns of differential expression across multiple conditions. This chapter provides an overview of the methodology and its implementation in EBARRAYS.

1 Introduction

Empirical Bayes (EB) methods are well-suited to high-dimensional inference problems and thus provide a natural approach to microarray data analysis. In contrast to methods that apply classical statistical inferences separately for different genes (e.g., t-tests, ANOVA), there is a kind of information sharing amongst genes in an EB analysis. This can be beneficial because experiments often involve tens of thousands of genes but only tens of microarrays, so the amount of information per gene can be relatively low. It may seem paradoxical that an inference about differential expression of a certain gene between two conditions, say, should be influenced by the expression levels of other genes, but for some time it has been recognized that related inference problems can be combined to gain advantage (e.g., Efron and Morris, 1973, 1977; Carlin and Louis, 1996). Roughly speaking, the data from other genes provide some information about the typical variability in the system. The general framework provided by EB analysis is quite flexible; probability distributions are specified in several layers and account for multiple sources of variation. Our calculations explore two particular specifications that may be effective for typical microarray studies.

We present here an R library called EBARRAYS that is designed to implement several EB methods for gene expression data analysis. The methodology is focused on specific questions concerning differential expression; it presumes that data have been sufficiently preprocessed so that they represent bona fide approximations to some relative gene expression
in the cells being sampled. These may be the probe-level intensity signals from an Affymetrix oligonucleotide system or intensity measurements from a spotted cDNA array. The methods that we discuss provide estimates of differential expression, hypothesis tests of differential expression between two conditions, and the assessment of patterns of differential expression among multiple conditions; in each case, we allow replicated expression profiles within each condition.

The first calculation is to estimate the magnitude of differential expression between two conditions. Although the common practice is to report a fold change on a raw or log scale, the EB framework leads to various shrinkage estimates of fold change that differ from the standard estimators, especially for genes of relatively low total abundance. When there is interest in rank ordering genes by the extent of differential expression, the EB ranking can be different from a ranking based on apparent fold change; the present methods may provide better estimates of the ranking because they implicitly account for differential variability in apparent fold-change estimates.

When comparing expression profiles from two groups of samples, an effective gene-specific inference summary is the posterior probability of differential expression. This transforms evidence from the scale of expression score or fold change to the familiar scale of probability. Underlying the calculation is a simple parametric mixture model. Some genes are differentially expressed and some are not, and we view the question as stochastic. Measured expression data have different distributions depending on this outcome, so the posterior probability of differential expression is readily computed once the mixture model is estimated. Posterior probability calculations carry over naturally to comparisons among multiple conditions. In an example that we consider from a study of rat mammary epithelium cells, there are four cellular conditions and four distinct patterns of differential expression of particular interest.

The work presented in Newton et al. (2001) may have been the first application of EB methodology for gene expression data analysis. There we considered both shrinkage estimation of fold change and the posterior probability of differential expression between two conditions. Our results were restricted to so-called “single-slide” data (rather than replicated profiles in each condition) and to one particular hierarchical model. These analyses are extended in Kendzierski et al. (2002); we review these extensions and describe a software implementation in the present chapter.
2 EB Methods

2.1 Canonical EB Example

Although empirical Bayes methods have been widely studied, they may not be familiar to many investigators analyzing microarray data, so we consider a canonical EB example in the present context. The expression profile of some cell type in some growth condition is a numerical quantification of the abundances (or relative abundances) of all mRNAs. Biological as well as other sources of variation cause measured expression levels to fluctuate around the hypothetical target profile \((\mu_1, \mu_2, \ldots, \mu_J)\), where \(\mu_j\) is the true expression level of gene \(j\) and \(J\) is the number of genes under study. To simplify matters, suppose that our measured profile is the single vector \((x_1, x_2, \ldots, x_J)\) (single array) and that a scale of measurement has been chosen so that all the \(x_j\)'s have the same variance. It would seem that the measured profile \((x_1, \ldots, x_J)\) is the obvious estimate of \((\mu_1, \ldots, \mu_J)\) in the absence of other information since each \(x_j\) approximates its respective \(\mu_j\). Indeed, under mild conditions, \((x_1, \ldots, x_J)\) is the maximum likelihood estimate and has the least variation among all unbiased estimates of the whole profile.

The obvious estimate has some problems when \(J\) is relatively large, as was first shown by Stein (1956). He observed that \(\sum_j x_j^2\) (the squared Euclidean length of the vector \((x_1, \ldots, x_J)\)) tends to be much greater than the squared length of the target profile \((\mu_1, \ldots, \mu_J)\). This fact implies that better estimates of the whole profile can be found. Interestingly, improved estimates can be found using an empirical Bayes argument. We treat the \(\mu_j\)'s themselves as random, work out a summary feature of their conditional distribution given the data, and then plug in estimates of remaining parameters using the whole dataset. For example, an improved estimator might have component \(j\) equal to \(w x_j + (1 - w) \bar{x}\), where \(\bar{x}\) is the average of expression measurements and \(w\) is some data-dependent weight in the unit interval.

Roughly speaking, optimal estimates of the whole profile involve sharing information among genes and are not obtained by simply combining optimal estimates of the individual components. The empirical Bayes argument provides an effective approach to statistical inference in high-dimensional problems. It can be used not only to improve the estimation of profiles but also to enable inference about differential gene expression, as we now show.

2.2 General Model Structure: Two Conditions

Our models attempt to characterize the probability distribution of expression measurements \(x_j = (x_{j1}, x_{j2}, \ldots, x_{jT})\) taken on a gene \(j\). (As we clarify below, the parametric specifications that we adopt allow either
that these \( x_{j,i} \) are recorded on the original measurement scale or that they have been log-transformed.) A baseline hypothesis might be that the \( I \) samples are exchangeable (i.e., that potentially distinguishing factors, such as cell-growth conditions, have no bearing on the distribution of measured expression levels). We would thus view measurements \( x_{j,i} \) as independent random deviations from a gene-specific mean value \( \mu_j \) and, more specifically, as arising from an observation distribution \( f_{\text{obs}}(\cdot | \mu_j) \). For example, a gene with a large \( \mu_j \) typically exhibits high expression measurements and high variability.

When comparing expression samples between two groups (e.g., cell types), the sample set \( \{1, 2, \ldots, I\} \) is partitioned into two subsets, say \( s_1 \) and \( s_2 \); \( s_k \) contains the indices for samples in group \( k \). The distribution of measured expression may not be affected by this grouping, in which case our baseline hypothesis above holds and we say that there is equivalent expression, \( \text{EE}_j \), for gene \( j \). Alternatively, there is differential expression, \( \text{DE}_j \); our formulation requires that there now be two different means, say \( \mu_{j1} \) and \( \mu_{j2} \), corresponding to measurements in \( s_1 \) and \( s_2 \), respectively. Arguably, the fold change of interest is \( \rho_j = \mu_{j1} / \mu_{j2} \) rather than the apparent fold change computed either by averaging separate fold-change estimates from pairs of arrays or by taking the ratio of average expression from each group. Empirical Bayes estimates of \( \rho_j \) are obtained by specifying a probability distribution on the gene effects \( \mu_{jk} \) themselves and then computing some measure \( \hat{\rho}_j \) of the center of the posterior distribution of each \( \rho_j \) given the data. In our calculations, we take the simplest assumption that the gene effects arise independently and identically from a system-specific distribution \( \pi(\mu) \). The precise form of \( \hat{\rho}_j \) depends on details of the model specification, but in the two cases that we have worked out the value \( \hat{\rho}_j \) is shrunk toward unity compared to the apparent fold-change estimate. This tends to provide more conservative inferences than the naïve approach and has been shown to reduce estimation error (Newton et al., 2001). Were we instead to treat the \( \mu_{jk} \)'s as fixed effects, we would not derive any such advantage and there would be no information sharing across genes.

In addition to reporting fold-change estimates, it may also be useful to gauge the significance of differential expression. Here, the empirical Bayes formulation has an appealing structure compared to the classical one-gene-at-a-time tests, which treat the gene effects as fixed and separate. Some unknown fraction \( p \) of the genes are differentially expressed (\( \text{DE}_j \)) and the remainder are equivalently expressed, and the state for each gene is considered to be a matter of chance. Putting this together with the framework above, we have that a gene \( j \) respecting equivalent expression (\( \text{EE}_j \)) presents data \( x_j = (x_{j1}, \ldots, x_{jI}) \) according to a distribution

\[
f_0(x_j) = \int \left( \prod_{i=1}^{I} f_{\text{obs}}(x_{ji} | \mu) \right) \pi(\mu) \, d\mu.
\] (1.1)
This predictive distribution is an average (i.e., a mixture) over the possible gene effects \( \mu_j \). Among other things, the averaging induces a positive dependence among measurements on the same gene, a property that we regularly observe in data. Alternatively, if gene \( j \) is differentially expressed, the data \( x_j = (x_{j1}, x_{j2}) \) are governed by the distribution

\[
f_1(x_j) = f_0(x_{j1}) f_0(x_{j2})
\]

owing to the fact that different mean values govern the different subsets \( x_{j1} \) and \( x_{j2} \) of samples. We have both continuous mixing over the unknown values of \( \mu_j \) and discrete mixing over the two patterns (DE\( j \) and EE\( j \)) for each gene. The marginal distribution of the data becomes

\[
p f_1(x_j) + (1 - p) f_0(x_j).
\]

With estimates of \( p, f_0 \), and thus \( f_1 \), the posterior probability of differential expression is calculated by Bayes’ rule as

\[
\frac{p f_1(x_j)}{p f_1(x_j) + (1 - p) f_0(x_j)}.
\]

To review, the distribution of data involves an observation component, a component describing variation of mean expression \( \mu_j \), and a discrete mixing parameter \( p \) governing the pattern of expression between conditions. The first two pieces combine to form a key predictive distribution \( f_0(\cdot) \) (see (1.1)), which enters both the marginal distribution of data (1.3) and the posterior probability of differential expression (1.4).

### 2.3 Multiple Conditions

Many studies take measurements from more than two cellular conditions, and this leads us to consider more patterns of mean expression than simply DE\( j \) and EE\( j \). For example, with three conditions, there are five possible patterns among the means, including equivalent expression across the three conditions (1), altered expression in just one condition (3), and distinct expression in each condition (1). We view a pattern of expression for a gene \( j \) as a grouping or clustering of conditions so that the mean level \( \mu_j \) is the same for all conditions grouped together. With microarrays from four cell conditions, there are 15 different patterns, in principle, but with extra information we might reduce the number of patterns to be considered. We discuss an application in Section 4 in which ten array sets are measured across four cell conditions, but the context tells us to look only at a particular subset of four patterns.

We always entertain the null pattern of equivalent expression among all conditions. Consider \( m \) additional patterns so that \( m + 1 \) distinct patterns
of expression are possible for a data vector \( x_j = (x_{j1}, \ldots, x_{jT}) \) on some gene \( j \). Generalizing (1.3), \( x_j \) is governed by a mixture of the form

\[
\sum_{k=0}^{m} p_k f_k(x_j),
\]

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

\[
P(k|x_j) \propto p_k f_k(x_j).
\]

The pattern-specific predictive density \( f_k(x_j) \) may be reduced to a product of \( f_0(\cdot) \) contributions from the different groups of conditions, just as in (1.2), and this suggests that the multiple-condition problem is really no more difficult computationally than the two-condition problem except that there are more unknown mixing proportions \( p_k \). (See Kendzierski et al. (2002) for details.) The posterior probabilities (1.6) summarize our inference about expression patterns at each gene. They can be used to identify genes with altered expression in at least one condition, to order genes within conditions, or to classify genes into distinct expression patterns.

We emphasize that our use of the term pattern refers to a grouping or clustering of conditions according to equal mean expression of a given gene. More generic concepts, such as a pattern of increasing means in a time-course experiment, are not so easily represented by groups of conditions having the same mean. Still, as we demonstrate in Section 3, there is a great flexibility provided by the present framework.

### 2.4 The Gamma Gamma and Lognormal Normal models

Following Kendzierski et al. (2002), we consider two particular specifications of the general mixture model described above. Each is determined by the choice of observation component and mean component, and each depends on a few additional parameters \( \theta \) to be estimated from the data.

In the Gamma-Gamma (GG) model, the observation component is a Gamma distribution having shape parameter \( \alpha > 0 \) and a mean value \( \mu_j \); thus, with scale parameter \( \lambda = \alpha / \mu_j \),

\[
f_{\text{obs}}(x|\mu_j) = \frac{\lambda^\alpha x^{\alpha-1} \exp\{-\lambda x\}}{\Gamma(\alpha)}
\]

for measurements \( x > 0 \). Note that the coefficient of variation in this distribution is \( 1/\sqrt{\alpha} \), taken to be constant across genes \( j \). Matched to this observation component is a marginal distribution \( \pi(\mu_j) \), which we take to be an inverse Gamma. More specifically, fixing \( \alpha \), the quantity \( \lambda = \alpha / \mu_j \)
has a Gamma distribution with shape parameter $\alpha_0$ and scale parameter $\nu$. Thus, three parameters are involved, $\theta = (\alpha, \alpha_0, \nu)$, and, upon integration, the key density $f_0(\cdot)$ has the form

$$f_0(x_1, x_2, \ldots, x_I) = K \frac{\left(\prod_{i=1}^{I} x_i\right)^{\alpha-1}}{\left(\nu + \sum_{i=1}^{I} x_i\right)^{I\alpha+\alpha_0}}.$$  

(1.7)

where

$$K = \frac{\nu^{\alpha_0} \Gamma(I\alpha + \alpha_0)}{\Gamma(I\alpha) \Gamma(\alpha_0)}.$$

In the lognormal normal (LNN) model, the gene-specific mean $\mu_j$ is a mean for the log-transformed measurements, which are presumed to have a normal distribution with common variance $\sigma^2$. Like the GG model, LNN also demonstrates a constant coefficient of variation: $\sqrt{\exp(\sigma^2) - 1}$ on the raw scale. A conjugate prior for the $\mu_j$ is normal with some underlying mean $\mu_0$ and variance $\tau_0^2$. Integrating as in (1.1), the density $f_0(\cdot)$ for an $n$-dimensional input becomes Gaussian with mean vector $\mu_0 = (\mu_0, \mu_0, \ldots, \mu_0)^T$ and exchangeable covariance matrix

$$\Sigma_n = (\sigma^2) \mathbf{I}_n + (\tau_0^2) \mathbf{M}_n,$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathbf{M}_n$ is an $n \times n$ matrix of ones.

The GG and LNN models characterize fluctuations in array data using a small number of parameters, and both involve the assumption of a constant coefficient of variation (CV). This property is often observed in microarray data, at least approximately (see Section 4). The models may provide useful results even if this assumption does not hold. We recommend that assumptions be checked on a case-by-case basis.

### 2.5 Model Fitting

For both the GG and LNN models, we use the method of maximum (marginal) likelihood to obtain estimates of the unknown parameters $\theta$. In the GG model, $\theta = (\alpha, \alpha_0, \nu)$ and in LNN, $\theta = (\mu_0, \sigma^2, \tau_0^2)$. We estimate parameters using the EM algorithm. (Any gene with at least one negative intensity value in one condition is omitted from this step of the calculation.) With data $\mathbf{x}_j$ governed by a mixture of the form (1.5), we introduce pattern indicators $\phi_{jl}$ defined as 1 if the expression pattern of gene $j$ is pattern $l$ and 0 otherwise. The complete data log likelihood is

$$l_c(\theta) = \sum_j \left\{ \sum_{k=0}^{m} \phi_{jk} \left[ \log f_k(\mathbf{x}_j) + \log(p_k) \right] \right\}.$$  

(1.8)
Using a current estimate $\theta_0$, the expectation given the observed data amounts to replacing $\phi_j$ with $\hat{\phi}_j$, the posterior probability of expression pattern $l$ for gene $j$ (see 1.6) (i.e., the E-step). For the M-step, we use the arithmetic mean of $\hat{\phi}_{jk}$ to estimate $p_k$; then optim in R provides updated estimates of $\theta$. The two steps are repeated until there is convergence in the estimates. Results should be checked from various starting configurations.

Certain initial values can be determined from simple summary statistics. Recall from subsection 2.4 that the coefficient of variation (CV) for the GG model is $1/\sqrt{\alpha}$, which gives $\alpha = \left(\frac{1}{\text{CV}}\right)^2$. Using the average coefficient of variation across genes as an estimate of CV gives an initial estimate of $\alpha$. Similar reasoning applies to estimating $\sigma^2$ in the LNN model since there CV is completely determined by $\sigma^2$.

3 Software

The empirical Bayes hierarchical modeling approach is implemented in EBARRAYS, an R library available from

http://www.biostat.wisc.edu/~kendzior

and written for R version 1.4.0. EBARRAYS can be used to calculate posterior probabilities of a predefined set of expression patterns under either GG or LNN assumptions. The main functions available in version 1.0 are:

cleanup reads in data and records data characteristics
em EM algorithm to fit the EB model
postprob posterior probabilities for expression patterns

These functions are supplemented by other functions that are not called directly by the user, including complete.loglik, which calculates the complete data log likelihood (1.8), and functions to evaluate the key predictive distribution $f_0$. The current package contains two possible forms for $f_0$, fogg for the GG model and f0lnn for the LNN model, and we note that other families could be added by the user.

Each analysis requires input files that contain the normalized intensities (datafile), identify the replicate samples (reppfile), and specify the patterns to be considered (patternfile). EBARRAYS assumes that input files are tab or single-space delimited ASCII text files. datafile contains intensities in $J$ rows and $I$ columns; row and column names should be provided. reppfile specifies which of the $I$ samples are considered replicates. It contains one row with $I$ columns ($r_1, r_2, \ldots, r_I$), where $r_1 = 1$ and for $k = 2, 3, \ldots, I$, $r_k = r_{k-1} + 1$ if samples $k$ and $k-1$ are considered replicates and $r_k = r_{k-1} + 1$ otherwise. patternfile specifies the patterns to be considered in the analysis. The $k$th row identifies which samples are assumed to have the same mean...
level of expression in pattern \( k \). Typically, the \((k, 1)\) element is 1 for every \( k \) and for \( l = 2, 3, \ldots, I \), \((k, l) = (k, l-1)\) if samples \( l \) and \( l-1 \) are assumed to have the same mean level of expression in pattern \( k \) and \((k, l) = (k, l-1)+1\) otherwise.

As an illustration, consider a dataset with \( I = 10 \) arrays taken from two conditions (five arrays in each condition ordered so that the first five columns contain data from the first condition). In this case, there are two, possibly distinct, levels of expression for each gene and two potential patterns (\( \mu_{j1} = \mu_{j2} \) and \( \mu_{j1} \neq \mu_{j2} \)). The replicate and pattern files are, respectively,

\[
1 \ 1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 2
\]

and

\[
1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1
\]

1 \ 1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 2

The exception to this convention is that zero columns can be used to identify arrays that are not used in model fitting or analysis. An example of this is given in Section 4. The initial data files (\textit{datafile}, \textit{repfile}, and \textit{patternfile}) are loaded into \texttt{R} and processed with the function \texttt{cleanup}:

\[
\texttt{cleanup(path,"datafile","repfile","patternfile")}
\]

\textit{path} is a character string specifying the directory containing the data. All output is written to this directory. \texttt{cleanup} returns a list. A few of the key components are:

- \texttt{int.data}: a data frame containing the intensity measurements.
- \texttt{gene.names, sample.names, gene.num, sample.num}: the names and total number of genes and samples.
- \texttt{all.pos}: a T/F vector indicating which genes have positive intensity measurements in every condition.

After initial data organization with \texttt{cleanup}, exploratory analyses can be carried out using standard \texttt{R} functions to ensure data quality and identify any potential problems. The function \texttt{em} can then be used to obtain parameter estimates under either \texttt{GG} or \texttt{LNN} model assumptions:

\[
\texttt{em(theta.init,p.init,data,family,num.iter,write.output)}
\]

\textit{Here, family} is a character string indicating whether the calculations are for the \texttt{GG} or \texttt{LNN} model, the two \texttt{.init} arguments are starting values for the EM iterations, \texttt{num.iter} indicates how many iterations of EM to run, and \texttt{write.output} allows the results to be written to a file. The output of \texttt{em} is a list including:
params.p.mat: a matrix containing estimates of the mixing proportions at each step of the EM algorithm.

params.theta.mat: a matrix containing estimates of theta at each step of the EM algorithm.

params.p.est: a vector containing final mixing proportion estimates.

params.theta.est: a vector containing final estimates of $\theta$.

If write.output is set to TRUE, the matrices and final estimates are written to files (named as above) in the directory pathname as specified in cleanup.

After fitting the mixture model with em, the function postprob can be used to calculate the posterior probability for each gene in each expression pattern. postprob requires output from cleanup, output from em, and a flag (TRUE or FALSE) indicating whether output should be written to a file.

```
postprob(data, em.out, write.output)
```

Output is a list with components postprob and table. Both are matrices with rows corresponding to genes. The $(j, k)$th entry of postprob is the posterior probability that gene $j$ is in pattern $k$. The table component includes this information and also appends the sample mean intensities averaged across conditions. If write.output is set to TRUE, these matrices are written to files postprobs and summary.postprobs in the directory pathname as specified in cleanup.

4 Application

In collaboration with Dr. M.N. Gould’s laboratory in Madison, we have been investigating gene expression patterns of mammary epithelial cells in a rat model of breast cancer. We use a subset of the data from this study to illustrate the mixture model calculations and the EBARRAYS library. For a more extensive analysis, see Kendziorski et al. (2002).

In summary, the experiment we consider used Affymetrix U34 chip sets and thus measured the expression of 26,379 rat genes. Each of $I = 10$ mRNA samples was probed using one chip set. The mRNA was derived from rat mammary epithelium cells of same-aged females living in a controlled environment. Four inbred strains of rats contributed mRNA, including two parental strains and two derived congenic lines. The parental strains differ greatly in their resistance to carcinogen-induced mammary cancer, with the Copenhagen (COP) strain almost completely resistant to certain carcinogens, while the inbred Wistar Furth (WF) strain was highly susceptible (Gould et al., 1989). By careful breeding, two intermediate
inbred lines (CI and CII) were produced that carry the homozygous WF/WF genotype throughout the genome except on a relatively small region of interest where the animals are homozygous COP/COP (Figure 1). Understanding expression patterns among these strains may be helpful in the study of breast cancer resistance.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>DE Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>COP</td>
<td>$\mu_{\text{COP}} = \mu_{\text{CI}} = \mu_{\text{CII}} = \mu_{\text{WF}}$</td>
</tr>
<tr>
<td>CI</td>
<td>$\mu_{\text{COP}} = \neq \mu_{\text{CI}} = \mu_{\text{CII}} = \mu_{\text{WF}}$</td>
</tr>
<tr>
<td>CII</td>
<td>$\mu_{\text{COP}} = \mu_{\text{CI}} = \neq \mu_{\text{CII}} = \mu_{\text{WF}}$</td>
</tr>
<tr>
<td>WF</td>
<td>$\mu_{\text{COP}} = \mu_{\text{CI}} = \mu_{\text{CII}} = \neq \mu_{\text{WF}}$</td>
</tr>
</tbody>
</table>

Figure 1. Schematic diagram showing animal lines (conditions) from which mRNAs were obtained (left) along with differential expression patterns (right). Genotypes shown in black (COP/COP) and white (WF/WF) are not drawn to scale (the homozygous COP/COP region is approximately 30 cM in congeneric line CI and 1.5 cM in CII). True expression intensities for each group are denoted by $\mu$. Note that differences in genotype do not imply differences in expression. Genes classified into the null pattern show equivalent expression across groups but differ in genotype.

As an initial processing step we ran all the data through DNA-Chip Analyzer (Li and Wong, 2001) which uses a statistical model for probe-level data to account for artifacts such as probe-specific biases. Corrected and normalized model-based estimates of gene expression were obtained for $J = 25,248$ genes (1131 were identified as outliers) and were stored in a datafile called data.txt. EBARRAYS uses the repfile to characterize the organization of replicates among the $I = 10$ samples across the four conditions. We note that, in column order, there is one COP sample, five CI samples, two CII samples, and two WF samples. To reflect this arrangement, the repfile (called reps.txt) is

$$1 \ 2 \ 2 \ 2 \ 2 \ 3 \ 3 \ 4 \ 4$$
The last bit of input is the *patternfile*, which tells EBARRAYS what patterns of mean expression will be considered in the analysis. Let us first ignore the derived congeneric lines and perform an analysis on the parental strains only. There are two possible expression patterns ($\mu_{\text{COP}} = \mu_{\text{WF}}$ and $\mu_{\text{COP}} \neq \mu_{\text{WF}}$). These are represented in the *patternfile* (called *patterns.txt*)

\[
\begin{align*}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2
\end{align*}
\]

Recall that a zero column indicates that the data in that condition are not considered in the analysis. An alternative approach would be to define a new data matrix containing intensities from the parental strains only and define the associated *patternfile* as

\[
\begin{align*}
1 & 1 & 1 \\
1 & 2 & 2
\end{align*}
\]

The corresponding *repfile* would be

\[
\begin{align*}
1 & 2 & 2
\end{align*}
\]

This may be useful in some cases, but in general we recommend importing the full data matrix and defining the pattern matrix as a $2 \times 10$ matrix with intermediate columns set to zero. Doing so facilitates comparisons of results among different analyses since attributes of the data, such as the number of genes that are positive across each condition, remain constant.

With EBARRAYS installed, we use the *library* function in R to connect it to the current session (see Parmigiani et al., Chapter 1, this volume). The illustration below assumes that the input files *data.txt*, *reps.txt*, and *patterns.txt* are in the local directory. First, we organize the data:

```r
> data <- cleanup(NULL,"data.txt","reps.txt","patterns.txt")
[1] "There are 25248 genes and 10 samples in this data set. The samples are arranged across 4 conditions."
[1] "Out of the 25248 genes, 24452 have positive intensities across each sample."
[1] "3 samples and 2 patterns will be considered in this analysis."
```

Preliminary data analysis can be done using standard R functions. For example, we can check to see if there is any relationship between the mean expression level and the coefficient of variation; recall that both GG and LNN models assume a constant CV. Figure 2 shows that the assumption is remarkably good for the three parental chip sets, with just a slight violation at very low mean expression. (Similar results were obtained using all ten samples).
Figure 2. Coefficient of variation (CV) (ranked) as a function of the mean (ranked) for the WF and COP data. The CV averaged across all genes is 0.215. Shaded vertical bars indicate nine subsets of 100 genes, each examined more closely in Figure 3.

Figure 3 shows a second diagnostic plot for nine subsets of 100 genes spanning the range of mean expression. Shown are qq plots against the best-fitting Gamma distribution. Again, the goodness of fit is striking. (Note that we only expect these qq plots to hold for equivalently expressed genes, so some violation is expected in general.)
Figure 3. Gamma qq plots for the nine subsets shown in Figure 2. Each subset contains 100 genes.

Using \texttt{em}, we can fit either the GG or the LNN model. Here we show how to do the GG fit, but we note that the LNN fit is done by the same approach. We recommend fitting both for the sake of comparison.

\begin{verbatim}
> emgg.out <- em(theta.init, p.init, data,"GG",20,TRUE)
[1] "GG model parameter estimates:
  12.480, 0.920, and 35.955"
[1] "Estimates of mixing proportions:
  0.99778, 0.00222"
\end{verbatim}

Posterior probabilities can then be obtained; highlighted below are a few genes of interest.
> ppgg.out <- postprob(data, emgg.out, TRUE)
> attributes(ppgg2.out)
  $names
[1] "postprob" "table"
> ppgg2.out$table[c(2008, 6516, 1987),]

|     | AVE.C1 | AVE.C2 | PO  | P1
|-----|--------|--------|-----|-----
| J00801.at | 3066.300 | 9082.8800 | 0.99286 | 0.00714
| L08100.at | 4367.490 | 14162.3000 | 0.98894 | 0.01016
| J00772.s.at | 391.968 | 678.8935 | 0.99897 | 0.00103

> ppgg2.out$postprob[c(2008, 6516, 1987),]

|     | PO  | P1
|-----|-----|-----
| J00801.at | 0.99286 | 0.00714
| L08100.at | 0.98984 | 0.01016
| J00772.s.at | 0.99897 | 0.00103

We identify 50 genes that are probably differentially expressed.

> sum(ppgg2.out$postprob[,2]>0.5)

[1] 50

Of these, 49 are also identified by the LNN model.

A nice feature of EBARRAYS is that comparisons among more than two groups can be carried out simply by changing the pattern matrix. For the four conditions, there are 15 possible expression patterns; however, if latent expression in each congenic matches one of the parents, only four expression patterns are possible (see Figure 1). This is the case that we will consider (thus, m + 1 = 4). The null pattern consists of equivalent expression across the four conditions. The three other patterns allow for differential expression between the parental strains, with the congenic lines exhibiting the same mean expression as one of the parents. Specifically, differential expression of the COP parent only is specified in pattern 1, between the congenics in pattern 2, and of the WF parent only in pattern 3.

The repfile is the same as before, but now the pattern matrix (patterns.txt) for the four group analysis is given by

```
 1 1 1 1 1 1 1 1 1
 1 2 2 2 2 2 2 2 2
 1 1 1 1 1 1 2 2 2
 1 1 1 1 1 1 1 1 2
```

The data are imported as in the comparison between two groups:

> data <- cleanup(NULL, "data.txt", "reps.txt", "patterns.txt")

[1] "There are 25248 genes and 10 samples in this data set."
[1] "The samples are arranged across 4 conditions."

[1] "Out of the 25248 genes, 24452 have positive intensities"
across each sample."

[1] "10 samples and 4 patterns will be considered in this analysis."

em and postprob are also called as before.

> emgf.out <- em(theta.init, p.init, data, "GG", 20, TRUE)
[1] "GG model parameter estimates: 
   16.720, 0.885, and 24.517"

[1] "Estimates of mixing proportions: 
   0.98492, 0.01228, 0.00186, and 0.00094"

> ppgf.out <- postprob(data, emgf.out, TRUE)

Under the GG model, 24,797 genes had posterior probability greater
than 0.5 of being in the null pattern; 250, 84, and 111 genes were classified
into patterns 1, 2, and 3, respectively. For six genes, no pattern had
posterior probability greater than 0.5. The LNN model identified slightly
more genes as differentially expressed. Specifically, 24,168 were classified
into the null pattern; 447, 343, and 280 were classified into patterns 1, 2,
and 3, respectively; there were ten genes with posterior probabilities less
than 0.5 for each pattern.

Of much interest in this analysis were genes showing differential
expression between the first three conditions (COP, CI, and CII) and the
WF condition. This corresponds to pattern 3 (see Figure 1). Genes with
high posterior probability of being in pattern 3 are easily identified, and
their probability distributions can be obtained:

> ppgf.out$postprob[c(2008, 6516, 1987),]

   PO   PI   P2   P3
J00801.at 0.04646 0.00021 0e+00 0.95332
L08100.at 0.00000 0.00000 0e+00 1.00000
J00772.s.at 0.04459 0.00013 2e-05 0.96527

The intensity values averaged within conditions show differences between
the average intensities in the first three conditions and that in the WF
condition C4. (Columns 5 through 8 contain the posterior probabilities as
shown above.)


     AVE_C1  AVE_C2  AVE_C3  AVE_C4
J00801.at 3066.300 9062.8800 4777.0160 995.295
L08100.at 4367.490 14162.300 4002.6490 1278.285
J00772.s.at 391.968 678.8935 325.8258 121.734
5 Discussion

The proposed empirical Bayes methodology accounts for replicate arrays, multiple conditions, and a range of modeling assumptions. The methodology is implemented in an R library called EBARRAYS. Functions in the library calculate posterior probabilities of patterns of differential expression across multiple conditions.

The hierarchical modeling approach accounts for patterns of differential expression for a given gene among cell types. It also considers differences among genes in their average expression levels and measurement fluctuations. Since properties of individual experiments affect these features, the distributions governing them are to some extent experiment-dependent. However, there are features of microarray data that are generally observed and well-described using certain parametric models, such as a constant coefficient of variation. Two parametric formulations have been implemented, and model checks indicate that the assumptions are reasonable for the data considered here (see Figures 2 and 3). We have observed similar results for other datasets as well, and we expect that simple parametric analyses will often be suitable for array data. Of course, they must be accompanied by diagnostic model checks. EBARRAYS is flexible enough to incorporate a user-supplied predictive distribution in place of GG or LNN if necessary.

The methodology proposed here is not the only mixture-model-based approach to expression data analysis. Other approaches have been developed to address the question of comparisons between two conditions. Efron et al. (2001) describe calculations in the context of a specific experimental design. After a long series of preprocessing 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 et 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. The method of Lonnstedt and Speed (2002) has features in common with ours.

Another general approach that applies to expression data from two conditions is to conduct a hypothesis test at each gene and then correct for multiple comparisons. Most of the test statistics currently used are t (or t-like) and differ primarily in the estimation of the variance. Dudoit et al. (2002) use a t-statistic with variance estimated by the within-gene sample variance and go on to address the multiple comparisons problem extensively using permutation analysis (see Dudoit and Yang, Chapter 3, this volume). Tusher et al. (2001) also use the within-gene sample variance, but adjust the denominator of their test statistic by adding a constant to account for the dependence between the relative difference in expression and absolute intensity (see Storey and Tibshirani, Chapter 12, this volume).
Methods that treat genes as separate fixed effects may have reduced efficiency compared to methods that treat the genes as arising from some population since they do not take full advantage of the level of information sharing among genes. Furthermore, classifying genes into expression patterns by the posterior probability is an optimal procedure in the context of the mixture model: it minimizes the expected number of errors. Interestingly, this goal is different from the goal in classical testing, which is to bound the type I error rate and then aim to maximize the power. Further investigation is required to more formally assess the advantages and disadvantages of these approaches.

References


1. Parametric Empirical Bayes Methods for Microarrays


