SUPPLEMENTARY MATERIAL

GaGa: a simple and flexible hierarchical model for microarray data analysis

David Rossell
Department of Biostatistics
M.D. Anderson Cancer Center, Houston, TX 77030, USA
rosselldavid@gmail.com

Contents

1 Bayesian procedure for gene differential expression analysis 2
2 Frequentist operating characteristics of the Bayesian procedure 3
3 Frequentist FDR for the Armstrong dataset 5
4 Goodness of fit in the Armstrong dataset 6
1 Bayesian procedure for gene differential expression analysis

Here we detail the Bayesian procedure that minimizes the Bayesian FNR subject to the Bayesian FDR being below a threshold $\alpha$. Let $\delta_1, \ldots, \delta_n$ denote the expression pattern that each gene follows, e.g., if there are only two groups to be compared $\delta_i = 0$ denotes the null hypothesis that both groups are equal and $\delta_i = 1$ denotes that they are different. Denote as $d_i = d_i(x)$ the pattern that gene $i$ is assigned to, i.e. $d_i = 0$ means that we declare the gene as equally expressed (EE) and $d_i \neq 0$ that we declare it differentially expressed (DE). The false negative (FNP) and false discovery proportions (FDP) can be written as:

\[
\begin{align*}
\text{FNP} &= \frac{\sum_{i=1}^{n} I(d_i = 0) I(\delta_i \neq 0)}{\sum_{i=1}^{n} I(d_i = 0)}, \\
\text{FDP} &= \frac{\sum_{i=1}^{n} I(d_i = 1) I(\delta_i = 0)}{\sum_{i=1}^{n} I(d_i = 1)},
\end{align*}
\]

where $I(\cdot)$ is the indicator function. That is, the FNP is the proportion of genes declared EE that are actually DE, and the FDP is the proportion of genes declared DE that are actually EE. The Bayesian FNR and Bayesian FDR are defined as the expected FNP and FDP, respectively, where the expectation is taken with respect to the posterior distribution of $\delta_1, \ldots, \delta_n$. \(^{(1)}\) For any fixed decisions $d_1, \ldots, d_n$, one can evaluate the Bayesian FNR and FDR simply as:

\[
\begin{align*}
\text{BFNR} &= \frac{\sum_{i=1}^{n} I(d_i = 0)(1 - v_{i0})}{\sum_{i=1}^{n} I(d_i = 0)}, \\
\text{BFDR} &= \frac{\sum_{i=1}^{n} I(d_i = 1)v_{i0}}{\sum_{i=1}^{n} I(d_i = 1)},
\end{align*}
\]

where $v_{i0} = P(\delta_i = 0|x)$ is the posterior probability that gene $i$ is EE. \(^{(2)}\) showed that, in a setup with only two hypotheses, the optimal rule to minimize BFNR subject to BFDR $< \alpha$ is to declare as DE all genes with $v_{i0}$ below a certain threshold $t$, i.e. $d_i = I(v_{i0} < t)$, where $t$ is the minimum value such that BFDR $\leq \alpha$. Note that BFNR and BFDR only take into account whether a gene was classified into pattern 0 or not. Therefore, when minimizing BFNR subject to BFDR $\leq \alpha$ it makes no difference whether a particular gene is assigned to pattern 1 or pattern 2, say. We propose the
obvious: given that a gene is declared DE, we assign it to the pattern with highest posterior probability, i.e. \( \delta_i = I(v_{i0} < t) \times \arg\max_{k \in \{1, \ldots, H-1\}} (v_{ik}) \). It is straightforward to see that, for any fixed BFNR and BFDR, this rule maximizes the expected number of genes correctly classified into their expression pattern.

2 Frequentist operating characteristics of the Bayesian procedure

In this section we review the algorithm that [3] proposed to estimate the frequentist FDR for any given procedure, i.e. the expected FDP in [1] when the procedure is applied under repeated sampling. As a first step, one obtains bootstrap samples from the original data, in such a way that the sample mean and variance of each gene are roughly preserved and that it represents a sample under the complete null hypothesis that no genes are DE. Then one repeatedly applies the procedure to each bootstrap dataset, obtaining an estimate of the number of false positives, and compares that to the number of genes found in the original dataset. More specifically, the algorithm is as follows:

Algorithm 1

1. Apply the procedure to the original dataset, and denote the number of genes declared to be DE as \( P \). Denote as \( \bar{X}_i \) and \( S_i \) the sample mean and standard deviation of the gene expression measurements for gene \( i = 1 \ldots n \).

2. For \( b = 1 \ldots B \), do:
   - Compute \( z_{ij} = (x_{ij} - \bar{X}_i)/S_i \), \( \forall i = 1 \ldots n, j = 1 \ldots J \).
   - For each gene, obtain a sample of size \( J \) with replacement from the collection of all \( z_{ij} \). Denote this sampled values as \( z_{ij}^{(b)} \). Then compute \( x_{ij}^{(b)} = S_i z_{ij}^{(b)} + \bar{X}_i \).
   - Apply the procedure to find differentially expressed genes to the bootstrap dataset. Since all discoveries are false positives, denote the number of genes declared DE as \( FP_b \).

3. Estimate the frequentist FDR as \( \hat{\text{FDR}} = \hat{\pi}_0 \frac{\sum_{b=1}^{B} FP_b}{B \times P} \), where \( \hat{\pi}_0 \) is an estimate of the proportion of EE genes.
Table 1: Hypothetical expression values for a gene, 10 arrays and 2 groups

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>2.8</td>
</tr>
<tr>
<td>2.8</td>
<td>3.2</td>
</tr>
<tr>
<td>3.1</td>
<td>2.0</td>
</tr>
<tr>
<td>2.5</td>
<td>2.1</td>
</tr>
<tr>
<td>2.7</td>
<td>2.1</td>
</tr>
<tr>
<td>2.0</td>
<td>2.5</td>
</tr>
<tr>
<td>2.1</td>
<td>2.9</td>
</tr>
<tr>
<td>2.3</td>
<td>10.5</td>
</tr>
<tr>
<td>2.9</td>
<td>2.2</td>
</tr>
<tr>
<td>2.4</td>
<td>2.4</td>
</tr>
</tbody>
</table>

There is an important remark to make here: there are other ways to simulate data under the null hypothesis. For instance, one could simply permute the group labels or bootstrap data within each gene, but this could be troublesome when the sample size is too small to provide an accurate representation of the null. Determining what sample size is large enough is not a simple matter.

For example, suppose that we have a gene for which most expression values are small, but for one of the groups it occasionally presents large expression values. Table 1 presents hypothetical values for a single gene and 2 groups. Note that there are \( \binom{10}{5} = 184,756 \) ways to permute the group labels, which at first sight may seem a large enough number. However, in group 2 there is an outlying value. For any permutation, whatever group the outlier is assigned to will tend to be declared to have higher expression levels than the other group, and it will be counted as a false positive. If one obtains a bootstrap sample, there is some probability that the effect of the outlier will be mitigated (the outlier may not be sampled at all, or be sampled the same number of times in both groups), but it will still tend to increase to false positive count. If one bootstraps residuals from other genes, such as Algorithm 1 does, the value of 10.5 may not be outlying at all anymore, since other genes may present values that are also far away from the mean.

Hence, we see that three reasonable strategies to sample under the null may result in three quite different sampling distributions and estimated FDR. The main issue is whether the outlying value should be considered to be an error or not. In our experience, in microarrays it is not unfrequent to encounter data such as that in Table 1. A possible biological explanation is that a small proportion of individuals from group 2 experience some kind of...
mutation, which causes the expression of a particular gene to raise considerably. If such a discovery is biologically meaningful one should not count it as a false positive, and hence methods based on permuting or bootstrapping each gene separately would not be appropriate. We agree with (3) that more research is needed regarding this topic, but we feel that unless the number of arrays is quite large it may be beneficial to use a resampling scheme that uses data from several genes at once.

3 Frequentist FDR for the Armstrong dataset

We now estimate the frequentist FDR of the Bayesian procedure outlined in Section 1 by applying Algorithm 1 to the Armstrong dataset.

Figure 1(a) displays the estimated frequentist FDR for target Bayesian FDRs ranging from 0 to 0.1, both for the GaGa and MiGaGa models and increasing amounts of data. For Bayesian FDR at the 0.05 level, the estimated frequentist FDR is always below 0.05. The only exception is the MiGaGa model applied to the full dataset, for which the frequentist FDR is estimated to be 6.4%.

Figure 2 provides the analogous estimates for the monotonically transformed data, which was analyzed with a GaGa model. For a Bayesian FDR of 0.05, the estimated frequentist FDR is below 0.05, as desired.
We conclude that the Bayesian procedure from Section 1 has reasonably good frequentist operating characteristics when applied to the Armstrong dataset.

4 Goodness of fit in the Armstrong dataset

In the original paper, Section 6.2.1, we evaluated some aspects of the overall goodness-of-fit. Figure 2(b) compares the marginal distribution of the monotonically transformed data with draws from the prior-predictive GaGa model, setting the hyper-parameters to their posterior mean. The monotonic transformation improves the fit of the GaGa model substantially (compare with Figure 4(a) in the original paper).

We now assess the fit for some genes individually. First, we select the two genes with the highest probability of being DE according to the Ga model. Figure 3(a) compares their observed expression values with draws from their posterior predictive distribution based on the Ga model. We see that, even though the model underestimates the variability for the MLL group, the two genes do actually seem to be differentially expressed. Figure 3(b) shows how draws from the GaGa model posterior predictive more appropriately capture
Figure 3: Observed expression values vs. predictive distribution. Large black symbols are actual observations, small gray symbols are draws from the posterior predictive. (a),(b): the two genes with highest probability of being DE according to the Ga model; (c),(d): two genes declared DE by the Ga model and declared EE by the GaGa model
the variability of the data. For these two genes in particular, however, the result of the inference is the same: both models declare probes 1914.at and 37809.at to be differentially expressed.

We now select two genes that are declared DE by the Ga model and EE by the GaGa model, and again compare the observed values to their posterior predictive distribution. Figure 3(c) reveals that the Ga model underestimates the variability in the data, while in panel (d) we see that GaGa represents it more satisfactorily. In this case the inference about the probes 1369.s.at and 2087.s.at from both models is radically different, for the GaGa model assigns a posterior probability < 0.01 that each gene is differentially expressed. The poor Ga fit to these two genes and the fact that no strong differences between groups are observed in Figure 3(c)-(d) suggest that the inference provided by the GaGa model is more realiable. Although not presented here, we observe a similar favorable behavior of the MiGaGa model with $M = 2$ components.

We conclude that the posterior distribution of the GaGa and MiGaGa models present an adequate fit to the data. This is in contrast with the prior-predictive plot in Figure 3(b) in the original paper, which suggested the GaGa fit to be of limited quality. This should not be too surprising, it merely reflects that the posterior distribution incorporates the information about bimodality present in the data.

References

