Statistica Sinica **18**(2008), 783-802

DETECTING DIFFERENTIALLY EXPRESSED GENES USING CALIBRATED BAYES FACTORS

Fang Yu[1], Ming-Hui Chen[2] and Lynn Kuo[2]

[1] *University of Nebraska Medical Center and* [2] *University of Connecticut*

Abstract: A common interest in microarray data analysis is to identify genes having changes in expression values between different biological conditions. The existing methods include using two-sample t-statistics, modified t-statistics (SAM), Bayesian t-statistics (Cyber-T), semiparametric hierarchical Bayesian models, and nonparametric permutation tests. All these methods essentially compare two population means. Unlike these methods, we consider using Bayes factors to compare gene expression levels that allows us to compare two population distributions. To adapt the use of Bayes factors to microarray data, we propose a new calibration approach that weighs two types of prior predictive error probabilities differently for each gene and, at the same time, controls the overall error rate for all genes. Moreover, a new gene selection algorithm based on the calibration approach is developed and its properties are examined. The proposed method is shown to have a smaller false discovery rate (FDR) and a smaller false non-discovery rate (FNDR) than several existing methods in several simulations. Finally, a data set from an affymetrix microarray experiment to identify genes associated with the mature osteoblast differentiation is used to further illustrate the proposed methodology.

Key words and phrases: Gene selection, Bayes factor, calibrating value, multilevel model, marginal likelihood.

1. Introduction

A common objective in microarray data analysis is to identify genes having different gene expression values between two conditions. A two-sample t-test based on the log transformed replicated data applied to each gene is often used. Significance Analysis of Microarray (SAM) (Tusher, Tibshirani and Chu (2001)) modifies the t test by adding a "fudge" factor to the standard error estimate of the two sample difference in the denominator of the t statistic. Cyber-T (Baldi and Long (2001)) uses similar test statistics except the denominator is replaced by a pooled variance estimate from neighboring genes. Newton, Noueiry, Sarkar and Ahlquist (2004) fit the data with gamma distributions with latent mean parameters distributed as a common mixture distribution of three nonparametric components. Methods based on linear models also exist to test whether the effects from the treatment relative to the control are zero

(MAANOVA, Kerr, Martin and Churchill (2000) and LIMMA, Smyth (2004)). These linear methods essentially conduct tests on equal means while controlling the variations of array, dye, etc.

The data set, however, may have a very complex structure. For example, it might be a mixture of normal distributions with several modes, so the selection of differentially expressed (DE) genes by comparing population means may lead to biased results. We use Bayes factors to select DE genes because they allow us to compare population means as well as the entire distributions. As pointed out by Kass and Raftery (1995), it may be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests in "nonstandard" statistical models that do not satisfy common regularity conditions. The Bayes factor approach directly measures the evidence of each gene being equally expressed (EE) versus DE. It has a direct interpretation on whether the null/alternative (EE/DE) hypothesis is true. In the literature, the Bayes factor value has been compared to the traditional p-value. The p-value approach tends to overstate the evidence against the null hypothesis, especially when a point null hypothesis is tested. Edwards, Lindman and Savage (1963), Berger and Sellke (1987) and Sellke, Bayarri and Berger (2001) provide more detailed discussions on the relationship between the Bayes factor and the p-value.

Liu, Parmigiani and Caffo (2004) discuss the use of Bayes factors to screen DE genes. When two distributions are different only in mean, the Bayes factor can be reduced to a test of equality of means. They consider the model where the replicated intensities are normally distributed with the same variance across two conditions. They assume that the distribution of Bayes factors is exchangeable among genes and then select the DE genes based on the ordered values of Bayes factors. No guidelines are given on the determination of the threshold value for each Bayes factor to select a DE gene.

In practice, it is unrealistic to assume the distribution of Bayes factors to be exchangeable across genes. Figure 1 gives the boxplots of the Bayes factors as a function of the variance of the intensities for each gene. It clearly shows that the distribution is not exchangeable across genes. Thus, new cut-off values of Bayes factors need to be developed. Garcia-Donato and Chen (2005) propose a new decision rule based on the sampling distribution of the Bayes factor instead of the observed Bayes factor value. This is reasonable because the uncertainty of the Bayes factor can be large depending on the observed data. They point out that the prior predictive distributions of the Bayes factor under each hypothesis are asymmetric. So they propose a calibration value to be the threshold that makes the two types of prior predictive error probabilities equal. However, the calibration value is based on the principle of prior equity, which may not be applicable here, as this does not put any control on the overall error rate for all genes.

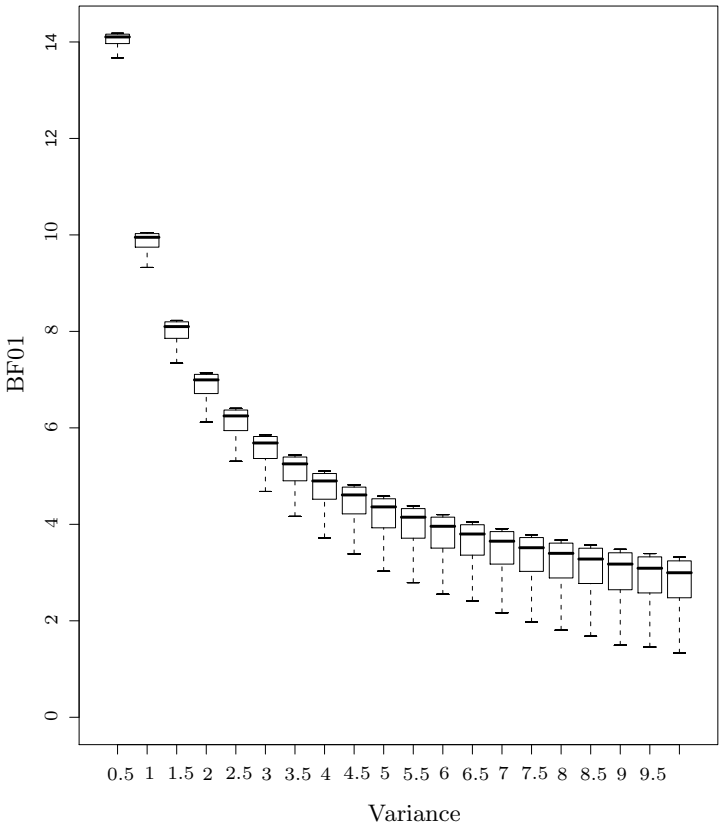

Figure 1. Boxplots of Bayes factors ordered by the variance of each gene ranging from 0.5 to 10 with an increment of 0.5.

In this paper, we propose an adjusted calibration value that weighs two types of prior predictive error probabilities differently, being more stringent on the EE genes, but also controls the overall error rate for all genes. Based on the adjusted calibration value, a novel gene selection algorithm is proposed. However, it is challenging to compute the calibration value because it requires a computationally intensive iterative procedure. Instead, we propose an alternative, but equivalent, gene selection algorithm. This alternative algorithm is attractive as (i) it only requires computation of the relative probability p^* of making a mistake under the null hypothesis for each gene; (ii) it directly provides evidence that a gene is DE; and (iii) the resulting relative probability is much easier to interpret than the original calibration value since p^* is analogous to the classical p-value.

The rest of the paper is organized as follows. In Section 2, we introduce the definition of the Bayes factor and describe different models considered in the

paper. They include a normal model with unequal variances. In Section 3, we propose an adjusted calibration method and examine its theoretical properties. We then develop a gene selection algorithm based on the calibration value. Further, the implementation issues of the proposed algorithm are discussed, and an efficient Monte Carlo algorithm is proposed when there is no closed form for the distribution of the Bayes factor. In Section 4, we carry out an extensive simulation study to investigate the performance of the proposed method. Section 5 provides an analysis of a data set from an affymetrix chip experiment. We conclude the paper with a brief discussion in Section 6.

2. The Bayes Factor Approach

We focus on the scenario in which the intensities are observed under two conditions (test or control) for replicated experiments. Let X_{1gj} denote the expression intensity of the g^{th} gene in the j^{th} sample under the first condition, and X_{2gj} the expression intensity under the second condition. There are a total of G genes with sample size n_{1g} under condition 1 and sample size n_{2g} under condition 2 for the g^{th} gene. Thus, the data on gene g can be summarized in two vectors that we label $\mathbf{X}_{1g} = (x_{1g1}, \ldots, x_{1gn_{1g}})$ and $\mathbf{X}_{2g} = (x_{2g1}, \ldots, x_{2gn_{2g}})$.

To detect whether the g^{th} gene is DE or not, we consider two hypotheses for each gene: H_{0g}: gene g is equally expressed (EE) and H_{1g}: gene g is differentially expressed (DE). The problem then becomes which hypothesis is more supported by the data. Assume the data have already been preprocessed with appropriate transformations and normalization. The two prior predictive distributions under the null hypothesis and the alternative hypothesis are given by $m_{0g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}) = \int f_{0g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}|\xi_{0g})\pi_{0g}(\xi_{0g})d\xi_{0g}$ and $m_{1g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}) = \int f_{1g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}|\xi_{1g})\pi_{1g}(\xi_{1g})d\xi_{1g}$, where $f_{0g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}|\xi_{0g})$ denotes the probability density of the g^{th} gene's data given the null H_{0g} hypothesis, and $\pi_{0g}(\xi_{0g})$ denotes the prior distribution of ξ_{0g} under H_{0g}. Similarly, $f_{1g}(\mathbf{X}_{1g}, \mathbf{X}_{2g}|\xi_{1g})$ and $\pi_{1g}(\xi_{1g})$ are defined under the alternative hypothesis. Then, given the observed data, the Bayes factor for H_{0g} against H_{1g} is given by

$$BF_{01}(g) = \frac{m_{0g}(\mathbf{X}_{1g}, \mathbf{X}_{2g})}{m_{1g}(\mathbf{X}_{1g}, \mathbf{X}_{2g})}. \tag{2.1}$$

The more evidence for the gene to be DE, the smaller the $BF_{01}(g)$.

One of the important issues for the Bayes factor based method is the specification of prior distributions. We use three examples to illustrate how prior distributions are constructed and the corresponding Bayes factors are computed, under various settings.

Example 1.(Equal Variance). Assume that we have equal replication numbers from both conditions for each gene, say n_g for gene g, and the observed expression measurements x_{1gj} and x_{2gj} come from independent normal distributions with $x_{1gj}|\mu_g, \sigma_g^2, \delta_g \sim \mathcal{N}(\mu_g - \delta_g/2, \sigma_g^2)$ and $x_{2gj}|\mu_g, \sigma_g^2, \delta_g \sim \mathcal{N}(\mu_g + \delta_g/2, \sigma_g^2)$ for $j = 1, \ldots, n_g$. The mean and variance parameters are assumed to come from conjugate priors: $\mu_g|\tau^2 \sim \mathcal{N}(0, \tau^2)$ and $\delta_g|\lambda^2 \sim \mathcal{N}(\mu_\delta, \lambda^2)$. Then the problem becomes that of testing whether $\delta_g = 0$ (H_{0g}). Let $\overline{x}_{1g}$ and $\overline{x}_{2g}$ denote the sample means of the observed intensities from conditions 1 and 2, respectively. Then the mean difference of the intensities across two conditions is $d_g = \overline{x}_{2g} - \overline{x}_{1g}$. Assuming that σ_g^2 is known, after some algebra, the Bayes factor is given by

$$BF_{01}(g) = \Big(\frac{n_g\lambda^2}{2\sigma_g^2} + 1\Big)^{\frac{1}{2}} \exp\Big\{ - \frac{n_g(d_g + \frac{2\sigma_g^2}{n_g\lambda^2}\mu_\delta)^2}{4\sigma_g^2(1 + \frac{2\sigma_g^2}{n_g\lambda^2})} + \frac{\mu_\delta^2}{2\lambda^2}\Big\}.$$

Observe that the distributions from different conditions under H_{1g} differ from each other only due to different means induced by the parameter δ_g. So the test presented here is the same as that for testing the equality of the means: $\mu_g + \delta_g/2 = \mu_g - \delta_g/2$. This is also the same as the hypotheses tested in SAM (Tusher et al. (2001)), Cyber-T (Baldi and Long (2001)), and the SHB method (Newton et al. (2004)). The Bayes factor can handle more general normal models such as unequal variances across conditions, as discussed in Example 2, or other distributions as in Example 3.

Example 2.(Unequal Variances). Assume that the intensities under two conditions have different variances. Then, the intensities under H_{1g} come from $x_{1gj}|\mu_{1g}, \sigma_{1g}^2 \sim \mathcal{N}(\mu_{1g}, \sigma_{1g}^2)$ and $x_{2gj}|\mu_{2g}, \sigma_{2g}^2 \sim \mathcal{N}(\mu_{2g}, \sigma_{2g}^2)$. Here we consider a prior predictive data-based conjugate prior (Chen and Ibrahim (2003)) for μ_{1g}, σ_{1g}^2, μ_{2g}, and σ_{2g}^2. Specifically, we take

$$\begin{aligned}
&\pi(\mu_{1g}, \mu_{2g}, \sigma_{1g}^2, \sigma_{2g}^2) \\
&\quad \propto \Big(\sigma_{1g}^2\Big)^{-\frac{a_{01g}}{2}} \exp\Big\{ - \frac{a_{01g}}{2\sigma_{1g}^2}[(\mu_{1g} - \overline{x}_{01g})^2 + s_{01g}^2]\Big\} \\
&\qquad \times \Big(\sigma_{2g}^2\Big)^{-\frac{a_{02g}}{2}} \exp\Big\{ - \frac{a_{02g}}{2\sigma_{2g}^2}[(\mu_{2g} - \overline{x}_{02g})^2 + s_{02g}^2]\Big\} \pi_0(\mu_{1g}, \mu_{2g}, \sigma_{1g}^2, \sigma_{2g}^2), \quad (2.2)
\end{aligned}$$

where $\overline{x}_{01g}$ and $\overline{x}_{02g}$ are the prior predictive means, s_{01g}^2 and s_{02g}^2 are the prior predictive variances, and $\pi_0(\mu_{1g}, \mu_{2g}, \sigma_{1g}^2, \sigma_{2g}^2)$ is an initial prior. The quantities a_{01g} and a_{02g} are positive scalar parameters that quantify one's belief in the prior predictors denoted by $(\overline{x}_{01g}, \overline{x}_{02g}, s_{01g}^2, s_{02g}^2)$. When $a_{01g} \to 0$ and

$a_{02g} \to 0$, (2.2) reduces to the initial prior. We specify the initial prior as $\pi_0(\mu_{1g}, \mu_{2g}, \sigma_{1g}^2, \sigma_{2g}^2) \propto \pi_0(\sigma_{1g}^2, \sigma_{2g}^2)$, where $\sigma_{1g}^2 \sim \mathcal{IG}(\alpha_{1g}, \beta_{1g})$ and, independently, $\sigma_{2g}^2 \sim \mathcal{IG}(\alpha_{2g}, \beta_{2g})$. Note that the prior (2.2) has a more general setting than the traditional conjugate normal-inverse gamma prior. For example, when $\pi_0 \propto 1$, (2.2) reduces to a conjugate prior. Another attractive feature of (2.2) is that it allows us to incorporate data information, where $(\overline{x}_{01g}, \overline{x}_{02g}, s_{01g}^2, s_{02g}^2)$ can be calculated from historical data. Also the initial prior is added in case there is no prior predictive information available. We choose the hyperparameters α_{1g}, α_{2g}, β_{1g}, and β_{2g}, so that the prior moments of σ_{1g}^2 and σ_{2g}^2 exist. We note that (β_{1g}, β_{2g}) are the variance-stabilized parameters, and play a role similar to the "fudge" factor in SAM. With a suitable choice of (β_{1g}, β_{2g}), a gene with small variances will not be declared DE because of a small difference in the means.

Under H_{0g}: $x_{1gj}|\mu_g, \sigma_{1g}^2 \sim \mathcal{N}(\mu_g, \sigma_{1g}^2)$ and $x_{2gj}|\mu_g, \sigma_{2g}^2 \sim \mathcal{N}(\mu_g, \sigma_{2g}^2)$. In this case, the prior predictive data-based conjugate prior reduces to $\pi(\mu_g, \sigma_{1g}^2, \sigma_{2g}^2) \propto (\sigma_{1g}^2)^{-(a_{01g}/2)} \exp\{-(a_{01g}/2\sigma_{1g}^2)[(\mu_g - \overline{x}_{01g})^2 + s_{01g}^2]\}\ (\sigma_{2g}^2)^{-a_{02g}/2} \exp\{-(a_{02g}/2\sigma_{2g}^2)[(\mu_g - \overline{x}_{02g})^2 + s_{02g}^2]\}\ \pi_0(\mu_g, \sigma_{1g}^2, \sigma_{2g}^2)$. We take $\pi_0(\mu_g, \sigma_{1g}^2, \sigma_{2g}^2) \propto \pi_0(\sigma_{1g}^2, \sigma_{2g}^2)$, and $\pi_0(\sigma_{1g}^2, \sigma_{2g}^2)$ is the same prior for $(\sigma_{1g}^2, \sigma_{2g}^2)$ under H_{1g}. After some messy algebra, the Bayes factor takes the form:

$$
\begin{aligned}
BF_{01}(g) \propto & \Big[\sum_j x_{1gj}^2 + a_{01g}(s_{01g}^2 + \overline{x}_{01g}^2) + \frac{\beta_{1g}}{2} - \frac{(n_{1g}\overline{x}_{1g} + a_{01g}\overline{x}_{01g})^2}{n_{1g} + a_{01g}}\Big]^{-\frac{1}{2}} \\
& \times \Big[\sum_j x_{2gj}^2 + a_{02g}(s_{02g}^2 + \overline{x}_{02g}^2) + \frac{\beta_{2g}}{2} - \frac{(n_{2g}\overline{x}_{2g} + a_{02g}\overline{x}_{02g})^2}{n_{2g} + a_{02g}}\Big]^{-\frac{1}{2}} \\
& \times \int \Big[1 + \frac{(n_{1g} + a_{01g})(\mu_g - \frac{n_{1g}\overline{x}_{1g} + a_{01g}\overline{x}_{01g}}{n_{1g} + a_{01g}})^2}{\sum_j x_{1gj}^2 + a_{01g}(s_{01g}^2 + \overline{x}_{01g}^2) + \frac{\beta_{1g}}{2} - \frac{(n_{1g}\overline{x}_{1g} + a_{01g}\overline{x}_{01g})^2}{n_{1g} + a_{01g}}}\Big]^{-\frac{n_{1g} + a_{01g} + 2\alpha_{1g}}{2}} \\
& \times \Big[1 + \frac{(n_{2g} + a_{02g})(\mu_g - \frac{n_{2g}\overline{x}_{2g} + a_{02g}\overline{x}_{02g}}{n_{2g} + a_{02g}})^2}{\sum_j x_{2gj}^2 + a_{02g}(s_{02g}^2 + \overline{x}_{02g}^2) + \frac{\beta_{2g}}{2} - \frac{(n_{2g}\overline{x}_{2g} + a_{02g}\overline{x}_{02g})^2}{n_{2g} + a_{02g}}}\Big]^{-\frac{n_{2g} + a_{02g} + 2\alpha_{2g}}{2}} d\mu_g. \quad (2.3)
\end{aligned}
$$

Example 3.(Non-normal Distributions). In this example, we use the same notation as before, except that x_{1gj} (x_{2gj}) denotes the raw intensity instead of the log-transformed intensity. Assume the raw intensities come from a gamma distribution, such that under H_{1g}, $x_{1gj} \sim \mathcal{G}(\alpha_g, \mu_g/\alpha_g)$, $x_{2gj} \sim \mathcal{G}(\alpha_g, \kappa_g\mu_g/\alpha_g)$, and $\mu_g \sim \mathcal{IG}(\alpha_0, \beta_0)$; under H_{0g}, everything is the same except $\kappa_g = 1$. Then the Bayes factor is given by

$$
BF_{01}(g) = \kappa_g^{\alpha_g n_{2g}} \Big[\frac{\beta_0 + \alpha_g \sum_{j=1}^{n_{1g}} x_{1gj} + \frac{\alpha_g}{\kappa_g} \sum_{j=1}^{n_{2g}} x_{2gj}}{\beta_0 + \alpha_g \sum_{i=1}^{2} \sum_{j=1}^{n_{ig}} x_{igj}}\Big]^{\alpha_0 + \alpha_g(n_{1g} + n_{2g})}.
$$

3. Threshold and Calibration Method

3.1. Constant threshold

As discussed before, the smaller the Bayes factor, the stronger the evidence against the null hypothesis. Thus, for each gene, after $BF_{01}(g)$ is computed, the threshold c can be specified so that gene g is declared to be DE if and only if $BF_{01}(g) < c$.

It is crucial to select a proper threshold for the Bayes factor in order to make a valid decision. For example, Jeffreys (1961), Kass and Raftery (1995) propose rules to define a series of constant values for the different levels of evidence of the Bayes factor against H_0. Applying the rule in Kass and Raftery (1995): if $BF_{01}(g) < 1$, the data show evidence that the gene g is DE; $BF_{01}(g) < 1/20$, the data show strong evidence to support the gene g to be DE. However, Vlachos and Gelfand (2003) and Garcia-Donato and Chen (2005) argue against these rules due to the asymmetric distributions of the Bayes factor under the two hypotheses. Consider the rule when the value 1 is chosen as the threshold. In Example 1, let $n_g = 3$, $\lambda^2 = 0.9$, $\mu_\delta = 3$, $\tau^2 = 0.5$, and $\sigma_g^2 = 0.1$. Then the probability of making mistake under H_{0g} is $Pr(BF_{01}(g) < 1|H_{0g}) = 0.03$, and the probability of making mistake under H_{1g} is $Pr(BF_{01}(g) > 1|H_{1g}) = 0.14$. Choosing a constant as a threshold, independent of g, may lead to a biased decision.

3.2. Ordering method

Liu et al. (2004) propose putting the calculated Bayes factors in nondecreasing order and choosing the top genes with small Bayes factors as the significant DE genes. This approach requires that the distribution of Bayes factors be exchangeable. Unfortunately, in most cases, the exchangeability condition is not satisfied. Even in a simple model with intensities from normal distributions, if the sample sizes are different among genes, the distribution of Bayes factors is not exchangeable across genes.

Example 1.(Continued). Define the new variable $Z_g = 2\sigma_g^2/(n_g\lambda^2)$. Consider the situation where the observed intensities on different genes may have different variances or sample sizes. Assume g_1 is a DE gene and g_2 is an EE gene. It is possible to have $\ln((Z_{g_1})^{-1} + 1) - (d_{g_1} + Z_{g_1}\mu_\delta)^2/[\lambda^2 Z_{g_1}(1 + Z_{g_1})] > \ln((Z_{g_2})^{-1} + 1) - (d_{g_2} + Z_{g_2}\mu_\delta)^2/(\lambda^2 Z_{g_2}(1 + Z_{g_2}))$. Thus the simple ordering method (Liu et al. (2004)) becomes inappropriate.

Figure 1 gives a visual illustration of the non-exchangeability for Example 1 with $n_g = 20$, $\lambda^2 = 10$, $\mu_\delta = 0$, and $p = 0.05$. While σ^2 changes from 0.5 to 10 with an increment of 0.5 at each step, we sample 10,000 Bayes factor values for each σ^2 value and then draw a boxplot for the sampled Bayes factors for each

σ^2. The plot shows that the centers of the Bayes factors monotonically decrease as a function of σ^2.

3.3. Adjusted calibration method

3.3.1. Calibration method

Let $\mathbf{X}$ denote the data. Given that c is a threshold value for the Bayes factor, then there are two possible mistakes: reject the null hypothesis when it is true (i.e., $BF_{01}(\mathbf{X}) < c$ when H_0 is true), and fail to reject the null hypothesis when it is false (i.e., $BF_{01}(\mathbf{X}) \geq c$ when H_1 is true). Consider the asymmetry of the two conditional distributions of the Bayes factor under H_0 and H_1. Garcia-Donato and Chen (2005) define the calibration value, a nonnegative number c that satisfies $Pr(BF_{01}(\mathbf{X}) \geq c|H_1) = Pr(BF_{01}(\mathbf{X}) < c|H_0)$. Although their method was designed to compare only one pair of models, it can be extended in our context by obtaining a threshold independently for each gene. Such calibration values are free of the ordering, exist, and are unique. As thresholds they ensure that the probability of wrongly choosing each gene to be DE is the same as the probability of wrongly declaring that gene to be EE. Unfortunately, this method fails to offer any adjustment for multiple comparisons over a large number of genes by considering the fact that the proportion of EE genes is usually much larger than the proportion of DE genes. Direct application of this method yields a large false positive rate (low specificity) when applied to microarray expression data.

3.3.2. Calibration method

Assume all genes have the same probability of being DE. Define $H_g = 1$ if the g^{th} gene is DE, and $H_g = 0$ if the g^{th} gene is EE, then $H_g \sim$ Bernoulli (p) for $g = 1, \ldots, G$ (Storey (2002)). On average, we expect to declare $p \times G$ genes to be DE, and propose a new calibration value to meet this expectation.

Definition 3.1. If for gene g with $Pr(H_g = 1) = p$, c_g is a non-negative value satisfying $Pr(BF_{01}(g) < c_g) = p$, then c_g is called the adjusted calibration value.

Here we declare gene g to be DE if $BF_{01}(g) < c_g$ and EE otherwise. It is clear from the definition that c_g is gene dependent. Although the distribution of $BF_{01}(g)$ may not be exchangeable, the distribution of the indicator $1_{\{BF_{01}(g)<c_g\}}$ is exchangeable. Furthermore, this adjusted calibration value has a correct proportion of genes being declared to be DE on average. Mathematically, we have $E\Big[G^{-1}\sum_{g=1}^{G} 1_{\{BF_{01}(g)<c_g\}}\Big] = G^{-1}\sum_{g=1}^{G} Pr(BF_{01}(g) < c_g) = p$. This adjusted calibration value also has several other attractive properties, formally stated in the following theorems.

Theorem 3.1. *The adjusted calibration value can be rewritten as*

$$Pr(BF_{01}(g) \geq c_g | H_g = 1) \times p = Pr(BF_{01}(g) < c_g | H_g = 0) \times (1-p). \quad (3.1)$$

Proof. Since $p = Pr(BF_{01}(g) < c_g)$, $Pr(H_g = 1) = p$, and $Pr(BF_{01}(g) < c_g) = Pr(BF_{01}(g) < c_g | H_g = 1) \times Pr(H_g = 1) + Pr(BF_{01}(g) < c_g | H_g = 0) \times Pr(H_g = 0)$, we have $p = Pr(BF_{01}(g) < c_g | H_g = 1)p + Pr(BF_{01}(g) < c_g | H_g = 0)(1-p)$, which leads to (3.1).

Theorem 3.1 shows that the adjusted calibration value has a similar form as that of Garcia-Donato and Chen (2005). When $p = 1/2$, the adjusted calibration value c_g is the same as theirs which is expected to select 50% of genes as DE. On the other hand, when p is close to 1, most genes will be declared DE genes; when p is close to 0, most genes are declared as not DE genes.

Theorem 3.2. *Let the density of the sampling distribution of $BF_{01}(g)$ under H_{ig} be $f_{BF,ig}$ for $i = 0, 1$. Assume that, for each gene, $f_{BF,0g} > 0$ and $f_{BF,1g} > 0$ for all intensities $\mathbf{X}_{1g}$ and $\mathbf{X}_{2g}$, then there exists a unique value c_g satisfying* (3.1).

Proof. Let $F_{BF,ig}$ denote the cumulative distribution function of $BF_{01}(g)$ under H_{ig} for $i = 0, 1$. Then (3.1) can be written as $p(1 - \int_0^{c_g} f_{BF,1g}(b)db) = (1-p)\int_0^{c_g} f_{BF,0g}(b)db$ or $pF_{BF,1g}(c_g) + (1-p)F_{BF,0g}(c_g) = p$. Since the function $g(t) = pF_{BF,1g}(t) + (1-p)F_{BF,0g}(t)$ is continuous and strictly increasing in $[0, \infty)$ with $\lim_{t\to\infty} g(t) = 1$ and $\lim_{t\to 0} g(t) = 0$, the adjusted calibration value exists and is unique.

Example 1.(Continued). Assuming the observed intensities come from the distribution structure defined in this example, we obtain

$$\begin{aligned} &Pr(BF_{01}(g) < c_g | H_g = 0) \\ &\quad = Pr\Big\{\chi^2_{1,\left(\frac{2\sigma_g^2\mu_\delta^2}{n_g\lambda^4}\right)} > 2\Big(1 + \frac{2\sigma_g^2}{n_g\lambda^2}\Big)\Big[\frac{1}{2}\ln(\frac{n_g\lambda^2}{2\sigma_g^2} + 1) + \frac{\mu_\delta^2}{2\lambda^2} - \ln c_g\Big]\Big\} \end{aligned}$$

and

$$\begin{aligned} &Pr(BF_{01}(g) \geq c_g | H_g = 1) \\ &\quad = Pr\Big\{\chi^2_{1,\left(1+\frac{2\sigma_g^2}{n_g\lambda^2}\right)\frac{\mu_\delta^2}{\lambda^2}} \leq \frac{4\sigma_g^2}{n_g\lambda^2}\Big[\frac{1}{2}\ln(\frac{n_g\lambda^2}{2\sigma_g^2} + 1) + \frac{\mu_\delta^2}{2\lambda^2} - \ln c_g\Big]\Big\}, \end{aligned}$$

where $\chi^2_{1,\nu}$ denotes a noncentral chi-square variate with 1 degree of freedom and non-centrality parameter ν. Thus, the adjusted calibration value c_g can be

calculated using the formula:

$$pPr\Big\{\chi^2_{1,\left(1+\frac{2\sigma_g^2}{n_g\lambda^2}\right)\frac{\mu_\delta^2}{\lambda^2}} \le \frac{4\sigma_g^2}{n_g\lambda^2}\Big[\frac{1}{2}\ln(\frac{n_g\lambda^2}{2\sigma_g^2}+1)+\frac{\mu_\delta^2}{2\lambda^2}-\ln c_g\Big]\Big\}$$
$$=\Big(1-p\Big)Pr\Big\{\chi^2_{1,\left(\frac{2\sigma_g^2\mu_\delta^2}{n_g\lambda^4}\right)} > 2\Big(1+\frac{2\sigma_g^2}{n_g\lambda^2}\Big)\Big[\frac{1}{2}\ln(\frac{n_g\lambda^2}{2\sigma_g^2}+1)+\frac{\mu_\delta^2}{2\lambda^2}-\ln c_g\Big]\Big\}.$$

For illustration, we take $n_g = 20$, $\lambda^2 = 5$, $\mu_\delta = 0$, $\sigma_g^2 = 0.5$, and $p = 0.05$. Then, the Bayes factor can be at most $(n_g\lambda^2/2\sigma_g^2+1)^{-1/2}\exp(\mu_\delta^2/2\lambda^2) = (n_g\lambda^2/2\sigma_g^2+1)^{-1/2} = 10$. Figure 2 displays the distributions of $BF_{01}(g)$ under H_{0g} and H_{1g}, respectively. In Figure 2, the intersection of the yellow/grey and red/cyan parts is the calibration value of Garcia-Donato and Chen (2005), which is around 3.1. Based on this calibration value, the probabilities of making mistake under both hypothesis (blue+yellow+gray for H_0, red for H_1) are both approximately 0.122. The intersection of the blue and green/yellow parts is the adjusted calibration value, which is around 0.4: the probability of wrongly declaring the gene to be EE (green+yellow+red, equals 0.201) is $0.95/0.05 = 19$ times the probability of wrongly declaring this gene as DE (blue, equals 0.011).

We note that in Examples 2 and 3, the model structures are so complex that none of the error probabilities evaluated from the Bayes factor have closed forms. In Example 2, even the Bayes factor does not have a closed form. A numerical integration algorithm is needed for calculating the adjusted calibration value.

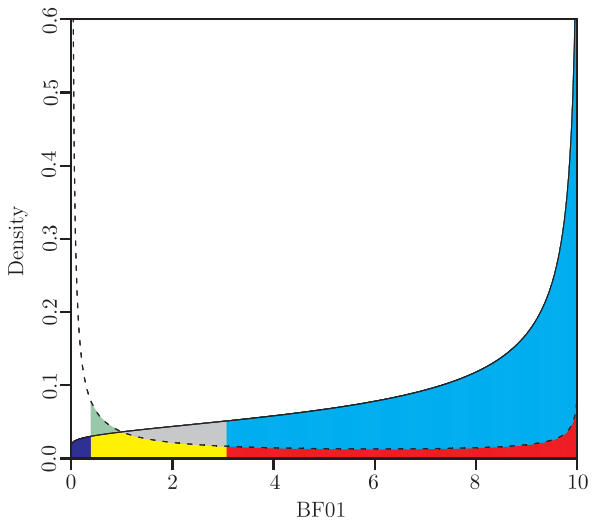

Figure 2. Density of BF_{01} on one gene under H_0 (solid line) and under H_1 (dashed line).

3.3.3. Gene selection algorithm

Let c_g denote the adjusted calibration value and let $BF^*_{01|\mathbf{X}}(g)$ be the Bayes factor based on the observed data for gene g. Consider the following gene selection algorithm: (i) compute $BF^*_{01|\mathbf{X}}(g)$; (ii) given p, compute the adjusted calibration value c_g via (3.1); and (iii) declare gene g to be DE if $BF^*_{01|\mathbf{X}}(g) < c_g$ and EE if $BF^*_{01|\mathbf{X}}(g) \geq c_g$. However, when the calibration distribution is analytically intractable, it is expensive to compute c_g in (ii). To overcome the computational difficulty of the above gene selection algorithm, we propose the following alternative gene selection algorithm: (i) compute $BF^*_{01|\mathbf{X}}(g)$; (ii) given the calculated Bayes factor $BF^*_{01|\mathbf{X}}(g)$, compute p^*_g such that

$$p^*_g = \frac{Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = 0)}{1 + Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = 0) - Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = 1)}; \tag{3.2}$$

and (iii) declare gene g to be DE if $p^*_g < p$ and EE if $p^*_g \geq p$.

From Theorem 3.3 below, we observe that the two algorithms are equivalent. In (i) and (ii), we evaluate the evidence against H_{0g}, and p^*_g in (3.2) is the *relative probability* of declaring gene g to be DE to the probability of declaring gene g to be EE. In practice, p^*_g behaves like the classical p-value because a small value of p^*_g measures evidence against H_{0g}, while a large value of p^*_g indicates evidence in favor of H_{0g}. The alternative algorithm is computationally attractive as it avoids the time-consuming iterative calculation of the threshold c_g.

Theorem 3.3. *Suppose the cumulative distribution function of the Bayes factor is continuous and strictly increasing. Given the probability $p = Pr(H_g = 1)$, $BF^*_{01|\mathbf{X}}(g) < c_g$ if and only if $p^*_g < p$.*

Proof. Since c_g is the adjusted calibration value satisfying (3.1), we have $p = h(c_g) = Pr(BF_{01}(g) < c_g|H_g = 0)/(1 + Pr(BF_{01}(g) < c_g|H_g = 0) - Pr(BF_{01}(g) < c_g|H_g = 1))$. Note that $h(c_g)$ is an increasing function of c_g and $p^*_g = h(BF^*_{01|\mathbf{X}}(g))$. Thus, under the assumption given in the theorem, $BF^*_{01|\mathbf{X}}(g) < c_g$ if and only if $p^*_g = h(BF^*_{01|\mathbf{X}}(g)) < h(c_g) = p$.

In practice, p is not known. We propose two methods for determining a guide value for it. In Method 1, we set the type I error at a pre-specified level α, calculate c_g using $Pr(BF_{01}(g) < c_g|H_g = 0) = \alpha$, and compute p_g via (3.2) by replacing $BF^*_{01|\mathbf{X}}(g)$ with c_g. We take the guide value to be the first quartile of the p_g values across all genes. In Method 2, we let SAM identify DE genes with the FDR at a pre-specified level γ, and let k_γ be the number of DE genes selected by SAM. Then the guide value is taken as the k^{th}_γ smallest ordered value of the p^*_g's obtained in (3.2). Based on empirical results obtained from a

simulation study in Section 4, and a data example in Section 5, both methods work reasonably well.

3.3.4. Computational development

When the distribution of the Bayes factor under each hypothesis is a standard distribution as in Example 1, the calculation of p_g^* is straightforward. However, in general, p_g^* is analytically intractable, as in Examples 2 and 3. Therefore, a numerical or Monte Carlo algorithm is needed for computing p_g^*. From (3.2), we see that the computation of p_g^* involves two integrals $Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = i) = \int_{BF_{01}(g)<BF^*_{01|\mathbf{X}}(g)} f_{BF,ig}(t)dt$ for $i = 0, 1$. As these integrals share the same structure, the following Monte-Carlo procedure can be used to compute them, here stated under H_{1g}. For $r = 1, 2, \ldots, R$,

S1. generate data $(\mathbf{X}^*_{1g,r}, \mathbf{X}^*_{2g,r})$ from its marginal likelihood:
  S1.1. generate $\xi_{1g,r}$ from its prior distribution $\pi_{1,g}(\xi_{1g})$,
  S1.2. given $\xi_{1g,r}$, generate data $(\mathbf{X}^*_{1g}, \mathbf{X}^*_{2g})_r$ from $f_{1g}(\mathbf{X}^*_{1g,r}, \mathbf{X}^*_{2g,r}|\xi_{1g,r})$,

S2. compute the Bayes factor $BF_{01}(g)(\mathbf{X}^*_{1g,r}, \mathbf{X}^*_{2g,r})$.

The probability $Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = 1)$ can be approximated by the proportion of $BF_{01}(g)(\mathbf{X}^*_{1g,r}, \mathbf{X}^*_{2g,r})$ less than $BF^*_{01|\mathbf{X}}(g)$, that is $Pr(BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)|H_g = 1) \approx \frac{1}{R}\sum_r 1_{\left\{BF_{01}(g)(\mathbf{X}^*_{1g,r}, \mathbf{X}^*_{2g,r})<BF^*_{01|\mathbf{X}}(g)\right\}}$, with equality attained when $R \to \infty$. The proposed computational algorithm is easy to implement, and has the additional advantage that we do not need to compute the data independent constants involved in the Bayes factor.

Theorem 3.4. *If $BF_{01}(g) = \kappa \times B(\mathbf{X}_{1g}, \mathbf{X}_{2g})$, where κ is the constant independent of the observed data and B is a function of the observed data, then the inequality: $BF_{01}(g) < BF^*_{01|\mathbf{X}}(g)$ does not involve κ.*

Thus, analytically intractable normalizing constants involved in the prior distributions under H_{0g} and H_{1g} need not be computed, and the calibration-based method is more practical than other Bayes factor-based methods, such as the one discussed in Section 3.2.

3.3.5. Protecting violation of exchangeability

As discussed in Section 3.2, in order for the ordering method (Liu et al. (2004)) to work, one requires that the distributions of Bayes factors be identical across genes. Fortunately, the adjusted calibration method can make a correct decision without exchangeability. For example, suppose gene g_1 is DE and gene g_2 is EE. For a given data set and a particular choice of the prior distribution, we may have $BF^*_{01|X}(g_2) < BF^*_{01|X}(g_1)$. Based on the ordering

method, gene g_2 is more likely to be declared DE. Based on (3.2), our calibration method will declare gene g_1 to be DE if $Pr(BF_{01}(g_1) \geq BF^*_{01|X}(g_1)|H_g = 1) \times p \geq Pr(BF_{01}(g_1) < BF^*_{01|X}(g_1)|H_g = 0) \times (1-p)$, and gene g_2 to be EE if $Pr(BF_{01}(g_2) \geq BF^*_{01|X}(g_2)|H_g = 1) \times p \leq Pr(BF_{01}(g_2) < BF^*_{01|X}(g_2)|H_g = 0) \times (1-p)$.

Example 1.(Continued). Based on the inequality above, the adjusted calibration method will protect against the violation of exchangeability if

$$\frac{p}{1-p} Pr\big[\chi^2_{1,\left(1+Z_{g_1}\right)\frac{\mu_\delta^2}{\lambda^2}} \leq \frac{(d_{g_1}+Z_{g_1}\mu_\delta)^2}{\lambda^2(1+Z_{g_1})}\big] + Pr\big[\chi^2_{1,Z_{g_1}\left(\frac{\mu_\delta^2}{\lambda^2}\right)} \leq \frac{(d_{g_1}+Z_{g_1}\mu_\delta)^2}{\lambda^2 Z_{g_1}}\big] \geq 1$$

and

$$\frac{p}{1-p} Pr\big[\chi^2_{1,\left(1+Z_{g_2}\right)\frac{\mu_\delta^2}{\lambda^2}} \leq \frac{(d_{g_2}+Z_{g_2}\mu_\delta)^2}{\lambda^2(1+Z_{g_2})}\big] + Pr\big[\chi^2_{1,\,Z_{g_2}\left(\frac{\mu_\delta^2}{\lambda^2}\right)} \leq \frac{(d_{g_2}+Z_{g_2}\mu_\delta)^2}{\lambda^2 Z_{g_2}}\big] \leq 1.$$

We note that when the exchangeability assumption is not violated, the adjusted calibration method makes the same decision as the ordering method because the adjusted calibration method uses the same percentile of the identical distribution as the threshold.

4. A Simulation Study

We conducted a simulation study to compare the performance of the adjusted calibration method (Cal. BF) to the six other methods: the ordering of Bayes factor (BF), SAM (Tusher et al. (2001)), QVALUE (Storey (2002)), SHB (Newton et al. (2004)), LIMMA (Smyth (2004)), and EBarrays (parametric empirical Bayes methods for microarray data, Kendziorski et al. (2003)).

To evaluate the performance of these seven methods, we use four error rates: false negative, false positive, conditional false discovery rate (cFDR), and false non-discovery rate (FNDR). Of the truly DE genes, the false negative rate is the proportion not detected as DE; of the truly EE genes, the false positive rate is the proportion declared to be DE. cFDR is the realized rate of false detections in the detected genes with a given positive size, and FNDR is the realized rate of false non-detections in the non-detected genes. Since the truth is known for all simulated data, the four error rates can be easily estimated.

Several simulations were conducted. In all simulations, the data were simulated so that 250 genes are in truth "differentially expressed" and 4,750 genes are in truth "not differentially expressed". In each simulation 500 data sets were generated from the same model and each method was applied to select 200, 250, and 300 genes to be DE.

Simulation I. We set $n_{1g} = n_{2g} = n_g = 10$. For each j, $j = 1, \ldots, 10$, we independently generated $x_{1gj} \sim N(\mu_g + 0.5, 0.3^2)$ and $x_{2gj} \sim N(\mu_g - 0.5, 1.2^2)$

Table 1. Method comparison based on simulation

Claimed DE	Method	Correctly Claimed DE	Correctly Claimed EE	False Neg.	False Pos.	cFDR	FNDR
200	Cal. BF	185.9 (3.5)	4735.9 (3.5)	0.257 (0.014)	0.003 (0.001)	0.071 (0.017)	0.013 (0.001)
	BF	173.4 (5.2)	4673.4 (5.2)	0.306 (0.021)	0.016 (0.001)	0.133 (0.026)	0.026 (0.001)
	SAM	156.0 (5.3)	4706.0 (5.3)	0.376 (0.021)	0.009 (0.001)	0.220 (0.027)	0.020 (0.001)
	QVALUE	135.0 (5.5)	4685.0 (5.5)	0.460 (0.022)	0.014 (0.001)	0.325 (0.027)	0.024 (0.001)
	SHB	157.3 (5.8)	4707.3 (5.8)	0.371 (0.023)	0.009 (0.001)	0.213 (0.029)	0.019 (0.001)
	LIMMA	179.6 (4.0)	4729.6 (4.0)	0.282 (0.016)	0.004 (0.001)	0.102 (0.020)	0.015 (0.001)
	EBarray	181.2 (3.9)	4731.2 (3.9)	0.275 (0.016)	0.004 (0.001)	0.094 (0.020)	0.014 (0.001)
250	Cal. BF	206.3 (4.6)	4706.3 (4.6)	0.175 (0.018)	0.009 (0.001)	0.175 (0.018)	0.009 (0.001)
	BF	184.2 (6.1)	4684.2 (6.1)	0.263 (0.024)	0.014 (0.001)	0.263 (0.024)	0.014 (0.001)
	SAM	173.7 (6.1)	4673.7 (6.1)	0.305 (0.025)	0.016 (0.001)	0.305 (0.025)	0.016 (0.001)
	QVALUE	151.6 (5.9)	4651.6 (5.9)	0.394 (0.023)	0.021 (0.001)	0.394 (0.023)	0.021 (0.001)
	SHB	166.3 (6.6)	4666.3 (6.6)	0.335 (0.026)	0.018 (0.001)	0.335 (0.026)	0.018 (0.001)
	LIMMA	198.4 (5.0)	4698.4 (5.0)	0.206 (0.020)	0.011 (0.001)	0.206 (0.020)	0.011 (0.001)
	EBarray	199.7 (4.9)	4699.7 (4.9)	0.201 (0.020)	0.011 (0.001)	0.201 (0.020)	0.011 (0.001)
300	Cal. BF	216.9 (4.6)	4666.9 (4.6)	0.132 (0.018)	0.017 (0.001)	0.277 (0.015)	0.007 (0.001)
	BF	189.4 (6.4)	4689.4 (6.4)	0.242 (0.026)	0.013 (0.001)	0.369 (0.021)	0.002 (0.001)
	SAM	185.9 (6.6)	4635.9 (6.6)	0.257 (0.027)	0.024 (0.001)	0.380 (0.022)	0.014 (0.001)
	QVALUE	164.3 (6.1)	4614.3 (6.1)	0.343 (0.024)	0.029 (0.001)	0.452 (0.020)	0.018 (0.001)
	SHB	172.3 (7.0)	4622.3 (7.0)	0.311 (0.028)	0.027 (0.001)	0.426 (0.023)	0.017 (0.002)
	LIMMA	209.0 (5.1)	4659.0 (5.1)	0.164 (0.021)	0.019 (0.001)	0.303 (0.017)	0.009 (0.001)
	EBarray	209.6 (5.3)	4659.6 (5.3)	0.162 (0.021)	0.019 (0.001)	0.301 (0.018)	0.009 (0.001)

for $g = 1, \ldots, 125$; $x_{1gj} \sim N(\mu_g - 0.5, 0.3^2)$ and $x_{2gj} \sim N(\mu_g + 0.5, 1.2^2)$ for $g = 126, \ldots, 250$; $x_{1gj}, x_{2gj} \sim N(\mu_g, 0.7^2)$ for $g = 251, \ldots, 5,000$; and $\mu_g \sim Unif(5, 11)$ for all g. The model in Example 2 was used to compute the Bayes factors, except that the variances of the intensities from different conditions under H_{0g} (EE) on each gene g satisfies $\sigma_{1g}^2 = \sigma_{2g}^2 = \sigma_g^2$. The prior parameters were specified as follows: $\bar{x}_{01g} = \bar{x}_{02g} = 0$, and $s_{01g} = s_{02g} = 0.5$ or 6 for every 50 genes alternatively. Both a_{01g} and a_{02g} were set to be $a_{0g}n_{0g}$, where $n_{0g} = n_g$ and $a_{0g} = 0.05$ when gene index g was a multiple of 8, and 0.005 otherwise.

Table 1 summarizes the average number of correctly claimed DE genes, the average number of correctly claimed EE genes, false negative, false positive, cFDR, and FNDR, with associated simulation standard errors (SE) given in parentheses. The results from Table 1 clearly show that the adjusted calibration method outperforms all six other methods based on all aspects. Among the six methods, LIMMA and EBarray performed much better than the other four. QVALUE did poorly because the q-values are calculated based on the p-values from independent two-sample t tests. Compared to the adjusted calibration method, a worse performance of the ordering of BF may be partially due to the fact that the exchangeability of the distribution of Bayes factors across genes does not hold and a relatively vague prior was specified. As expected, SHB does

not perform well because SHB assumes a constant CV (coefficient of variation) within each biological condition. A potential reason why SAM does not perform well is the setting of different variances for log-intensities from different biological conditions. Compared to the other methods, the proposed adjusted calibration method also has the smallest simulation standard errors.

We also considered some other choices for a_{0g}. For example, a_{0g} was set to be 0.1 when the gene index g is a multiple of 8, and 0.005 otherwise. Then, the average numbers of correctly claimed DE genes (SE) and the average numbers of correctly claimed EE genes (SE) are 206.5 (4.2) and 4706.5 (4.2) for Cal. BF, and 150.1 (6.9) and 4650.1 (6.9) for BF if 250 genes were selected to be DE. The results based on this choice of prior for Cal. BF are very similar to those given in Table 1, while the ordering method is quite sensitive to the prior. We also observed that when the same, but small, a_{0g} is specified for all genes, Cal. BF and ordering methods work equally well. In practice, we recommend choosing a value of a_{0g} which leads to a moderately informative prior or a relatively vague prior. Although Cal. BF is quite robust to the choice of a_{0g} in this simulation, we recommend doing a sensitivity analysis for several values of a_{0g}.

Simulation II. We again set $n_{1g} = n_{2g} = n_g = 10$. We then independently generated $\exp(x_{1gj}) \sim \mathcal{G}(\gamma_{1g}, \exp(\mu_g + 0.5)/\gamma_{1g})$ and $\exp(x_{2gj}) \sim \mathcal{G}(\gamma_{2g}, \exp(\mu_g - 0.5)/\gamma_{2g})$ for $g = 1, \ldots, 125$; $\exp(x_{1gj}) \sim \mathcal{G}(\gamma_{1g}, \exp(\mu_g - 0.5)/\gamma_{1g})$ and $\exp(x_{2gj}) \sim \mathcal{G}(\gamma_{2g}, \exp(\mu_g + 0.5)/\gamma_{2g})$ for $g = 126, \ldots, 250$; $\exp(x_{1gj})$, $\exp(x_{2gj}) \sim \mathcal{G}(\gamma_g, \exp(\mu_g)/\gamma_g)$ for $g = 251, \ldots, 5,000$; and $\mu_g \sim Unif(5, 11)$, $\gamma_{1g} \sim Unif(0.3, 1.1)$, $\gamma_{2g} \sim Unif(2.2, 6.2)$, and $\gamma_g \sim Unif(1.2, 3.2)$ for all g. The same model and prior as in Simulation I were used to compute the Bayes factors. We conducted this simulation to examine the robustness of the log-normal distribution for all seven methods.

The results are summarized in Table 2. From Tables 1 and 2, we can see that all methods performed worse in Simulation II than in Simulation I. SHB is most robust among all methods, which may be partially due to the fact that gamma distributions are assumed for the intensities on raw scale in SHB. Compared to SAM, QVALUE, LIMMA, and EBarrays, the Bayes factor-based methods are much more robust to the specification of log-normal distributions. Despite a fitted model different from the generation model, the adjusted calibration method still outperforms the six other methods.

Simulation III. We set $n_{1g} = n_{2g} = n_g$. We generated n_g randomly between 5 and 40. Then, given n_g, we independently generated $x_{1gj} \sim N(\mu_g + 0.5, 0.6^2)$ and $x_{2gj} \sim N(\mu_g - 0.5, 1.0^2)$ for $g = 1, \ldots, 125$; $x_{1gj} \sim N(\mu_g - 0.5, 0.6^2)$ and $x_{2gj} \sim N(\mu_g + 0.5, 1.0^2)$ for $g = 126, \ldots, 250$; and $x_{1gj}, x_{2gj} \sim N(\mu_g, 0.8^2)$ for $j = 1, \ldots, n_g$ and $g = 251, \ldots, 5,000$. We took $\mu_g \sim Unif(5, 11)$ for all g. The same model as in Simulation I was used to compute the Bayes factors. The prior

Table 2. Sensitivity Analysis

Claimed DE	Method	Correctly Claimed DE	Correctly Claimed EE	False Neg.	False Pos.	cFDR	FNDR
200	Cal. BF	166.8 (4.6)	4716.8 (4.6)	0.333 (0.018)	0.007 (0.001)	0.166 (0.023)	0.017 (0.001)
	BF	165.4 (4.9)	4665.4 (4.9)	0.338 (0.020)	0.018 (0.001)	0.173 (0.025)	0.028 (0.001)
	SAM	123.9 (4.1)	4673.9 (4.1)	0.504 (0.016)	0.016 (0.001)	0.380 (0.021)	0.026 (0.001)
	QVALUE	109.9 (5.1)	4659.9 (5.1)	0.560 (0.020)	0.019 (0.001)	0.450 (0.026)	0.029 (0.001)
	SHB	148.7 (5.8)	4698.7 (5.8)	0.405 (0.023)	0.011 (0.001)	0.256 (0.029)	0.021 (0.001)
	LIMMA	127.7 (4.7)	4677.7 (4.7)	0.489 (0.019)	0.015 (0.001)	0.362 (0.023)	0.025 (0.001)
	EBarray	140.0 (4.5)	4690.0 (4.5)	0.440 (0.018)	0.013 (0.001)	0.300 (0.022)	0.023 (0.001)
250	Cal. BF	185.0 (5.2)	4685.0 (5.2)	0.260 (0.021)	0.014 (0.001)	0.260 (0.021)	0.014 (0.001)
	BF	177.1 (5.7)	4677.1 (5.7)	0.292 (0.023)	0.015 (0.001)	0.292 (0.023)	0.015 (0.001)
	SAM	129.7 (4.3)	4629.7 (4.3)	0.481 (0.017)	0.025 (0.001)	0.481 (0.017)	0.025 (0.001)
	QVALUE	119.6 (4.9)	4619.6 (4.9)	0.522 (0.020)	0.027 (0.001)	0.521 (0.020)	0.027 (0.001)
	SHB	162.6 (6.3)	4662.6 (6.3)	0.350 (0.025)	0.018 (0.001)	0.350 (0.025)	0.018 (0.001)
	LIMMA	135.8 (4.7)	4635.8 (4.7)	0.457 (0.019)	0.024 (0.001)	0.457 (0.019)	0.024 (0.001)
	EBarray	148.7 (4.9)	4648.7 (4.9)	0.405 (0.019)	0.021 (0.001)	0.405 (0.019)	0.021 (0.001)
300	Cal. BF	196.6 (5.3)	4646.6 (5.3)	0.214 (0.021)	0.022 (0.001)	0.345 (0.018)	0.011 (0.001)
	BF	183.3 (6.0)	4683.3 (6.0)	0.267 (0.024)	0.014 (0.001)	0.389 (0.020)	0.004 (0.001)
	SAM	134.2 (4.8)	4584.2 (4.8)	0.463 (0.019)	0.035 (0.001)	0.553 (0.016)	0.025 (0.001)
	QVALUE	126.9 (5.0)	4576.9 (5.0)	0.492 (0.020)	0.036 (0.001)	0.577 (0.017)	0.026 (0.001)
	SHB	171.9 (6.4)	4621.9 (6.4)	0.312 (0.026)	0.027 (0.001)	0.427 (0.021)	0.017 (0.001)
	LIMMA	142.0 (4.9)	4592.0 (4.9)	0.432 (0.020)	0.033 (0.001)	0.527 (0.016)	0.023 (0.001)
	EBarray	155.0 (4.9)	4605.0 (4.9)	0.380 (0.020)	0.031 (0.001)	0.483 (0.016)	0.020 (0.001)

parameters were specified as follows: $\bar{x}_{01g} = \bar{x}_{02g} = 0$, $s_{01g} = s_{02g} = 0.5$ $a_{01g} = a_{02g} = a_0 n_{0g}$, where $n_{0g} = n_g$ and $a_0 = 0.4$ for every gene. Under this simulation setting, although the prior is exchangeable across genes, the distribution of Bayes factors is still not exchangeable due to the different sample sizes. We conducted this simulation to study the effect of sample size on the performance of the adjusted calibration method and the ordering of BF.

For this simulation, we report only the average number of correctly claimed DE genes (SE) and the average number of correctly claimed EE genes (SE), for brevity. These numbers are 160.9 (4.1) and 4710.9 (4.1) for Cal. BF, and 140.7 (4.2) and 4690.7 (4.2) for BF if 200 genes were selected to be DE; 176.9 (5.2) and 4676.9 (5.2) for Cal. BF, and 146.9 (4.3) and 4646.9 (4.3) for BF if 250 genes were selected to be DE; and 187.7 (5.5) and 4637.7 (5.5) for Cal. BF, and 151.4 (4.2) and 4601.4 (4.2) for BF if 300 genes were selected to be DE. Again, the adjusted calibration method performs much better than the ordering of BF.

The Guide Value. Under the setting of Simulation I, we also investigated performance of the two proposed methods for producing a guide value of p. Using Method 1, the average numbers of claimed DE genes and the average numbers of correctly claimed DE genes by Cal. BF were 93.55 and 93.20 for the type I

error $\alpha = 0.01$, and 127.65 and 126.65 for $\alpha = 0.02$, respectively. Using Method 2, with a control of FDR at $\gamma = 0.05$, the average number of DE genes claimed by the SAM was 82.75, and the average numbers of correctly identified DE genes were 79.55 and 82.45 by SAM and Cal. BF, respectively. For $\gamma = 0.10$, the average number of DE genes claimed by SAM was 119.35 and, in this case, SAM and Cal. BF correctly identified 108.45 and 118.55 DE genes, respectively. Thus, both proposed methods worked quite well.

5. Analysis of Microarray Data

We consider a data set from Kalajzic et al. (2005), where mouse calvarial cultures at day 7 and 17 underwent Affymetrix microarray analyses to understand the gene expression profile of osteoblast lineage at defined stages of differentiation. They demonstrated the feasibility of generating more homogeneous populations of cells at distinct stages of osteoprogenitor maturation by utilizing collα1 promoter-GFP transgenic mouse lines. They also argued the needs for doing this cell separation for valid microarray interpretations. Two statistical methods: SAM (Tusher et al. (2001)) and SHB (Newton et al. (2004)) were applied to several data sets corresponding to different stages of bone cells differentiation to select differentially expressed genes. We use only one of the data sets for illustration and comparison. We focus on detecting genes that are regulated at mature osteoblast, that is, day 17 cultures between cells with $2.3\texttt{GFP}^{\texttt{pos}}$ and cells with $2.3\texttt{GFP}^{\texttt{neg}}$.

Note that SAM uses the pooled estimate of equal variances from two conditions in modified t-statistics. SHB fits the raw intensities by gamma distributions with a fixed shape parameter. Therefore it assumes that the data have equal coefficients of variation for each condition across genes. We can explore several models for the adjusted calibration method. Table 3 lists the summary statistics for the ratios of the sample standard deviations across two conditions for all genes. We see that the median of the ratios is 2.844, and the largest ratio value is 1.289e+05. Therefore, unequal variances of intensities between the two conditions seems evident. We adopt the model in Example 2 to derive the Bayes factor and the adjusted calibration method to select DE genes. The values of a_{01g} and a_{02g} were chosen to be $a_0 n_g$, where $a_0 = 0.5$ and $n_g = 3$ and, in the initial prior $\pi_0(\sigma_{1g}^2, \sigma_{2g}^2)$, we set $\alpha_{1g} = \alpha_{2g} = 1.0$ and $\beta_{1g} = \beta_{2g} = 1.0$. Thus, a moderately informative prior was used in the analysis.

Table 3. Summary statistics of the ratios between the two sample standard deviations across genes.

Min	Q1	Median	Mean	Q3	Max
1.000	1.608	2.844	42.39	6.773	1.289e+05

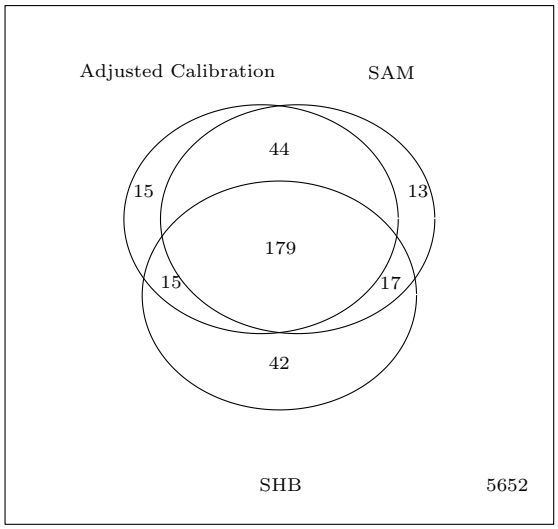

Figure 3. Number of DE genes selected by the adjusted calibration, SAM and SHB methods.

We matched the number of the selected DE genes by calibration and SHB to that of SAM. This was 253 genes, controlling the median of the false discovery rates at 5% as determined by SAM. Figure 3 shows the Venn diagram of these selected genes by the three methods. The results show 325 of 5,977 genes selected by one of the three methods, with 179 genes picked by all three. There were 255 genes picked by at least two methods, among which, 223 genes were picked by both calibration and SAM; 194 genes were picked by both calibration and SHB; 196 genes were picked by both SAM and the SHB method.

The 15 genes selected by calibration, missed by SAM or SHB, deserve our attention. Among them, *activated leukocyte cell adhesion molecule* (Alcam) plays a critical role in the differentiation of mesenchymal tissues in multiple species (Bruder et al. (1998)), the function of *chemokine c-c motif receptor 5* (Ccr5) in osteoblasts was investigated in Yano et al. (2005), and the expression of *very low density lipoprotein receptor* (Vldlr) was supported by different human osteoblast cell lines (Niemeier et al. (2005)). This example illustrates the potential advantages of the adjusted calibration method in screening for expressed genes in data.

We conducted a sensitivity analysis on the choice of the scale parameter a_0 in a_{01g} and a_{02g}. Instead of $a_0 = 0.5$, we considered $a_0 = 0.3$ and $a_0 = 0.7$. There are 230 and 220 genes selected by both SAM and the calibration method with

these choices. Of them, there are 215 DE genes in common between $a_0 = 0.3$ and $a_0 = 0.5$, and 217 DE genes in common between $a_0 = 0.7$ and $a_0 = 0.5$. That is, there are 96.41% and 97.31% genes selected in common, respectively.

Finally, we mention that if we set the type I error $\alpha = 0.01$ to determine the guide value discussed in Section 3.3.3, the calibration method selected 292 DE genes, of which 242 genes overlapped with the DE genes selected by SAM with a control of FDR at 0.05.

6. Discussion

Although only normal distributions or gamma distributions were used in our illustrative examples, the proposed calibration method can easily be extended to other types of distributions, such as a mixture of two or three normals. The theory and the gene selection algorithm discussed in Section 3 remain valid. The only complication for such extension is that the Bayes factor may become analytically intractable. Due to the recent advance in Bayesian computation, there are many efficient Monte Carlo methods available for computing Bayes factors. A hybrid of the computational algorithm discussed in Subsection 3.3.4 and the Monte Carlo methods given in Chen, Shao and Ibrahim (2000) can be developed for computing p_g^* in (3.2).

Acknowledgement

The authors wish to thank the Editors, an associate editor and a referee for helpful comments and suggestions which have led to an improvement of this article. The authors would also like to thank David Rowe and Ivo Kalajzic at the University of Connecticut Health Center for helpful discussions, providing the microarray data, and a reference for ALCAM in Section 5. The research was partially supported by NIH grant #P20 GM5764-01 for Yu and Kuo and by NIH grants #GM 70335 and #CA 74015 for Chen.

References

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p value and evidence. *J. Amer. Statist. Assoc.* **82**, 112-122.

Bruder, S. P., Ricalton, N . S., Boynton, R. E., Connolly, T. J., Jaiswal, N., Zaia, J. and Barry, F. P. (1998). Mesenchymal stem cell surface antigen SB-10 corresponds to activated leukocyte cell adhesion molecule and is involved in osteogenic differentiation. *J. Bone and Mineral Research* **13**, 655-663.

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statist. Sinica* **13**, 461-476.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70**, 193-242.

Garcia-Donato, G. and Chen, M.-H. (2005). Calibrating Bayes factor under prior predictive distributions. *Statist. Sinica* **15**, 359-380.

Jeffreys, H. (1961). *Theory of Probability*. Third edition. Clarendon Press, Oxford.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Kalajzic, I., Staale, A., Yang. W-P, Wu, Y. Johnson, S. E., Feyen, J. H. M., Krueger, W., Maye, P., Yu., F., Zhao, Y., Kuo, L., Gupta, R. R., Achenie, L. E. K., Wang, H.-W. Shin, D.-G. and Rowe, D. W. (2005). Expression profile of Osteoblast lineage at defined stages of differentiation. *J. Biological Chemistry* **280**, 24618-24626.

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist. Medicine* **22**, 3899-3914.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computational Biology* **7**, 819-837.

Liu, D.-M., Parmigiani, G. and Caffo, B. (2004). Screening for differentially expressed genes: are multilevel models helpful? Johns Hopkins University Tech Report.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155-176.

Niemeier, A., Kassem, M., Toedter, K., Wendt, D., Ruether, W., Beisiegel, U. and Heeren, J. (2005). Expression of LRP1 by human osteoblasts: a mechanism for the delivery of lipoproteins and vitamin K_1 to bone. *J. Bone and Mineral Research* **20**, 283-293.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.* **3**, No. 1, Article 3.

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55**, 62-71.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64**, 479-498.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116-5121.

Vlachos, P. K. and Gelfand, A. E. (2003). On the calibration of Bayesian model choice criteria. *J. Statist. Plann. Inference* **111**, 223-234.

Yano, S., Mentaverri, R., Kanuparthi, D., Bandyopadhyay, S., Rivera, A., Brown, E. M. and Chattopadhyay, N. (2005). Functional expression of beta-chemokine receptors in osteoblasts: role of regulated upon activation, normal T cell expressed and secreted (RANTES) in osteoblasts and regulation of its secretion by osteoblasts and osteoclasts. *Endocrinology* **146**, 2324-2335.

Department of Biostatistics, College of Public Health, 984350 Nebraska Medical Center, Omaha, NE 68198, U.S.A.

E-mail: fangyu@unmc.edu

Department of Statistics, University of Connecticut, 215 Glenbrook road, Storrs, CT, 06269-4120, U.S.A.

E-mail: mhchen@stat.uconn.edu

Department of Statistics, University of Connecticut, 215 Glenbrook road, Storrs, CT, 06269-4120, U.S.A.

E-mail: lynn@stat.uconn.edu

(Received May 2006; accepted January 2007)