BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 3 2007, pages 328–335
doi:10.1093/bioinformatics/btl612

Gene expression

Flexible empirical Bayes models for differential gene expression

Kenneth Lo* and Raphael Gottardo

Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, Canada V6T 1Z2

Received on October 1, 2006; revised on November 21, 2006; accepted on November 26, 2006
Advance Access publication November 30, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Inference about differential expression is a typical objective when analyzing gene expression data. Recently, Bayesian hierarchical models have become increasingly popular for this type of problem. The two most common hierarchical models are the hierarchical Gamma–Gamma (GG) and Lognormal–Normal (LNN) models. However, to facilitate inference, some unrealistic assumptions have been made. One such assumption is that of a common coefficient of variation across genes, which can adversely affect the resulting inference.

Results: In this paper, we extend both the GG and LNN modeling frameworks to allow for gene-specific variances and propose EM based algorithms for parameter estimation. The proposed methodology is evaluated on three experimental datasets: one cDNA microarray experiment and two Affymetrix spike-in experiments. The two extended models significantly reduce the false positive rate while keeping a high sensitivity when compared to the originals. Finally, using a simulation study we show that the new frameworks are also more robust to model misspecification.

Availability: The **R** code for implementing the proposed methodology can be downloaded at http://www.stat.ubc.ca/~c.lo/FEBarrays

Contact: c.lo@stat.ubc.ca

Supplementary information: The supplementary material is available at http://www.stat.ubc.ca/~c.lo/FEBarrays/supp.pdf

1 INTRODUCTION

As a natural development following the success of genome sequencing, DNA microarray technology has emerged for the sake of exploring the functioning of genomes (Schena *et al*., 1995). By exploiting the ability of a single-strand nucleic acid molecule to hybridize to a complementary sequence, researchers can simultaneously measure the expression levels of thousands of genes within a cell. A common task with microarray is to determine which genes are differentially expressed under two different conditions.

In recent years, there has been a considerable amount of work on the detection of differentially expressed genes. An early statistical treatment can be found in Chen *et al*. (1997) A common approach is to test a hypothesis for each gene using variants of t or F-statistics and then try to correct for multiple testing (Tusher *et al*., 2001; Efron *et al*., 2001; Dudoit *et al*., 2002). Due to the small number of replicates, variation in gene expression can be poorly estimated. Tusher *et al*. (2001) and Baldi and Long (2001) suggested using a modified t statistic where the denominator has been regularized by adding a small constant to the gene-specific variance estimate. Similar to an empirical Bayes approach this results in shrinkage of the empirical variance estimates towards a common estimate. Lönnstedt and Speed (2002) proposed an empirical Bayes normal mixture model for gene expression data, which was later extended to the two condition case by Gottardo *et al*. (2003) and to more general linear models by Smyth (2004) and Cui *et al*. (2005), though Smyth (2004) and Cui *et al*. (2005) did not use mixture models but simply empirical Bayes normal models for variance regularization. In each case, the authors derived explicit gene-specific statistics and did not consider the problem of estimating p the proportion of differentially expressed genes. Newton *et al*. (2001) developed a method for detecting changes in gene expression in a single two-channel cDNA slide using a hierarchical gamma–gamma (GG) model. Kendziorski *et al*. (2003) extended this to replicate chips with multiple conditions, and provided the option of using a hierarchical lognormal–normal (LNN) model. Both models are implemented in an **R** package called EBarrays (Empirical Bayes microarrays) and from now on we use the name EBarrays to refer to the methodology. Both EBarrays model specifications rely on the assumption of a constant coefficient of variation across genes. In this paper, we extend both models by releasing this assumption and introduce EM type algorithms for parameter estimation, thus extending the work of Lönnstedt and Speed (2002) and Gottardo *et al*. (2003) as well.

The structure of the paper is as follows. The extended forms of the two EBarrays hierarchical models and the estimation procedures are presented in Section 2. In Section 3, the performance of the extended models is examined on three experimental datasets and compared to five other baseline and commonly used methods. Section 4 presents a simulation study to further compare our empirical Bayes approach to the other methods. Finally, in Section 5 we discuss our results and possible extensions.

2 A BAYESIAN FRAMEWORK FOR IDENTIFYING DIFFERENTIAL EXPRESSION

2.1 A hierarchical model for measured intensities

In a typical microarray experiment, two conditions are compared for gene expression. Let us denote by X_{gr} and Y_{gr} the intensities of gene g from the rth replicate in the two conditions, respectively.

*To whom correspondence should be addressed.

Measurements between the two conditions are assumed to be independent. The proposed model is an extension of the EBarrays framework (Newton *et al*., 2001; Kendziorski *et al*., 2003). Extensions to the original two types of model formulation are considered in turn below.

GG. Here, a Gamma distribution is used to model the measured intensities of a given gene. Explicitly, the probability density of X_{gr} (resp. Y_{gr}) with shape and rate parameters a_g and θ_{gx} (resp. θ_{gy}) is given by

$$p(x \mid a_g, \theta_{gx}) = \frac{1}{\Gamma(a_g)} \theta_{gx}^{a_g} x^{a_g - 1} \exp(-x\theta_{gx}) \quad \text{for } x > 0. \tag{1}$$

To borrow strength across genes we assume an exchangeable Gamma(a_0, ν) prior for the rate parameters, and a Lognormal(η, ξ) prior for the shape parameters. The Gamma prior is used for simplicity as it is conjugate to the sampling distribution (Newton *et al*., 2001) while the Lognormal prior is suggested by a histogram plot of the empirical shape parameters estimated by the method of moments (see Supplementary material). The hyperparameters a_0, ν, η and ξ are assumed unknown and will be estimated as part of our approach.

The proposed model extends the EBarrays GG model by placing a prior on the shape parameter. In the original GG model, the shape parameter a was assumed to be constant and common to all genes whereas now it is gene specific. However, strength is borrowed across genes through the prior distribution. By 'borrowing strength', we mean that information from all genes is used when estimating a_g, which comes from the hyperparmeters through the prior.

LNN. The second formulation is an extension of the EBarrays LNN framework. The intensities are assumed to be lognormally distributed, i.e. the log-transformed intensities are from a normal distribution, and we write $\log X_{\text{gr}} \sim \text{N}(\mu_{gx}, \tau_{gx}^{-1})$ and $\log Y_{gr} \sim \text{N}(\mu_{gy}, \tau_{gy}^{-1})$, respectively. A conjugate prior is imposed on the mean μ_{gx} (resp. μ_{gy}) and precision τ_{gx} (resp. τ_{gy}). Explicitly, we set $\mu_{gx} \mid \tau_{gx} \sim \text{N}(m, k\tau_{gx}^{-1})$ and $\tau_{gx} \sim \text{Gamma}(\alpha, \beta)$ respectively. In the original LNN model, the precision τ was assumed to be constant and common to all genes. Our proposed formulation extends the EBarrays model by releasing the assumption of a constant coefficient of variation $\sqrt{\exp(\tau^{-1}) - 1}$, which is equivalent to the assumption of a constant variance τ^{-1} on the log scale. Note that our proposed formulation is also the framework of Gottardo *et al*. (2003). However, in this paper we use an EM based algorithm to estimate the unknown parameters, including the proportion of differentially expressed genes.

On assuming a prior on both μ_{gx} (resp. μ_{gy}) and τ_{gx} (resp. τ_{gy}) common to all genes, strength is borrowed across genes through both means and variances of the distributions when making inferences. Again, we mean that information from all genes is used when estimating both μ_{gx} (resp. μ_{gy}) and τ_{gx} (resp. τ_{gy}). In particular, this is essential for variances—due to the small number of replicates variance estimates can be very noisy. Similar ideas have been used in Smyth (2004) and Cui *et al*. (2005), where the authors concentrated on variance regularization.

2.2 A mixture model for differential expression

We use a mixture model to identify differentially expressed genes. We assume that a priori $\theta_{gx} = \theta_{gy}$ (resp. $\mu_{gx} = \mu_{gy}$) with probability $1 - p$ and $\theta_{gx} \neq \theta_{gy}$ (resp. $\mu_{gx} \neq \mu_{gy}$) with probability p. For the former case, the model specification is just as stated in Section 2.1, while the latter case is modeled through setting the gene-specific parameters common to both conditions.

Let us denote by z_g the indicator variable equal to one if there is real change in expression for gene g and zero otherwise. Then one can define the posterior probability of change, $\Pr(z_g = 1 \mid \mathbf{x}_g, \mathbf{y}_g, p, \boldsymbol{\psi})$, where $\mathbf{x}_g = (x_{g1}, x_{g2}, \ldots, x_{gR_x})'$ and $\mathbf{y}_g = (y_{g1}, y_{g2}, \ldots, y_{gR_y})'$ and $\boldsymbol{\psi}$ denotes the vector of unknown hyperparameters. Applying the Bayes rule, we obtain

$$\begin{aligned} \hat{z}_g &= \Pr(z_g = 1 \mid \mathbf{x}_g, \mathbf{y}_g, p, \boldsymbol{\psi}) \\ &= \frac{p p_A(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi})}{p p_A(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi}) + (1-p) p_0(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi})}, \end{aligned} \tag{2}$$

where $p_A(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi})$ and $p_0(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi})$ denote the joint marginal density of the measured intensities of gene g under both the alternative (differential expression) and null (no differential expression) models respectively given $\boldsymbol{\psi}$. The marginal density for the extended LNN model can be computed explicitly and is given in Appendix. For the extended GG model only θ_g can be integrated out, and the corresponding 'conditional' marginal density is given in Appendix. In the next section we describe an approximate estimation procedure to deal with this difficulty.

2.3 Parameter estimation using the EM-algorithm

Here we start with the extended LNN model as the estimation procedure is straightforward. The vector of unknown parameters $\boldsymbol{\Phi} = (\boldsymbol{\psi}', p)'$, where $\boldsymbol{\psi} = (m, k, \alpha, \beta)'$, can be estimated by maximizing the integrated likelihood using the EM-algorithm (Dempster *et al*., 1977). The estimation of p is important since it calibrates the posterior probability of change for multiple testing, as seen in (2). Such estimation is also part of some multiple testing procedure such as Storey's q-value (Storey, 2003). Estimation of the parameter p can be difficult (Smyth, 2004; Bhowmich *et al*., 2006), and as suggested by Newton *et al*. (2001) we place a Beta(2,2) prior over p, which avoids numerical issues when p gets close to 0 or 1. Given the large number of genes, the prior on p has essentially no effect on the final estimation, and thus on the number of genes called differentially expressed.

Treating the z_g's as missing data, the complete data log-likelihood is given by

$$\begin{aligned} l_c(\boldsymbol{\Phi} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_g [&z_g \log p_A(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi}) + (1 - z_g) \log p_0(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi}) \\ &+ (1 + z_g) \log p + (2 - z_g) \log (1 - p)]. \end{aligned} \tag{3}$$

During the E-step, the expectation is obtained by replacing z_g by $\hat{z}_g$ as given by (2) while the M-step consists of maximizing the conditional expectation with respect to the parameter vector $\boldsymbol{\Phi} = (\boldsymbol{\psi}', p)'$. At convergence, the estimated parameters can be substituted into (2) to compute the posterior probability of change for each gene.

Because the prior of the extended GG model is not conjugate to the sampling distribution, only the marginal density conditional on

a_g is analytically available for each gene. We refer to it as the conditional marginal density. To incorporate information about the prior for the a_g's, we propose to estimate the hyperparameters η and ξ beforehand through an empirical Bayes approach using the method of moments (see Appendix for details), and add $\log[\pi(a_g \mid \eta, \xi)]$ to the log conditional density as a penalty term. Again, treating the z_g's as missing data, the corresponding modified complete data log-likelihood can be written as

$$\begin{aligned}\tilde{l}_c(\Phi \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_g \Big\{ & z_g \log p_A(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi}, a_g) \\ & + (1 - z_g) \log p_0(\mathbf{x}_g, \mathbf{y}_g \mid \boldsymbol{\psi}, a_g) \\ & + (1 + z_g) \log (p) + (2 - z_g) \log (1 - p) \\ & + \log \pi(a_g \mid \eta, \xi) \Big\},\end{aligned} \tag{4}$$

where $\boldsymbol{\psi} = (a_0, \nu)'$. The vector of parameters to be estimated becomes $\boldsymbol{\Phi} = (a_1, a_2, \ldots, a_G, \psi', p)'$.

Similar to the extended LNN model, we can use the EM algorithm to maximize the modified marginal likelihood. During the E-step to obtain the conditional expectation of the modified complete data log-likelihood z_g in (4) is replaced by $\hat{z}_g$ as in (2). The M-step consists of maximizing $\tilde{l}_c$ given the current z_g's. Such maximization can be difficult given the high-dimensionality of $\boldsymbol{\Phi}$ and here we suggest to exploit the conditional structure of the model during the maximization step, namely that given $\boldsymbol{\psi}$ and p, the genes are conditionally independent and each a_g can be maximized over separately. Let us split the unknown parameters into two groups, namely, $\boldsymbol{\Phi}_1 = (a_1, a_2, \ldots, a_G)'$ (gene-specific shape parameters) and $\boldsymbol{\Phi}_2 = (\boldsymbol{\psi}', p)'$ (global parameters). Then the M-step would consist of iteratively maximizing over $\boldsymbol{\Phi}_1$ given $\boldsymbol{\Phi}_2$ and $\boldsymbol{\Phi}_2$ given $\boldsymbol{\Phi}_1$. Here, we decided to maximize over $\boldsymbol{\Phi}_1$ only once during the first iteration to reduce the computational burden, and then take EM-iterations with respect to $\boldsymbol{\Phi}_2$ only until convergence. It turns out that the estimates obtained were very similar to the ones obtained when maximizing over both $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$, while significantly reducing the computing time.

Details about the estimation of (η, ξ) and initialization of the EM algorithm can be found in Appendix.

3 APPLICATION TO EXPERIMENTAL DATA

3.1 Data description

To illustrate our methodology we use three publicly available microarray datasets: one cDNA experiment and two Affymetrix spike-in experiments. All three have the advantage that in each case the true state (differentially expressed or not) of all or some of the genes is known.

The HIV-1 data. The expression levels of 4608 cellular RNA transcripts were measured 1 h after infection with human immunodeficiency virus type 1 (HIV-1) using four replicates on four different slides. 13 HIV-1 genes have been included in the set of RNA transcripts to serve as positive controls, i.e. genes known in advance to be differentially expressed. Meanwhile, 29 non-human genes have also been included and act as negative controls, i.e. genes known to be not differentially expressed. Another dataset was obtained by repeating the four aforementioned experiments but with an RNA preparation different from that for the first dataset. For easy reference, in this paper we label the two datasets as HIV-1A and HIV-1B, respectively. See van't Wout *et al*. (2003) for more details of the HIV-1 data. The data were lowess normalized using a global lowess normalization step (Yang *et al*., 2002).

The HGU95A Spike-in data. This dataset was obtained from a spike-in study by Affymetrix used to develop and validate the MAS 5.0 (Affymetrix Manual, 2001) platform. The concentrations of 14 spiked-in human gene groups in 14 groups of HGU95A GeneChip© arrays were arranged in a Latin square design. The concentrations of the 14 groups in the first array group are 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 pM, respectively. Each subsequent array group rotates the spike-in concentrations by one group such that each human gene was spiked-in at a particular concentration level on exactly one array group, and each concentration level came with exactly one spiked-in gene group in each array group. There are three technical replicates in each array group. The third array group has been removed from the analysis as one of its replicates was missing. We use a set of 16 spiked-in genes in our list in recognition of the extras reported by Hsieh *et al*. (2003) and Cope *et al*. (2004). Analysis is performed on each set of probe summary indices computed using gcRMA (Wu *et al*., 2004), RMA (Irizarry *et al*., 2003b), MAS 5 and dChip (Li and Wong, 2001), respectively.

The HGU133A Spike-in data. This dataset was obtained from another spike-in study done with HGU133A arrays. A total of 42 spiked-in genes were organized in 14 groups, and the concentration used were 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512 pM. The arrangement of the spike-in concentrations was similar to the Latin square design stated above. Again, there are three technical replicates in each array group. For more information see Irizarry *et al*. (2003a). In addition to the original 42, we claim that another 20 genes should also be included in the spiked-in gene list as they consistently show significant differential expression across the array groups in the exploratory data analysis. Similar observations have been made by Sheffler *et al*. (2005). Moreover, the probe sets of three genes contain probe sequences exactly matching those for the spiked-ins. These probes should be hybridized by the spike-ins as well. As a result, our expanded spiked-in gene list contains 65 entries in total.

3.2 Results

We compare our proposed methods—extended GG (eGG) and extended LNN (eLNN) models—to five other methods, namely, EBarrays GG and LNN models, the popular Significance Analysis of Microarrays (SAM) (Tusher *et al*., 2001), Linear Models for Microarray data (LIMMA) (Smyth, 2004), and a fully Bayesian approach named BRIDGE (Gottardo *et al*., 2006a). The results have been organized in Tables 1–3.

In the analysis of the HIV-1 data, we obtain the number of genes called differentially expressed (DE) for each method. Among those genes called DE, we look at the number of true positives (TP), i.e. genes known to be DE in advance, and the number of false positives (FPs), i.e. genes known to be not DE. Gottardo *et al*. (2006b) showed that one of the HIV genes, which was expected to be highly differentially expressed had a very small estimated log ratio and did not properly hybridize in the second experiment (HIV-1B).

Table 1. Analysis of differential expression with the HIV-1 data

Method	DE	TP*	FP*
HIV-1A			
GG	24	13	0
LNN	18	13	1
eGG	13	13	0
eLNN	14	13	0
LIMMA	13	13	0
SAM	13	13	0
BRIDGE	14	13	0
HIV-1B			
GG	18	11	1
LNN	18	11	1
eGG	12	11	0
eLNN	12	11	0
LIMMA	11	11	0
SAM	13	11	0
BRIDGE	11	11	0

The FDR is controlled at 0.1.
*The numbers of TP and FP are based on the controls, namely, the 13 (resp. 12 in the second experiment) HIV-1 and the 29 non-human genes of which the states are known in advance, only. They do not represent the true numbers of TP and FP in the entire data.

We removed the corresponding gene from the list of known differentially expressed genes. Thus there are 13 genes known to be DE in the first experiment and 12 in the second. To compare the performance between the seven methods, we intend to control the false discovery rate (FDR) at a fixed level of 0.1. The FDR cutoffs can be selected using a direct posterior probability calculation as described in Newton *et al*. (2004). For the HIV-1A dataset, when the FDR is controlled at 0.1, all methods can identify the 13 positive controls. Meanwhile, EBarrays LNN has made one FP. Similar result is observed when the HIV-1B dataset is considered. All methods detect 11 out of the 12 positive controls but both versions of EBarrays (GG and LNN) have made one FP. Concluded from the HIV-1 datasets, along with LIMMA, SAM and BRIDGE our proposed eGG and eLNN methods appear to perform the best as they recognize the most positive controls and do not get any FP.

For the HGU95A spike-in data, after removing the array group with one missing replicate, we have a set of 13 array groups. To evaluate the different methods we compare the first array group to the other array groups, leading to 12 comparisons. Since dChip may return negative probe summary indices, which cannot be processed by the aforementioned methods, those genes with negative summary indices were filtered out. This excluded 5.5 spike-ins on average. This time, since we know the actual status of each gene, we can check the true FDR of each method against the desired FDR. In addition, we look at the number of false negative (FNs) as a power assessment.

Unlike the results on the HIV-1 data, SAM does not show a competitive performance. A large number of FN (>11) have been observed with SAM for both gcRMA and RMA summary indices, considering that there are only 16 entries in our spiked-in gene list. eLNN and LIMMA have the actual FDR closest to the desired FDR in general, though they have a relatively large number

Table 2. Analysis of differential expression with the HGU95A spike-in data

Method	FN	FDR
gcRMA		
GG	2.42	0.22
LNN	1.83	0.22
eGG	1.58	0.28
eLNN	5.83	0.09
LIMMA	4.33	0
SAM	11.25	0.05
BRIDGE	3.6	0.06
RMA		
GG	2.42	0.28
LNN	2.42	0.25
eGG	2.25	0.2
eLNN	3.25	0.15
LIMMA	3.08	0.08
SAM	12.58	0.23
BRIDGE	2.33	0.17
MAS5		
GG	6.5	0.7
LNN	5.42	0.84
eGG	4.33	0.53
eLNN	7.08	0.26
LIMMA	5.58	0.27
SAM	5.83	0.27
BRIDGE	12.08	0
dChip		
GG	3.25	0.7
LNN	3.58	0.74
eGG	2.83	0.43
eLNN	6.08	0.34
LIMMA	4.83	0.3
SAM	3	0.45
BRIDGE	4.00	0.34

The FDR is controlled at 0.1. The values of FN and FDR shown are the averages across the 12 comparisons.

of FN cases regarding MAS 5 and dChip summary indices. The actual FDRs for EBarrays GG and LNN methods are too high compared to the other methods, and our proposed extended versions have lowered the rates by a wide margin while keeping relatively small FN rates.

The HGU133A spike-in data have a set of 14 array groups, and therefore 13 comparisons have been made. A total of 14 out of 65 spiked-in genes on average have been filtered from the analysis with dChip due to negative summary indices. The relative performance of the six methods is similar to that for the HGU95A data. It is worth mentioning that eGG is the only method that can sustain the FN cases to a low number for all four types of probe summary indices, though its FDR is higher than the desired one. SAM has considerably more FN cases than the other methods for gcRMA and RMA, while its FDR is close to the desired one. Similarly, eLNN and LIMMA exhibit good FDR performance but with better FN rates. Again, the FDRs for EBarrays GG and LNN methods are at quite a high-level, while their extended versions (eGG and eLNN) have significantly reduced the rates while keeping relatively small FN rates.

Table 3. Analysis of differential expression with the HGU133A spike-in data

Method	FN	FDR
gcRMA		
GG	5.85	0.2
LNN	5.92	0.2
eGG	6.46	0.23
eLNN	13.08	0.07
LIMMA	10.38	0.08
SAM	22.23	0.12
BRIDGE	6.01	0.11
RMA		
GG	4.38	0.14
LNN	4.46	0.13
eGG	5.23	0.06
eLNN	6.69	0.09
LIMMA	6.15	0.03
SAM	17.15	0.1
BRIDGE	4.53	0.08
MAS5		
GG	15.77	0.89
LNN	15.85	0.87
eGG	9.23	0.59
eLNN	15.77	0.23
LIMMA	13.85	0.31
SAM	13.77	0.28
BRIDGE	18.46	0.25
dChip		
GG	9.31	0.48
LNN	9.69	0.58
eGG	6.69	0.44
eLNN	11.31	0.3
LIMMA	9.38	0.26
SAM	5.08	0.28
BRIDGE	6.92	0.51

The FDR is controlled at 0.1. The values of FN and FDR shown are the averages across the 13 comparisons.

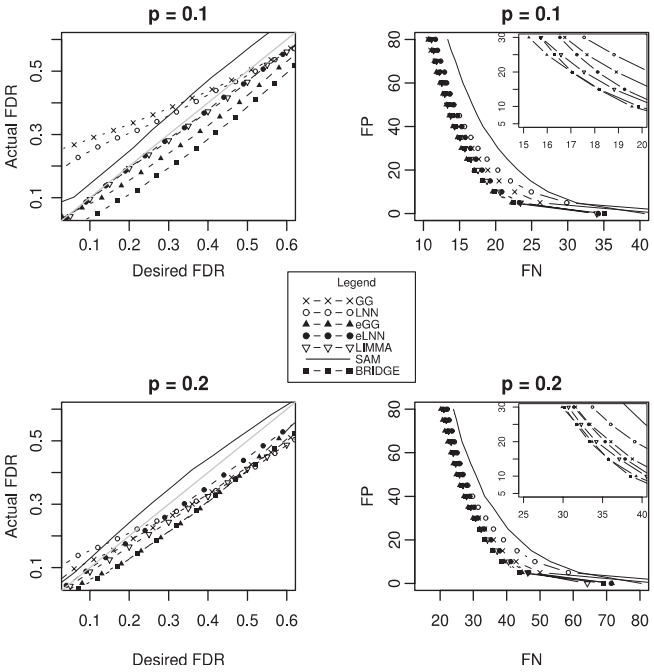

Fig. 1. Simulation results generated from the extended GG model.

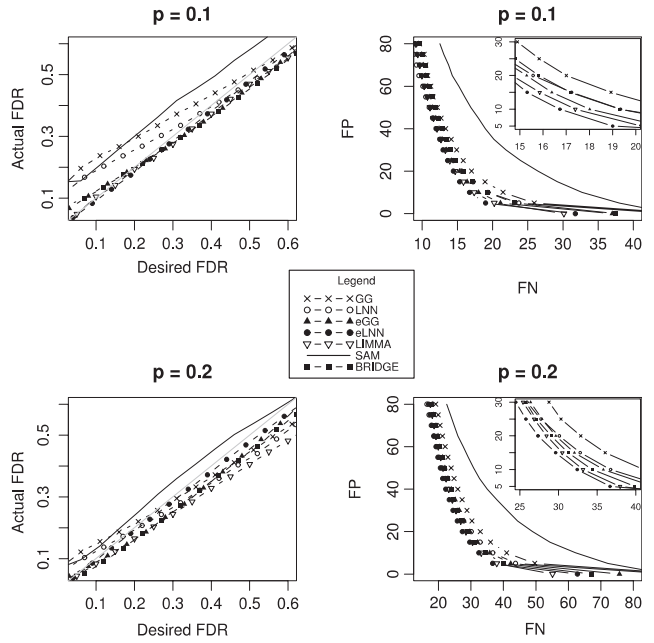

Fig. 2. Simulation results generated from the extended LNN model.

4 SIMULATION STUDIES

4.1 Data generation

We now use a series of simulation to study the performance of our empirical Bayes framework under different model specifications compared to the original EBarrays framework and the methods presented in Section 3.2. In order to do so, we generated data from the following models: EBarrays GG ($a = 5$, $a_0 = 0.8$, $\nu = 15$), EBarrays LNN ($m = 5$, $\sigma^2 = 2$, $\tau^{-1} = 0.25$, σ^2 being the variance parameter of the prior of μ_{gx} or μ_{gy}), extended GG ($\eta = 2$, $\xi = 1, a_0 = 1, \nu = 20$) and extended LNN ($m = 5, k = 12, \alpha = 2, \beta = 0.5$). The values of the parameters are set in the proximity of the estimates from the HIV-1 data. We fixed the number of genes to 500, the number of replicates to three in each group and generated 100 datasets under each of the above models for two different values of $p = \{0.1,0.2\}$.

4.2 Results

The seven methods mentioned in Section 3.2 are applied to each simulated dataset to make inference about differential expression. Results are summarized graphically in two ways: a plot of the actual FDR against the desired FDR, and a plot of the number of FP against the number of FN. The curves show the average results across the 100 simulated datasets. For each dataset, results are collected by setting the cutoffs for the posterior probabilities or p-values at different points in turn in detecting differential expression.

As expected, the EBarrays GG and LNN models perform quite poorly compared to the eGG and eLNN models when the variance is

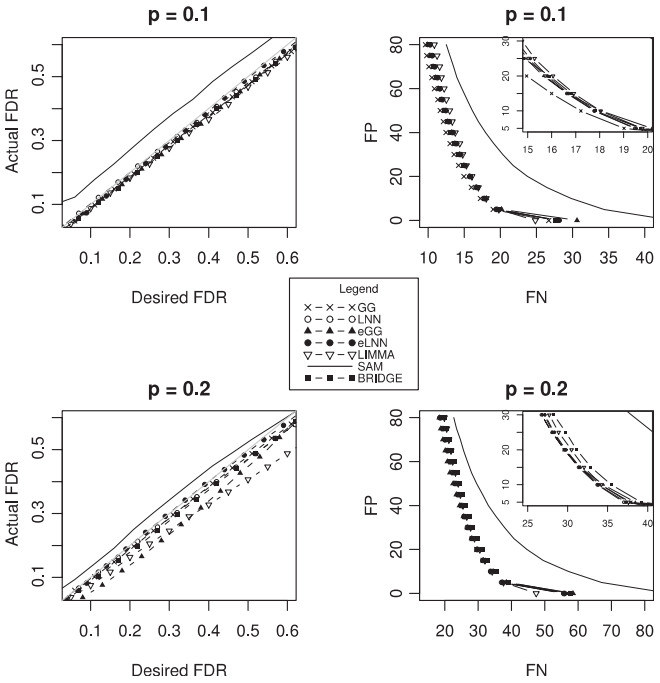

Fig. 3. Simulation results generated from the EBarrays GG model.

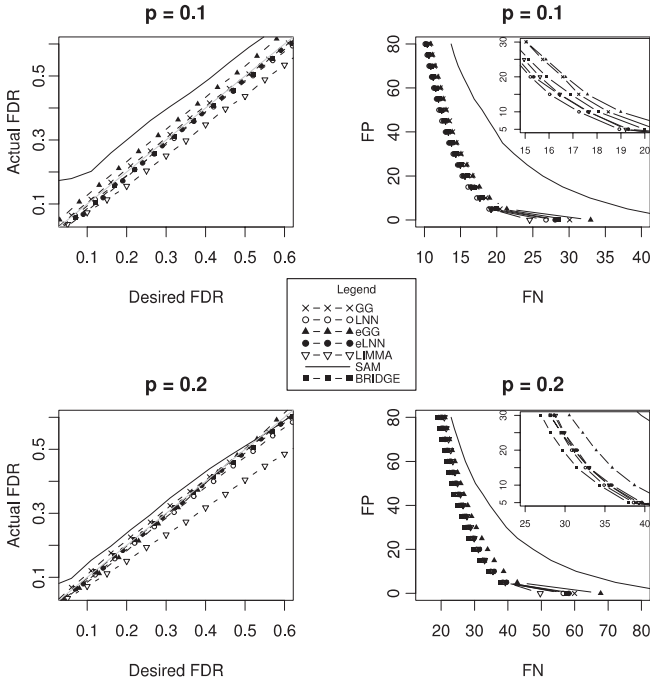

Fig. 4. Simulation results generated from the EBarrays LNN model.

not constant and clearly under estimate the FDR (Figs 1 and 2). On the other hand, the eGG and eLNN models are comparable to EBarrays when the variance is constant, showing that strength borrowing across genes is working well (Figs 3 and 4). Finally, both GG and eGG (resp. LNN and eLNN) appear to perform relatively well under LNN and eLNN (resp. GG and eGG) model specifications respectively. This confirms previous simulation studies (Kendziorski *et al*., 2003).

Overall, SAM is not performing very well and tend to under estimate the FDR by a large amount. Meanwhile, LIMMA and BRIDGE consistently show good performance for data generated from the four models, suggesting that they are good candidates for identifying differential expression under a wide variety of settings.

5 DISCUSSION

We have extended the EBarrays empirical Bayes framework for differential gene expression by releasing the constant coefficient of variation assumption, and introducing two algorithms that can be used for parameter estimation. Using both experimental and simulated data we have shown that the extended framework clearly improves the original framework. In addition, it appears that the eLNN model performs better than the eGG one as shown with the spike-in data, and that it is comparable to BRIDGE, a more computational fully Bayesian approach. This is not the case for the original EBarrays framework, where the GG model generally performs better. This confirms previous findings of Gottardo *et al*. (2006a) and suggests that EBarrays GG is more robust to the model misspecification of a constant coefficient of variation compared to the LNN formulation. However, when the EBarrays model formulations are extended and the constant coefficient of variation assumption is released, the LNN model seems more appropriate.

In spite of the complications accompanying the model enhancements relative to the original EBarrays framework, the proposed methodology remains to be highly competitive in terms of processing time. In the analysis with the HGU133A data of >20 000 genes, it takes about 5 min to complete the eGG or eLNN analysis of one comparison between the two array groups each with three replicates on the **R** platform.

In this paper, we have compared our approach with five alternatives, but there are many other methods for detecting differentially expressed genes with gene expression data. We chose these five because they are either obvious baseline methods or widely used; they are also representative of other methods. More comparisons between statistical tests can be found in Cui and Churchill (2003). Among explicit adjustments for multiple testing, we considered only the FDR control method as it is interpretable under each method.

For simplicity and ease of comparison, we assumed that we were in a situation with only two conditions of interest. However, the methodology could easily be extended to the multiple condition case (Kendziorski *et al*., 2003) or more complex ANOVA type designs (Cui and Churchill, 2003; Smyth, 2004).

ACKNOWLEDGEMENTS

The authors thank Adrian Raftery, Ka Yee Yeung and Roger Bumgarner for helpful discussion, and the two referees and the associate editor for suggestions that clearly improved an earlier draft of the article. This work was supported by a grant from the National Sciences and Engineering Research Council of Canada.

Conflict of Interest: none declared.

REFERENCES

Affymetrix Manual (2001), *Affymetrix Microarray Suite User Guide version 5.0*, Santa Clara, CA.

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

Bhowmick,D. *et al.* (2006) A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics*, **7**, 630–641.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, **2**, 364–374.

Cope,L.M. *et al.* (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Cui,X. and Churchill,G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Cui,X. *et al.* (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B*, **39**, 1–38.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111–139.

Durbin,B.P. *et al.* (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Gottardo,R. *et al.* (2003) Statistical analysis of microarray data: a Bayesian approach. *Biostatistics*, **4**, 597–620.

Gottardo,R. *et al.* (2006a) Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, **62**, 10–18.

Gottardo,R. *et al.* (2006b) Quality control and robust estimation of cDNA microarray with replicates. *J. Am. Stat. Assoc.*, **11**, 30–40.

Hsieh,W.P. *et al.* (2003) Who are those strangers in the latin square? In: Johnson,K.F. and Lin,S.M. (Eds.), *Methods of Microarray Data Analysis III: Papers from CAMDA'02*, Kluwer, Boston, pp. 199–208.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Irizarry,R.A. *et al.* (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A. *et al.* (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **2**, 249–264.

Kendziorski,C.M. *et al.* (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist. Med.*, **22**, 3899–3914.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Lönnstedt,I. and Speed,T. (2002) Replicated microarray data. *Statist. Sinica*, **12**, 31–46.

Newton,M.A. *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Newton,M. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Schena,M. *et al.* (1995) Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, **270**, 467–470.

Sheffler,W. *et al.* (2005) A learned comparative expression measure for Affymetrix GeneChip DNA microarrays. *Proc. Comput. Syst. Bioinform. Conf.*, 144–154.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Statist.*, **31**, 2013–2035.

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.

van't Wout,A.B. *et al.* (2003) Cellular gene expression upon human immunodeficiency virus type 1 infection of $CD4^+$ T-cell lines. *J. Virol.*, **77**, 1392–1402.

Wu,Z. *et al.* (2004) A model based background adjustment for oligonucleotide expression data. *J. Am. Stat. Assoc.*, **99**, 909–917.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

APPENDIX

Marginal densities of measured intensities

Under the extended GG model, the joint marginal densities of measured intensities of a given gene g are developed without integrating a_g away, i.e. they are conditional on a_g. Denote by $\mathrm{G}(x;a,b)$ the Gamma density function with shape a and rate b. The explicit forms of the conditional marginal densities are given by

$$\begin{aligned} p_A(\mathbf{x}_g,\mathbf{Y}_g\,|\,\boldsymbol{\psi},a_g) &= \int_0^\infty \prod_{l=1}^{n} \mathrm{G}(x_{gl};\,a_g,\theta_{gx})\;\mathrm{G}(\theta_{gx};\boldsymbol{\psi})d\theta_{gx} \\ &\quad\times \int_0^\infty \prod_{i=l}^{n} \mathrm{G}(y_{gl};\,a_g,\theta_{gy})\;\mathrm{G}(\theta_{gy};\boldsymbol{\psi})d\theta_{gy} \\ &= \left\{\frac{\Gamma(na_g+a_0)}{\Gamma^n(a_g)\Gamma(a_0)}\right\}^2 \frac{\nu^{2a_0}\left(\prod_l x_{gl}y_{gl}\right)^{a_g-1}}{\left[(\nu+\sum_l x_{gl})(\nu+\sum_l y_{gl})\right]^{na_g+a_0}} \end{aligned} \tag{5}$$

and

$$\begin{aligned} p_0(\mathbf{x}_g,\mathbf{y}_g\,|\,\boldsymbol{\psi},a_g) &= \int_0^\infty \prod_{l=1}^{n} \mathrm{G}(x_{gl};\,a_g,\theta_g)\prod_{l=1}^{n}\mathrm{G}(y_{gl};\,a_g,\theta_g)\cdot\mathrm{G}(\theta_g;\boldsymbol{\psi})d\theta_g \\ &= \frac{\Gamma(2na_g+a_0)}{\Gamma^{2n}(a_g)\Gamma(a_0)}\frac{\nu^{a_0}\left(\prod_l x_{gl}y_{gl}\right)^{a_g-1}}{\left(\nu+\sum_l x_{gl}+\sum_l y_{gl}\right)^{2na_g+a_0}}, \end{aligned} \tag{6}$$

where $\boldsymbol{\psi}=(a_0,\,\nu)'$.

The joint marginal densities of measured intensities under the extended LNN model are developed in a similar fashion, this time by integrating μ_g and τ_g away. Denote by $\mathrm{LN}(x;a,b)$ the Lognormal density function with mean and variance parameters a and b, respectively, and by $\mathrm{N}(x;a,b)$ the normal density function. The marginal densities are developed as follows:

$$\begin{aligned} p_A(\mathbf{x}_g,\mathbf{y}_g\,|\,\boldsymbol{\psi}) &= \int_0^\infty\int_{-\infty}^\infty \prod_l \mathrm{LN}(x_{gl};\,\mu_{gx},\tau_{gx}^{-1})\;\mathrm{N}(\mu_{gx};m,k\tau_{gx}^{-1}) \\ &\quad\times\mathrm{G}(\tau_{gx};\alpha,\beta)d\mu_{gx}d\tau_{gx} \\ &\quad\times\int_0^\infty\int_{-\infty}^\infty \prod_l \mathrm{LN}(y_{gl};\,\mu_{gy},\tau_{gy}^{-1})\;\mathrm{N}(\mu_{gy};m,k\tau_{gy}^{-1}) \\ &\quad\times\mathrm{G}(\tau_{gy};\alpha,\beta)d\mu_{gy}d\tau_{gy} \\ &= \frac{\beta^{2\alpha}\Gamma^2(\frac{n}{2}+\alpha)}{\left(\prod_l x_{gl}y_{gl}\right)(2\pi)^n(kn+1)\Gamma^2(\alpha)} \\ &\quad\times\left\{\beta+\frac{1}{2k}\left[\frac{-(k\sum_l \log x_{gl}+m)^2}{kn+1}+k\sum_l(\log x_{gl})^2+m^2\right]\right\}^{-(\frac{n}{2}+\alpha)} \\ &\quad\times\left\{\beta+\frac{1}{2k}\left[\frac{-(k\sum_l \log y_{gl}+m)^2}{kn+1}+k\sum_l(\log y_{gl})^2+m^2\right]\right\}^{-(\frac{n}{2}+\alpha)} \end{aligned} \tag{7}$$

and

$$\begin{aligned} p_0(\mathbf{x}_g,\mathbf{y}_g\,|\,\boldsymbol{\psi}) &= \int_0^\infty\int_{-\infty}^\infty \prod_l \mathrm{LN}(x_{gl};\mu_g,\tau_g^{-1})\prod_l \mathrm{LN}(y_{gl};\mu_g,\tau_g^{-1}) \\ &\quad\times \mathrm{N}(\mu_g;m,k\tau_g^{-1})\;\mathrm{G}(\tau_g;\alpha,\beta)d\mu_g d\tau_g \\ &= \frac{\beta^\alpha\Gamma(n+\alpha)}{\left(\prod_l x_{gl}y_{gl}\right)(2\pi)^n(2kn+1)^{\frac{1}{2}}\Gamma(\alpha)} \\ &\quad\times\left\{\beta+\frac{1}{2k}\left(\frac{-\left[k(\sum_l \log x_{gl}+\sum_l \log y_{gl})+m\right]^2}{2kn+1}\right.\right. \\ &\quad\left.\left.+k\left[\sum_l(\log x_{gl})^2+\sum_l(\log y_{gl})^2\right]+m^2\right)\right\}^{-(n+\alpha)}, \end{aligned} \tag{8}$$

where $\boldsymbol{\psi}=(m,k,\alpha,\beta)'$.

Estimation of η and ξ for the prior of a_g

As mentioned in Section 2.3, to make use of the modified complete data log-likelihood (4) in the extended GG model we need to provide estimates of the hyperparameters for the Lognormal(η,ξ) prior of a_g beforehand. Here we propose to use the method of moments (MMs) to estimate η and ξ. First we would like to come up with simple estimates of the a_g's. On noting that the coefficient of variation is given by $1/\sqrt{a_g}$ for each gene, a robust empirical estimate of a_g may be provided by

$$\tilde{a}_g=\frac{\mathrm{med}(\mathbf{x}_g,\mathbf{y}_g)^2}{\mathrm{mad}(\mathbf{x}_g,\mathbf{y}_g)^2},$$

where med and mad stand for median and median absolute deviation, respectively. Note that a robust counterpart to mean and SD is adopted since there are usually relatively few replicates. With these crude estimates of a_g's, we can then obtain the estimates of η and ξ:

$$\hat{\eta}=\mathrm{med}(\{\log\tilde{a}_g\})\quad\text{and}\quad\hat{\xi}=\mathrm{mad}(\{\log\tilde{a}_g\})^2. \tag{9}$$

Again, a robust version of MM is proposed here.

Initialization of the EM algorithm

We need to initialize the parameters to be estimated before the EM type algorithm described in Section 2.3 can be applied. Similar to the estimation for η and ξ above, robust MM estimates of (a,a_0,ν) are obtained for the extended GG model. Similar measure is taken for (m,α,β) if the data are modeled under the extended LNN framework, while k is empirically chosen to be 30. After the crude estimation step, updated estimates of the aforementioned parameters are obtained on maximizing the corresponding marginal null log-likelihood under either model formulation. This step is taken in order to bring the initial estimates closer to the estimates returned by the EM algorithm. Using these initial estimates together with p set as 0.5, the most likely value under the Beta (2,2) prior, initial estimates of z_g's are obtained, which are then used to update the parameter estimates in the EM algorithm.