# TESTING FOR DIFFERENTIAL EXPRESSION IN SIMULATED AND REAL CDNA MICROARRAY DATA USING FREQUENTIST AND BAYESIAN METHODS

*Timo Erkkilä*[*]*, Matti Nykter, Harri Lähdesmäki, Miika Ahdesmäki, and Olli Yli-Harja*

Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
[*]timo.p.erkkila@tut.fi

## ABSTRACT

In this paper, we test and compare different data models for finding differential expression in cDNA microarray measurements. We use Bayesian hierarchical error model (HEM) and its variants that are derived by changing the functional form of the original HEM variance. In addition to heterogeneous variance, we use the HEM with exponential and constant variance functions. The standard $t$-test for finding differential expression is our reference test. For both approaches, false discovery rates (FDR) are estimated. With data simulations, we test the accuracy of variance models and FDR estimators. The fit of exponential variance function to real data is observed as well. The parameters for the Bayesian models are estimated using Gibbs sampling.

## 1. INTRODUCTION

For different microarray technologies, data models that try to explain sources of variation of measured expression have been proposed (see [1] and [2]). One usual set-up for microarray measurements is to measure gene expression profiles of two or more biological samples under different conditions, or expression profiles of two or more different cell lines. The differences between the biological samples are then assumed to be characterized by the measured expression profiles that represent expressions proportional to the underlying mRNA concentrations [3].

Biological and technical variations are the primary sources of variation in the measured data. Other sources, *e.g.* , environmental conditions, degradation of mRNA [4], and labeling of samples [5], also have a significant role. Thus, without taking the high amount of uncertainty into account, one may not be able to accurately identify, say, differences between conditions, while increasing FDR. Throughout the study, we (a) try to take the nature of variation in the measurements into account, and (b) estimate FDR in finding differentially expressed genes.

In Section 2, we introduce the HEM variants, and give the formulas for calculating FDR estimates. In Sections 2.1 and 2.2, we introduce the used data simulation methods, prior distributions, and Gibbs sampling.

## 2. METHODS

In this study, we assume that there are no missing values in the data. We can therefore use the following labeling for our data sets: $i \in \{1, \ldots, I\}$ corresponds to gene index, $j \in \{1, 2\}$ corresponds to biological condition, and the replicates are denoted as $k \in \{1, \ldots, K\}$.

Before using the expression data, we transform it into $\log_2$-domain [3]. This is done for two reasons: the data may contain multiplicative biases for different microarray slides and, more importantly, the models we use assume data to contain $\log$-normally distributed components.

We fit the HEM to cDNA microarray expressions of 2-color cDNA microarrays; one color channel is for a biological sample under condition $j = 1$, and the other channel for a biological sample under condition $j = 2$. When biological replicates are missing from the experiment, *i.e.* , when only technical replicates are available, the HEM takes the form

$$y_{ijk} = x_{ij} + e_{ijk} \sim N(x_{ij}, \sigma_{ij}^2) \tag{1}$$

where

$$x_{ij} = \mu + g_i + c_j + r_{ij}. \tag{2}$$

In Eq. 1, $y_{ijk}$ is the observed data and in Eq. 2, $\mu$ is the grand mean over all slides, $g_i$ is the gene effect, $c_j$ is the condition effect, and $r_{ij}$ is the interaction effect of gene $i$ and condition $j$. The term $e_{ijk}$ models the error of the whole experiment process. Thus, the model is similar to the standard 2-way ANOVA, except that HEM uses prior knowledge for estimating unknown parameters and does not, in general, assume constant variance. Different ways to stabilize the variation of expression values have been proposed [6], but one may also give the variance a functional form; in this study, we have used the following functions:

$$\sigma^2(x_{ij}) = \begin{cases} \sigma_{ij}^2 & ,\text{heterogeneous} \\ a^2 + be^{-cx_{ij}} & ,a^2, b, c, x_{ij} > 0 \\ a^2 & ,b, c = 0 \end{cases} \tag{3}$$

where the word *heterogeneous* refers to the original HEM. In the exponential function, $x_{ij}$ is the true expression of the gene $i$ and condition $j$, and the variance is assumed to

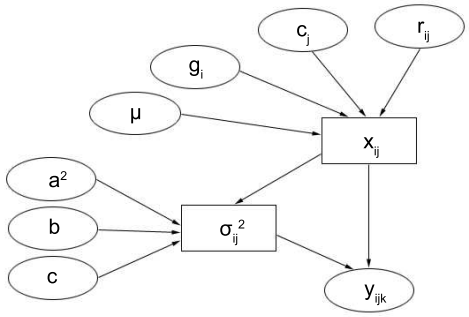



Figure 1. Model graph with functional variance. Ellipses represent stochastic nodes and rectangles represent function nodes. For all the ellipses, a prior distribution is assigned.

be intensity-dependent. The parameters $a^2$,$b$, and $c$ need to be estimated from the data, and if $b$ is negligible compared to $a^2$, or if $c$ is close to zero, the variance shrinks to constant, which is the special case of the model. The assumption of functional variance complicates the dependency graph of parameters in the model, which can be seen in Fig. 1. Without the functional variance, the variance node itself would be stochastic, and no direct relationship to $x_{ij}$ would exist. The edge between the variance and true expression, in fact, makes the model unhierarchical. We simplify the dependency by approximating the relationship using sample mean over the replicates $k$

$$\overline{x}_{ij} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk} \tag{4}$$

instead of $x_{ij}$. The parameters of the HEM and its variants can be solved using, for instance, Gibbs sampling (see 2.2 for further information).

After Bayesian parameter estimation for the models, we use hypothesis testing methods to find differential expression in the data set. For HEM models, we use the $H$-score [7], a modified version of $F$-statistic, and for the $t$-test, we use $p$-values. The $H$-score for gene $i$ is

$$H_i = \frac{1}{2}\sum_{j=1}^{2} \frac{(\hat{x}_{ij} - \hat{x}_{i\cdot})^2}{\hat{\sigma}^2(\hat{x}_{ij})} \tag{5}$$

where the hat over the letter denotes a Bayesian posterior mean parameter estimate, and the dotted subscript denotes averaging over that index. Since no null data is available, *i.e.* , a reference data with no differential expression from which to compute null $H$-scores is missing, we simulate such data by permuting the original data set, so that indices $i \in \{1, \ldots, I\}$ are preserved. $H$-scores of the null data set is computed with

$$H^0_{ip} = \frac{1}{2}\sum_{j=1}^{2} \frac{(\overline{x}_{ijp} - \overline{x}_{i\cdot p})^2}{\hat{\sigma}^2(\overline{x}_{ijp})} \tag{6}$$

where $p$ is the permutation index, $p \in \{1, \ldots, P\}$, and

$$\hat{\sigma}^2(\overline{x}_{ijp}) = \begin{cases} \overline{\sigma}^2_{ijp} & ,\text{heterogeneous} \\ \hat{a}^2 + \hat{b}e^{-\hat{c}\overline{x}_{ijp}} & ,\text{exponential} \\ \hat{a}^2 & ,\text{constant} \end{cases} . \tag{7}$$

We use $P = 100$ permutations, and for each permutation, we use the standard sample estimators to calculate $\overline{x}_{ijp}$ (and $\overline{\sigma}^2_{ijp}$, if we are assuming heterogeneous variance), as Bayesian sampling after each permutation would drastically increase the computation time. It is noteworthy, that we have modified the $H_0$-score calculation in Eq. 6 to take into account the functional forms for variance. for both the $H$-score and $H_0$-score, it is crucial to use similar variance estimators, to reduce FDR estimator bias. See Fig. 3(a) for illustration of estimation bias: the actual scores are calculated using the assumed functional model for the variance, whereas the null scores are calculated using a sample variance estimator for the permuted data set. The FDR estimators for the HEM variants and the $t$-test (as proposed in the $R$ implementation document of HEM and in [8], respectively) are

$$\widehat{FDR}_{HEM}(H_j) = \frac{\hat{\pi}_0 R^0(H_j)}{R(H_j)} \tag{8}$$

and

$$\widehat{FDR}_T(p_j) = \frac{\hat{\pi}_0 S^0(p_j)}{S(p_j)} \tag{9}$$

where

$$\begin{array}{l} R^0(H_j) = \frac{1}{P}\sum_{p=1}^{P} \#_i\{H^0_{ip} : H^0_{ip} > H_j\} \\ R(H_j) = \#_i\{H_i : H_i > H_j\} \\ S^0(p_j) = Ip_j \\ S(p_j) = \#_i\{p_i : p_i < p_j\} \end{array} . \tag{10}$$

The $\#_i$ denotes the number of values, that fulfill the terms inside the braces for $i \in \{1, \ldots, I\}$. The point estimates $p_{\lambda_n}$ of $\pi_0$ are also calculated as in [7] and [8] using the percentiles $\lambda_n = 0.01n, n \in \{1, \ldots, 100\}$, but the estimator for $\pi_0$, the proportion of non-significant genes, is calculated using weighted average. We use the cumulative distribution function of $N(0.1, 0.3)$ to generate weights for each percentile $\lambda_n$:

$$c_{\lambda_n} = \Phi\left(\frac{\lambda_n - 0.1}{0.3}\right), \quad n \in \{1, \ldots, 100\} \tag{11}$$

The weight matrix is a diagonal matrix $C = \text{diag}\,(c_{\lambda_1}, \ldots, c_{\lambda_{100}})$, and $\mathbf{p} = [p_{\lambda_1}, \ldots, p_{\lambda_{100}}]^T$. The estimator $\hat{\pi}_0$ is therefore

$$\hat{\pi}_0 = \left(\mathbf{1}^T C \mathbf{1}\right)^{-1} \mathbf{1}^T C \mathbf{p}. \tag{12}$$

The reason for using $C$, that gives more weight as $n$ increases, is to compensate the bias and variance of each point estimate; when $n$ increases, the bias of point estimate $p_{\lambda_n}$ decreases, whereas the variance increases [8].

### 2.1. Simulations

In the simulation study, we generate cDNA microarray data with outliers using methods proposed in [9]. The data consists of $I = 5000$ genes, $J = 2$ conditions, and $K = 10$ replicates. The distributions of the simulator are

$$\begin{array}{rl} \forall i: & z_i \sim Exp(\lambda') \\ \forall i: & o_i \sim Ber(1-\pi_0) \\ \forall o_i = 1: & s_i \sim Rademacher \\ \forall o_i = 1: & b_i \sim Beta(\alpha', \beta') \\ \forall i,j,k: & y_{ijk} \sim N(x_{ij}, \sigma^2(x_{ij})) \end{array} , \quad (13)$$

the functions of the simulator are

$$\begin{array}{rl} \forall o_i = 1: & \sqrt{t_i} = 10^{s_i b_i} \\ \forall o_i = 1: & z_{i1} = z_i \sqrt{t_i} \\ \forall o_i = 1: & z_{i2} = z_i / \sqrt{t_i} \\ \forall o_i = 0: & z_{i1} = z_{i2} = z_i \\ \forall i,j: & x_{ij} = \log_2(z_{ij}) \\ \forall i,j: & \sigma^2(x_{ij}) = a^2 + be^{-cx_{ij}} \end{array} , \quad (14)$$

and the parameters for the functions and distributions are set to

$$\begin{array}{c} \lambda' = 1000, \pi_0 = 0.96, \\ \alpha' = 1.7, \beta' = 4.8 \\ a^2 = 0.2, b = 1.0, c = 0.4. \end{array} . \quad (15)$$

So, the simulation of measurements in short: Generate $I$ measurements from an exponential distribution. With probability $1 - \pi_0$, a measurement $i$ is assigned as differentially expressed. With probability $0.5$ it is an overexpression, and $t$ is the shifting value between the conditions $j = 1$ and $j = 2$. The expressions are the $\log_2$-transformed, and variance is generated from the exponential function, using the $\log_2$-transformed measurements. Finally, for each replicate $k$, normally distributed noise is added.

### 2.2. Prior specification and Gibbs sampling

We have built the Gibbs samplers with WinBUGS (Bayesian inference Using Gibbs Sampling for Windows) [10]. The software is suitable for generating Gibbs samplers for models, where the parameter dependencies form a directed acyclic graph (DAG). WinBUGS can be downloaded from `http://www.mrc-bsu.cam.ac.uk/bugs/`.

The used prior and hyperprior distributions for all parameters are tabulated in Table 1. The priors are on the left-hand side and the used parameters for the distributions are on the right-hand side. The chosen parameter and distribution values are similar as in [7]. We use Gibbs sampling posterior mean to calculate the estimates for each parameter in the model, we generate a $600$-point sample after a $300$-point burn-in period. We noticed, that in this study such amount of iterations is sufficient for the Markov chains to converge.

| | |
|---|---|
| $\mu \sim U(0, \mu_{max})$ | $\mu_{max} = 50$ |
| $g_i \sim N(0, \sigma_g^2)$ | $\sigma_g = 1$ |
| $c_j \sim N(0, \sigma_c^2)$ | $\sigma_c = 1$ |
| $r_{ij} \sim N(0, \sigma_r^2)$ | $\sigma_r = 1$ |
| $e_{ijk} \sim N(0, \sigma^2(x_{ij}))$ | Eq. 3 |
| $\sigma_{ij}^{-2}, a^{-2} \sim \Gamma(\alpha, \beta)$ | $\alpha = 1, \beta = 0.125$ |
| $b, c \sim U(0, t_{max})$ | $t_{max} = 5$ |

Table 1. Prior and hyperprior distributions for the HEM.

## 3. RESULTS AND CONCLUSION

The simulation study consisted of data generation with known parameters, Bayesian and frequentist parameter estimation, visualization of FDR estimation accuracy for all models (Fig. 3(a) and 3(b)), and visualization of accuracy (ROC curves) for finding differential expression. The simulations show, that if such exponential variance structure exists, the functional form of variance in HEM can be modified to better fit the data (Fig. 2), thus resulting in more accurate differential expression detection (Fig. 3(c)). The approximation of dependency between the variance function and true expression could reduce the accuracy of the variance fit drastically, if the amount of replicates was small. Also, after each perumation for calculating the $H_0$-score, the using of Bayesian estimates instead of sample estimates would increse the performance of FDR estimation for the HEM variants.

## 4. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland, (application number 213462, Finnish Programme for Centres of Excellence in Research 2006-2011).

## 5. REFERENCES

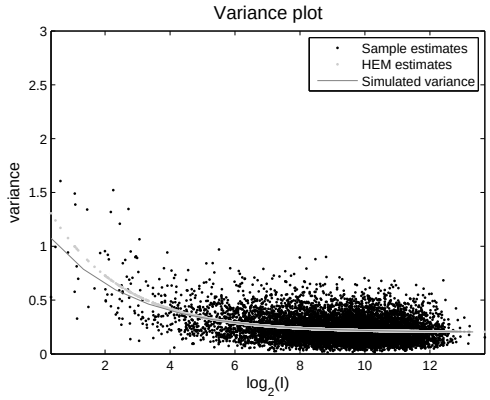
[1] A. Lewin, S. Richardson, C. Marshall, A. Glazier, and T. Aitman, "Bayesian modeling of differential gene expression.," *Biometrics*, vol. 62, no. 1, pp. 1–9, Mar 2006.

[2] K. Lo and R. Gottardo, "Flexible empirical bayes models for differential gene expression.," *Bioinformatics*, vol. 23, no. 3, pp. 328–335, Feb 2007.

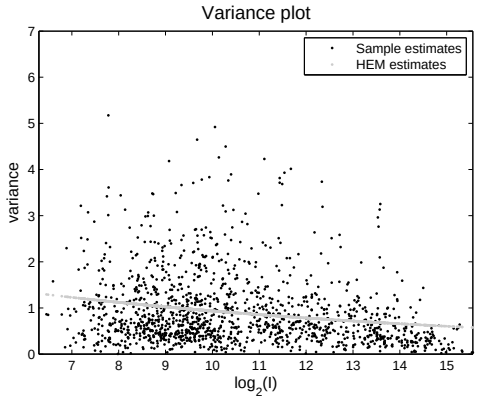
[3] J. Quackenbush, "Microarray data normalization and transformation.," *Nat Genet*, vol. 32 Suppl, pp. 496–501, Dec 2002.

[4] H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, G. Marcucci, U. Marcucci, and K. Kornacker, "Chipping away at the chip bias: Rna degradation in microarray analysis.," *Nat Genet*, vol. 35, no. 4, pp. 292–293, Dec 2003.

[5] K. K. Dobbin, E. S. Kawasaki, D. W. Petersen, and R. M. Simon, "Characterizing dye bias in microarray experiments.," *Bioinformatics*, vol. 21, no. 10, pp. 2430–2437, May 2005.




(a) Exponential variance fit using simulated data. The solid grey line is the simulated variance, black dots are the sample variance estimates, and light gray dots are the HEM estimates.




(b) Exponential variance fit using 4 replicates of E-MEXP-1385 *mus musculus* data. The black dots are sample variance estimates, and the light gray dots are the HEM estimates.

Figure 2. Variance plots as functions of intensity.

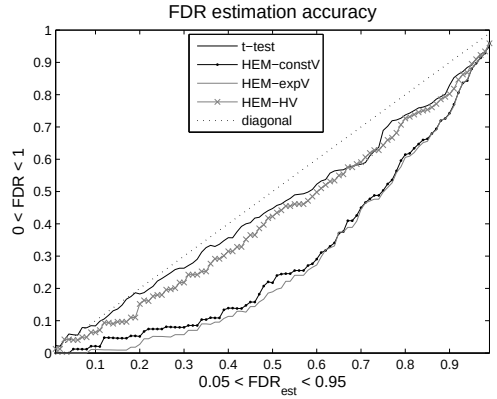

[6] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression.," *Bioinformatics*, vol. 18 Suppl 1, pp. S96–104, 2002.

[7] H. J. Cho and J. K. Lee, "Bayesian hierarchical error model for analysis of gene expression data.," *Bioinformatics*, vol. 20, no. 13, pp. 2016–2025, Sep 2004.

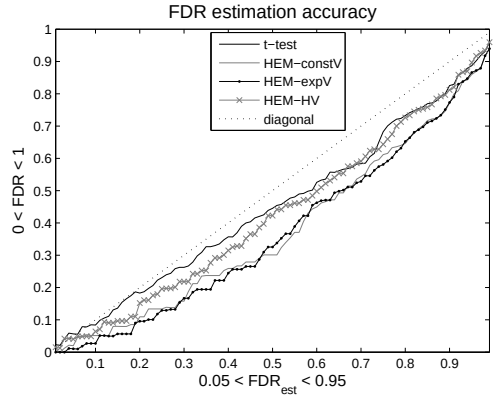
[8] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies.," *Proc Natl Acad Sci U S A*, vol. 100, no. 16, pp. 9440–9445, Aug 2003.

[9] Y. Balagurunathan, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Simulation of cdna microarrays via a parameterized random signal model.," *J Biomed Opt*, vol. 7, no. 3, pp. 507–523, Jul 2002.

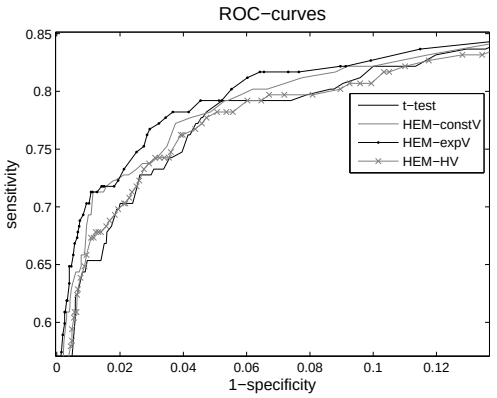
[10] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, vol. 10, pp. 325–337, 2000.


(a) FDR estimation accuracy using the $H_0$-score without variance correction; the FDR estimations of HEM models with functional variance perform poorly.




(b) Variance functionality taken into account; the bias is reduced.




(c) Increasing the accuracy of variance estimation increases the accuracy of finding differentially expressed genes. The HEM with exponential variance function performs somewhat better than the other models.

Figure 3. FDR estimation accuracy plots and ROC curves.