# Bayesian variable selection method for modeling dose-response microarray data under simple order restrictions

Martin Otava[1], Adetayo Kasim[2], Ziv Shkedy[1], Dan Lin[1] and Bernet S. Kato[3]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Center for Statistics, Universiteit Hasselt, Belgium
[2] Wolfson Research Institute, Durham University, United Kingdom
[3] National Heart and Lung Institute, Imperial College London, United Kingdom

E-mail for correspondence: martin.otava@uhasselt.be

**Abstract:** Bayesian modeling of dose-response microarray data offers the possibility to jointly establish the dose-response relationships between gene expression and increasing doses of therapeutic compound, and to determine the nature of the relationships wherever it exist. Moreover, correction for multiplicity adjustment for Bayesian modeling of dose-response microarray data can be based on the direct posterior probability of the null model. The posterior probabilities are obtained by translating the inequality constraints for monotone relationship into Bayesian variable selection problem.

**Keywords:** Microarray Data; Bayesian Analysis; Dose-Response Relationship; False Discovery Rate; Direct Posterior Probability.

## 1 Introduction

Dose-response microarray experiments are a growing area in biomedical and pharmaceutical research to study the relationship between increasing doses of a therapeutic compound and the activity of entire genome at once. The primary goal of such an experiment is to identify genes with significant dose-response relationship under the monotone constraints (Lin et al., 2012). Secondly, it is necessary to determine the nature of the relationship wherever it exists. Denote the mean gene expression of a gene under the placebo dose as $\mu_0$. Similarly, we consider an increasing doses of a therapeutic compound and $\mu_i$, $i = 1, \ldots, K$ be an the mean gene expression under dose $i$. Therefore, the primary interest is to test the null hypothesis

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \ldots = \mu_K, \tag{1}$$

TABLE 1. The set of seven possible monotonic dose-response models for an experiment with three dose levels. The mean response of dose level $i$ is denoted as $\mu_i$. The model $g_0$ represents the null model of no dose effect.

| Model | Non-decreasing profile | Non-increasing profile |
|---|---|---|
| $g_1$ | $\mu_0 = \mu_1 = \mu_2 < \mu_3$ | $\mu_0 = \mu_1 = \mu_2 > \mu_3$ |
| $g_2$ | $\mu_0 = \mu_1 < \mu_2 = \mu_3$ | $\mu_0 = \mu_1 > \mu_2 = \mu_3$ |
| $g_3$ | $\mu_0 < \mu_1 = \mu_2 = \mu_3$ | $\mu_0 > \mu_1 = \mu_2 = \mu_3$ |
| $g_4$ | $\mu_0 < \mu_1 = \mu_2 < \mu_3$ | $\mu_0 > \mu_1 = \mu_2 > \mu_3$ |
| $g_5$ | $\mu_0 = \mu_1 < \mu_2 < \mu_3$ | $\mu_0 = \mu_1 > \mu_2 > \mu_3$ |
| $g_6$ | $\mu_0 < \mu_1 < \mu_2 = \mu_3$ | $\mu_0 > \mu_1 > \mu_2 = \mu_3$ |
| $g_7$ | $\mu_0 < \mu_1 < \mu_2 < \mu_3$ | $\mu_0 > \mu_1 > \mu_2 > \mu_3$ |

against the alternative hypotheses

$$\begin{array}{c} H_a^{up} : \mu_0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K, \\ \text{or} \\ H_a^{dn} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_K \end{array} \qquad (2)$$

with at least one strict inequality. The choice between $H_a^{up}$ and $H_a^{dn}$ depends on the direction of the ordered constraints. Note that the determination of the nature of the dose-response relationship is related to the further decomposition of the alternative hypotheses into their basic hypotheses. This process results in $2^K - 1$ hypotheses under each of the monotone directions. For a dose-response microarray experiments with one control dose and $K = 3$ (i.e. three increasing doses of a therapeutic compound), the alternative hypotheses can be decomposed into further basic hypotheses as shown in Table 1. Note that each alternative hypothesis corresponds to a monotone model. In particular the null hypothesis corresponds to the null model for which $\mu_0 = \mu_1 = \mu_2 = \mu_3$.

Bayesian modeling of dose-response microarray data offers a framework to simultaneously establish a dose-response relationship and to determine the nature of the relationship by providing posterior probability for each of the models $g_i$, $i = 1, \ldots, K$, given the data. The posterior probability of the null model is particularly interesting, because it is also a probability of false positives findings, i.e. of genes that are wrongly assigned to the alternative hypotheses. Hence, posterior probability allows for adjustment for false discovery rate (Newton et al., 2007), to identify few important genes in a pool of potential false positives. However, the estimation of the required parameters to obtain posterior probability for the models requires estimation under equality constraints between two or more parameters which could not be estimated with the standard approach of Gelfand et al. (1991). Therefore, the Bayesian variable selection approach offers elegant solution how to identify the relationship and correct for multiplicity simultaneously using conditional false discovery rate.

## 2 Methodology

The Bayesian inequality models (Klugkist and Hoijtink, 2005) cannot be used in our framework because of the equality constraints specified in the models. The equality constraints would cause that standard estimation approach assigns zero probabilities to each of our models except $g_7$. Therefore, we propose the following parametrization. We consider the following linear model,

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim N(0, \sigma^2), \;\; i = 0, \cdots, K, \quad j = 0, 1, 2, \cdots, n_i, \tag{3}$$

where $\boldsymbol{Y} = (Y_{01}, Y_{02}, \ldots, Y_{Kn_K})$ are gene expression levels and $n_i$ represents the number of observations at the $i$th dose level. Reparameterize the mean response such that

$$\mathrm{E}(Y_{ij}) = \mu_i = \begin{cases} \mu_0, & i = 0, \\ \mu_0 + \sum_{\ell=1}^{i} \delta_\ell, & i = 1, \ldots, K., \end{cases} \tag{4}$$

with the constraints that $\delta_\ell \geq 0$ for an upward trend or $\delta_\ell \leq 0$ for a downward trend. The difference in the mean structures of the different models therefore depends on which of the components in $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_K)$ are set to be equal to zero. The problem of model estimation is equivalent to decision which columns in the full design matrix of model 4 are selected or deleted. This is related to the Bayesian variable selection (BVS) approach (George and McCulloch, 1993), which is used to determine an optimal model from a priori set of $R$ known plausible models. In our setting the BVS model allows us to calculate the posterior probability of each model, $p(g_r|\text{data})$ and in particular the posterior probability of the null model, $p(g_0|\text{data})$. Let $z_i$, $i = 1, \ldots, K$ be an indicator variable such that

$$z_i = \begin{cases} 1, & \delta_i \;\; \text{is included in the model,} \\ 0, & \delta_i \;\; \text{is not included in the model,} \end{cases} \tag{5}$$

and let $\theta_i = \delta_i \cdot z_i$. Hence, we can reformulate the mean structure in (4) (O'Hara and Sillanpää, 2009) in terms of $\theta_i$ and $z_i$ as

$$\mathrm{E}(Y_{ij}) = \mu_0 + \sum_{\ell=1}^{i} \theta_\ell = \mu_0 + \sum_{\ell=1}^{i} z_\ell \delta_\ell, \;\; i = 1, \ldots, K. \tag{6}$$

For $K$ dose levels experiment, the vector $\boldsymbol{z} = (z_1, \ldots, z_K)$ defines uniquely each one of the $2^K$ plausible models. For example for $K = 3$ and $\boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 0)$ we obtain $\mathrm{E}(Y_{ij}|\boldsymbol{z}) = (\mu_0, \mu_0 + \delta_1, \mu_0 + \delta_1, \mu_0 + \delta_1)$, which corresponds to the mean of model $g_3$). We assume that $z_i$ and $\delta_i$ are independent, and use truncated normal prior distribution for $\delta_i$ and

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi_i), \\ \pi_i &\sim \text{U}(0, 1). \end{aligned} \tag{7}$$

As pointed out by O'Hara and Sillanpää (2009) the posterior inclusion probability of $\delta_i$ into the model equals the posterior mean of $z_i$. The posterior probability of each model can be straightforwardly obtained by using the transformation of $\boldsymbol{z}$ instead of the entire vector $\boldsymbol{z}$ itself. Denote $M_R = 1 + \boldsymbol{z}\boldsymbol{c}$, where $\boldsymbol{c} = (1, 2, \dots, 2^{K-1})^T$, then $M_R$ has unique value for each of the plausible models (for example: $M_R = 2$ only for the model $g_3$). Thus, the posterior probability of $M_R = r$, $r = 1, \dots, R$, defines uniquely the posterior probability of the $r$th model,

$$p(M_R = r|\text{data}) = p(g_r|\text{data}), \tag{8}$$

and in particular, the posterior probability of the null model is given by,

$$p(M_R = 1|\text{data}) = p(g_0|\text{data}). \tag{9}$$

Assume that there are $m = 1, \dots, M$ genes in the experiment and the aim is to find the differentially expressed ones with respect to dose. In our framework, the problem is translated to the determination if the gene follows any other model than $g_0$. Assume that the genes satisfying $p_m(g_0|\text{data}) \leq \alpha$ for given threshold $\alpha$ are considered differentially expressed. Hence, according to Newton et al. (2007), $p_m(g_0|\text{data})$ represents probability of such statement being false. Let $I_m$ be an indicator variable of $p_m(g_0|\text{data}) \leq \alpha$. Since $p_g(g_0|\text{data})$ is also the probability that the considering the $m$th gene differentially expressed is incorrect, the expected number of false discoveries (cFD) is

$$\text{cFD}(\alpha) := \text{E}(\text{cFD}) = \sum_{m=1}^{M} p_m(g_0|\text{data}) I_m. \tag{10}$$

Newton et al. (2007) defined the conditional (on the data) false discovery rate as

$$\text{cFDR}(\alpha) = \frac{\text{cFD}(\alpha)}{N(\alpha)}, \tag{11}$$

where $N(\alpha)$ is the number of genes declared differentially expressed for a given threshold $\alpha$. Note that $\text{cFDR}(\alpha)$ is interpreted as the average error that is made by considering any gene as differentially expressed. Hence, the value of $\alpha$ is selected is such a way that $\text{cFDR}(\alpha)$ does not exceed a pre-specified threshold $\tau$.

## 3 Results

We apply the direct posterior probability approach discussed above for multiplicity adjustment. The framework enables adjustment for false discovery rates among the significant genes. We use the `R2WINBUGS` package to fit a gene specific model and to obtain the posterior probability of the null model. For each gene an MCMC simulation of 20000 iterations (from which

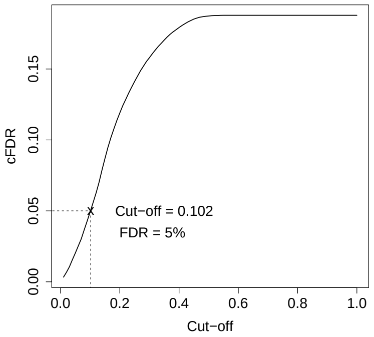

FIGURE 1. *Adjustment for multiplicity. The relationship between the conditional false discovery rate (cFDR) and the cut-off values.*

5000 are used as burn-in period) was used to fit the BVS model. Figures 1 and 2 show the relationship between false discovery rate (cFDR), number of significant genes and cut-off value $\alpha$. Figure 1 shows that an increase in cut-off values results in an increase in false discovery rate. However, the false discovery rate reaches its maximum of 0.2 at the cut-off of about 0.5. Figure 2 also shows an increase in the number of significant genes with an increase in cut-off values. The implication of the finding is that, as expected, the higher the cut-off value, the larger the number of significant genes and consequently, the higher the proportion of false positives among the significant genes. Similar to the frequentists practice, one may wish to control for false discovery rate at 1% or 5%, which corresponds to cut-off values of 0.029 and 0.102, respectively. Based on these cut-off values, the corresponding numbers of significant genes are 609 and 3295 genes, respectively.

## 4 Discussion

There are two main challenges in Bayesian analysis of dose-response microarray data. The first is the presence of strictly equality relationship between differences in gene expressions at different doses of a therapeutic compound and the second is the question how to adjust for multiplicity. The BVS method is useful as an approach to circumvent the first problem by replacing strict equality between doses by a common parameter. The BVS model estimates equal means for two successive dose levels, $i$ and $i-1$ whenever the corresponding binary variable for the $i$th dose level $z_i = 0$. Further, the posterior probability of the null model can be estimated and can be used for multiplicity adjustment. In summary, the BVS methodology offers the tools how to handle the differentially expressed genes finding

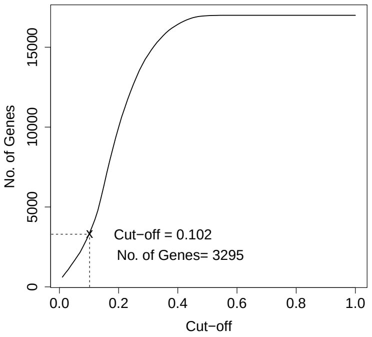

FIGURE 2. *Adjustment for multiplicity. The relationship between number of significant genes and the cut-off values.*

in elegant and efficient way.

## References

Gelfand, A.E., Smith, A.F.M., and Lee, T.M. (1992). Bayesian Analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, **87**, 523 − 532.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881 − 889.

Klugkist, I. and Hoijtink, H. (2007). The Bayes Factor for Inequality and About Equality Constrained Models. *Computational Statistics and Data Analysis*, **51**, 6367 − 6379.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L. , (editors)(2012). *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R - Order Restricted Analysis of Microarray Data.* Springer.

Newton, M.A., Wang, P., and Kendziorski, C. (2007). Hierarchical mixture models for expression profiles. In: *Do, K.M., Müller, P. and Vannucci, M. (Editors): Bayesian Inference for gene expression and proteomics*, Cambridge university press, pp. 40 − 52,

O'Hara, R. B. and Sillanpää, M. J. (2009). Review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85 − 118.