Estimating the False Discovery Rate for interval null-hypotheses using nonparametric deconvolution

M.A. van de Wiel[1] and K.I. Kim[1]

[1] Technische Universiteit Eindhoven, PO Box 513 5600 MB Eindhoven, The Netherlands,

Keywords: Gene expression data, multiple testing, deconvolution

1 Introduction

Microarrays are a widespread technique to measure expression of thousands of genes simultaneously. When applied to two different conditions (e.g. disease versus healthy) one is often interested in those genes that express significantly different under the two conditions. It is generally recognized that multiple testing corrections are necessary and the False Discovery Rate (FDR) is a useful criterion to do so.

As opposed to common procedures in literature, we do not base the selection criterion on statistical significance only, but also on biological significance. Therefore, we select only those genes which are significantly more differentially expressed than some cut-point c. We use the Bayesian interpretation of FDR: the probability that the parameter of interest lies in the null domain given that the test criterion exceeds a threshold. We show how to improve the simple estimator by using nonparametric deconvolution. We study the performance of the method using simulations and apply it to real data.

2 Method

We focus on paired measurements, which are very common when using two-channel microarrays in which, for example, diseased tissue is directly hybridized together with healthy tissue from the same individual. The method can, however, be adjusted to deal with unpaired measurements as well. We assume technical bias has been filtered out by so-called normalisation methods. For each gene g we compute the average difference between the measurements (on log scale) of the two tissues over the individuals (Y_g). This random variable includes biological noise. We use the following simple model:

$$Y_g = \mu_g + \epsilon_g, \tag{1}$$

where μ_g is the true differential expression for gene g and $\epsilon_g \sim F$ is an error term.

In terms of hypothesis testing we would like to test simultaneously: $H_{0g} : \mu_g \in A_0, g = 1, \ldots, G$ versus its negation. In literature $A_0 = \{0\}$ is mostly used. We however prefer to use an interval hypothesis $A_0 = \{\mu_g : \mu_g < c\}$, because then rejection of H_{0g} ensures that μ_g is further than c away from '0'. Biologists already use (non-statistical) rules to select genes which are more than a certain distance apart to ensure biological significance. The criterion $C_g(t)$ equals 1 if the statistic applied for testing H_{0g} exceeds threshold t. For given cut-point c, the Bayesian interpretation of the false discovery rate is the mean false positive probability (see also Broët et al. (2004)):

$$\mathrm{FDR}(t) = \frac{1}{G}\sum_g \mathrm{P}(\mu_g < c | C_g(t) = 1) = \frac{1}{G}\sum_g \frac{\mathrm{P}(\mu_g < c)\mathrm{P}(C_g(t) = 1 | \mu_g < c)}{\mathrm{P}(C_g(t) = 1)} = \frac{1}{G}\sum_g \frac{\pi_0 \pi(t)}{p(t)}.$$

Threshold $t = t'$ is chosen such that $\mathrm{FDR}(t') < 0.05$ (say) and genes are selected when $C_g(t') = 1$.

Estimation of π_0 and $\pi(t)$ depends critically on estimation of f_μ: the density of μ_g, whereas $p(t)$ may directly be estimated as the fraction of genes for which $C_g(t) = 1$. A naive estimator of $\mathrm{FDR}(t)$ is obtained by simply ignoring the error process in (1) and estimating f_μ by the (smoothed) empirical density of Y_g. In a parametric Bayesian setting one could assume a parent density f on the mean values $\mu_1, \ldots, \mu_G$, but there is no natural choice of f. This approach is taken in Scott and Berger (2004), who also show that particular choices of f give completely different results.

The crux of the advanced FDR estimator is the recovering of the actual density of μ_g by deconvolution. We assume μ_g to be i.i.d. with density f_μ and to be independent of $\epsilon_g \sim f_\epsilon$. Then, $Y_1, \ldots, Y_G$ are identically distributed random variables with density f_Y. Therefore, if $f_\epsilon = N(0, \sigma)$, we have for the characteristic functions, using (1),

$$\phi_Y(\omega) = \phi_\mu(\omega) \exp(-\sigma^2\omega^2/2), \quad \text{so} \quad \phi_\mu(\omega) = \phi_Y(\omega) \exp(\sigma^2\omega^2/2). \tag{2}$$

So, if we knew $\phi_Y(\omega)$, we could find f_μ by Fourier inversion of $\phi_\mu(\omega)$. However, we only have data from f_Y. We used the approach by Delaigle and Gijbels (2002) to deal with this problem. This is a kernel density estimation method, where the kernel is defined on the level of the characteristic function.

3 Results

Extensive simulations have been performed to evaluate the accuracy of the naive and advanced FDR estimate. We show results for $f_\mu = \frac{1}{16}N(-2.5, 0.7) + \frac{7}{8}N(0, 0.9) + \frac{1}{16}N(2.5, 0.7)$ and $\sigma = 0.848$. The test statistic is based on a simple Z-test.

We found that after deconvolution, the FDR estimator was very good over a wide range of thresholds in case of moderate Gaussian error and improved upon the naive estimator. Other error distributions have been studied and accuracy results varied according to smoothness properties and size of the error. The method has been successfully applied to real data which resulted in a biologically and statistically significant collection of genes which was validated with independent techniques.

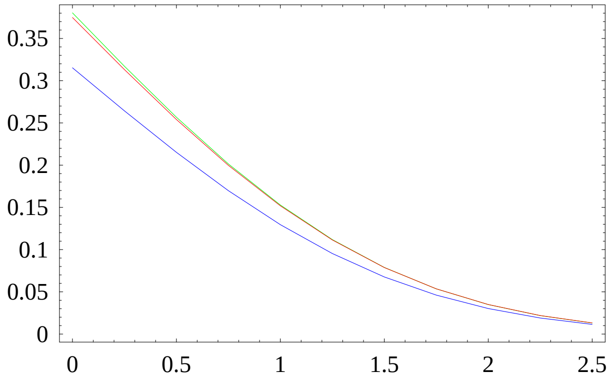

FIGURE 1. Threshold (x) versus FDR (y). True = green, Naive = blue, Deconv = red

Bibliography

Broët, P., A. Lewin, S. Richardson, C. Dalmasso, and H. Magdelenat (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562–2571.

Delaigle, A., and I. Gijbels (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**(4), 869–886.

Scott, J.G., and J.O. Berger (2004). An exploration of aspects of Bayesian multiple testing. Technical Report 2003, Duke University, www.isds.duke.edu/˜berger/papers/multcomp.pdf.