Statistical Science
2008, Vol. 23, No. 1, 41–44
DOI: 10.1214/07-STS236A
Main article DOI: 10.1214/07-STS236
© Institute of Mathematical Statistics, 2008

Comment: Microarrays, Empirical Bayes and the Two-Groups Model

Kenneth Rice and David Spiegelhalter

Through his various examples, Professor Efron makes a convincing case that cutting-edge science requires methods for detecting multiple "non-nulls." These methods must be straightforward to implement, but perhaps more importantly statisticians need to be able to justify them unambiguously. Efron's Empirical Bayes approach is certainly computationally efficient, but we feel the rationale for making each of his steps is unattractively ad hoc. This concern is practical, not philosophical; Efron's criterion for choice of tuning parameters seems to be that they look "believable." In less expert hands, this approach seems to introduce a lot of leeway for practitioners to simply "tune" away until they get the results they want.

In an attempt to address this problem, we will describe an approach developed in a fully model-based framework. As with locfdr, the calculations are fast, but our whole analysis derives from clear up-front statements about what the analysis is trying to achieve, and the modeling assumptions made. The results look reassuringly similar to Professor Efron's. We hope this will be helpful for understanding the current paper, and in making a contribution to this general field.

We begin by following Efron in placing the local false discovery rate, $\mathrm{fdr}(z)$, as the primary focus of the analysis, and exploit the fact that it can offer a neat parameterization of the two-part model. If the marginal, "mixture" density for the z-values is

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z)$$

and $\mathrm{fdr}(z) = p_0 f_0(z)/f(z)$, then

$$f_1(z) = \frac{p_0}{1-p_0}\,\frac{1-\mathrm{fdr}(z)}{\mathrm{fdr}(z)}\,f_0(z).$$

We observe that, because f_1 is a density, we only need to know f_0 and fdr in order to find its normalized form, and in turn this tells us the value of p_0. Thus, for a given f_0, specifying fdr sets up everything else we require for model-based analysis.

Naturally, the analysis we report will depend on the functional form assumed for fdr, and Efron implicitly assumes a rather flexible form of fdr, through a seventh-order polynomial-smoothed density estimate. However, this approach does not rule out an $\widehat{\mathrm{fdr}}$ with multiple peaks. Thinking of the schools example, we would not want to be the statistician explaining how two "bad" schools may have $z_1 < z_2 < 0$, but yet $\widehat{\mathrm{fdr}}(z_1) > 0.2$ while $\widehat{\mathrm{fdr}}(z_2) < 0.2$. Put more simply, Efron's method can report that School 1 has worse performance, but only School 2 is called an outlier. We find it more straightforward to a priori justify our choice of fdr by careful consideration of its role in the reported inference.

In our experience, the search for non-null "discoveries" is based around two ideas; first, we will not discover anything near the center of f_0 (effectively Efron's "zero assumption," also termed "purity" by Genovese and Wasserman, 2004). A second sensible assumption is that the evidence for z being "null" will decrease monotonically as we move out from the center. One way to satisfy this is with a logistic-linear form for fdr, giving a two-component normal mixture for f_1, but we get closer to the spirit of Efron's analysis by assuming that fdr is unity inside a central region, and then follows a half-normal decline, that is,

$$\mathrm{fdr}^H(z) = \begin{cases} e^{-(z+k_a)^2/2}, & z < -k_a,\\ 1, & -k_a \le z \le k_b,\\ e^{-(z-k_b)^2/2}, & z > k_b. \end{cases}$$

Following the observation above, taking the null component f_0 to be standard Normal, now defines the following marginal distribution $f^H(z)$:

$$f^H(z) = p_0 (2\pi)^{-1/2} \cdot \begin{cases} e^{-|z|k_a + k_a^2/2}, & z < -k_a,\\ e^{-z^2/2}, & -k_a \le z \le k_b,\\ e^{-|z|k_b + k_b^2/2}, & z > k_b, \end{cases}$$

Kenneth Rice is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232, USA (e-mail: kenrice@u.washington.edu). David Spiegelhalter is Senior Scientist, MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 OSR, United Kingdom (e-mail: David.Spiegelhalter@mrc-bsu.cam.ac.uk).