Statistical Science
2008, Vol. 23, No. 1, 29–33
DOI: 10.1214/07-STS236C
Main article DOI: 10.1214/07-STS236

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

T. Tony Cai

Professor Efron is to be congratulated for his innovative and valuable contributions to large-scale multiple testing. He has given us a very interesting article with much material for thought and exploration. The two-group mixture model (2.1) provides a convenient and effective framework for multiple testing. The empirical Bayes approach leads naturally to the local false discovery rate (Lfdr) and gives the Lfdr a useful Bayesian interpretation. This and other recent papers of Efron raised several important issues in multiple testing such as theoretical null versus empirical null and the effects of correlation. Much research is needed to better understand these issues.

Virtually all FDR controlling procedures in the literature are based on thresholding the ranked p-values. The difference among these methods is in the choice of the threshold. In multiple testing, typically one first uses a p-value based method such as the Benjamini–Hochberg procedure for global FDR control and then uses the Lfdr as a measure of significance for individual nonnull cases. See, for example, Efron (2004, 2005). In what follows I will first discuss the drawbacks of using p-value in large-scale multiple testing and demonstrate the fundamental role played by the Lfdr. I then discuss estimation of the null distribution and the proportion of the nonnulls. I will end with some comments about dealing with the dependency.

In the discussion I shall use the notation given in Table 1 to summarize the outcomes of a multiple testing procedure.

With the notation given in the table, the false discovery rate (FDR) is then defined as $\mathrm{FDR} = \mathrm{E}(N_{10}/R|R > 0)\Pr(R > 0)$.

1. THE USE OF *p*-VALUES: VALIDITY VERSUS EFFICIENCY

In the classical theory of hypothesis testing the p-value is a fundamental quantity. For example, the decision of a test can be made by comparing the p-value with the prespecified significance level α. In the more recent large-scale multiple testing literature, p-value continues to play a central role. As mentioned earlier, nearly all FDR controlling procedures separate the nonnull hypotheses from the nulls by thresholding the ordered p-values.

A dual quantity to the false discovery rate is the false nondiscovery rate $\mathrm{FNR} = \mathrm{E}(N_{01}/S|S > 0)\Pr(S > 0)$. In a decision-theoretical framework, a natural goal in multiple testing is to find, among all tests which control the FDR at a given level, the one which has the smallest FNR. We shall call an FDR procedure *valid* if it controls the FDR at a prespecified level α, and *efficient* if it has the smallest FNR among all FDR procedures at level α. The literature on FDR controlling procedures so far has focused virtually exclusively on the validity; the efficiency issue has been mostly untouched.

In a recent article, Sun and Cai (2007) considered the multiple testing problem from a compound decision point of view. It is demonstrated that p-value is in fact not a fundamental quantity in large-scale multiple testing; the local false discovery rate (Lfdr) is. Thresholding the ordered p-values does not in general lead to efficient multiple testing procedures. The reason for the inefficiency of the p-value methods can be traced back to Copas (1974) where a weighted classification problem was considered. Copas (1974) showed that if a symmetric classification rule for dichotomies is admissible, then it must be ordered by the likelihood ratios, which is equivalent to being ordered by the Lfdr. Sun and Cai (2007) showed that, under mild conditions, the multiple testing problem is in fact equivalent to the weighted classification problem. I will discuss below some of the findings in Sun and Cai (2007) and draw connections to the present paper by Professor Efron.

The local false discovery rate, defined in (2.7), was first introduced in Efron et al. (2001) as the a posteriori probability of a gene being in the null group given the z-score z. The results in Sun and Cai (2007) show that the Lfdr is a fundamental quantity which can be used directly for optimal FDR control. By using the Lfdr

T. Tony Cai is Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: tcai@wharton.upenn.edu).