Statistics and Probability Letters 87 (2014) 86–93

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

# Monotone false discovery rate

Joong-Ho Won [a,*], Johan Lim [b], Donghyeon Yu [b], Byung Soo Kim [c], Kyunga Kim [d]

[a] Korea University, Seoul, Republic of Korea
[b] Seoul National University, Seoul, Republic of Korea
[c] Yonsei University, Seoul, Republic of Korea
[d] Sookmyung Women's University, Seoul, Republic of Korea

ARTICLE INFO

*Article history:*
Received 11 November 2012
Received in revised form 16 December 2013
Accepted 18 December 2013
Available online 4 January 2014

*Keywords:*
Adaptive decision rule
False discovery rate
Empirical Bayes methods
Mode matching
Isotonic regression

ABSTRACT

This paper proposes a procedure to obtain monotone estimates of both the local and the tail false discovery rates that arise in large-scale multiple testing. The proposed monotonization is asymptotically optimal for controlling the false discovery rate and also has many attractive finite-sample properties.



## 1. Introduction

The advance of modern high-throughput technologies in many scientific disciplines such as genomics and brain imaging has dramatically increased both the size and the dimension of the data and made data analysis a major challenge. In particular, it is often required to test thousands or millions of hypotheses simultaneously when analyzing large-scale, high-dimensional data. Unlike the case of testing a single hypothesis, type I error in multiple hypothesis testing is not uniquely defined. Traditional approaches, e.g., the family-wise error rate (FWER), are far too conservative and produce many false negatives in high-dimensional settings. For this reason, the concept of a false discovery rate (FDR), or the expected proportion of false positives among declared positives, is introduced and now widely accepted.

The FDR is originally proposed by Benjamini and Hochberg (1995), who developed a stepwise procedure to control the FDR. Storey (2002) proposes to estimate the FDR of a fixed rejection region and introduces the $q$-value, which is the minimum FDR level to reject the null hypothesis given observed data. Both the Benjamini–Hochberg procedure and the $q$-value assume independence among the summarizing statistics. Unfortunately the independence assumption rarely holds in practice; hence often discrepancy appears between the theoretical and the observed distributions of the summarizing statistics. For this reason, Efron has recently introduced an empirical Bayes (EB) procedure based on a two-group mixture model (Efron, 2004, 2007a,b). The EB procedure uses the $z$-values instead of the $p$-values and fits them using the two-group mixture model. The EB framework introduces two variants of the FDR: the local FDR, denoted by "fdr", is the ratio of the null sub-density to the marginal mixture density of the two-group model; the tail FDR, denoted by "Fdr", is the ratio of the null sub-survival function (tail probability) to the marginal survival function. The EB procedure estimates the null and the marginal mixture

* Corresponding author.
*E-mail address:* wonj@korea.ac.kr (J.-H. Won).

0167-7152/$ – see front matter 
http://dx.doi.org/10.1016/j.spl.2013.12.011

distributions from the data. Hence it takes into account the dependence among test statistics. The estimated null distribution is referred to as the empirical null.

The main theme of this paper is monotonicity in the FDR. Monotonicity is desirable in many settings as it maintains the order of the observed test statistics. In particular, we show that the monotonicity condition for the local FDR implies the monotone likelihood ratio condition (MLRC) of Sun and Cai (2007),[1] under which the local FDR yields the optimal oracle decision rule. We then show that a monotone estimate of the local FDR results in a data-driven decision rule that is, under some regularity conditions, asymptotically optimal. Furthermore, we prove that a monotone estimate of the local FDR satisfies the MLRC in finite-sample settings, which by itself is desirable in practice.

Despite many attractive features of monotonicity, unfortunately, few existing procedures to estimate fdr or Fdr take monotonicity into account. Broberg (2005) investigates the use of the monotone FDR in the setting that the theoretical null distribution of $p$-values is uniform on $[0, 1]$. In this setting, monotonicity of fdr (resp. Fdr) is equivalent to that of the marginal density function (resp. the marginal survival function). Monotonicity is enforced by estimating the marginal density function (resp. the marginal survival function) under appropriate constraints, either parametrically or non-parametrically. A similar procedure is employed by Strimmer (2008). For more flexible EB procedures (Efron, 2007a,b), however, one has to estimate both the null and the marginal distributions. We undertake to see how to impose monotonicity in this setting.

We begin with a review of the empirical Bayes theory of the false discovery rate. In Section 3, attractive statistical properties of the monotone FDR are discussed. We show that monotonicity in the local FDR is equivalent to that in the likelihood ratio of the components of the two-group mixture model and implies that of the tail FDR. After proving the claims made above, we propose a procedure that ensures monotonicity in the estimates of the local and the tail FDRs and that naturally leads to an adaptive decision rule using the monotonized estimates. In Section 4, we conduct a numerical study to demonstrate that the monotonized FDR can improve the performance of the FDR estimates. In Section 5, we illustrate that the proposed procedure can improve real-world data analyses. Section 6 concludes the paper.

## 2. Empirical Bayes theory of false discovery rates

This section reviews the empirical Bayes theory of false discovery rate inference, largely developed by Efron (2004, 2007a,b).

Suppose we have a collection of $N$ hypotheses and their corresponding "summarizing statistics" $T_1, \ldots, T_N$. Assume that $T_i$s have a common marginal distribution whose density is of the two-group mixture form:

$$f(t) = p_0 f_0(t) + p_1 f_1(t), \tag{1}$$

where $f_0(t)$ and $f_1(t)$ are the null and the non-null densities, respectively; $p_0$ is the proportion of the null group, and $p_1 = 1 - p_0$. We define the null sub-density as $p_0 f_0(t)$. The local false discovery rate (denoted by fdr) and the right tail FDR (denoted by Fdr) at $t$ are, respectively, defined as

$$\mathrm{fdr}(t) = \frac{p_0 f_0(t)}{f(t)} \quad \text{and} \quad \mathrm{Fdr}(t) = \frac{p_0 S_0(t)}{p_0 S_0(t) + p_1 S_1(t)}, \tag{2}$$

where $S_0(t)$ and $S_1(t)$ are the survival functions of the null and the non-null groups, respectively. Note that the tail FDR corresponds to one-sided hypotheses toward the positive side, and the other direction or the left tail counterpart can be similarly defined.

Knowledge of the null density $f_0(t)$ plays a crucial role in the inference regarding fdr and Fdr. The null distribution of the test statistics for single hypothesis testing is often known theoretically, e.g., standard normal, Student's $t$, or chi-square. However, in multiple hypothesis testing, the observed test statistics often do not follow the theoretical null distribution. This phenomenon may be due to failed assumptions, unobserved covariates, correlations among the samples or among the test statistics (Efron, 2007b).

To remedy this problem, several authors advocate a family of empirical Bayes procedures, referred to as the empirical null method (Efron, 2007a,b; Schwartzman, 2008). This method estimates the null distribution from the data itself. For $N$ sufficiently large, the components of the mixture density (1) can be estimated under a certain set of assumptions. These assumptions include that $f_0(t)$ is unimodal, and that the most of the probability mass around the peak of $f(t)$ is due to the null sub-density $p_0 f_0(t)$. Therefore, a reliable estimation of $f_0(t)$ and $p_0$ is very important for accurate inference of the FDRs discussed above.

To estimate $f(t)$, $f_0(t)$, and $p_0$, Efron (2007b) proposes two methods, named "central matching" and "MLE fitting". First, central matching is a two-step procedure. At step 1, the mixture density $f(t)$ is modeled as a semi-parametric exponential family, e.g., $f(t) = c_\beta \exp\{\sum_{j=1}^{7} \beta_j t^j\}$, where $c_\beta$ is a normalization constant. Subsequently the $N$ test statistics are binned into $K$ bins with equal width $\Delta$ centered at $t_1, t_2, \ldots, t_K$. Let $y_k$ be the count in bin $k$. Then the parameters $\{\beta_j\}$ are fitted to $\{y_k\}$ using Lindsey's method (Lindsey, 1974). At step 2, $f_0(t)$ is fit to the estimated $f(t)$ around $t = 0$. Assuming $f_0(t)$ is a normal density, the parameters (mean and variance) for $f_0(t)$ are estimated by least squares. Second, MLE fitting undertakes

[1] Sun and Cai (2007) call the condition "SMLR", without spelling out what it abbreviates. Later they refer to the same condition as "MLRC", while identifying that "SMLR" stands for the "symmetric monotone likelihood ratio" (personal communications with Wenguang Sun, 2013).