ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A scoring criterion for rejection of clustered *p*-values

Cai Qingyun

International College, Xiamen University, China

ARTICLE INFO

Article history:
Received 20 April 2015
Received in revised form 4 February 2016
Accepted 8 February 2016
Available online xxxx

Keywords:
CNV dataset
FDR
Cross-sample cancer study
Importance sampling
Multiple comparison
Sequential analysis

ABSTRACT

In dealing with the multiplicity problem of large dataset, clusters or families of hypotheses are often the units of interest. A scoring method is motivated in adopting a rejection space for *p*-values that are classified into spatial or labeled groups. A score that measures the benefits/costs of making a true/false discovery is computed and rejection space that maximizes the number of rejections with positive score is adopted. Renewal and boundary-crossing theories are used to compute the exceedance probability of the score. Level of strong group type I error control is validated using Monte Carlo and importance sampling methods. It is shown that the scoring method maintains detection power and achieves robustness against model deviation. The scoring method is applied on a copy number variation tumor dataset and short intervals of the chromosome with biological relevance are identified.

1. Introduction

Maintaining strong control when correcting for multiple comparisons on a large number of *p*-values, for example by applying the Bonferroni correction, can result in severe detection power loss. An important advancement of the statistical community has made collectively over the past two decades is on how excessive small *p*-value proportions can be used for declaring second-order significance. Benjamini and Hochberg (1995) proposed the use of false discovery rate (FDR) control and showed that Simes (1986) procedure achieves this control. Genovese and Wasserman (2002) and subsequently many other researchers, see for example Chi (2007) and references therein, studied the two-groups model that considering the probability of a null hypothesis being true or false, and showed that there is successful detection when there is a minimum threshold fraction of signals. With respect to *p*-value dependencies, FDR tests are less sensitive or conservative, see Sarkar and Chang (1997), Storey et al. (2004), Wu (2008), and also less sensitive to model deviation.

With the development of modern technology, multiplicity problem gets more complex in large dataset. For instance, microarray studies could involve testing tens of thousands of probes in the identification of differentially expressed genes. Key statistical components of microarray analysis are examined in Allison et al. (2006). In addition to multiple testing, one of the other components that are intensively used in the analysis of gene expression is the classification of genes to categories or clusters on the basis of prior information or some sort of similarity metric. Similarly, in multiple hypothesis testing, one approach to address multiplicity issue is to utilize the classification or structuring information of the dataset. Hierarchical FDR-controlling procedure is introduced in Yekutieli et al. (2006) to hierarchically test families of hypotheses that are arranged in a tree structure. The approach is applied to identify differentially expressed genes (see also Yekutieli, 2008 and Reiner-Benaim et al., 2007). Benjamini and Heller (2007) introduced a powerful testing procedure that using clusters as the testing units of spatial signals and it has two main advantages. Firstly, FDR can be controlled on clusters. Secondly, it increases signal-to-noise ratio when in many cases signals tends to be appeared in clusters.

E-mail address: caiqy@xmu.edu.cn.

http://dx.doi.org/10.1016/j.csda.2016.02.003
0167-9473/

Please cite this article in press as: Cai, Q., A scoring criterion for rejection of clustered *p*-values. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.02.003

In this paper, we propose a score function that measures the benefits (true discoveries) and costs (false discoveries) of rejecting p-values with labeling or location information. The method searches for positive score and adopts the rejection critical vector that maximizes the number of rejections on the multi-group scenario. Each coordinate of the critical vector represents a critical value for one group and null hypotheses with p-values less than the critical value are rejected. There are two parameters in the score function that investigator needs to determine, a benefit/cost ratio when making a true/false discovery and a parameter that determines the level of strong group Type I error control. Efron (2005, 2007, 2008) and Sun and Cai (2007) highlighted the benefits of estimating false discovery rate (referred as the local fdr) at the test statistic values. It has a different flavor from the multiple comparison methods mentioned earlier and has the positive characteristic of rejection at a value being chiefly determined by frequency of values lying close by. Though we convert these values to p-values before defining the score function, there is a nice connection between scoring method in the one group setting and local fdr. This will be elaborated in Section 2 where we introduce and motivate the score function for multiple grouping p-values. In Section 3, we discuss how the parameters of the score function should be chosen and compute the intercept parameter using analytical formulas, as well as Monte Carlo and importance sampling methods. In Section 4, we studied a cross-sample tumor dataset to identify copy number variation (CNV) prone regions. We end the paper with a discussion in Section 5, on the detection characteristics and robustness of the scoring method, followed by a preliminary numerical study for dependency case.

2. Scoring rejection spaces

Consider a large dataset with identification tags or location information on the p-values that can be used to partition them into m_0 groups. Let m_i be the number of p-values in group i for $i = 1, \ldots, m_0$. Define the following,

$$\begin{aligned}
R_i(t_i) &= \text{ number of rejections in group } i \text{ for a candidate critical value } t_i;\\
R(\mathbf{t}) &= \sum_{i=1}^{m_0} R_i(t_i), \text{ total number of rejections, where } \mathbf{t} = (t_1, \ldots, t_{m_0});\\
V_i(t_i) &= \text{number of rejected true null hypotheses in group } i;\\
V(\mathbf{t}) &= \sum_{i=1}^{m_0} V_i(t_i), \text{ total number of rejected true null hypotheses;}\\
S_i(t_i) &= \text{ number of rejected false null hypotheses in group } i;\\
S(\mathbf{t}) &= \sum_{i=1}^{m_0} S_i(t_i), \text{ total number of rejected false null hypotheses;}\\
\#(\mathbf{t}) &= \text{ number of groups containing rejected } p\text{-values.}
\end{aligned}$$

We define a score function that counts the number of rejections, with two parameters that adjusted to the costs of rejecting true null hypotheses and true null groups (a true null group contains only true null hypotheses). Denote the two specified penalty parameters as $\widetilde{\lambda}$ and η ($\widetilde{\lambda} > 0$ and $\eta > 0$). The score function is defined as follows.

$$\widetilde{sc}(\mathbf{t}) = S(\mathbf{t}) - \widetilde{\lambda} V(\mathbf{t}) - \eta \#(\mathbf{t}).$$

The critical vector $\mathbf{t}$ that maximizes the number of rejections over the set $\{\mathbf{t} : \widetilde{sc}(\mathbf{t}) \geq 0\}$ is chosen and those null hypotheses in ith group with p-values less than the ith entry of the critical vector are rejected. Under the assumption that p-values of true null hypotheses are independent and follow Uniform(0, 1) distribution, $V(\mathbf{t})$ can be estimated as $\sum_{i=1}^{m_0} m_i t_i$. A substitution of the score function is thus

$$sc(\mathbf{t}) = R(\mathbf{t}) - \lambda \sum_{i=1}^{m_0} m_i t_i - \eta \#(\mathbf{t}), \tag{2.1}$$

$\lambda = \widetilde{\lambda} + 1 > 1$ and $sc(\mathbf{0}) = 0$. To simplify the exposition, we do not make use of the estimates of the proportion of false null hypotheses. See Storey (2002, 2003) for the successful applications of these estimates in FDR control.

The score function is related to existing FDR methodology in one group setting. The FDR controlling procedure proposed in Benjamini and Hochberg (1995) is known as the Benjamini–Hochberg (BH) procedure. It rejects p-values no larger than the critical value $t = \max\{i : p_{(i)} \leq i\alpha/m\}$, where $m = \sum_{i=1}^{m_0} m_i$ is the total number of p-values, $p_{(1)}, \ldots, p_{(m)}$ are the order p-values and α is the desired controlling level. It is shown that the procedure controls FDR at $\pi_0\alpha$, i.e. FDR $= E[V(t)/R(t)] = \pi_0\alpha \leq \alpha$, with π_0 is the proportion of true null hypotheses. Consider in the scoring method, $m_0 = 1$, $\eta = 0$ and $\lambda = \alpha^{-1}$. Since

$$\{t : \widetilde{sc}(t) \geq 0\} = \{t : V(t)/R(t) \leq \alpha\},$$

the BH procedure is the same as rejecting all p-values smaller than the critical value t that maximizing $R(t)$ over the set $\{t : \widetilde{sc}(t) \geq 0\}$. The score function also has close connection with local fdr. Let $f_0(z)$ and $f_1(z)$ be the densities of true and false

Please cite this article in press as: Cai, Q., A scoring criterion for rejection of clustered p-values. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.02.003