Methodol Comput Appl Probab (2017) 19:429–443
DOI 10.1007/s11009-016-9491-x

A Double Application of the Benjamini-Hochberg Procedure for Testing Batched Hypotheses

Qingyun Cai[1] · Hock Peng Chan[2]

Received: 5 April 2014 / Revised: 30 August 2015
Accepted: 23 February 2016 / Published online: 5 March 2016

Abstract The Benjamini-Hochberg procedure (BH) controls the false discovery rate (FDR), and on a large dataset optimizes signal discovery subject to this control. However it applies a common p-value rejection threshold that precludes it from taking advantage of index information of the null hypotheses, making it suboptimal for detecting clustered signals. We propose a double application of the BH procedure on two-level hierarchical and related datasets, the first application to identify p-value batches, and a second application on each identified batch for null hypotheses rejections. We propose a mixture model on two tiers to model signal clustering, and show that on this model, the double application reduces FDR and maintains the power of BH. We show that the doubly applied BH satisfies an average FDR control. Benjamini and Bogomolov (J R Stat Soc Ser B 76:297–318, 2014) considered a more general class of procedures and error criterions, and showed average FDR control under dependency assumptions different from ours. Their proof is also technically different. We end the paper with a description of Yekutieli's (J Am Stat Assoc 103:309–316, 2008) procedure on hierarchical datasets, and a proposed hybrid of the double BH procedure and Yekutieli's procedure that combines the strengths of both.

Keywords Cluster detection · FDR control · Hierarchical datasets · Multiple hypothesis testing · PPV · Sensitivity

Mathematics Subject Classification (2010) 62L10

✉ Qingyun Cai
caiqy@xmu.edu.cn

[1] International College, Xiamen University, Xiamen, China

[2] Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

1 Introduction

Benjamini and Hochberg (BH) (1995) initiated the study of false discovery rate (FDR) control when they showed that FDR control is achieved by Simes' procedure (Simes 1986), which relaxes the p-value rejection threshold when there is a large proportion of small p-values. Based on this principle, when there is a large cluster of small p-values, the threshold should also be relaxed, but only locally around the cluster. This is the motivation behind the double application of the BH procedure, the first time globally on batches of p-values, the second time locally within each batch identified in the first application.

Large sample estimation in FDR control, brought forth by the introduction of the two-groups mixture model, has played a critical role in the understanding and improvements of the BH procedure. The incorporation of false null proportion estimation into the BH procedure by Storey (2002, 2003), results in increased signal discovery within the stated FDR constraint. Genovese and Wasserman (2002) applied large sample theory to show that the BH procedure is an asymptotically optimal distribution-free method and Chi (2007) provided limiting asymptotics of the BH procedure under the mixture model. Efron and Zhang (2011) used the mixture model to estimate local false discovery rates and applied them to DNA copy number datasets. A separate line of development is in adaptations of the BH procedure and FDR measures to hierarchical and spatial datasets (Benjamini and Heller 2007; Pacifico et al. 2004; Yekutieli 2008). Related to this is the recent interest in the use of the hidden Markov model to model and analyze signal clustering in spatial datasets (Chi 2011; Sun and Cai 2009; Wu 2008). Multi-stage designs have also been considered recently (Zehetmayer et al. 2008).

We investigate here the hierarchical datasets studied in Yekutieli (2008) (and applied in Reiner-Benaim et al. 2007), focusing on the two-level special case and introducing a mixture model on two tiers to understand the effect that signal clustering has on the occurrence of false discoveries in these datasets. The set of p-values is partitioned into disjoint subsets, and we refer to the p-values in a subset as a batch of p-values. The BH is applied twice, firstly between batches, followed by within batches. The mixture model on two tiers is introduced in Section 2, and on it we perform a large sample comparison of the operating characteristics of the doubly applied BH versus the standard BH. From this comparison, we conclude that repeating the BH increases its positive predictive value (i.e. reduces false discoveries) and maintains sensitivity (power). A recent paper by Benjamini and Bogomolov (2014) considered a wide class of procedures on two-level datasets, of which the double BH can be considered a special case. However no power analysis was done there. One of the main contributions of this paper is coming up with a mixture model on two tiers to help us differentiate and analyze competing procedures on two-level datasets.

In Section 3 the doubly applied BH is shown to satisfy an average FDR constraint. The average is over all batches identified in the first application of BH. Benjamini and Bogomolov (2014) considered a wider class of procedures and provided a more general error control result. Our dependency assumptions are however quite different, and the techniques used in our proof are also different. In Section 4 we combine the strengths of the double BH and Yekutieli's (2008) procedure, on a hybrid procedure that is more relaxed for within-batch p-value rejection. This is useful for applications in which identification of signals within batches is of interest. The theoretical predictions of Sections 2–4 are validated on a simulation study in Section 4. In Section 5 the application of the doubly applied BH is illustrated on a copy number dataset. The main paper ends with a short discussion in Section 6. The proofs are consolidated in the appendices.

2 Methodology

We start out by first considering ordered p-values $p_{(1)} \leq \cdots \leq p_{(n)}$ of n null hypotheses. The BH procedure for identifying signals, i.e. false null hypotheses, at a control level of α, is as follows:

BH procedure

1. Let $q = \min_{1 \leq j \leq n}(np_{(j)}/j)$ be the BH-adjusted p-value.
2. (a) If $q > \alpha$, then let $R = 0$.
   (b) If $q \leq \alpha$, then let $R = \max\{j : np_{(j)}/j \leq \alpha\}$.
3. Reject the null hypotheses corresponding to the R smallest p-values (i.e. those with p-values not more than $\alpha R/n$).

Let n_0 be the number of p-values belonging to true null hypotheses and assume that these p-values are independent and uniformly distributed. If $n_0 = n$, then $P(R > 0) = \alpha$, see Seeger (1968) or Simes (1986), i.e. the family-wise Type I error rate is controlled weakly at level α. Let V be the number of true null rejected by BH and let $R \vee 1 = \max(R, 1)$. It was shown in Benjamini and Hochberg (1995) that when the p-values are independent, the BH procedure guarantees that

$$\text{FDR} := E\left(\frac{V}{R \vee 1}\right) = \frac{\alpha n_0}{n} \leq \alpha. \tag{1}$$

Since $n_0 \leq n$, FDR is indeed controlled at level α.

In this paper, we consider a slightly more complicated two-level set-up, with the n null hypotheses pre-allocated to m batches, k_i null hypotheses in batch i for $1 \leq i \leq m$. Hence $n = k_1 + \cdots + k_m$. In the discussion in Section 6, we shall touch on extensions to datasets with even more complicated indexing. Let R_i be the number of rejected null hypotheses in batch i, and V_i the corresponding number of rejected true null, when the BH is applied. Hence $R = \sum_{i=1}^m R_i$ and $V = \sum_{i=1}^m V_i$. On this two-level set-up, our interest is in the following alternative, the focus of this paper.

Double BH (dBH) procedure

1. For $i = 1, \ldots, m$: Let $p_{i,(1)} \leq \cdots \leq p_{i,(k_i)}$ be the ordered p-values in batch i. The BH-adjusted p-value of the ith batch is defined to be

$$p_i^* = \min_{1 \leq j \leq k_i}(k_i p_{i,(j)}/j). \tag{2}$$

2. Let $p^*_{(1)} \leq \cdots \leq p^*_{(m)}$ be the ordered BH-adjusted batch p-values. We define the Simes-type combined p-value to be

$$p^* = \min_{1 \leq i \leq m}(mp^*_{(i)}/i).$$

3. (a) If $p^* > \alpha$: Let $S^* = 0$. Reject no null hypotheses and end here.
   (b) If $p^* \leq \alpha$: Let $S^* = \max\{i : mp^*_{(i)}/i \leq \alpha\}$ and $\alpha^* = \alpha S^*/m$.
   For $i = 1, \ldots, m$:
      i If $p_i^* > \alpha^*$: Do not reject any null hypothesis from batch i and let $R_i^* = 0$.
      ii If $p_i^* \leq \alpha^*$: Let

$$R_i^* = \max\{j : k_i p_{i,(j)}/j \leq \alpha^*\}.$$

      Reject the null hypotheses corresponding to the R_i^* smallest p-values in batch i.

It can be easily checked that the dBH rejects null hypotheses from exactly S^* batches. Let there be k_{i0} true null hypotheses from batch i (hence $n_0 = \sum_{i=1}^m k_{i0}$), and let V_i^* of them be rejected by dBH. Let $R^* = \sum_{i=1}^m R_i^*$ and $V^* = \sum_{i=1}^m V_i^*$.

If $n_0 = n$, then by induction $P(R^* > 0) = \alpha$, so as in BH, we have weak Type I error rate control. In addition, as will be shown in Section 3, dBH controls the average FDR (instead of the FDR for BH) at level α. That is,

$$\text{aveFDR} := E\left(\frac{1}{S^* \vee 1} \sum_{i:R_i^*>0} \frac{V_i^*}{R_i^*}\right) \le \alpha. \tag{3}$$

We call this average FDR control because the number of terms summed in Eq. 3 is S^*. The implication of such a control will be discussed in Section 3. When there is only $m = 1$ batch, Eq. 3 reduces to Eq. 1, the usual FDR control. The identified batches with p-values $p_i^* \le \alpha$ are known as selected family of hypotheses in Benjamini and Bogomolov (2014), where more general classes of procedures and error criteria are considered.

2.1 The Two-Groups Mixture Model

Let F_0 be the Uniform(0,1) distribution, representing the p-value distribution of a true null hypothesis, and let F_1 be the p-value distribution of a false null hypothesis. Let P_π denote the probability measure under which n_0, the number of true null, is distributed as Binomial($n, 1-\pi$). Under P_π, the n p-values are independent and identically distributed (i.i.d.) from the mixture distribution $F_\pi = (1-\pi)F_0 + \pi F_1$. For ease of exposition, we shall assume that F_1 is continuous, and we shall also assume that

$$\lim_{x \to 0^+} \frac{F_1(x)}{x} = \infty. \tag{4}$$

For $0 < \alpha < 1$ and $\pi > 0$, let $x_{\alpha,\pi}$ be the largest solution in x in

$$\frac{(1-\pi)x + \pi F_1(x)}{x} = \alpha^{-1} \left(\text{equivalently } \frac{F_1(x)}{x} = \frac{\alpha^{-1}-1}{\pi} + 1\right). \tag{5}$$

The existence of $x_{\alpha,\pi}$ is guaranteed by Eq. 4 since $\lim_{x\to 1} F_1(x)/x = 1$ and F_1 is continuous. By large sample theory (Glivenko-Cantelli Lemma), $x_{\alpha,\pi}$ is the weak limit of the p-value rejection threshold of BH cf. Genovese and Wasserman (2002), that is,

$$\alpha R/n \xrightarrow{p} x_{\alpha,\pi} \text{ as } n \to \infty, \tag{6}$$

provided that there exists $\epsilon_0 > 0$ such that

$$\frac{F_1(x_{\alpha,\pi} - \epsilon)}{x_{\alpha,\pi} - \epsilon} > \frac{\alpha^{-1}-1}{\pi} + 1 \text{ for } 0 < \epsilon < \epsilon_0. \tag{7}$$

The *sensitivity* of a multiple comparison procedure refers to the proportion of signals picked up by the procedure, while its *positive predictive value* (ppv) refers to the proportion of hypotheses it rejects that are true discoveries. In other words, 1-ppv is the false discovery proportion. By Eqs. 5–7, details in Appendix A, as $n \to \infty$,

$$\text{sensitivity(BH)} = \frac{R-V}{n-n_0} \xrightarrow{p} F_1(x_{\alpha,\pi}), \tag{8}$$

$$1 - \text{ppv(BH)} = \frac{V}{R \vee 1} \xrightarrow{p} \alpha(1-\pi). \tag{9}$$

2.2 The (Two-Groups) Mixture Model on two Tiers

We introduce now the mixture model on two tiers, which extends the two-groups mixture model to a two-level hierarchical dataset. The mixture model on two tiers is a signal clustering model that we will use to analyze multiple comparison procedures. Consider a dataset with $n = mk$ null hypotheses indexed into m batches, with each batch containing exactly k null hypotheses. We define a batch to be defective if it has a positive expected proportion of false null hypotheses, and non-defective otherwise. The set of defective batches, or more specifically the set of their indices, is denoted by D. Let π be the expected proportion of false null hypotheses within the whole dataset. We assume that a batch is defective with positive probability λ ($\geq \pi$), and within each defective batch, p-values are generated as i.i.d. $F_{\pi/\lambda}[= (1-\pi/\lambda)F_0 + (\pi/\lambda)F_1]$. For non-defective batches, p-values are generated as i.i.d. F_0. Since each p-value is distributed marginally as F_π, when $m \to \infty$ and $k \to \infty$, (8) and (9) hold. The terminologies of "batch" and "defective/non-defective" are borrowed from the quality control literature, to avoid clashes with the use of "group", which has quite different meanings in the FDR literature.

Efron (2008) introduced the separate-class model which contains two classes of null hypotheses having different expected proportions of false null. Our consideration here is in a sense a special case with one of the classes having zero expected proportion of false null. However in Efron (2008) it is known which class each null hypothesis belongs to, whereas here it is not. This is an important distinction and therefore to avoid confusing comparisons of procedures, we do not label this model as a special case of the separate-class model.

Theorem 1 *Consider the mixture model on two tiers with false null proportion π and defective batch proportion λ. Under (7), as $m \to \infty$ and $k \to \infty$,*

$$\textit{sensitivity(dBH)} = \frac{R^*-V^*}{n-n_0} \xrightarrow{p} F_1(x_{\alpha,\pi}), \tag{10}$$

$$1 - \textit{ppv(dBH)} = \frac{V^*}{R^* \vee 1} \xrightarrow{p} \frac{\alpha(\lambda-\pi)}{1-\alpha+\alpha\lambda} \tag{11}$$

$$[< \alpha(1-\pi) \textit{ when } \lambda < 1].$$

What is interesting and somewhat surprising in the comparison of BH versus dBH in Eqs. 8 and 10, is that dBH has the same asymptotic sensitivity as BH. That is, dBH provides no improvement in sensitivity despite the clustering of signals. The comparison of Eq. 9 with Eq. 12 however shows that applying dBH results in higher ppv, that is, fewer false discoveries.

The lack of sensitivity improvement of dBH over BH is due to the strictness of dBH in rejecting null hypotheses within batches. A within-batch control level of $\alpha^* = \alpha S^*/m$ (as opposed to α), where S^* (as defined in the dBH procedure) is the number of batches identified in the first application of dBH, is applied. Recall that a batch is identified if its batch p-value p_i^* does not exceed α^*. Let W^* of the identified batches be non-defective and recall that V_i^* is the number of false discoveries in batch i for $1 \le i \le m$.

Theorem 2 *Under the assumptions of Theorem 1, as $m \to \infty$ and $k \to \infty$,*

$$\frac{W^*}{S^* \vee 1} \overset{p}{\to} \alpha(1-\lambda), \tag{12}$$

$$\frac{1}{S^* \vee 1} \sum_{i \in D} \frac{V_i^*}{R_i^* \vee 1} \overset{p}{\to} \alpha(\lambda - \pi). \tag{13}$$

Equation 12 says that the dBH controls the

$$\text{batch FDR} := E\left(\frac{W^*}{S^* \vee 1}\right)$$

asymptotically. Equation 13 says that the average FDR in the defective batches is so small that when added to Eq. 12, the control at level α is still satisfied asymptotically. In contrast when BH is applied, batch FDR is in general not controlled. This makes batch-based inference difficult. We shall illustrate this point in Example 1 below. The proofs of Theorems 1 and 2 are given in Appendices B and C respectively. Note that whereas Benjamini and Bogomolov (2014) in their Theorem 1 and Section 5 have already shown such batch FDR and average FDR control, Theorem 2 provides stronger conclusions as the decomposition in Eqs. 12 and 13 for the mixture model on two tiers allows us to appreciate the relative contributions to the average FDR from the defective and non-defective batches.

3 Batch and Average FDR Control

In this section, we shall only assume that the true null p-values are independent, as opposed to having a more restricted mixture model. We shall show in Theorem 3 that the average FDR control of dBH at level α, as hinted by Eqs. 12 and 13, holds in general. Benjamini and Bogomolov (2014, Theorem 3) extended (14) below to a wider class of testing procedures and error criterions, albeit with the requirement that the set of all the p-values is positive regression dependent on the subset of true null hypotheses. Our Theorem 3 requires only that the true null p-values are i.i.d. Whereas Benjamini and Bogomolov applied the techniques developed in Benjamini and Yekutieli (2001, 2005), we extend the reverse-time martingale argument of Storey et al. (2004) to prove Theorem 3, see Appendix D.

Theorem 3 *If the true null hypotheses have i.i.d. Uniform*(0, 1) *p-values, then the average FDR*

$$E\left(\frac{1}{S^* \vee 1} \sum_{i:R_i^*>0} \frac{V_i^*}{R_i^*}\right) \quad \left[= E\left(\frac{W^*}{S^* \vee 1}\right) + E\left(\frac{1}{S^* \vee 1} \sum_{i \in D} \frac{V_i^*}{R_i^* \vee 1}\right)\right]$$
$$= \frac{\alpha}{m} \sum_{i=1}^{m} \frac{k_{i0}}{k_i} \quad (\le \alpha), \tag{14}$$

where k_{i0} refers to the number of true null hypotheses in batch i.

Example 1 Consider $m = 100$ batches, each containing $k = 1000$ null hypotheses. Let the p-value of the jth null hypothesis in batch i be given by

$$p_{ij} = \begin{cases} \Phi(Z_{ij} - 4) & \text{if } i = 1, \\ \Phi(Z_{ij}) & \text{if } i > 1, \end{cases}$$

where Z_{ij} are i.i.d. standard normal random variables and Φ is the distribution function of the standard normal. Thus only batch 1 contains false null hypotheses. From Fig. 1 left, we see that when the BH is applied at control level $\alpha = .1$, there are many batches containing rejected null hypotheses. In contrast, we see from Fig. 1 right that dBH applied at the same control level rejects only null hypotheses from batch 1.

4 Yekutieli's Procedure for Hierarchical Datasets

Yekutieli (2008) proposed an extension of BH for application on multi-level hierarchical datasets. For comparison purposes, we focus the discussion of his procedure on the batch set-up, a two-level special case.

YEK procedure

1. Let $p_{(1)} \le \cdots \le p_{(n)}$ be the ordered p-values of the n null hypotheses in the m batches and let the BH-adjusted p-value $q = \min_{1 \le j \le n}(np_{(j)}/j)$.
2. (a) If $q > \alpha$, then reject no null hypotheses and end here.
   (b) If $q \le \alpha$, then for $i = 1, \ldots, m$:
      i Let $p_{i,(1)} \le \cdots \le p_{i,(k_i)}$ be the ordered p-values of the null hypotheses in batch i.
      ii As in Eq. 2, let the BH-adjusted batch p-value $p_i^* = \min_{1 \le j \le k_i}(k_i p_{i,(j)}/j)$.
      iii If $p_i^* > \alpha$, then let $R_i^Y = 0$ and reject no null hypotheses in batch i.
      iv If $p_i^* \le \alpha$, then define

$$R_i^Y = \max\{j : k_i p_{i,(j)}/j \le \alpha\},$$

and reject the null hypotheses correponding to the R_i^Y smallest p-values in batch i.

Step 1 of YEK plays the important role of ensuring weak Type I error rate control at level α. YEK is excellent for detecting clustered signals. This is formalized in the following.

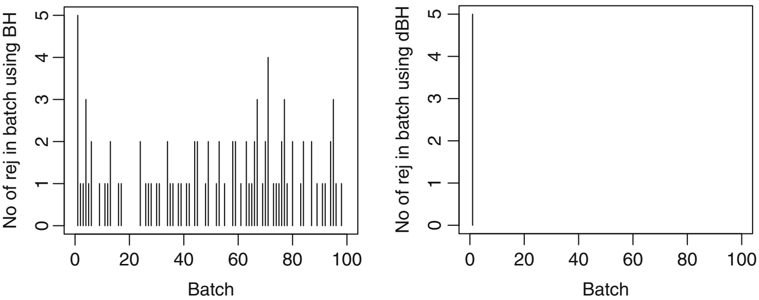

Fig. 1 Number (truncated to 5) of rejected null hypotheses when BH and dBH are applied at control level $\alpha = .1$. There are 100 batches of 1000 null hypotheses each and only batch 1 contains false null

Note that by Eq. 5, assumption (15) below is satisfied if $F_1(x)/x$ is monotone decreasing on $(x_{\alpha,\pi/\lambda} - \epsilon_0, x_{\alpha,\pi/\lambda}]$.

Theorem 4 *Consider the mixture model on two tiers with false null proportion π and defective batch proportion λ. Assume that there exists $\epsilon_0 > 0$ such that*

$$\frac{F_1(x_{\alpha,\pi/\lambda} - \epsilon)}{x_{\alpha,\pi/\lambda} - \epsilon} > \frac{\alpha^{-1} - 1}{\pi/\lambda} + 1 \textit{ for } 0 < \epsilon < \epsilon_0. \tag{15}$$

Then as $m \to \infty$ and $k \to \infty$,

$$\text{sensitivity(YEK)} \xrightarrow{p} F_1(x_{\alpha,\pi/\lambda}), \tag{16}$$

$$1 - \text{ppv(YEK)} \xrightarrow{p} \alpha(1 - \pi/\lambda). \tag{17}$$

In the comparison of Eq. 10 with Eq. 16, we see that the asymptotic sensitivity of dBH is less than that of YEK, for $x_{\alpha,\pi/\lambda} > x_{\alpha,\pi}$ when $\lambda < 1$. However YEK does not provide batch FDR control, whereas Theorem 3 says that dBH does. The lack of batch control by YEK is illustrated by the following.

Example 2 Consider $m = n$ with $k_i = 1$ for $1 \le i \le m$. If the BH-adjusted p-value $q \le \alpha$, then all null hypotheses with p-values not exceeding α, as opposed to $\alpha R/n$ for BH, are rejected by YEK. FDR is not controlled by YEK and since each p-value constitutes a batch, batch FDR is not controlled as well.

The nice thing about the dBH versus YEK comparison above is that it teaches us how to combine the strength of dBH (batch FDR control) with that of YEK (higher sensitivity). The resulting hybrid is described below.

YEK–dBH procedure

Execute steps 1 and 2 of dBH, as laid out in Section 2. Replace step 3 of dBH by the following.

3. (a) If $p^* > \alpha$: Let $S^* = 0$. Reject no null hypotheses and end here.
   (b) If $p^* \le \alpha$: Let $S^* = \max\left\{i : mp^*_{(i)}/i \le \alpha\right\}$ and $\alpha^* = \alpha S^*/m$.
   For $i = 1, \ldots, m$:
      i If $p^*_i > \alpha^*$, then do not reject any null hypotheses from batch i and let $R^{*Y}_i = 0$.
      ii If $p^*_i \le \alpha^*$, then let

$$R^{*Y}_i = \max\{j : k_i p_{i,(j)}/j \le \alpha\},$$

and reject the null hypotheses corresponding to the R^{*Y}_i smallest p-values in batch i.

The set of batches containing rejected null that is identified by dBH and YEK–dBH is identical. Hence for dBH and YEK–dBH, the number of non-defective batches that have been identified for null hypotheses rejections is W^*, and their batch FDR are identical as well. On the mixture model on two tiers, the asymptotic sensitivity and false discovery proportion of YEK and YEK–dBH are identical, see Eqs. 16–19.

Theorem 5 *(a) The YEK–dBH procedure controls the batch FDR at level α. That is, under the assumption that the true null hypotheses have i.i.d. Uniform(0,1) p-values,*

$$E\left(\frac{W^*}{S^* \vee 1}\right) \le \alpha.$$

(b) On the mixture model on two tiers, with false null proportion π and defective batch proportion λ, if Eq. 15 *holds, then*

$$\text{sensitivity(YEK–dBH)} \overset{p}{\to} F_1(x_{\alpha,\pi/\lambda}), \tag{18}$$

$$1 - \text{ppv(YEK–dBH)} \overset{p}{\to} \alpha(1 - \pi/\lambda). \tag{19}$$

Example 3 Consider $n = 32^2$ null hypotheses that are divided into $m = 32$ batches, with each batch containing $k = 32$ null hypotheses. Let m_1 of the batches contain k_1 false null hypotheses each. This is the set of defective batches denoted by D. Let the remaining batches contain only true null hypotheses. The false null p-values are distributed as $2\Phi(-|Z + 3|)$, where $Z \sim N(0, 1)$.

Figure 2 summarizes the performances of BH, dBH, YEK and YEK–dBH on a simulation study involving 1000 Monte Carlo repetitions for each plotted point. Figure 2 top left shows that YEK and YEK–dBH have comparable sensitivity. Figure 2 top right and bottom left show that the FDR and batch FDR of YEK–dBH are significantly smaller than that of YEK, when there is signal clustering. Figure 2 also shows that of the four procedures only BH controls FDR (top right), that dBH and YEK–dBH both control batch FDR (bottom left), and that dBH controls average FDR (bottom right), at the specified control level of $\alpha = .2$.

5 FDR Analysis of Copy Number Aberrations

We illustrate here the application of dBH on the copy number aberrations (CNA) of chromosome 1 of 207 glioblastoma subjects, in a dataset taken from The Cancer Genome Atlas (TCGA) Project (The Cancer Genome Atlas 2008). The aberrations, differences in copy numbers from the usual number of two, occur due to mutations of the chromosomes. Readings from $n = 42075$ probes are taken from chromosome 1 of each subject, the objective being to identify mutations that affect the onset of glioblastoma. The multi-subject aspect of the study enables differentiation between passenger mutations, which are random events not contributing to the onset of cancer, from driver mutations, which reside in ‘cancer genes’ and confer growth advantage to the cancer cell (Greenman et al. 2007; Haber and Settleman 2007). For example, neurofibromin 1 is a known human glioblastoma suppressor gene, and its influence on the development of glioblastoma is detectable by negative CNA close to this gene in multiple glioblastoma subjects. On the other hand the influence of AKT3, a protein which promotes tumor cell survival and development, is detectable by positive CNA in multiple tumor subjects.

Efron and Zhang (2011) proposed the estimation of local false discovery rates (fdr) on multi-subject copy number datasets, by assuming a two-groups mixture model. The fdr was estimated at each probe using information from the readings of all subjects at that probe. The maximum fdr estimate was used as a summary statistic, with critical thresholds simulated via block bootstrap. Positive and negative CNA were dealt with separately due to their different scientific implications. Their study on the dataset was successful in identifying

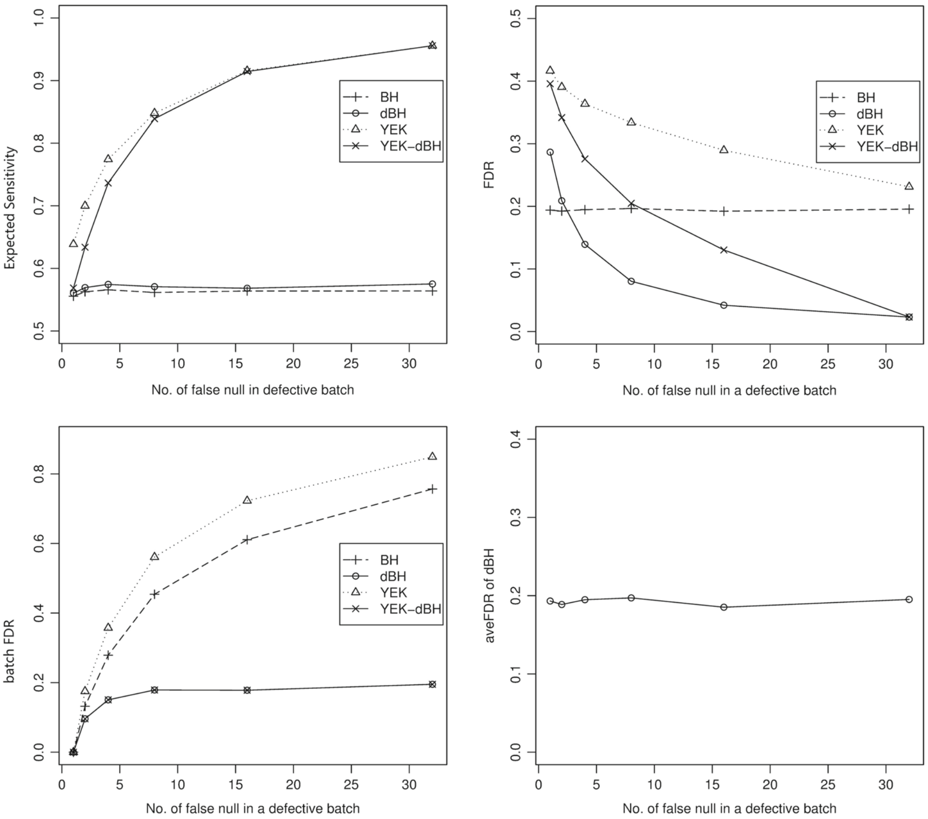

Fig. 2 Analysis of four multiple comparison procedures on 32 batches of 32 null hypotheses each, with m_1 defective batches each containing k_1 false null hypotheses, and the remaining batches containing only true null. Clustered signals are associated with small m_1 and large k_1. The plots above are based on $(m_1, k_1) =$ (32, 1), (16,2), (8,4), (4,8), (2,16), (1,32), with control level set to $\alpha = .2$

a short interval of probes 8850–8900 housing the cancer genes FAF1 and CDKN2C. The number of subjects there with significant negative CNA was estimated to be around 10.

5.1 Application of BH

Unlike in Efron and Zhang (2011), we apply the more powerful higher-criticism test (cf. Donoho and Jin 2004) to summarize the information on each probe across subjects. Let $\ell = 207$ be the number of subjects and Z_{ih}, $1 \leq h \leq \ell$, the z-scores at probe i. The higher-criticism test statistic at probe i is

$$\mathrm{HC}_i = \sup_{z:\bar{\Phi}(z)\geq \ell^{-1}} \frac{\#\{h : Z_{ih} \geq z\} - \ell\bar{\Phi}(z)}{\sqrt{\ell\Phi(z)[1-\Phi(z)]}}, \tag{20}$$

where $\bar{\Phi}(z) = 1 - \Phi(z)$ is the upper tail probability of the standard normal. Let p_i be the one-sided p-value of HC_i, which is estimated from 10^6 Monte Carlo repetitions.

We apply BH on $p_1, \ldots, p_n$ at control level $\alpha = .01$. The number of rejected subjects at probe i, when p_i is significantly small, is given by the value of $\#\{h : Z_{ih} \geq z\}$ when the argument in (20) is maximized. Figure 3 top left plots the number of rejections on the whole chromosome, and Fig. 3 bottom left the number on the segment 8600–9000. We see that like in Efron and Zhang (2011), the interval 8850–8900 is highlighted with around 10 subjects estimated to have significant negative CNA. The figure on the bottom left also indicates that at probes 8900+, there may be up to 30 subjects with significant negative CNA. This feature is missing in the plots given in Efron and Zhang (2011).

5.2 Application of dBH

We apply dBH on the p-values by grouping them into $m = 210$ batches of 200 consecutive p-values each, except for the last batch, which has 275 p-values. The batch division here is based on the notion that a gene that is linked to cancer is likely to affect all probes nearby, thereby resulting in the clustering of probe signals. Therefore edge effects aside, the batching captures the idea of first localizing the source of the link within a stretch of chromosomes (identifying a batch), before pin-pointing it more precisely (rejecting within an identified batch). As in Section 5.1, a control level of $\alpha = .01$ is applied. For easy comparisons, we place the outputs for dBH to the right of the outputs of BH in Fig. 3. We

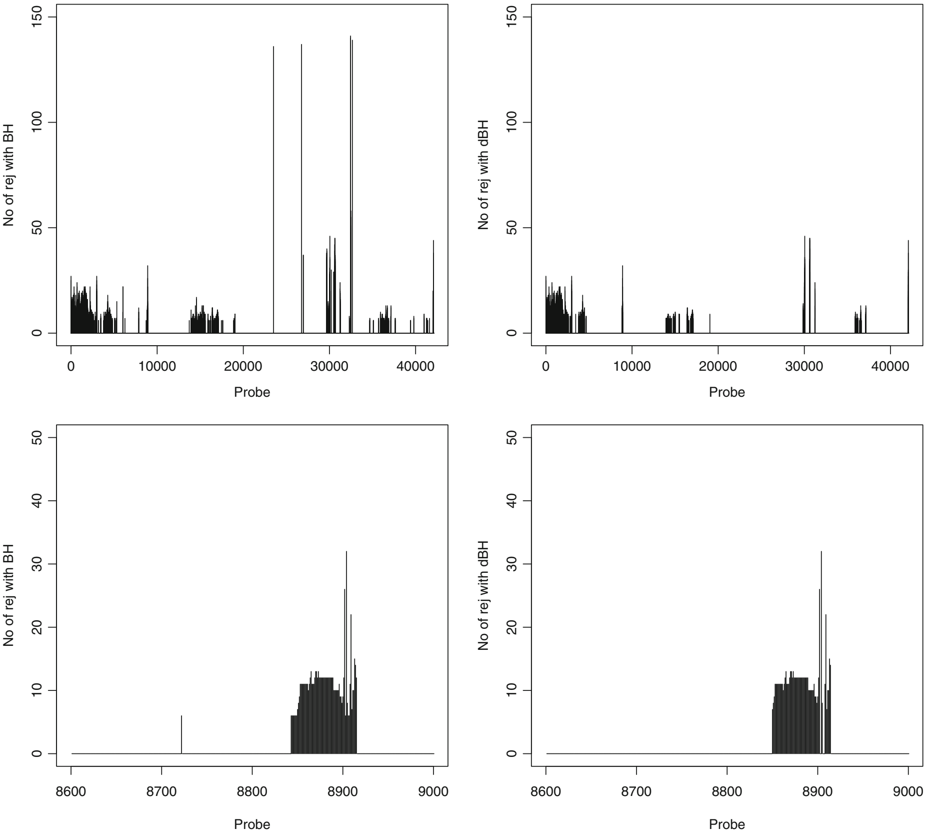

Fig. 3 Number of subjects with significant negative CNA on the full chromosome using BH and dBH. The cancer genes FAF1 and CDKN2C are located within probes 8850–8900

see from the comparisons that using BH twice results in less noisy output, in general on the full chromosome, and in particular on the shorter segment. The identified signal in the segment of biological interest (probes 8850–8900+) is not reduced when dBH is applied. These observations are consistent with the theoretical expectations arising from Eqs. 8–12.

5.3 Comparison of BH vs dBH

We see from Fig. 3 that significant probes come in clustered stretches. One possible explanation for this is that proximity to the gene of interest is critical in the presence of a signal. The dBH procedure takes advantage of this feature by re-organizing the one-dimensional spatial dataset into a two-level hierarchical dataset, with BH applied to first highlight short segments of probes, which are subsequently investigated by a second round of BH. In the application of just one round of BH, the clustering information is wasted and the follow-up investigation, with scattered false discoveries, is tricky.

6 Discussion

We show here that on two-level hierarchical datasets, the dBH procedure achieves both batch FDR and average FDR control. To understand the operating characteristics of dBH, we consider further the mixture model on two tiers to show that at the same control level, dBH improves upon BH by reducing the number of false discoveries while maintaining sensitivity. To achieve higher sensitivity, we propose a hybrid of dBH and Yekutieli's procedure that combines the strengths of both. It should be emphasized however that (as our simulation studies show) FDR control is not guaranteed on the procedures proposed here.

In principle we can deal with hierarchical datasets containing three or more levels by applying BH repeatedly. Further studies are needed on this multi-level procedure to understand the more complex dynamics involved. A potential pay-off would be a procedure that can detect signals of multiple scales, when applied on spatial datasets.

Appendix

In the proofs below, we write $U_n \overset{p}{\sim} V_n$ to mean that $U_n/V_n \overset{p}{\to} 1$ as $n \to \infty$.

A Proofs of Eqs. 8 and 9

Since $\alpha R/n$ is the p-value rejection threshold when there are n null hypotheses, by Eq. 6 and the weak law of large numbers,

$$\text{asymptotic sensitivity} = \lim_{n\to\infty} P_n\{\text{hypothesis rejected}|\text{false null}\} = F_1(x_{\alpha,\pi}), \tag{21}$$

and Eq. 8 holds. Since $x_{\alpha,\pi}$ is the solution of Eq. 5,

$$\begin{aligned}\text{asymptotic ppv} &= \lim_{n\to\infty} P_n\{\text{hypothesis false null}|\text{rejected}\} \\ &= \frac{\pi F_1(x_{\alpha,\pi})}{\pi F_1(x_{\alpha,\pi}) + (1-\pi)x_{\alpha,\pi}} = 1-\alpha(1-\pi),\end{aligned} \tag{22}$$

and Eq. 9 holds as well.

B Proof of Theorem 1

The asymptotic sensitivity of dBH is the asymptotic within-batch sensitivity of a defective batch. Since the within-batch false null proportion in a defective batch is π/λ, and the within-batch control level is α^*, by the arguments in Eq. 21,

$$\text{sensitivity(dBH)} \overset{p}{\sim} F_1(x_{\alpha^*,\pi/\lambda}),$$

and Eq. 10 follows from

$$x_{\alpha^*,\pi/\lambda} \overset{p}{\rightarrow} x_{\alpha,\pi} \text{ as } m \rightarrow \infty, k \rightarrow \infty. \tag{23}$$

To show Eq. 23, first note that by Eq. 4,

$$p_i^* \overset{p}{\rightarrow} 0 \text{ for } i \in D,$$

keeping in mind that the distribution of p_i^* is identical over $i \in D$. For $i \notin D$, p_i^* is distributed as Uniform(0,1). Hence the set of batch p-values is asymptotically a mixture of random variables having point mass at 0, with Uniform(0,1) random variables, in the ratio $\lambda : (1-\lambda)$. By the arguments leading to Eq. 6, $\alpha S^*/m (= \alpha^*)$ converges in probability to the solution in x of

$$\frac{F^*(x)}{x} = \frac{\alpha^{-1} - 1}{\lambda} + 1,$$

where F^* is the distribution function of the point mass at 0. In other words,

$$(\alpha^*)^{-1} \overset{p}{\rightarrow} \frac{\alpha^{-1} - 1}{\lambda} + 1. \tag{24}$$

By Eqs. 5 and 24,

$$\frac{F_1(x_{\alpha^*,\pi/\lambda})}{x_{\alpha^*,\pi/\lambda}} = \frac{(\alpha^*)^{-1} - 1}{\pi/\lambda} + 1 \overset{p}{\rightarrow} \frac{\alpha^{-1} - 1}{\pi} + 1 = \frac{F_1(x_{\alpha,\pi})}{x_{\alpha,\pi}}. \tag{25}$$

We then apply the regularity condition (7) to check that Eq. 23 follows from Eq. 25.

Because $\sum_{i \notin D} R_i^* = o_p(n)$, the asymptotic ppv of dBH is equal to the asymptotic within-batch ppv of a defective batch. The within-batch control level is α^* and the false null proportion of a defective batch is π/λ. Hence by the arguments in Eq. 22, the asymptotic ppv of dBH is $1 - \alpha^*(1 - \pi/\lambda)$. Replacing α^* by its asymptotic value $\frac{\alpha\lambda}{1-\alpha+\alpha\lambda}$, see Eq. 24, provides us with the weak convergence part of Eq. 12. The inequality part of Eq. 12 can be easily verified.

C Proof of Theorem 2

For $i \notin D$, the p-values $p_{i1}, \ldots, p_{ik_i}$ are i.i.d. Uniform(0,1) and hence the BH-adjusted batch p-value p_i^* is Uniform(0,1), cf. Simes (1986). Moreover, since $\{p_{ij} : i \notin D, 1 \le j \le k_i\}$ are independent, $\{p_i^* : i \notin D\}$ are independent as well. The relation (12) is then (9) applied on the batch p-values, with λ the probability that a batch is defective (false null), S^* the number of identified (rejected) batches, and W^* the number of incorrectly identified (rejected true null) batches.

To show Eq. 13, we first note that by Eq. 4, p^* is asymptotically 0 and hence by step 3(b) of dBH,

$$P_{\pi,\lambda}(S^* > 0,\ \alpha^* = \alpha S^*/m) \rightarrow 1. \tag{26}$$

Since $\#D \sim \text{Binomial}(m, \lambda) \overset{p}{\sim} m\lambda$, by Eq. 26,

$$\frac{\#D}{S^* \vee 1} \overset{p}{\sim} \frac{\#D}{S^*} \overset{p}{\sim} \frac{\alpha\lambda}{\alpha^*}. \tag{27}$$

We are able to conclude (13) after multiplying (27) with

$$\frac{1}{\#D} \sum_{i \in D} \frac{V_i^*}{R_i^* \vee 1} \overset{p}{\sim} \alpha^*(1 - \pi/\lambda). \tag{28}$$

To see (28), note that the within-batch control level is α^* and the false null proportion of a (defective) batch $i \in D$ is π/λ. So by the arguments in Eq. 22,

$$1 - (\text{ppv of batch } i) = \frac{V_i^*}{R_i^* \vee 1} \overset{p}{\sim} \alpha^*(1 - \pi/\lambda), \quad i \in D,$$

and Eq. 28 follows.

D Proof of Theorem 3

The first equality of (14) follows from $V_i^* = R_i^*$ for $i \notin D$ and $\sum_{i \notin D} \mathbf{1}_{\{R_i^* > 0\}} = W^*$. To obtain the inequality part, assume without loss of generality that $\{p_{ij} : 1 \le i \le m, 1 \le j \le k_{i0}\}$ are the [i.i.d. Uniform(0,1)] true null p-values, and that the false null p-values $\{p_{ij} : 1 \le i \le m, k_{i0} + 1 \le j \le k_i\}$ are known and positive. For $i = 1, \ldots, m$ and $u = r\ell$, with $1 \le \ell \le k_i$ and $1 \le r \le m$, define

$$V_i(u) = \#\{j \le k_{i0} : p_{ij} \le \alpha u/(mk_i)\}. \tag{29}$$

Double BH can be viewed as proceeding in the following manner.

1. Initialize with $r = m$.
2. Apply BH separately on each batch at control level $\frac{\alpha r}{m}$. Let $S_r (\le S_{r+1}$ when $r < m)$ be the number of batches having one or more rejections.
3. (a) If $S_r = r$, then let $S^* = r$ and end here.
   (b) If $S_r < r$, then decrease r by 1 and repeat step 2.

For $1 \le i \le m$, the number of null hypotheses rejected in batch i at step S^* is R_i^*, and the corresponding p-value rejection threshold there is $\tau_i/(mk_i)$, where

$$\tau_i = (S^* \vee 1)(R_i^* \vee 1). \tag{30}$$

By Doob's optional sampling theorem and Eq. 29,

$$E\left(\frac{V_i(\tau_i)}{\tau_i}\right) = E\left(\frac{V_i(mk_i)}{mk_i}\right) = \frac{\alpha k_{i0}}{mk_i}. \tag{31}$$

Sum (31) over i to get

$$E\left(\sum_{i=1}^{m} \frac{V_i(\tau_i)}{\tau_i}\right) = \frac{\alpha}{m} \sum_{i=1}^{m} \frac{k_{i0}}{k_i} \le \alpha. \tag{32}$$

Substitute Eq. 30 into Eq. 32 and replace $V_i(\tau_i)$ by V_i^* for $1 \le i \le m$ to obtain the second equality and inequality part of Eq. 14.

E Proof of Theorem 4 and 5

Since $\pi > 0$, by Eq. 4 and the weak law of large numbers, with asymptotic probability 1, $q \leq \alpha$ and within-batch rejection proceeds. For YEK, the within-batch control level is α and within a defective batch, the false null proportion is π/λ. The asymptotic sensitivity of YEK is the asymptotic within-batch sensitivity of a defective batch, which is $F_1(x_{\alpha,\pi/\lambda})$ [(16) shown]. The asymptotic ppv of YEK is likewise the asymptotic within-batch ppv of a defective batch, which is $\alpha(1-\pi/\lambda)$ [(17) shown]. Since $\sum_{i \notin D} R_i^Y = o_p(n)$, the rejections from non-defective batches are asymptotically negligible.

Theorem 4(a) follows easily from Eq. 14. In contrast to dBH, the within-batch control level of YEK–dBH is α (as opposed to α^*), and hence as in YEK, the asymptotic sensitivity and ppv of YEK–dBH are is the same as that of YEK [(18) and (19) shown].

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a powerful and practical approach to multiple testing. J R Stat Soc Ser B 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Benjamini Y, Yekutieli D (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. J Am Stat Assoc 100:71–93

Benjamini Y, Heller R (2007) False discovery rates for spatial signals. J Am Stat Assoc 102:1272–1281

Benjamini Y, Bogomolov M (2014) Selective inference on multiple families of hypotheses. J R Stat Soc Ser B 76:297–318

Chi Z (2007) On the performance of FDR control: constraints and a partial solution. Ann Stat 35:1409–1431

Chi Z (2011) Effects of statistical dependence on multiple testing under a hidden Markov model. Ann Stat 39:439–473

Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat 32:962–994

Efron B (2008) Simultaneous inference: when should hypothesis testing problems be combined? Ann Appl Stat 2:197–223

Efron B, Zhang NR (2011) False discovery rates and copy number variation. Biometrika 98:251–271

Genovese C, Wasserman L (2002) Operating characteristics and extensions of the false discovery rate procedure. J R Stat Soc Ser B 64:499–517

Greenman C et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446:153–158

Haber DA, Settleman J (2007) Drivers and passengers. Nature 446:145–146

Pacifico MP, Genovese C, Verdinelli I, Wasserman L (2004) False discovery control for random fields. J Am Stat Assoc 99:1002–1014

Reiner-Benaim A, Yekutieli D, Letwin NE, Gregory EI, Lee NH, Kafkafi N, Benjamini Y (2007) Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay. Bioinformatics 23:2239–2246

Seeger P (1968) A note on a method for the analysis of significances en masse. Technometrics 10:586–593

Simes RJ (1986) An improved bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Storey JD (2002) A direct approach to false discovery rate. J R Stat Soc Ser B 64:479–498

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31:2013–2035

Storey JD, Taylor JE, Siegmund DO (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66:187–205

Sun W, Cai TT (2009) Large-scale multiple testing under dependence. J R Stat Soc Ser B 71:393–424

The Cancer Genome Atlas (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061–1068

Wu WB (2008) On false discovery control under dependence. Ann Stat 36:364–380

Yekutieli D (2008) Hierarchical false discovery rate-controlling methodology. J Am Stat Assoc 103:309–316

Zehetmayer S, Bauer P, Posch M (2008) Optimized multi-stage designs controlling the false discovery or the family-wise error rate. Stat Med 37:4145–4160