Journal of Statistical Planning and Inference 199 (2019) 219–235

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Group sequential BH and its adaptive versions controlling the FDR

Sanat K. Sarkar [a,*], Aiying Chen [b], Li He [c], Wenge Guo [d]

[a] Temple University, United States
[b] Sanofi Pasteur, United States
[c] Merck Research Laboratories, United States
[d] New Jersey Institute of Technology, United States

ARTICLE INFO

Article history:
Received 17 November 2017
Received in revised form 30 May 2018
Accepted 4 July 2018
Available online 19 July 2018

Keywords:
False discovery rate
Multiple testing
Group sequential design
BH procedure
Adaptive procedure
Positive dependence

ABSTRACT

This paper considers the problem of simultaneous testing of multiple hypotheses in a multi-stage group sequential setting subject to control over the false discovery rate (FDR). A multi-stage group sequential form of the BH procedure is developed, and a proof of its FDR control for *p*-values satisfying a positive dependence condition both between and within stages is given. This group sequential BH is adapted to the proportion of true nulls in two different ways, resulting in the proposal of two adaptive group sequential BH. While one of these adaptive procedures is theoretically shown to control its FDR when the *p*-values are positively dependent between but independent within stages, the other one's FDR control is assessed through simulations. Comparative performance studies of the proposed procedures in terms of FDR control, power, and proportion of sample saved carried out through extensive simulations provide evidence of superior performance of the proposed adaptive procedures.

1. Introduction

In many modern scientific investigations, such as those in gene and protein expression studies where thousands of genes are tested for possible association with some disease condition, and in pharmacogenetics research where genetic contributions are studied in evaluating safety and efficacy of drugs, questions are investigated through large-scale, long-term follow up studies. For economic benefits and safety reasons, data are often accrued sequentially in these studies allowing interim analyses to be performed for making early decisions. Statistical analyses of data in these studies often involve simultaneous testing of a large number of hypotheses, making multiple testing in a sequential framework involving multiple stages a frequently arising statistical problem. This brings newer challenge for developing large-scale multiple testing method that is applicable to a sequential setting with multiple stages and controls an appropriate error rate such as the false discovery rate (FDR), which is the expected proportion of false rejections out of the total number of rejections.

The notion of FDR has been introduced by Benjamini and Hochberg (1995), along with a powerful and easy-to-use method, known as the BH method, that controls the FDR when multiple hypotheses are simultaneously tested in a non-sequential or single-stage setting. Given a set of m null hypotheses $H_1, \ldots, H_m$ to be simultaneously tested using their respective p-values $P_1, \ldots, P_m$, the level α BH method is a step-up method with the critical constants $\lambda_i = i\alpha/m$, $i = 1, \ldots, m$; that is, it rejects H_i for all i such that $P_i \le P_{(R)}$, where $R = \max\{i : P_{(i)} \le \lambda_i\}$, provided the maximum exists, otherwise, it rejects none, with $P_{(1)} \le \cdots \le P_{(m)}$ being the ordered values of the P_i's. Benjamini and Hochberg (1995)

* Corresponding author.
E-mail addresses: sanat@temple.edu (S.K. Sarkar), aiying.chen@sanofipasteur.com (A. Chen), li.he@merck.com (L. He), wenge.guo@njit.edu (W. Guo).

https://doi.org/10.1016/j.jspi.2018.07.001
0378-3758/

showed that the FDR of their method is less than or equal to $m_0\alpha/m$, where m_0 is the number of true null hypotheses, and hence the FDR is controlled at α, when the p-values are independent. Later on, Benjamini and Yekutieli (2001), Finner and Roters (2001), Sarkar (2002), and Storey et al. (2004) proved that the BH method's FDR under independence of the p-values is actually exactly equal to $m_0\alpha/m$. Benjamini and Yekutieli (2001) and Sarkar (2002) further showed that the FDR of the BH method is less than or equal to $m_0\alpha/m$ under a form of positive dependence condition that is shared by p-values in many multiple testing situations; see also Finner et al. (2009) and Sarkar (2008). The BH method, as a single-stage multiple testing method, has gained much popularity because of its applicability to a wide variety of scientific investigations and other desirable theoretical properties. Benjamini and Hochberg (2000) suggested the idea of adapting it to the data in an attempt to tighten its FDR control by estimating m_0 from the data and incorporating the estimate into it. They, of course, did not offer any theoretical proof of its ultimate FDR control. Several such adaptive versions of the BH method utilizing a wide variety of estimators for π_0, with theoretical proofs of their FDR control being offered only under independence, have been put forward in the literature. See, for example, Storey (2002), Storey et al. (2004), Benjamini et al. (2006), Gavrilov et al. (2009), Blanchard and Roquain (2009), and Sarkar (2008).

Papers dealing with FDR control in multi-stage statistical experiments involving simultaneous testing of multiple hypotheses do exist in the literature prior to this work.

Benjamini and Yekutieli (2005) introduced a two-stage procedure in the context of quantitative trait locus (QTL) mapping analysis, where the BH procedure is applied at each stage using the data available only at that stage and the hypotheses rejected at the first stage are further tested at the second stage. With m_0 being the number of true nulls and α_1 and α_2 being chosen differently for the first and second stages respectively, it controls the FDR at level $m_0\alpha_1\alpha_2/m$ under the same positive dependence condition for which the BH procedure is known to control the FDR.

Zehetmayer and coauthors in a sequence of papers considered the problem of controlling the FDR in the context of gene association or gene expression studies under a sequential setting when the test statistics are normally and independently distributed. Zehetmayer et al. (2005) considered a two-stage design where the total number of observations is fixed with certain fraction of these observations allocated to the first stage and the remaining observations distributed among the hypotheses whose first stage p-values are less than or equal to a prefixed futility boundary. They defined a sequential p-value and derived an estimation based approach to controlling the FDR following Storey (2002). Specifically, an estimate of the FDR is derived using the sequential p-values and a rejection threshold is chosen so that this estimate is less than or equal to a nominal level α. The hypotheses whose sequential p-values are less than or equal to this rejection threshold are then rejected. This approach was later extended by Zehetmayer et al. (2008) to control the FDR under multi-stage designs with fixed stagewise sample sizes as well as under multi-stage designs where the overall number of observations is fixed and at each stage a pre-specified fraction of observations is evenly distributed among the selected hypotheses according to some futility boundaries. Zehetmayer and Posch (2012) further assessed the following selection methods used at the first stage in a two-stage design — (i) the hypotheses whose first stage p-values are less than or equal to some prefixed futility boundary are selected; (ii) a prefixed number of most significant hypotheses are selected; and (iii) the hypotheses rejected by the BH procedure in Benjamini and Yekutieli (2005) at some fixed level α_1 are selected — with simulation studies showing that the FDR is controlled in all these scenarios. Zehetmayer et al. (2015) further proposed a sample size reassessment procedure controlling the FDR under a two-stage design. Based on the data available at the first stage, they derived an asymptotic expression of a selected power measure and determined a sample size for the second stage so that the power of the FDR controlling procedure is at some specified value. Their simulation results showed that their sample size reassessment procedure controls the FDR despite the data dependent choice of sample size.

Victor and Hommel (2007) also considered extending the BH procedure. Focusing on a two-stage adaptive design, they derived a method using a generalized definition of sequential p-value allowing for both early rejection and early acceptance at the interim analysis. They noted, however, that unlike in single testing where a futility boundary can often be determined based on sample size considerations, the choice of futility boundaries in multiple testing is often challenging since the joint distribution of the underlying test statistics is rarely known.

By utilizing sequential p-values, Malek et al. (2017) proposed a sequential conversion method to transform a fixed sample multiple testing procedure controlling some type I error rate, such as the FDR, to a sequential multiple testing procedure that still controls the same error rate. Specifically, their method applies the fixed sample multiple testing procedure, for example, the BH method, to the sequential p-values at each stage and allows an early rejection once sufficient evidence has accumulated against the null hypothesis.

Sarkar et al. (2013) extended the BH method and its adaptive version from single- to a two-stage adaptive design setting. More specifically, they considered screening the null hypotheses sequentially at the first stage as being rejected or accepted subject to certain boundaries on the FDR across all hypotheses and testing the remaining null hypotheses at the second stage having combined their p-values from the two stages using some combination function. These methods were theoretically proved to control the FDR under the assumption that the pairs of first and second-stage p-values across all hypotheses are independent and those which correspond to the null hypotheses are identically distributed as a pair (p_1, p_2) satisfying the p-clud property of Brannath et al. (2002). Bartroff and Song (2013) further considered the problem of developing a multi-stage FDR controlling procedure and developed such a procedure by appropriately adjusting the BH critical values at each stage, and assumed independence of the p-values across the hypotheses to prove its FDR control. However, in most studies involving group sequential design, the p-values are rarely independent across hypotheses, just as in the case of fixed sample design, and the underlying dependence structure can often be characterized by assuming a positive dependence condition.

Therefore, there seems to be an urgent need for developing a multiple testing procedure in a group sequential framework that controls the FDR under positive dependence across hypotheses and effectively identifies the false null hypotheses.

In this paper, we propose such a procedure, which we call the Group Sequential BH (GSBH) procedure. It extends the original BH method from single to multiple stages in a group sequential setting and uses the alpha spending function approach of Lan and DeMets (1983), which is widely used to set group sequential boundaries that control the Type I error rate while allowing flexibility in the number and the timing of the interim looks. With $t_k \in [0, 1]$ being the fraction of information observed up till kth interim look, an alpha spending function $\alpha(t_k)$ describes the rate at which the total error rate α is spent and it satisfies $\alpha(0) = 0$ at the beginning of trial and $\alpha(1) = \alpha$ at the end of trial (Proschan et al. (2006)). In GSBH, we consider allocating to each analysis stage a nominal error level based on some alpha spending function. Then, at each stage, we apply, at the corresponding nominal level, a step-up procedure based on the p-values that are associated with the active hypotheses (those not rejected in the previous stages) and BH type critical values. We show that the proposed GSBH procedure controls the FDR at level $m_0\alpha/m$ under Assumptions 1 and 2 defined in Section 2. These two assumptions considered together can be viewed as an adaptation of the positive dependence condition typically assumed for the BH method to control its FDR, from single to multiple stages. We also propose two adaptive versions of the GSBH procedure. While the first of these adaptive GSBH procedures uses at each stage a standard estimate of m_0 obtained from the data available at the first stage, the second does the same at each stage but based on the data available up to that stage. We offer a theoretical proof of the first adaptive GSBH procedure's control of FDR when the p-values are independent across hypotheses but positively dependent across stages. For the second proposed adaptive GSBH, however, we provide simulation evidence of its FDR control under the same assumption about the p-values.

We carry out extensive simulation studies to compare the performance of the three proposed procedures against the BH, which is the ideal one to use had the data been available across all stages. The performance comparisons are made in terms of the FDR control, the average power, the expected proportion of saved samples, and the FNR (the expected proportion of false hypotheses that are accepted out of the total number of accepted hypotheses) in situations where the underlying test statistics are either independent or positively dependent. We perform these studies using the following two alpha spending functions, $\alpha_{P0}(t) = \alpha \log(1 + (e - 1)t)$ and $\alpha_{OF}(t) = 2(1 - \Phi(z_{\alpha/2}/\sqrt{t}))$, where Φ and $z_{\alpha/2}$ are respectively the cumulative distribution function and the upper $\alpha/2$th percentile of the standard normal distribution, α is the total error rate, and $t \in [0, 1]$ is the information fraction. These two alpha spending functions approximate, respectively, the rejection boundaries of Pocock (1977) and O'Brien and Fleming (1979) for group sequential tests with equal group sizes. The Pocock (1977) rejection boundaries are constant across stages, while O'Brien and Fleming (1979) rejection boundaries change across stages allowing easier rejections at later stages. The outputs of these two alpha spending functions are presented in Table 2 for a four-stage design with equal allocation of the total sample size across the 4 stages and with the total error rate α chosen as 0.025. A real data set is also used to demonstrate applications of our proposed procedures under this same design.

The paper is organized as follows. With some concepts and assumptions given in Section 2, we present our proposed group sequential BH procedure and its adaptive versions in Section 3. The numerical findings from simulation studies are given in Section 4, and a real data application is presented in Section 5. Some concluding remarks are made in Section 6 and proofs of all results are given in the Appendix.

2. Notations and assumptions

Suppose that m null hypotheses $H_i, i = 1, \ldots, m$, are simultaneously tested in a K-stage group sequential design. Let $\mathbf{P}^{(k)} = \{P_i^{(k)}, i = 1, \ldots, m\}$ be the set of p-values corresponding to the m null hypotheses at stage k based on all the iid observations taken up to that stage. Let, for $k = 1, \ldots, K$, I_k be the index set for the active hypotheses at the beginning of the kth stage (that is, the hypotheses not rejected in the previous stages), with $I_1 = \{1, \ldots, m\}$, and I_0 be the index set of the true null hypotheses. For $k = 1, \ldots, K$, denote the p-values corresponding to the active hypotheses at stage k as $\mathbf{P}^{((k),(I_k))} = \{P_j^{((k),(I_k))}, j \in I_k\}$, with their ordered values as $P_{(1)}^{((k),(I_k))} \leq \cdots \leq P_{(|I_k|)}^{((k),(I_k))}$ and the corresponding hypotheses as $H_{(1)}^{((k),(I_k))}, \ldots, H_{(|I_k|)}^{((k),(I_k))}$, where $|I_k|$ is the cardinality of the set I_k. It is important to note that I_k is random for $k = 2, \ldots, K$, and it depends on the p-values in the first $k - 1$ stages.

The following assumption is made for the marginal distributions of the p-values associated with each null hypothesis:

Assumption 1. For each $i \in I_0$ and $k = 1, \ldots, K$, $P_i^{(k)} \sim U(0, 1)$.

We make the following assumption on the dependence structure of the p-values to propose the GSBH:

Assumption 2. For each $k = 1, \ldots, K$, the sets of p-values $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(k)}$ are jointly positively regression dependent on subset (PRDS) of p-values associated with the null hypotheses; that is, for any coordinatewise nondecreasing function φ of all the individual p-values in $(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(k)})$, the conditional expectation $E\{\varphi(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(k)}) \mid P_i^{(j)} \leq u\}$ is nondecreasing in u, for each $i \in I_0$ and $j = 1, \ldots, k$.

Assumption 2 can be viewed as an extension from single stage (when $K = 1$) to multiple stages of the positive dependence condition assumed for the BH method to control its FDR under dependence [Sarkar (2008), and Finner et al. (2009)]. It is a slightly weaker form of, and we loosely define it here as, the PRDS condition originally assumed in Benjamini and Yekutieli

(2001) and Sarkar (2002). In the single-stage case, it is satisfied by p-values generated from test statistics having multivariate distributions, including multivariate normal with non-negative correlations and multivariate t with the associated normal statistics having non-negative correlations [Karlin and Rinott (1980), Sarkar and Chang (1997), and Sarkar (1998)], arising in many multiple testing situations.

In the current setting of group sequential multiple testing involving multiple stages, we will show in the example below that this assumption is satisfied by p-values generated from normally distributed test statistics with non-negative correlations.

Example. Given a set of m normally distributed random variables $\mathbf{X}$ with mean vector $\mu = (\mu_1, \ldots, \mu_m)'$ and known covariance matrix Σ having non-negative correlations, consider testing $H_{i0} : \mu_i = 0$ against $H_{i1} : \mu_i > 0$ simultaneously for $i = 1, \ldots, m$ in a K-stage group sequential design based on cumulative samples of sizes $n_1 \leq \cdots \leq n_K$ up to each of the K stages. With $\bar{\mathbf{X}}^{(k)}$ denoting the sample mean vector at the kth stage, let $\mathbf{Z}^{(k)} = \sqrt{n_k}\bar{\mathbf{X}}^{(k)}$ be the vector of marginal test statistics associated with all the hypotheses at the kth stage, for $k = 1, \ldots, K$. For any fixed $k = 1, \ldots, K$, $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(k)}$ are jointly distributed under any alternatives as mk-dimensional multivariate normal with the mean vector $(\sqrt{n_1}, \ldots, \sqrt{n_k})' \otimes \mu$ and the covariance matrix $A \otimes \Sigma$, where $A = ((a_{jj'}))$ is a $k \times k$ matrix with $a_{jj'} = \min(\sqrt{n_j}, \sqrt{n_{j'}})/\max(\sqrt{n_j}, \sqrt{n_{j'}})$, for $j, j' = 1, \ldots, k$. Since the off-diagonal entries of $A \otimes \Sigma$ are all non-negative, the mk random variables in $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(k)}$ are jointly PRDS, similar to what is known in single-stage multiple testing involving multivariate normal test statistics; that is, for any coordinatewise nondecreasing function ϕ of the mk random variables in $(\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(k)})$, the following property holds:

$$E\{\phi(\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(k)}) \mid Z_i^{(j)} \leq z\} \text{ is nondecreasing in } z, \tag{1}$$

with $Z_i^{(j)}$ being the ith component of $\mathbf{Z}^{(j)}$, for each i such that $\mu_i = 0$ and $j = 1, \ldots, k$. In other words, the p-values generated from these test statistics satisfy Assumption 2, since this assumption remains invariant under comonotone transformations of the underlying random variables.

The adaptive version of the GSBH will be developed under the following assumption:

Assumption 2a. For each $i \in I_0$ and $k = 1, \ldots, K$, the sets of p-values $P_i^{(1)}, \ldots, P_i^{(k)}$ are positively dependent through stochastic ordering (PDS); that is, for any coordinatewise nondecreasing function ψ of $P_i^{(1)}, \ldots, P_i^{(k)}$, the conditional expectation $E\{\psi(P_i^{(1)}, \ldots, P_i^{(k)}) \mid P_i^{(j)} \leq u\}$ is nondecreasing in u, for each $j = 1, \ldots, k$.

The notion of PDS was originally introduced in Block et al. (1985), but in a slightly stricter form (with the expectation of ψ being taken conditional on $P_i^{(j)} = u$). Assumption 2a is a weaker version of Assumption 2.

3. The proposed procedures

This section presents our three proposed group sequential multiple testing procedures controlling the FDR given an alpha spending function, the first one of which is a multi-stage group sequential version of the BH method while the other two are adaptive versions of it.

Definition 3.1 (*Group Sequential BH (GSBH) Method*)**.** Given an alpha spending function $\alpha(t)$ and information fractions $t_1 \leq \cdots \leq t_K$, the K-stage group sequential BH procedure proceeds as follows based on K sequences of critical constants $\{\lambda_i^{(k)} = \frac{i}{m}\alpha_k, i = 1, \ldots, m\}$, with $\alpha_k = \alpha(t_k) - \alpha(t_{k-1})$, $k = 1, \ldots, K$:

- Stage 1: Let $R_1 = \max\{1 \leq i \leq m : p_{(i)}^{((1),(I_1))} \leq \lambda_i^{(1)}\}$ be the number of rejections of a step-up procedure based on $\mathbf{P}^{((1),(I_1))}$ and the critical constants $\lambda_j^{(1)}$, $j = 1, \ldots, |I_1|$. Reject the hypotheses $H_{(i)}^{((1),(I_1))}$ for $i \leq R_1$. If $R_1 = |I_1|$, then reject all hypotheses and stop. Otherwise, all the remaining hypotheses are tested in stage 2.

  $\vdots$

- Stage k: Let $R_k = \max\{j \leq |I_k| : p_{(j)}^{((k),(I_k))} \leq \lambda^{(k)}_{\sum_{i=1}^{k-1} R_i + j}\}$ be the number of rejections of a step-up procedure based on $\mathbf{P}^{((k),(I_k))}$ and the critical constants $\lambda^{(k)}_{\sum_{i=1}^{k-1} R_i + j}$, $j = 1, \ldots, |I_k|$. Reject the hypotheses $H_{(j)}^{((k),(I_k))}$ with all $j \leq R_k$. If $R_k = |I_k|$, then reject all hypotheses and stop testing. Otherwise, continue to the next stage.
- Continue until all the hypotheses are rejected or the final stage is reached.

Let V_k and R_k denote the numbers of false rejections and total rejections, respectively, at stage k, for $k = 1, \ldots, K$, in the above K-stage group sequential procedure. Then, the overall FDR of this procedure is given by

$$FDR = E\left\{\frac{V_1 + V_2 + \cdots + V_K}{(R_1 + R_2 + \cdots + R_K) \vee 1}\right\},$$

where $a \vee b = \max(a, b)$. The following theorem, one of our main results, establishes the overall FDR control of the GSBH method at α under some positive dependence conditions.

Theorem 3.1. *The FDR of the GSBH method satisfies the following inequality:*

$$FDR \leq \pi_0 \alpha, \text{ where } \pi_0 = \frac{m_0}{m},$$

under Assumptions 1 and 2.

The proof of this theorem is given in Appendix.

Next, we propose two adaptive versions of the GSBH method, each of which is developed as a repeated application of a single-stage adaptive BH method utilizing an estimate of π_0 using the available data. Several versions of such single-stage adaptive BH method have been put forward in the literature; see, e.g., Blanchard and Roquain (2009), Sarkar (2008), and He and Sarkar (2013) for a list of those methods. However, we will use the following one that is commonly used as an adaptive BH method to control FDR in the context of single-stage multiple testing: Apply the α-level BH method using adaptive p-values $Q_i = \hat{\pi}_0 P_i$, where $\hat{\pi}_0 = [W(\eta) + 1]/m(1-\eta)$, and $W(\eta) = \sum_{i=1}^{m} \mathbb{1}(P_i > \eta)$, for some tuning parameter $0 < \eta < 1$. This, like all other single-stage adaptive BH methods proposed in the literature, is theoretically known to control the FDR under independence of the p-values.

Definition 3.2 (*Adaptive Group Sequential BH (AdaptGSBH) Method*)**.** The AdaptGSBH method for K-stage group sequential multiple testing proceeds as the GSBH method with $\mathbf{P}^{(k)}$ replaced by $\mathbf{Q}^{(k)} = \{Q_i^{(k)} = \hat{\pi}_0 P_i^{(k)}, i = 1, \ldots, m\}$, where $\hat{\pi}_0 = [m - \sum_{i=1}^{m} \mathbb{1}(P_i^{(1)} \leq \eta) + 1]/m(1-\eta)$, for some fixed $0 < \eta < 1$.

Theorem 3.2. *The AdaptGSBH controls the FDR if the p-values are independent across hypotheses and Assumption 2a holds.*

A proof of this theorem is given in Appendix.

The reason we have considered using the data available only in stage 1 to estimate π_0 while developing AdpatGSBH is that it is amenable to a theoretical proof of the corresponding adaptive GSBH's control of FDR, at least under independence. However, a more intuitive and meaningful approach to estimating π_0 at the kth stage would be to use all the observations available up to that stage to obtain the following:

$$\hat{\pi}_0^{(k)} = \frac{m - \sum_{i \in I_k} \mathbb{1}(P_i^{(k)} \leq \eta) - \sum_{j=1}^{k-1} \sum_{i \in I_j \setminus I_{j+1}} \mathbb{1}(P_i^{(j)} \leq \eta) + 1}{m(1-\eta)}, \tag{2}$$

for $k = 1, \ldots, K$. Therefore, we propose in the following an adaptive version of GSBH based on this sequence of estimates as an alternative to AdpatGSBH.

Proposition 3.1. *An alternative to AdaptGSBH for K-stage group sequential multiple testing proceeds as GSBH with* $\mathbf{P}^{(k)}$ *replaced by* $\mathbf{Q}^{(k)} = \{Q_i^{(k)} = \hat{\pi}_0^{(k)} P_i^{(k)}, i = 1, \ldots, m\}$, *where* $\hat{\pi}_0^{(k)}$ *is defined in (2) for some fixed* $0 < \eta < 1$.

Of course, a theoretical justification of this alternative to AdaptGSBH under the same conditions assumed in Theorem 3.2 seems extremely complicated, and so we rely on numerical validation of its FDR control while examining its performance relative to its competitors in the next section.

4. Simulation studies

This section presents the results of two simulation studies we conducted to investigate how well our proposed GSBH procedure and its two adaptive versions, the AdaptGSBH based on $\hat{\pi}_0$ and its alternative based on the sequence of estimators $\hat{\pi}_0^{(k)}, k = 1, \ldots, K$, perform. For convenience, these two adaptive GSBH procedures are referred to as ad.GSBH1 and ad.GSBH2 respectively in our simulation studies. Since the BH procedure would be the ideal one to use had the data been available for all the hypotheses across all stages, we have included it in our simulation studies as a benchmark. The performance of each of these procedures is measured in terms of its FDR control and its power, which is defined using two different quantities — False Non-Discovery Rate (FNR), the expected proportion of false acceptances among all acceptances, and Average Power, the expected proportion of false nulls that are rejected. We considered testing $m(=50)$ hypotheses simultaneously in a four-stage group sequential design with simulated samples equally allocated to each stage, and chose the tuning parameter η as 0.5 for the two adaptive GSBH procedures.

In the first study, we assessed the relative performances of the three proposed procedures against BH under independence in two scenarios — (i) across different values of π_0, and (ii) across different values of μ, the common mean assumed for each of the alternative hypotheses. More specifically, we did the following:

(1) Generated $N = 120$ multivariate normal random variables independently from $N_m(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ with $\pi_0 m$ proportion of the μ_i's being equal to 0 and the rest being equal to a common value μ, and $\Sigma = I_m$, first for each $\pi_0 \in \{0.1, 0.2, \ldots, 0.9\}$ with $\mu = 0.2$ and then for each $\mu \in \{0.1, 0.2, \ldots, 1\}$ with $\pi_0 = 0.5$.

(2) Allocated $N/K = 30$ samples to each of the four stages and calculated the cumulative standardized sample mean for each stage based on the samples up to that stage before converting them to p-values to test $H_i : \mu_i = 0$ against $K_i : \mu_i > 0$ simultaneously for $i = 1, \ldots, m$ using the four procedures, the GSBH, ad.GSBH1, ad.GSBH2, and BH. Specifically, to apply the GSBH (and its adaptive procedures), we started by applying to the first stage p-values (and Q values) a stepup procedure

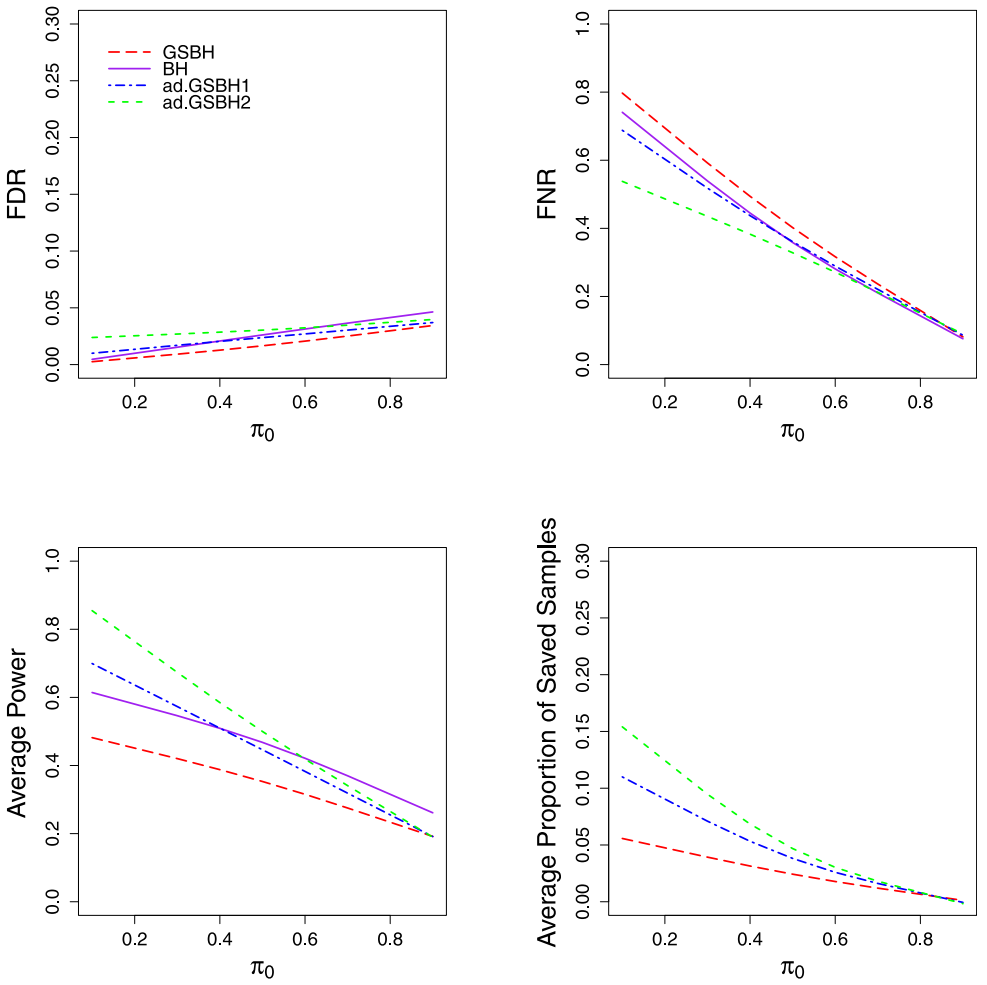

Fig. 1. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under independence with increasing proportion of true null hypotheses π_0 using the O'Brien–Flemming alpha spending function.

based on the first-stage critical values, proceeded to the second stage to test the hypotheses that are not rejected at the first stage using a stepup procedure based on the second-stage critical values and continued in the same fashion until all the hypotheses are rejected or the final stage has been reached. On the other hand, the BH procedure is applied to the p-values at the last stage.

(3) Determined the false discovery proportion (FDP, defined as the proportion of true null hypotheses that are rejected, and as 0 if there is no rejection), false non-discovery proportion (FNP, defined as the proportion of false null hypotheses that are accepted and as 0 if there is no acceptance), and the proportion of correctly rejected false nulls, for each of the four procedures, and the proportion of the samples that are saved for the first three procedures.

(4) Repeated the steps (1)–(4) 2000 times before calculating the averages of FDP, FNP, and proportion of correctly rejected false nulls to simulate the values of FDR, FNR and Average Power, respectively, for each of the four procedures, and the average of the proportion of the saved samples for the three group sequential procedures.

The results in this study are displayed in Figs. 1–4. Figs. 1 and 3 show the results for O'Brien–Flemming alpha spending function, respectively, across different values of π_0 when $\mu = 0.2$ and across different values for μ when $\pi_0 = 0.5$, while Figs. 2 and 4 do the same for Pocock alpha spending function. As seen from these graphs, the FDR is controlled by all our proposed procedures under independence. Although this is expected for GSBH and ad.GSBH1 due to Theorems 3.1 and 3.2, it is an interesting finding for ad.GSBH2 since a theoretical justification of its FDR control under independence is yet to be established. It is also interesting to note that the performance of GSBH can be significantly improved by their adaptive versions when π_0 is small, a scenario when such adaptation is much needed, or when μ is large. These adaptive versions, especially ad.GSBH2, can even outperform BH in those scenarios. In terms of sample saving, the adaptive versions of GSBH, especially ad.GSBH2, is more efficient than GSBH. The adaptive procedures do lose their edges over their non-adaptive version, GSBH, as π_0 increases.

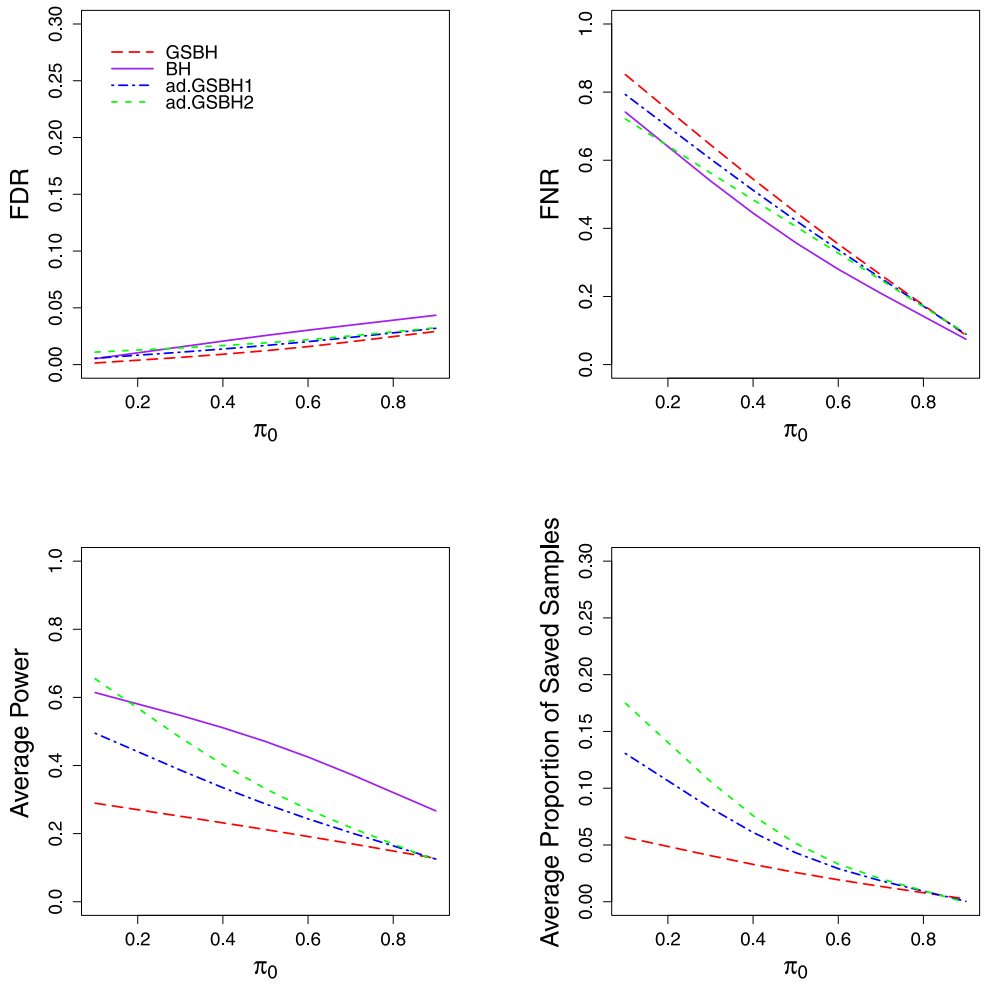

Fig. 2. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under independence with increasing proportion of true null hypotheses π_0 using the Pocock alpha spending function.

Our second simulation study focused on evaluating how our proposed procedures perform with increasing strength of positive dependence under three specific dependence structures: Uniform pairwise dependence, auto-regressive of order one [AR(1)] dependence, and block dependence. Specifically, we generated $N = 120$ independent multivariate normal random variables each from the $N_m(\boldsymbol{\mu}, \Sigma)$ distribution, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$, with $\pi_0 m$ proportion of the μ_i's being equal to 0 and the rest being equal to 0.2, and Σ is chosen as $\Sigma = \rho \mathbf{1_m} \mathbf{1'_m} + (1-\rho)\mathbf{I_m}$ to represent uniform pairwise dependence, as $\Sigma = \left(\rho^{|i-j|}\right)$ to represent AR(1) dependence, and as $\Sigma = I_{\frac{m}{s}} \otimes \left[\rho \mathbf{1_s} \mathbf{1'_s} + (1-\rho)\mathbf{I_s}\right]$ to represent block dependence (with block size $s = 5$). We then followed the steps (2)–(5) as in simulation study 1 with the step (1) in it replaced by the data generating scheme mentioned above for each $\rho \in \{0.1, 0.2, \ldots, 0.9\}$.

Figs. 5, 7 and 9 show the results for increasing strength of positive dependence, respectively, under equal correlation structure, AR(1) dependence, and block dependence based on O'Brien–Flemming alpha spending function. Figs. 6, 8 and 10, respectively, show the corresponding results based on Pocock alpha spending function. As seen from these figures, GSBH controls the FDR under all the positive dependence scenarios we have considered, as expected. However, the adaptive versions of the GSBH can fail to control the FDR when positive dependence among the tests, especially under uniform pairwise dependence structure, is strong. In terms of average power and FNR, the adaptive versions of GSBH can be more powerful than BH when correlation is high, although they may not control the FDR in those cases. Finally, with increasing correlation ρ, the group sequential procedures can achieve more sample saving, with the adaptive procedures being relatively more efficient.

Comparing the two alpha spending functions we used, we observe that using Pocock alpha spending function can generally lead to more sample saving; whereas, using O'Brien–Flemming alpha spending function can lead to tighter FDR control, higher average power and lower FNR. This can be explained by the fact that, when the samples are equally allocated to each stage, Pocock alpha spending function equally allocates the significance level to each stage, while O'Brien–Flemming

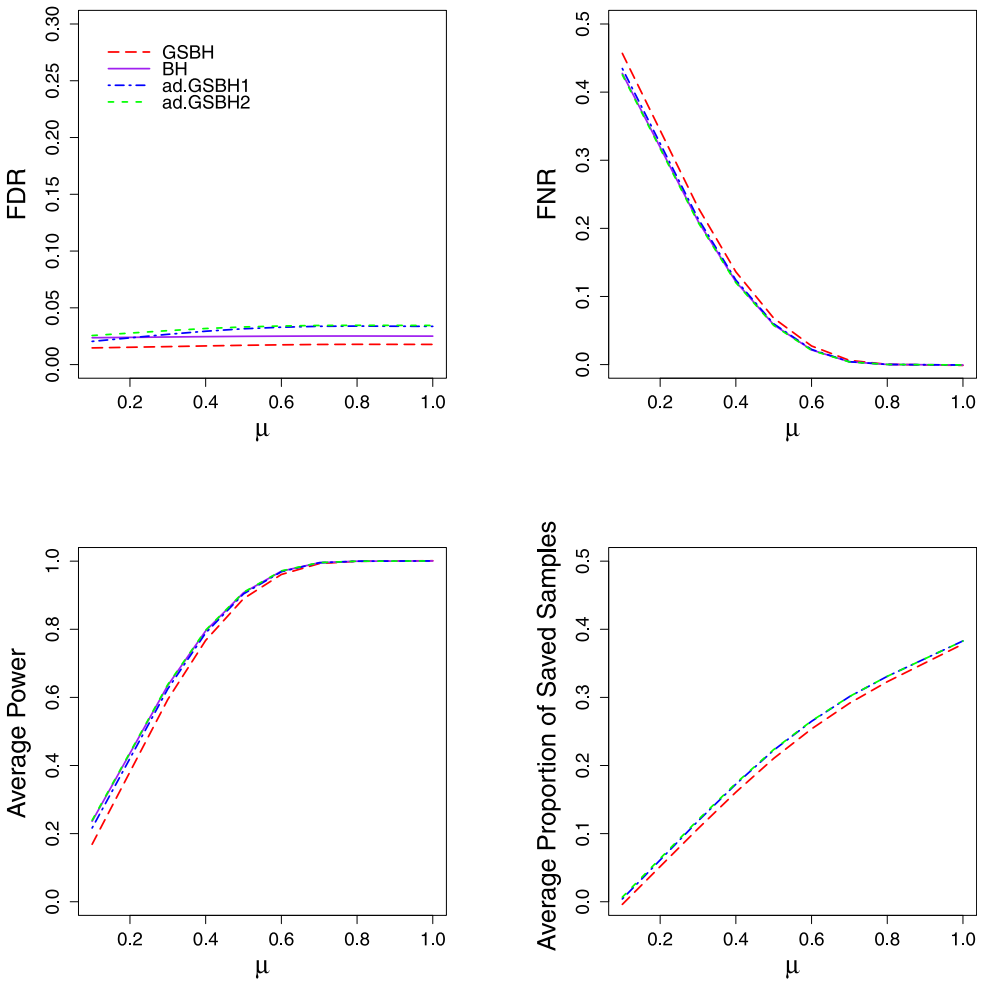

Fig. 3. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under independence with increasing magnitude of the common alternative mean μ using the O'Brien–Flemming alpha spending function.

alpha spending function spends more of the significance level at later stages, as seen from the definitions of these two alpha spending functions.

5. Real data application

We apply our proposed procedures to a real dataset from an experiment in Tian et al. (2003). In the original data, gene expression in multiple myeloma was measured in 36 patients without and 137 patients with bone lytic lesions. The original data set was generated with Affymetrix Human U95A chips, with each chip consisting of 12,625 probe sets. It was reanalyzed in Zehetmayer et al. (2008) and Sarkar et al. (2013). In this paper, we only use the gene expression measurements of 36 patients with bone lytic lesions and a control group of the same size without lesions. We consider a group sequential framework with three and four stages by allocating 12 and 9 subjects per group to each stage, respectively. In each stage, and for each of the $m = 12{,}625$ sets, we perform a two-sample T test to obtain a p-value to compare expression levels between the two groups based on all the data accumulated up to that stage. We apply our proposed GSBH and its adaptive versions, using both O'Brien–Flemming and Pocock alpha spending functions, and also the BH procedure to be used as a benchmark. The allocation of the total error rate α (=0.025) according to these two alpha spending functions is shown in Table 2 for the four stage design. Application of our proposed GSBH (or its adaptive versions) involves at each stage that of the BH procedure at the error level allocated to that stage by the chosen alpha spending function, to the p-values (or Q values) corresponding to the active hypotheses at that stage. The BH procedure makes 69 discoveries. The number of discoveries by the GSBH and its adaptive procedures are listed in Table 1. As seen from this table, the GSBH and its adaptive procedures based on O'Brien–Flemming alpha spending function lead to many more rejections than those using the Pocock

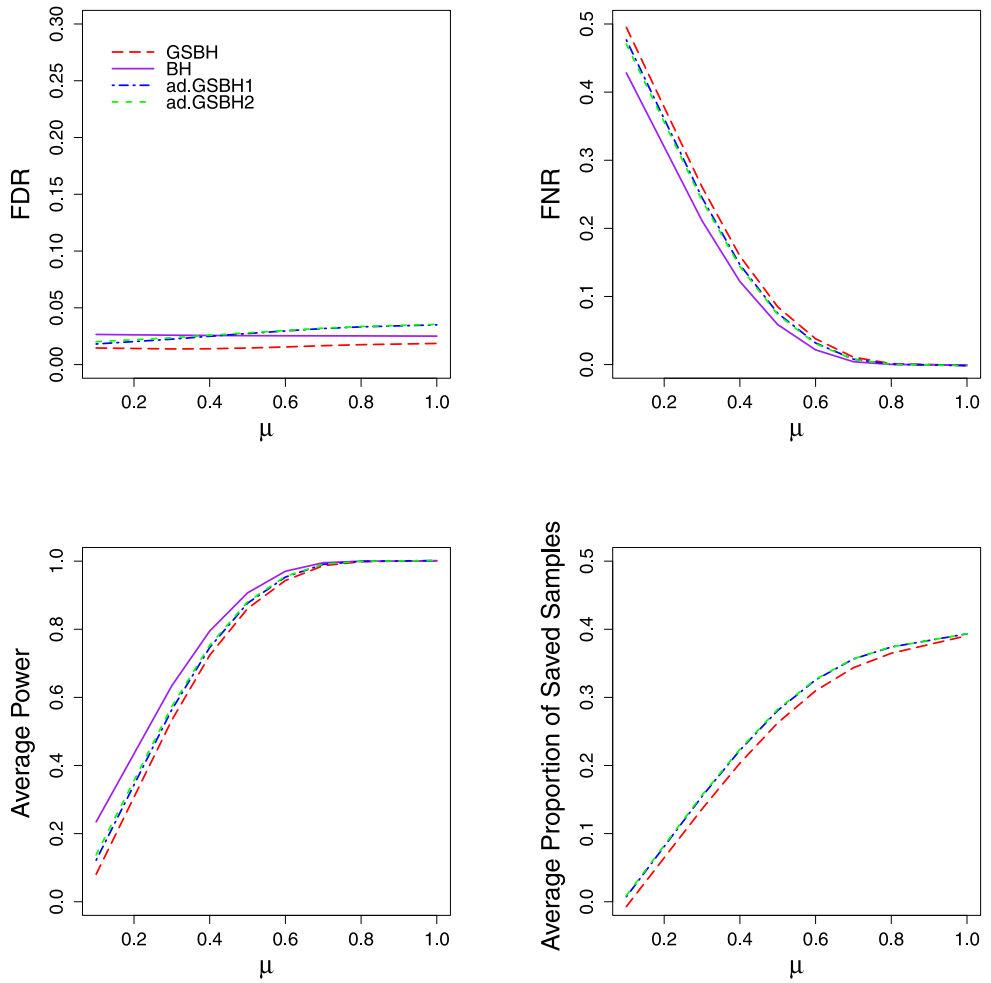

Fig. 4. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under independence with increasing magnitude of the common alternative mean μ using the Pocock alpha spending function.

Table 1
The number of discoveries out of 12,625 probe sets in a 3-stage and 4-stage group sequential designs by GSBH, ad.GSBH1, and ad.GSBH2 procedures at $\alpha = 0.025$, using O'Brien–Flemming and Pocock alpha spending functions.

	Pocock			O'Brien–Flemming		
	GSBH	ad.GSBH1	ad.GSBH2	GSBH	ad.GSBH1	ad.GSBH2
3-stage design						
Stage 1	6	6	6	0	0	0
Stage 2	7	8	8	6	6	6
Stage 3	10	10	10	49	60	61
Total discoveries	23	24	24	55	66	67
Proportion of saved samples (%)	0.050	0.053	0.053	0.016	0.016	0.016
4-stage design						
Stage 1	1	1	1	0	0	0
Stage 2	18	18	18	10	12	12
Stage 3	6	7	7	7	7	7
Stage 4	8	8	8	36	44	46
Total discoveries	33	34	34	53	63	65
Proportion of saved samples (%)	0.089	0.091	0.091	0.053	0.061	0.061

alpha spending function; whereas, more saving in sample size is achieved by the Pocock alpha spending function. When using O'Brien–Flemming alpha spending function, our ad.GSBH2 procedure is almost as powerful as the BH procedure, making 67 and 65 discoveries, respectively, in a 3- and 4-stage designs.

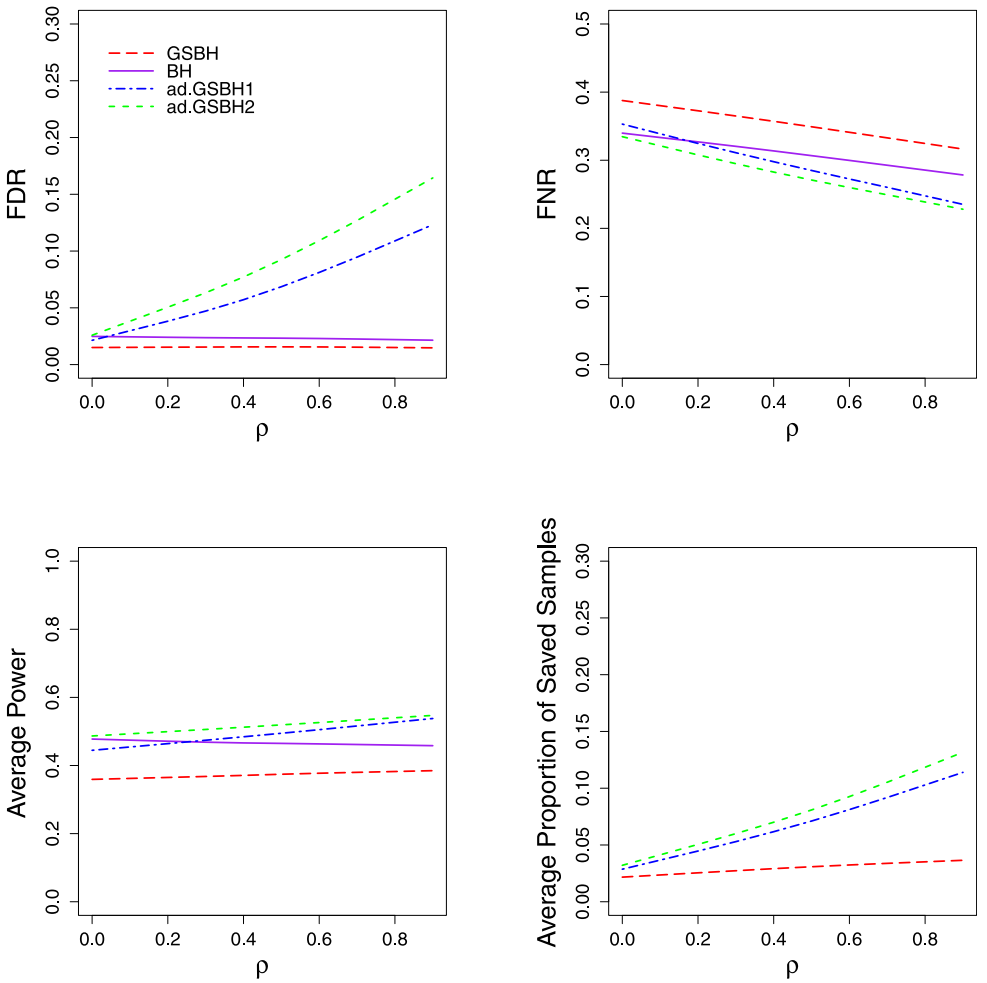

Fig. 5. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under uniform pairwise dependence with increasing correlation ρ using the O'Brien–Flemming alpha spending function.

Table 2
Allocation of the total error rate α (=0.025) by the O'Brien–Flemming and Pocock alpha spending functions for the four stage design with equal sample size allocation.

Stage	O'Brien–Flemming	Pocock
1	0.000007	0.008934
2	0.001525	0.015503
3	0.009649	0.020700
4	0.025000	0.025000

6. Concluding remarks

We have extended in this paper both the BH and its adaptive version incorporating a commonly used type of estimate of the number true nulls as well as the associated assumptions on the p-values, from single to multiple stages for testing multiple hypotheses controlling the FDR in a group sequential setting. The proposed procedures are quite flexible, easy-to-use and robust in the sense of controlling the FDR even if K is random or infinity, or if the trial is terminated due to some reasons. The use of error spending function allows one to achieve early rejections at either earlier or later stages of the design. Furthermore, simulation studies and real data application indicate that the choice of alpha spending function has a significant impact on the sample size saving and power of the proposed procedure. Depending on whether sample size reduction or rejection power is more critical, one can make choice of the appropriate error spending function to meet that goal.

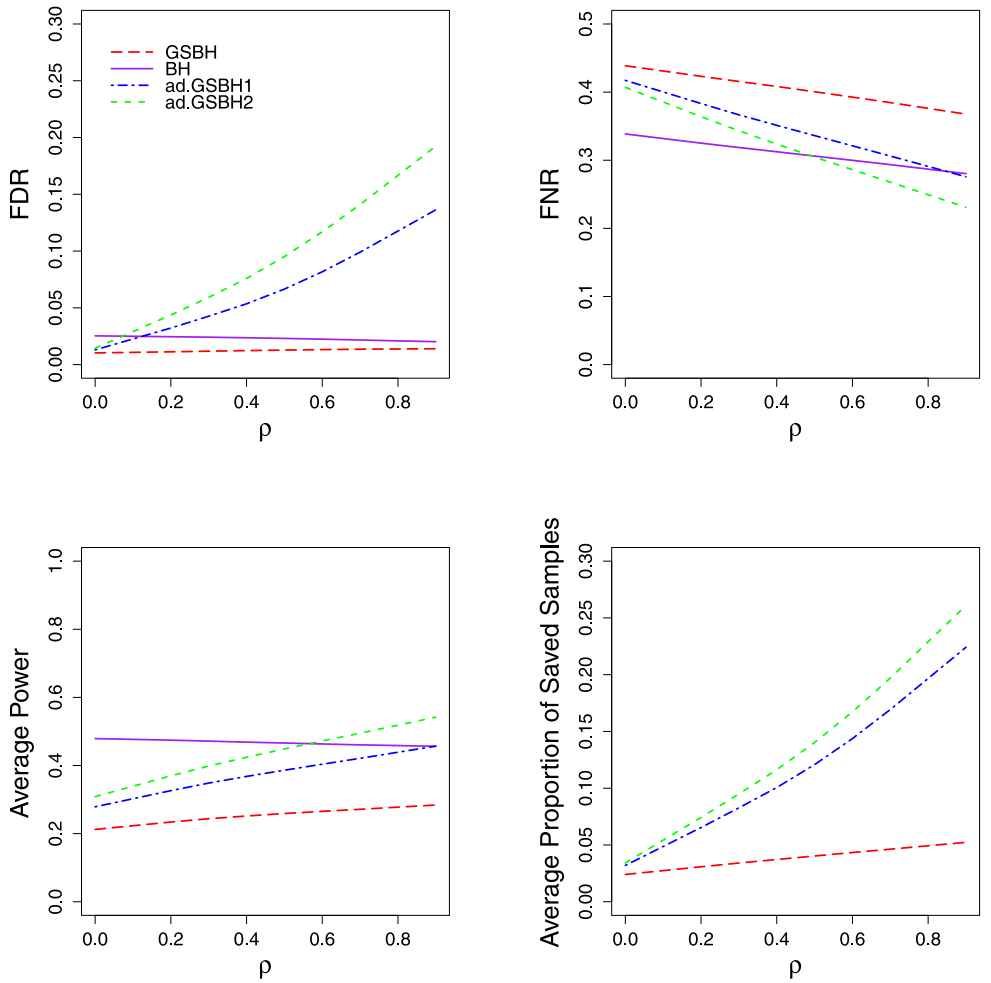

Fig. 6. Simulated FDR, FNR, average power, and average proportion of saved samples of the GSBH and its two adaptive versions. The results are calculated under uniform pairwise dependence with increasing correlation ρ using the Pocock alpha spending function.

It is important to note that our proposed FDR controlling methods in group sequential setting are different from those that exist in the literature (that are mentioned in Introduction), especially those in Benjamini and Yekutieli (2005), Zehetmayer et al. (2005) and Zehetmayer and Posch (2012). In those other methods, only early acceptance at interim stages is allowed, since in the gene expression or gene association studies they evaluated, π_0 is close to 1 and so early rejection may not be possible. In contrast, we focus on developing a generalization of the BH procedure that maintains the FDR control under the some positive dependence condition by assessing early rejection options at interim analyses and stopping each test once it has reached some significance level.

Allowing early acceptance for futility in a group sequential design can result in a reduction in sample size and total costs, especially in the context of multiple hypothesis testing. However, as noted in Victor and Hommel (2007), constructing appropriate futility boundaries when testing multiple hypotheses to control the FDR in a multi-stage group sequential setting can be very challenging and unfavourable futility boundaries may have serious consequence with respect to the performance of the study design. Method on constructing futility boundaries and adjusting the proposed generalization of the BH procedure to incorporate early acceptance like in Sarkar et al. (2013) should be further investigated, but we leave this as a future research project.

Some other questions, left unresolved here, are also worth addressing in future research. We have validated the dependence assumptions on the p-values only for normally distributed test statistics. It would be worthwhile to validate them for test statistics arising in other commonly encountered multiple testing situations. We believe that a wide class of adaptive GSBH methods containing the ones proposed here can be constructed based on other types of estimate of the number of true nulls following Sarkar (2008).

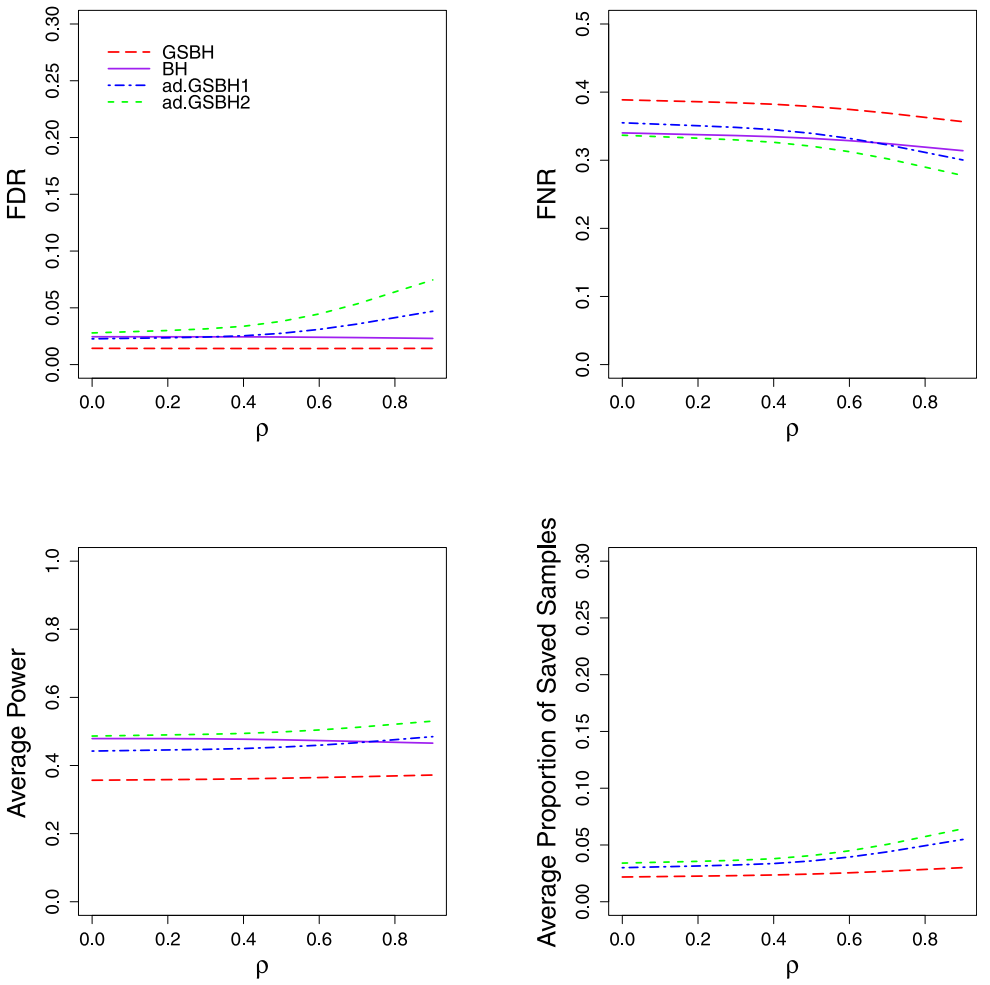

Fig. 7. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under AR(1) dependence with increasing correlation ρ using the O'Brien–Flemming alpha spending function.

Acknowledgements

The research of the first author is supported by the NSF Grant DMS-1309273 and the research of the fourth author is supported by the NSF Grant DMS-1309162. We thank a reviewer whose comments have improved the presentation of the paper.

Appendix

A.1. Proof of Theorem 3.1

Proof. Before proving the theorem, let us state the following two lemmas containing some results from Sarkar (2008) that will facilitate our proof.

Lemma A.1. *Given a set of test statistics $T_1, \ldots, T_m$ and their ordered values $T_{(1)} \le \cdots \le T_{(m)}$ associated with m hypotheses, let $R = \max\{1 \le i \le m : T_{(i)} \le c_i\}$ denote the number of rejections when a step-up test with critical values $c_1 \le \cdots \le c_m$ is applied to these hypotheses. Let $R^{(-i)}$ and $\tilde{R}^{(-i)}$ denote the numbers of rejection in the step-up tests with the critical values $c_2 \le \cdots \le c_m$ and $c_1 \le \cdots \le c_{m-1}$, respectively, applied to the $m-1$ hypotheses corresponding to $(T_1, \ldots, T_m) \setminus \{T_i\}$. Then, we have the following results:*

(i) $\mathbb{1}(T_i \le c_r, R = r) = \mathbb{1}(T_i \le c_r, R^{(-i)} = r-1)$, *for* $r = 1, \ldots, m$;
(ii) $\mathbb{1}(T_i > c_{r+1}, R = r) \le \mathbb{1}(T_i > c_{r+1}, \tilde{R}^{(-i)} = r)$, *for* $r = 0, \ldots, m-1$.

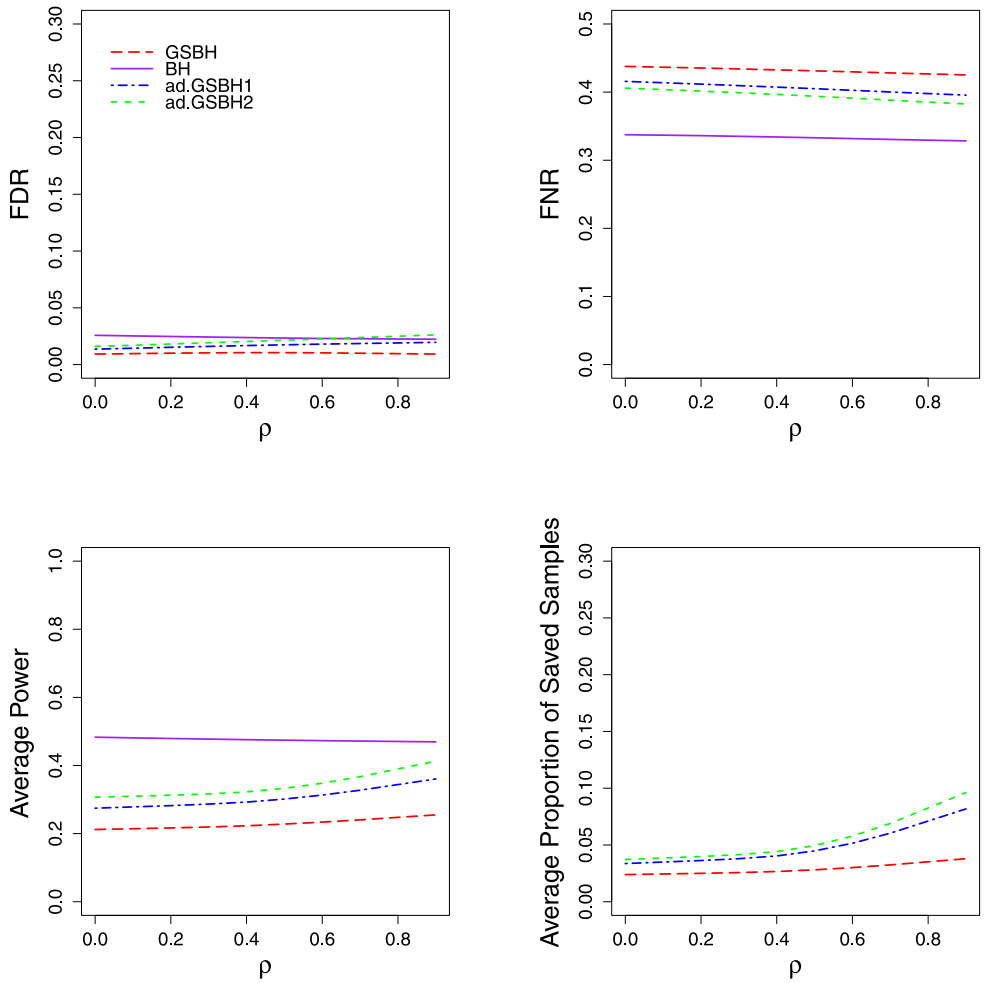

Fig. 8. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under AR(1) dependence with increasing correlation ρ using the Pocock alpha spending function.

Lemma A.2 (*Sarkar (2008, Lemma 4.2 (ii))*)**.** *Consider two random variables U and R, where $U \sim U(0, 1)$ and R is discrete defined on $\{1, \ldots, n\}$. Then, given an increasing set of constants $c(1) \leq \cdots \leq c(m)$ such that $c(r)/r$ is non-increasing in $r = 1, \ldots, m$, we have*

$$E\left\{\frac{\mathbb{1}(U \leq c(R))}{R}\right\} \leq c(1),$$

if U and R are positively dependent in the sense that $P(R \leq r \mid U \leq u)$ is non-decreasing in u.

The theorem is now proved using the above lemmas.

$$\begin{aligned} FDR &= E\left\{\frac{V_1 + V_2 + \cdots + V_K}{[(R_1 + R_2 + \cdots + R_K) \vee 1]}\right\} \\ &\leq E\left\{\frac{V_1}{R_1 \vee 1}\right\} + E\left\{\frac{V_2}{[(R_1 + R_2) \vee 1]}\right\} + \cdots + E\left\{\frac{V_K}{[(R_1 + R_2 + \cdots + R_K) \vee 1]}\right\}, \end{aligned}$$

where

$$E\left\{\frac{V_k}{[(R_1 + \cdots + R_k) \vee 1]}\right\} \leq \pi_0 \alpha_k, \tag{3}$$

which is already known in the literature for $k = 1$ (see, for example, Sarkar (2008) for a proof) and will be proved below for each $k = 2, \ldots, K$.

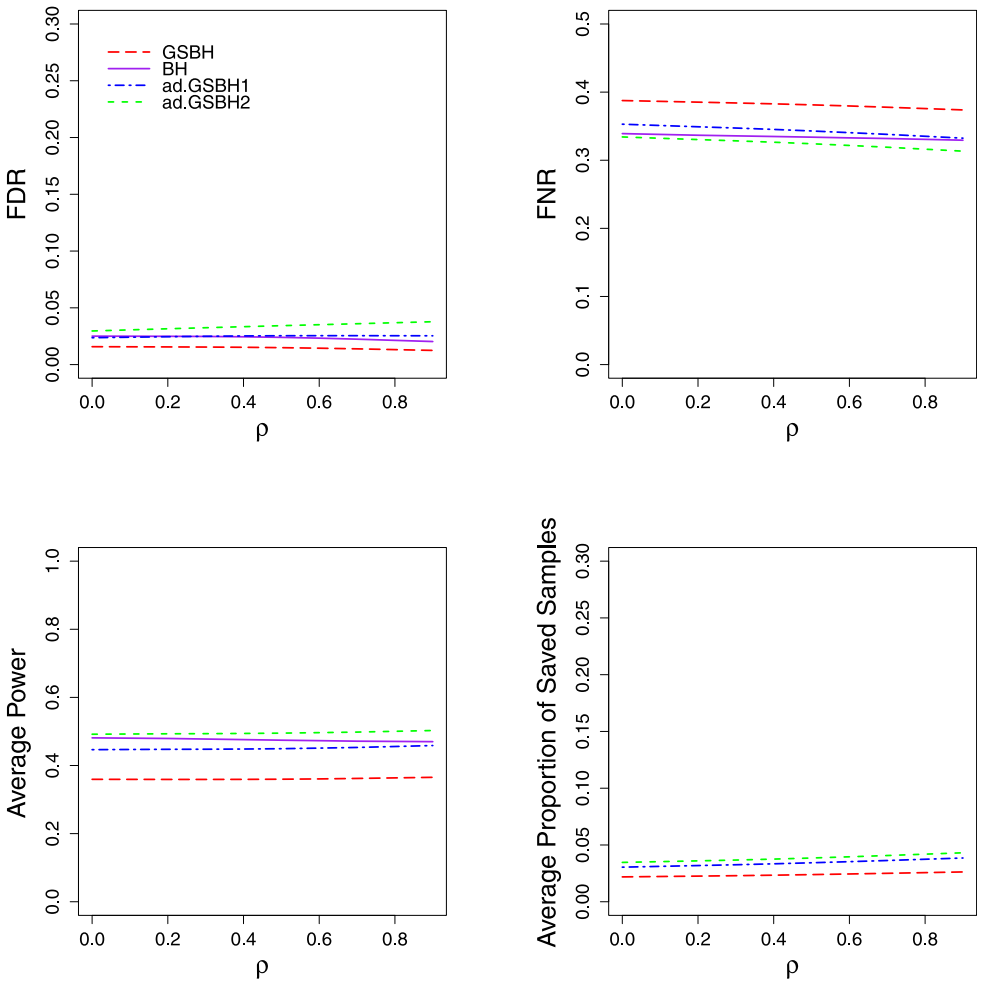

Fig. 9. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under block dependence with increasing correlation ρ using the O'Brien–Flemming alpha spending function.

The GSBH operates at the kth stage as a step-up procedure with the critical constants $\lambda^{(k)}_{\sum_{j=1}^{k-1}R_j+i}$, $i=1,\ldots,|I_k|$ applied to the set of hypotheses $\{H_i, i\in I_k\}$ not rejected in the previous $k-1$ stages. Therefore, we see that

$$E\left\{\frac{V_k}{[(R_1+\cdots+R_k)\vee 1]}\right\}=\sum_{i\in I_0}E\left\{\frac{\mathbb{1}\left(P_i^{(k)}\le\lambda^{(k)}_{\sum_{j=1}^{k}R_j},\, i\in I_k\right)}{\left((\sum_{j=1}^{k}R_j)\vee 1\right)}\right\},$$

where

$$\begin{aligned}&\left\{P_i^{(k)}\le\lambda^{(k)}_{\sum_{j=1}^{k}R_j},\, i\in I_k\right\}\\ &=\left\{P_i^{(1)}>\lambda^{(1)}_{R_1+1},\, P_i^{(2)}>\lambda^{(2)}_{R_1+R_2+1},\ldots,P_i^{(k-1)}>\lambda^{(k-1)}_{\sum_{j=1}^{k-1}R_j+1},\, P_i^{(k)}\le\lambda^{(k)}_{\sum_{j=1}^{k}R_j}\right\}.\end{aligned}\tag{4}$$

Now, since $R_j<m$ (otherwise, the GSBH would have stopped before the jth stage), for $j=1,\ldots,k-1$, and $R_k>0$ (otherwise, the left-hand side of (3) would be zero, making the inequality trivially true), we can use Lemma A.1(i) and (ii) to express the expectation on the right-hand side of (4) more explicitly in terms of the p-values corresponding to H_i, for each $i\in I_0$, across the first k stages and the rest as follows:

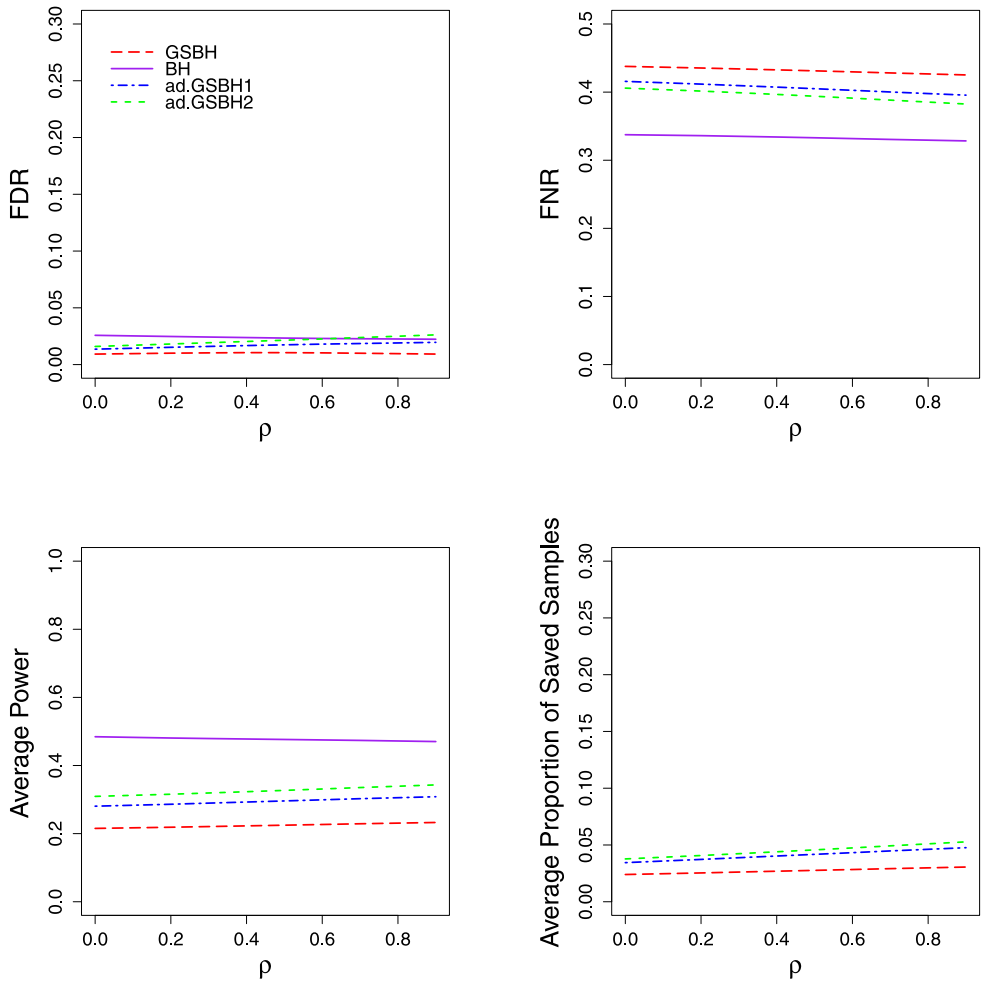

Fig. 10. Simulated FDR, FNR, average power, and average proportion of saved samples for GSBH and its two adaptive versions. The results are calculated under block dependence with increasing correlation ρ using the Pocock alpha spending function.

$$
\begin{aligned}
& E\left\{\frac{\mathbb{1}\left(P_i^{(k)} \le \lambda^{(k)}_{\sum_{j=1}^{k} R_j}, i \in I_k\right)}{(\sum_{j=1}^{k} R_j) \vee 1}\right\} \\
\le\ & E\left\{\frac{\mathbb{1}\left(P_i^{(1)} > \lambda^{(1)}_{\tilde{R}_1^{(-i)}+1}, \ldots, P_i^{(k-1)} > \lambda^{(k-1)}_{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+1}, P_i^{(k)} \le \lambda^{(k)}_{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k{}^{(-i)}+1}\right)}{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)} + R_k{}^{(-i)} + 1}\right\} \\
\le\ & E\left\{\frac{\mathbb{1}\left(P_i^{(k)} \le [\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)} + R_k{}^{(-i)} + 1]\alpha_k/m\right)}{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)} + R_k{}^{(-i)} + 1}\right\}, \qquad (5)
\end{aligned}
$$

where $\tilde{R}_j^{(-i)}$ is the number of rejection in the step-up test with the first $|I_j| - 1$ critical values at the jth stage applied to the $|I_j| - 1$ hypotheses corresponding to $(P_1^{(j)}, \ldots, P_{|I_j|}^{(j)}) \setminus \{P_i^{(j)}\}$, for $j = 1, \ldots, k-1$, and $R_k^{(-i)}$ is the number of rejection in the step-up test with the last $|I_k| - 1$ critical values at the kth stage, applied to the $|I_k| - 1$ hypotheses corresponding to $(P_1^{(k)}, \ldots, P_m^{(k)}) \setminus \{P_i^{(k)}\}$. Since $\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)} + R_k{}^{(-i)} + 1$ is a decreasing function of $(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(k)})$, we can apply Lemma A.2 to the last expectation in (5) to claim that it is less than or equal to α_k/m under Assumptions 1 and 2, and thus complete our proof of the theorem. ■

A.2. Proof of Theorem 3.2

Proof. For notational convenience, we will continue to use the same notations as we have used for the GSBH to denote the critical values and the numbers of rejection, even though they are now associated with the Q_i's. Going through the same arguments as used to get up to (5) using Lemma A.1(i) and (ii) to prove Theorem 3.1, we first note that

$$
\begin{aligned}
& E\left\{\frac{V_k}{[(R_1+\cdots+R_k)\vee 1]}\right\} \\
&= \sum_{i\in I_0} E\left\{\frac{\mathbb{1}\left(Q_i^{(k)} \le \lambda^{(k)}_{\sum_{j=1}^k R_j}, i\in I_k\right)}{\left(\sum_{j=1}^k R_j\right)\vee 1}\right\}, \\
&\le \sum_{i\in I_0} E\left\{\frac{\mathbb{1}\left(P_i^{(k)} \le \frac{m(1-\eta)}{m-\sum_{i'\in I_1}\mathbb{1}(P_{i'}^{(1)}\le\eta)+1}\lambda^{(k)}_{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1}\right)}{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1}\right\} \\
&= \sum_{i\in I_0} E\left\{\frac{\mathbb{1}\left(P_i^{(k)} \le \frac{(1-\eta)\alpha_k[\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1]}{m-\sum_{i'\in I_1}\mathbb{1}(P_{i'}^{(1)}\le\eta)+1}\right)}{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1}\right\} \\
&\le \sum_{i\in I_0} E\left\{\frac{\mathbb{1}\left(P_i^{(k)} \le \frac{(1-\eta)\alpha_k[\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1]}{m-\sum_{i'\in I_1\setminus\{i\}}\mathbb{1}(P_{i'}^{(1)}\le\eta)}\right)}{\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1}\right\}. \qquad (6)
\end{aligned}
$$

Now, under independence of the p-values across hypotheses we note that (i) $P_i^{(k)}$ is independent of $\sum_{i'\in I_1\setminus\{i\}}\mathbb{1}(P_{i'}{}^{(1)}\le\eta)$, and (ii) conditionally given the p-values associated with the hypotheses other than H_i, while $\sum_{i'\in I_1\setminus\{i\}}\mathbb{1}(P_{i'}^{(1)}\le\eta)$ is fixed, $\sum_{j=1}^{k-1}\tilde{R}_j^{(-i)}+R_k^{(-i)}+1$ is a stochastically decreasing function of the p-values associated with H_i arising in the first $k-1$ stages. Therefore, by taking the expectation in (6) by conditionally fixing the p-values associated with the hypotheses other than H_i, for each $i\in I_0$, we get the following inequality under Assumption 2a and by applying Lemma A.2:

$$
E\left\{\frac{V_k}{[(R_1+\cdots+R_k)\vee 1]}\right\} \le \alpha_k \sum_{i\in I_0} E\left\{\frac{1-\eta}{m-\sum_{i'\in I_1\setminus\{i\}}\mathbb{1}(P_{i'}{}^{(1)}\le\eta)}\right\}, \qquad (7)
$$

which is less than or equal to α_k (see, for instance, Sarkar 2008), as desired. This proves the theorem. ■

References

Bartroff, J., Song, J., 2013. Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. arXiv preprint. arXiv:1311.3350.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Benjamini, Y., Hochberg, Y., 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Statist. 25 (1), 60–83.

Benjamini, Y., Krieger, A.M., Yekutieli, D., 2006. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93 (3), 491–507.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29 (4), 1165–1188.

Benjamini, Y., Yekutieli, D., 2005. Quantitative trait loci analysis using the false discovery rate. Genetics 171 (2), 783–790.

Blanchard, G., Roquain, É., 2009. Adaptive false discovery rate control under independence and dependence. J. Mach. Learn. Res. 10, 2837–2871.

Block, H.W., Savits, T.H., Shaked, M., 1985. A concept of negative dependence using stochastic ordering. Statist. Probab. Lett. 3 (2), 81–86.

Brannath, W., Posch, M., Bauer, P., 2002. Recursive combination tests. J. Amer. Statist. Assoc. 97 (457), 236–244.

Finner, H., Dickhaus, T., Roters, M., 2009. On the false discovery rate and an asymptotically optimal rejection curve. Ann. Statist. 37 (2), 596–618.

Finner, H., Roters, M., 2001. On the false discovery rate and expected type I errors. Biom. J. 43 (8), 985–1005.

Gavrilov, Y., Benjamini, Y., Sarkar, S.K., 2009. An adaptive step-down procedure with proven FDR control under independence. Ann. Statist. 37 (2), 619–629.

He, L., Sarkar, S.K., 2013. On improving some adaptive BH procedures controlling the FDR under dependence. Electron. J. Stat. 7, 2683–2701.

Karlin, S., Rinott, Y., 1980. Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. J. Multivariate Anal. 10 (4), 467–498.

Lan, K.G., DeMets, D.L., 1983. Discrete sequential boundaries for clinical trials. Biometrika 70 (3), 659–663.

Malek, A., Katariya, S., Chow, Y., Ghavamzadeh, M., 2017. Sequential multiple hypothesis testing with type I error control. In: Artificial Intelligence and Statistics. pp. 1468–1476.

O'Brien, P.C., Fleming, T.R., 1979. A multiple testing procedure for clinical trials. Biometrics 35 (3), 549–556.

Pocock, S.J., 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika 64 (2), 191–199.

Proschan, M.A., Lan, K.G., Wittes, J.T., 2006. Statistical Monitoring of Clinical Trials: A Unified Approach. Springer Science & Business Media.
Sarkar, S.K., 1998. Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. Ann. Statist. 26 (2), 494–504.
Sarkar, S.K., 2002. Some results on false discovery rate in stepwise multiple testing procedures. Ann. Statist. 30 (1), 239–257.
Sarkar, S.K., 2008. On methods controlling the false discovery rate. Sankhyā 70 (2), 135–168.
Sarkar, S.K., Chang, C.-K., 1997. The Simes method for multiple hypothesis testing with positively dependent test statistics. J. Amer. Statist. Assoc. 92 (440), 1601–1608.
Sarkar, S.K., Chen, J., Guo, W., 2013. Multiple testing in a two-stage adaptive design with combination tests controlling FDR. J. Amer. Statist. Assoc. 108 (504), 1385–1401.
Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3), 479–498.
Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66 (1), 187–205.
Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., Shaughnessy Jr., J.D., 2003. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N. Engl. J. Med. 349 (26), 2483–2494.
Victor, A., Hommel, G., 2007. Combining adaptive designs with control of the false discovery rate–a generalized definition for a global p-value. Biom. J. 49 (1), 94–106.
Zehetmayer, S., Bauer, P., Posch, M., 2005. Two-stage designs for experiments with a large number of hypotheses. Bioinformatics 21 (19), 3771–3777.
Zehetmayer, S., Bauer, P., Posch, M., 2008. Optimized multi-stage designs controlling the false discovery or the family-wise error rate. Stat. Med. 27 (21), 4145–4160.
Zehetmayer, S., Graf, A.C., Posch, M., 2015. Sample size reassessment for a two-stage design controlling the false discovery rate. Stat. Appl. Genet. Mol. Biol. 14 (5), 429–442.
Zehetmayer, S., Posch, M., 2012. False discovery rate control in two-stage designs. BMC Bioinformatics 13 (1), 81.