JOURNAL OF COMPUTATIONAL INNOVATION AND ANALYTICS
http://e-journal.uum.edu.my/index.php/jcia

How to cite this article:
Ramos, M. L. F., & Park, D. H. (2022). Power Boosting for Ordered Multiple Hypotheses with Application to Genome-Wide Association Studies. *Journal of Computational Innovation and Analytics, 1*(1), 1-17. https://doi.org/10.32890/jcia2022.1.1.1

POWER BOOSTING FOR ORDERED MULTIPLE HYPOTHESES WITH APPLICATION TO GENOME-WIDE ASSOCIATION STUDIES

[1]Mark Louie F. Ramos & [2]DoHwan Park
[1&2] Department of Mathematics and Statistics
University of Maryland Baltimore County, MD, United States
[1]Department of Mathematics and Physics
University of Santo Tomas, MLA, Philippines

[1]Corresponding author: markram1@umbc.edu

Received: 4/8/2021 Revised: 15/8/2021 Accepted: 1/12/2021 Published: 27/1/2022

ABSTRACT

A method for addressing the multiplicity problem is proposed in the setting where the hypotheses test sites may be arranged in some order based on a notion of proximity, such as SNPs of a chromosome in genetic association studies. It is shown that this method is able to control family-wise error rate in the weak sense and numerical evidence shows that this method controls false discovery rate in the strong sense under sparsity. The method is applied to some genome-wide association studies data with asthma and it is argued that this Power Boosting method may be combined with existing error-rate controlling methods in order to improve true positive rates at controllable and possibly negligible cost to the nominal level of error-rate control.

Keywords: Multiple Testing, False Discovery Rate, Family-Wise Error Rate.

INTRODUCTION

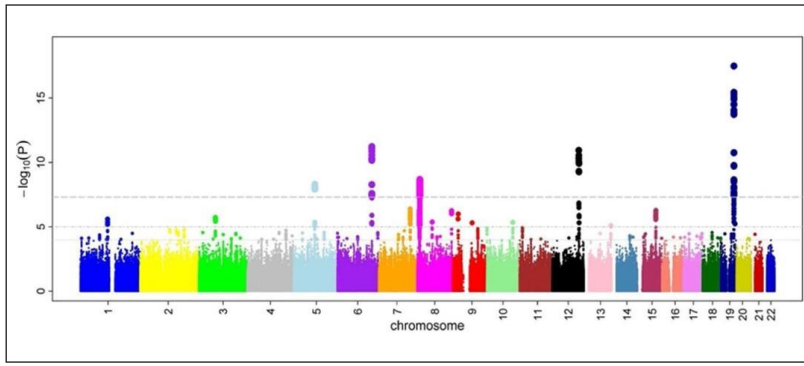
In multiple testing, maintaining a practicable balance between type 1 error rate control and statistical power is a common issue (Aslam & Albassam, 2020; Verhoeven et al., 2005; Yang et al., 2005). An individual test that is calibrated to have sufficient power for detecting a minimum effect size cannot be relied upon to maintain that same power if the common corrections for multiplicity, such as the Bonferroni correction for family-wise error rate (FWER) or the Benjamini-Hochberg procedure for false discovery rate (FDR) (Benjamini & Hochberg, 1995), are to be applied to the results of several, individually powerful enough tests. This is because these methods maintain control over their respective type 1 error rates by imposing penalties on the individual significance levels based on the nominal significance level. For example, if there are N hypotheses and it is desired to maintain level α control over type 1 error, then application of the Bonferroni correction would mean that an individual null hypothesis can only be rejected if its corresponding p-value is less than α/N. For large N, it can become practically impossible to reject any hypothesis despite there being clearly strong evidence for rejection when considering a single hypothesis.

However, when some prior assumptions can be made about the context in which the multiplicity problem exists, this can be used in order to develop customized methods that may offer better power. For example, if it can be assumed that there is no negative dependency among the test statistics, then Sidak's procedure (Sidak, 1967) offers a more powerful alternative to Bonferroni. This study is interested in the setting where the test sites are arranged in some meaningful way, such as proximity. Furthermore, the setting of interest is such that there is no dependence among the test statistics of adjacent test sites coming from the null distribution, but positive dependence exists among test statistics of adjacent test sites coming from the alternative distribution. As example of this is the study of single nucleotide polymorphisms (SNPs), where it is of interest to determine which of typically thousands of variants in each chromosome is associated with some variable of interest. The chromosomes do not have a meaningful ordering, but the variants within each chromosome do, and it is known that association detected in a variant in one position makes it more likely that a similar association can be detected in neighboring positions. Figure 1, which was taken from a study on the association

between SNPs and microcirculation (Ikram, Xueling, Jensen, Cotch, & Hewitt, 2010), illustrates this situation. In this study, multiplicity correction was done using Bonferroni adjustment for 1 million tests, although the study itself conducted 2.2 million tests, thus resulting in a significance level of 0.11 instead of 0.05. This study considers if this or similar situations can be improved by considering that significant test sites tend to be in close proximity with one another.

Figure 1

Association between SNPs in each chromosome and microcirculation. SNPs of each chromosome are arranged in a meaningful order. The Y-axis are log-transformed p-values between SNPs in each position and microcirculation. Statistically significant associations tend to come from adjacent positions. The figure is taken from Ikram et al. (2010)



OVERVIEW OF MULTIPLICITY CORRECTION APPROACHES

The multiplicity problem refers to the issue that arises when a statistical test is simultaneously applied to numerous test sites (Miller, 1981). Basing inference on the outcomes of these individual tests inflates the probability of type 1 error, making the resulting detections unreliable. In order two address this, concepts of type 1 error for multiple hypotheses were developed. The most prominent among these are the family-wise error rate and the false discovery rate(Dmitrienko et al., 2010). These are briefly reviewed as follows.

For N total hypotheses such that there are a total of R rejections and V false positives, the family-wise error rate is defined as

$$FWER = Pr(V > 0) \tag{1}$$

On the other hand, the false discovery rate is defined in Benjamini and Hochberg (1995) as

$$FDR = E(\frac{V}{R} \mid R > 0)Pr(R > 0) \tag{2}$$

A method that controls FWER likewise controls FDR. In addition, suppose there are two methods, Method 1 and Method 2 with corresponding false discovery rates FDR_1 and FDR_2, then combining the two methods such that a hypothesis is rejected if it can be rejected in either method will have a false discovery rate that is less than or equal to the sum of the individual methods' FDRs. That is,

$$FDRcombined \leq FDR1 + FDR2 \tag{3}$$

The same relationship exists for FWER.

Both classical and modern procedures that control either error rate typically impose some penalty on the individual level test sites (Qu et al., 2010; Kirsch et al., 2012; Noble, 2009; Efron et al., 2001). However, not all multiplicity correction procedures are designed this way. Other procedures are designed to avoid imposing penalties on the nominal significance level and instead rely upon some predefined structure of the hypotheses in order to maintain type 1 error rate control. One example of this is the fixed sequence procedure, in which the hypotheses are pre-ordered in some fashion and tested sequentially at the nominal level α until the first null hypothesis that is retained (Dmitrienko et al., 2010). It can be shown that this procedure is able to preserve all of the significance levels for the first test, but comes with the drawback of having zero power for the next hypotheses once one of them is retained. Nonetheless, this example shows that type 1 error rate control can be achieved not just by decreasing the individual test level α. This is the primary idea on which this proposed method is hinged.

As shown in Equation 3, it is possible to combine two FDR controlling approaches. However, if both are controlling FDR by adjusting the

individual level α, it is not meaningful to combine them since the combined procedure will just reduce to the more conservative method between the two. Furthermore, combining two methods, in general, presents the challenge of dividing the nominal level α between them. Since the nominal level α must remain conserved between the two methods, a major drawback of combining two methods is that splitting the α level between the two tests would result in weakening the power of both tests. This issue is addressed by making the partition in such a way that one test will receive most of the nominal α level while the other will receive a much smaller, practically negligible portion. For example, the main test can have an alpha level of $\alpha_1 = 0.04999$, while the power-boost method combined with it will have $\alpha_2 = 0.0001$. In this way, the resulting test will still have a nominal level of $\alpha = 0.05$ at negligible cost to the power of the main test.

PROPOSED METHOD

Let H_1, H_2, ... H_N be ordered hypotheses based on some idea of nearness. Such that H_i is nearest to H_{i+1} and H_{i-1}. Define a block of k hypotheses as H_j, H_{j+1}, ... H_{j+k-1}. For some pre-determined k that depends only on N, reject each block of k or more hypotheses when each of the hypotheses in the block can be rejected at the nominal level α. This means that each hypothesis is tested at the nominal level α instead of a multiplicity corrected significance level such as α/N for Bonferroni correction. Instead, the multiplicity problem is addressed by imposing the restriction that the hypotheses with p values less than α must be together in a sufficiently sized block in order for them to be rejected.

The rationale for this method is grounded on the reality that in many multiple testing scenarios where the test sites can be so ordered, observation of individually significant hypotheses that are clustered together is typically considered as stronger practical evidence than if the same number of individually significant hypotheses are scattered apart. This is the basis of cluster inference (Lee & Steigerwald, 2020), but this method differs from cluster inference approaches in that only k, the minimum number of hypotheses that should be significant at the nominal level, needs to be pre-determined, and this is done based only on the number of test sites N. That is, for any N, it is desired to either compute or estimate the error rate for each k in order to select the minimum k that is suitable for the nominal level α.

CONTROL OVER FWER IN THE WEAK SENSE

Control over FWER in the weak sense means that FWER is controlled in the setting where all of the hypotheses are true negatives.

Proposition 1: For any $N > 3$ and any nominal level $\alpha < 0.5$, there exists a k, $1 < k < N$ such that FWER is controlled in the weak sense.

Proof of Proposition 1

Consider the more relaxed method that rejects k or more hypotheses if each of them can be rejected at the nominal level α. This method is actually one of the earliest FWER controlling methods (Wilkinson, 1951). Clearly, it is enough to show that FWER is controlled in the weak sense for this more relaxed method to prove that FWER is also controlled for the proposed method in the weak sense. The binomial expansion of $[\alpha+(1-\alpha)]^N$ shows the sum of the probability mass function for the number of false positives among N hypotheses in the weak condition.

$$1 = [\alpha + (1-\alpha)]^N = \sum_{u=0}^{N} {}_nC_u\, \alpha^u (1-\alpha)^{N-u} \tag{4}$$

Since the method either rejects k or more null hypotheses together or none at all, then $FWER = Pr(V > 0) = Pr(V \geq k)$ and the family-wise error rate for a specific $k = k'$ is given by

$$FWER_{k'} = \sum_{u=k'}^{N} {}_NC_u \alpha^u (1-\alpha)^{N-u} \tag{5}$$

Thus, for any $\alpha < 0.5$, the smallest k can always be chosen such that

$$\sum_{r=0}^{k'-1} {}_{k'-1}C_r \alpha^r (1-\alpha)^{k'-1-r} > 1-\alpha \tag{6}$$

And so,

$$1 = \sum_{u=k'}^{N} {}_NC_u \alpha^u (1-\alpha)^{N-u} + \sum_{r=0}^{k'-1} {}_{k'-1}C_r \alpha^r (1-\alpha)^{k'-1-r} \tag{7}$$

$$FWER_{k'} = \sum_{u=k'}^{N} {}_{N}C_{u}\alpha^{u}(1-\alpha)^{N-u} = 1 - \sum_{r=0}^{k'-1} {}_{k'-1}C_{r}\alpha^{r}(1-\alpha)^{k'-1-r} < \alpha \quad (8)$$

Thus, FWER is also controlled for the proposed method.

Growth of k relative to N

Having established that there is a k that controls FWER for any N in the weak sense, it is next demonstrated that the k needed for any given N is reasonably small relative to N. This is done numerically. An algorithm was constructed to compute the exact value of *FWER* at given N, k, and α by extracting the entire sample space. For example, at $N = 2$, $k = 2$, the sample space is $\{WW, RR, WR, RW\}$, where R means the hypothesis for the test site is rejected, and W means that it is retained. In this example, FWER is $Pr(RR) = \alpha^2$, since this is the only situation where a false rejection will be made using the power boosting method. This also illustrates that in the power boosting method, the probability of making at least one false rejection (FWER) is the same as the probability of making at least k false rejections.
However, finding the exact FWER is computationally intensive at large N, so another algorithm was constructed to simulate multiple test outcomes under the weak setting and use this to estimate FWER. The steps taken to for this estimation is provided as follows.

1. Generate N test sites. The value of each test site is either 1 (rejection) with probability α or 0 (failure to reject) with probability $1 - \alpha$.
2. Check if there are k or more adjacent test sites that each have a value of 1. If this is true, then at least one false rejection has occurred. If so, add this to a counter variable R.
3. Repeat Steps 1 and 2 10000 times. The estimate of FWER is $R/10000$

Table 1 shows the computational results for selected N, k, and α. Some exact computations are also shown to demonstrate that the estimated FWER does not differ much from the exact computation. As seen in Table 1, k does grow at a much slower rate than N. When testing 100,000 hypotheses simultaneously, the number of adjacent tests that need to be significant at $\alpha = 0.05$ is only 5. For a million hypotheses, $k = 6$ is sufficient. Also, for smaller values of α, the difference between

the orders of N and k is increased such that for a million hypotheses and $\alpha = 0.01$, $k = 4$ becomes sufficient. This shows how the method may be combined with another FWER controlling method. For example, with 5000 hypotheses, k can be set to 4, having an FWER of 0.0001. Then, the other method can be calibrated to have an FWER of 0.0099, such that overall FWER is still controlled at a 0.01 level, and the penalty for using the power boosting method as part of the combination is practically negligible. However, this application depends on the ability of the method to control FWER in the strong sense as well as the weak sense. If FWER is not controlled in the strong sense, then a less stringent error rate such as the FDR may be considered, where it must then be shown that the method controls this error rate in the strong sense.

Table 1

Estimation of FWER under the weak sense

$\alpha = 0.05$			
N	*k*	Estimated FWER	Exact FWER
3	2	0.0047	0.0049
20	2	0.0446	0.0445
400	3	0.0487	
5000	4	0.0290	
100000	5	0.0300	
1000000	6	0.0100	
$\alpha = 0.01$			
N	*k*	Estimated FWER	Exact FWER
3	2	0.0002	0.0002
20	2	0.0020	0.0019
400	3	0.0009	
5000	4	0.0001	
100000	4	0.0000	
1000000	4	0.0070	

Failure to control FWER in the strong sense

Control over FWER in the strong sense means that FWER is controlled in every possible configuration of true positives and true

negatives. Unfortunately, the power boosting method fails to control FWER in the strong sense. To see this, one just needs to consider the setting where the first k hypotheses are true positives, and the rest of the hypotheses from H_{k+1} to H_N are true negatives. In this case, if it is assumed that the tests are powerful enough to detect the true positives without any multiplicity correction, then the first k hypotheses will always be rejected. However, since the method rejects for any block of k or more hypotheses that can be rejected at the nominal α, then H_{k+1} will be falsely rejected at a rate of α, which means that the FWER in this situation will at least be α.

A consequence of this is that the power boosting method is not suitable to combine with other FWER controlling approaches to control FWER. Nonetheless, since the method controls FWER in the weak sense, then FDR is also controlled in the weak sense. Furthermore, the FWER for a specific k at a given N is the same as the FDR under the weak sense, and so Table 1 would be identical for FDR. Thus, what remains to be determined is whether or not there exists a k for every N that controls FDR in the strong sense.

CONTROL OVER FDR IN THE STRONG SENSE

The practicability of using the power boosting method to augment existing FDR controlling approaches depends on the extent to which it can control FDR in the strong sense.

Maximum FDR

A direct way of assessing control over FDR in the strong sense is to identify the configuration of true positives and true negatives for which FDR is the largest. Obviously, if FDR is controlled (less than the preset α) at the configuration where it is largest, then it is controlled in every other possible configuration. For any configuration of true positives and true negatives among N hypotheses, the FDR may be calculated directly. For example, suppose $N = 2$ and let $k = 2$. Let X be a true negative and let Y be a true positive. Then the possible configurations are listed as $\{XX, YY, XY, YX\}$. At a nominal level α and assuming that the individual-level hypothesis test has sufficient power

to always detect a true positive, the FDR at $k = 2$ for configuration XX is α^2 and the FDR for configurationXY or YX is $\frac{\alpha}{2}$. Since α is always selected to be small, then the maximum FDR is obtained at configuration XY (or YX). Thus in this example, it is shown that FDR is controlled in the strong sense. This illustrates that for any N and any k, it is theoretically possible to find the maximum FDR by calculating the FDR for each configuration of true positives and true negatives. However, this approach will quickly become intractable for larger N. Also, in many situations, configurations, where the majority of the test sites are true positives, are unrealistic. Thus, it is reasonable to have some assumption of sparsity, where most of the test sites are true negatives, and only a small proportion are true positives. This assumption is commonly used in the development of procedures for multiple testing across various contexts (Ghosh & Chakraborty, 2017; Bogdan et al., 2011; Frommlet & Bogdan, 2013).

Assumption on the Maximum number of True Positives

Consider the assumption that at most only a certain proportion of the test sites can be true positives. Let M be the maximum number of test sites that are true positives such that $M << N$.
Proposition 2: Let $M << N$. If there exists a $k_1 \leq M$ for which FDR is controlled in the weak sense then there exists another k_2 such that $k_1 \leq k_2 \leq M$ for which FDR is also controlled in the strong sense.
Evidence for Proposition 2 is presented numerically as follows.

1. Generate N test sites with M sites from the alternative distribution and $N - M$ sites from the null distribution. For those from the null distribution, the value of each test site is either 1(rejection) with probability α or 0(failure to reject) with probability $1 - \alpha$. For members of the alternative distribution, the value is 1. That is, it is assumed that the test is always individually powerful enough to identify a true alternative without multiplicity correction.
2. Check if there are blocks of k or more adjacent test sites that each have a value of 1. If this is true, count the number of test sites across all such blocks from the null distribution (#False Positives) and the number of test sites in all the blocks (#Positives). The false discovery proportion is computed as FDP=(#False Positives)/(#Positives)
3. Repeat Steps 1 and 2 10000 times. The estimate of FDR is the average of the FDPs.

Setting $N = 100$ and $M \leq 5$, simulations were conducted to estimate FDR for different patterns of true positives and true negatives. Even in this relatively small-scale setting, in terms of the total number of test sites, the total number of possible configurations exceed 79 million. However, most of these configurations are essentially duplicates of one another for purposes of computing FDR. As such, only a select few configurations were considered. For example, the configuration 0XXXXX0 represents the configuration that all five true positives are together. The 0's in each end represent blocks of null test sites. There are many configurations like this, such as when there are 50 null sites, and then five true positives, and then 50 null sites. However, such a configuration is redundant to the configuration that just changes the division of test sites to the left and right, such as a configuration where there are 20 null sites, followed by the five true positive sites, and then 80 null sites. As such, only one set of FDR estimations for this configuration and all configurations similar to it as described, is needed.

Results are shown in Table 2. As shown here, FDR is controlled at the nominal level ($\alpha = 0.05$) for $k = 5$ in all the configurations considered. The same is true for $k = 4$, but is no longer true for $k = 3$ or $k = 2$. More importantly, Table 2 shows that the maximum FDR estimate for $k = 5$ is 0.02063 while for $k = 4$ is 0.03027. While in both cases, some conservation of the nominal level α is observed, the conservation is not small enough to justify adding the power-boosting method to another FDR controlling approach. Nonetheless, the amount of α conserved scales well with the number of hypotheses with the maximum proportion of true positives held constant. This is illustrated in Table 3, where $N = 1000$ and $M = 50$. In this setting, FDR seemed negligible for $k = 50$ in any setting. As such, it was possible to decrease k. Table 3 shows that for $k = 20$, the maximum FDR is less than 0.01, indicating that there is sufficient conservation to append the power boost to another FDR controlling method for this setting. Doing so will enable the combined test to reject a subset of 20 or more adjacent test sites if each is found to have a p-value less than 0.05 prior to any controlling adjustment, and the combined method will still be able to control FDR at a nominal level of 0.059, only slightly higher than the level if only one FDR procedure is used. In exchange for this, the combined test can detect as significant, a group of test sites that may seem obviously interesting to the practitioner, but would have failed detection otherwise because their individual signals are not strong enough to be detected by the α adjusting method.

APPLICATION TO GWAS DATA

Genome-wide association studies (GWAS) attempt to associate specific genetic variations with particular diseases. This typically involves the conduct of millions of hypothesis tests. Data from was sourced from the UK Biobank (*GWASBot*, n.d.). The dataset associates genetic variants in the human genome with asthma. The genetic correlation was estimated using LD-score regression, and *p*-values were computed based on a Normal distribution (*UKBB*, n.d.).

Table 2

FDR for selected configurations N=100 M=5. X represents a test site belonging to the alternative distribution, while a 0 represents a block of test sites belonging to the null distribution. k represents the different block sizes for which FDR is computed

		k			
Pattern	Number of True Positives	5	4	3	2
0X0X0X0X0X0	5	0.00329	0.00686	0.07499	0.25865
0XXXXX0	5	0.01753	0.01752	0.02178	0.07451
0XXXX0X0	5	0.01783	0.01915	0.02419	0.09120
0XXXX00X0	5	0.02063	0.02007	0.02753	0.10216
0XXX00XX0	5	0.00330	0.02823	0.04664	0.08772
0XXX00XX0	5	0.00344	0.03027	0.04552	0.08956
0XX0XX0X0	5	0.01266	0.02623	0.04799	0.10170
0XXXX0	4	0.02014	0.02085	0.02476	0.08744
0XXX0	3	0.00380	0.02607	0.03102	0.10790
0XX0	2	0.00037	0.00499	0.04472	0.13304
0XX0XX0	4	0.01177	0.01332	0.05270	0.09726
0XX00XX0	4	0.00054	0.00866	0.07360	0.10436
0XXX00X0	4	0.00261	0.02673	0.03270	0.11532
0X00XX00X0	4	0.00041	0.00504	0.05185	0.16487
0X00XX0	3	0.00027	0.00469	0.04791	0.14973
0X00X0	2	0.00010	0.00130	0.02082	0.26076
NA	0	0.00000	0.00100	0.00940	0.21140

Table 3

FDR for selected configurations N=1000 M=50. X represents a block of test sites belonging to the alternative distribution, while a 0 represents a block of test sites belonging to the null distribution. In each setting, the size of X and 0 are varied

		k		
Pattern	Number of True Positives	10	15	20
0XXX0	50	0.00209	0.00197	0.00208
0X0X0	50	0.00581	0.00939	0.00899
0X0X0	50	0.01692	0.01328	0.00096
0X0X0	50	0.01864	0.01400	0.00109
0X0X0	50	0.00578	0.00119	0.00000
0X0X0	50	0.00518	0.00215	0.00000
0X0X0	50	0.00339	0.00064	0.00000
0X0X0	50	0.00424	0.00425	0.00582
0X0X0	50	0.00423	0.00829	0.00506
0X0X0	50	0.00500	0.00862	0.00598
0X0X0	50	0.00830	0.00608	0.00008
0X0X0	50	0.00439	0.00359	0.00333
0X0X0	50	0.00398	0.00363	0.00548
0X0X0	50	0.00371	0.00469	0.00307
0X0X0	50	0.00613	0.00767	0.00165
0XXX0	20	0.00466	0.00549	0.00533
0X0X0	20	0.00524	0.00462	0.00534
0X0X0	20	0.00939	0.00023	0.00023

For simplicity, this demonstration is limited to Chromosome 2 and Chromosome 6. Each with about 2 million variants. For each variant, the null hypothesis is that the variant is not associated with asthma. The purpose is to identify all variants for which there is evidence of significant association with asthma. The variants in each chromosome are meaningfully ordered according to physical proximity, and it is known that strong association in one position increases the likelihood of strong associations in neighboring positions. These qualities make it suitable to use the method for this setting.

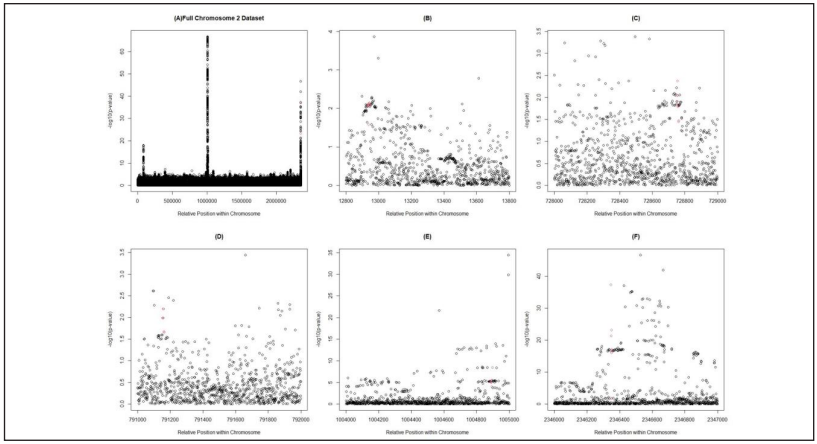
From numerical estimation, it is determined that $k = 10$ at nominal $\alpha = 0.05$ is sufficient to control FWER at > 0.0001 in the weak sense when there are 2 million hypotheses. This means that when identifying

10 or more consecutive variants as significant, if each of them has a *p*-value less than 0.05, the probability of at least one false discovery is less than 0.0001. Furthermore, given the sheer sparsity of this setting, where it is expected that there are relatively very few variants that are truly associated with asthma, this is assumed to be a sufficient choice for *k* for controlling FDR at the same nominal level.

For Chromosome 2, the method identified 54 variants in 5 blocks that are significant at nominal $\alpha = 0.05$. Results are illustrated in Figure 2. The significant blocks were identified around the same positions where variants with the highest negative log-transformed *p*-values were found, yet all but the block in Figure 2 (F) are not likely to be identified as significant by a testing procedure that controlled for multiplicity by adjusting individual *p*-values. Most importantly, this Power Boosting method is not supposed to stand alone. It can be added onto another method that is calibrated at a nominal significance level of 0.05 with negligible impact to the error rate since the nominal significance level at $k = 10$ for the Power Boost is less than 0.0001. It is notable to observe that only the block in Figure 2(F) is located at a spike in the dataset, whereas there are three visible spikes in Figure 2(A). That is, the correlation of *p*-values around the spikes where not strong enough to satisfy the requirement for $k = 10$ that is necessary for error rate control.

Figure 2

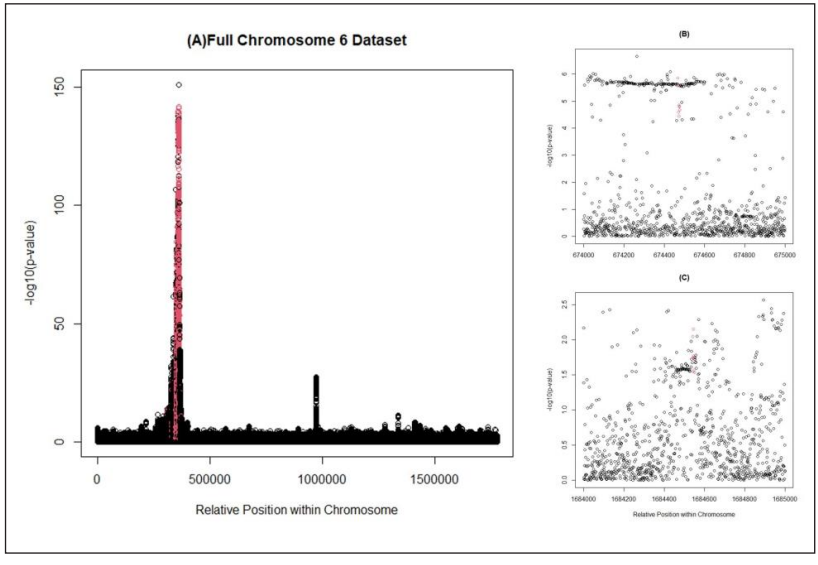
Results for Chromosome 2. Significant variants are colored red. (A) is the entire dataset while (B) to (F) show snapshots around locations where significant blocks were found



In contrast to Chromosome 2, much stronger block relationships were found in Chromosome 6. Application of the method yielded 10131 significant variants spread across at least 10 blocks. Most importantly, many of those blocks contain hundreds of consecutive variants. This is illustrated in Figure 3. As shown in Figure 3(A), most of the blocks are found in the exact same position as the highest spike, establishing that this spike is interesting not just because of the significant *p*-values found around it, but because these *p*-values are also so consistent that there are blocks of hundreds of consecutive test sites where each hypothesis in the block is significant at the nominal level. It is notable to compare this and the result in Chromosome 2, because while both have this high peak, the signal was not as strongly consistent across consecutive positions for Chromosome 2 as it is for Chromosome 6. Once again, it is emphasized that these discoveries for Chromosome 6 were achieved at a nominal level that is low enough that it can be appended onto any other FDR controlling method.

Figure 3

Results for Chromosome 6. Significant variants are colored red. (A) is the entire dataset while (B) to (C) show snapshots around locations where significant blocks were found



CONCLUSION

A power boosting method for type 1 error rate control was demonstrated. This method is applicable in settings where the hypotheses can be reasonably assumed to follow some practical ordering, such that it is possible to determine which test sites are nearest to each test site. It was shown that this method is able to control FWER in the weak sense. More importantly, numerical evidence is provided that under an assumption of sparsity, the method is able to control FDR in the strong sense and at an adjustable level of conservativeness, making it possible to append the method onto another FDR controlling procedure in order to provide extra power at a minimal cost. This extra power is demonstrated through the application of the method on a GWAS dataset.

ACKNOWLEDGMENT

We thank the Associate Editor and the anonymous reviewer for their many constructive comments and suggestions that have improved the paper. This research was supported by the National Research Foundation of Korea, funded by the Korea government (no. NRF-2019H1D3A2A02102167).

REFERENCES

Aslam, M., & Albassam, M. (2020). Presenting post hoc multiple comparison tests under neutrosophic statistics. *Journal of King Saud University Science*, *32*(6), 2728-2732.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bogdan, M., Chakrabarti, A., Frommlet, F., & Ghosh, J. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, *39*(3), 1551–1579.

Dmitrienko, A., Bretz, F., Westfall, P., Troendle, J., Wiens, B., Tamhane, A., & Hsu, J. (2010). *Multiple testing problems in pharmaceutical statistics.* Chapman and Hall.

Efron, B., Tibshirani, R., Storey, J., & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151–1160.

Frommlet, F., & Bogdan, M. (2013). Some optimality properties of fdr controlling rules under sparsity. *Electronic Journal of Statistics*, *7*, 1328–1368.

Ghosh, A., & Chakraborty, A. (2017). Use of em algorithm for data reduction under sparsity assumption. *Computational Statistics*, *32*, 387–407.

Gwasbot. (n.d.). https://twitter.com/SbotGwa status/1422180670240067585. (Accessed: 2021-08-3)

Ikram, M., Xueling, S., Jensen, R., Cotch, M., & Hewitt, A. (2010). Four novel loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS Genet.*, *6*(10), e1001184.

Kirsch, A., Mitzenmacher, M., Pietracaprina, A., Pucci, G., Upfal, E., & Vandin, F. (2012). An efficient rigorous approach for identifying statistically significant frequent itemsets. *Journal of the ACM*, *59*(3), 12:1– 12:22.

Lee, C., & Steigerwald, D. (2020). Inference for clustered data. *The Stata Journal*, *18*(2), 447-460.

Miller, R. (1981). *Simultaneous statistical inference*. New York: Springer.

Noble, W. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*(12), 1135–1137.

Qu, H., Tien, M., & Polychronakos, C. (2010). Statistical significance in genetic association studies. *Investigative Medicine*, *33*(5), 266–270.

Sid´ak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633.

Ukbb. (n.d.). https://ukbb-rg.hail.is/rg$_s$*ummary*$_2$0002$_1$111.*html.*(*Accessed* : $2021 - 08 - 3$)

Verhoeven, K., Simonsen, K., & McIntyre, L. (2005). Implementing false discovery rate control: increasing your power. *OIKOS*, *108*(3), 643–647.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, *48*(3), 156–158.

Yang, Q., Cui, J., Chazaro, I., Cupples, A., & Demissie, S. (2005). Power and type i error rate of false discovery rate approaches in genome-wide association studies. *BMC Genetics*, *6*(1), S134.