Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 429–448
http://www.ssca.org.in/journal.html

r-Power for Multiple Hypotheses Testing under Dependence

Swarnita Chakraborty, Adebowale Sijuwade and Nairanjana Dasgupta
Department of Mathematics and Statistics, Washington State University, United States

Received: 28 March 2024; Revised: 13 September 2024; Accepted: 17 September 2024

Abstract

In an era of "big data" the challenge of managing large-scale multiplicity in statistical analysis has become increasingly crucial. The concept of r-power, introduced by Dasgupta *et al.* (2016), presents an innovative approach to addressing multiplicity with a focus on the reliability of selecting a relevant list of hypotheses. This manuscript advances the r-power conversation by relaxing the original assumption of independence among hypotheses to accommodate a block diagonal correlation structure. Through analytical exploration and validation via simulations, we unveil how the underlying dependence structure influences r-power. Our findings illuminate the nuanced role that dependence plays in the reliability of hypothesis selection, offering a deeper understanding and novel perspectives on managing multiplicity in large datasets. Furthermore, we highlight the practicality and applicability of our results in the context of a Genome-Wide Association Study (GWAS).

Key words: r-power; Multiplicity; Multiple hypotheses testing; Dependence; False positives; Genome-wide association study.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Multiple hypotheses testing has always been a concern in scientific research due to the challenge of increasing false discoveries with growing multiplicity. However, the increasing prevalence of large-scale testing has brought this topic front and center. Despite the progress made, researchers continue to seek the "holy grail" that balances statistical power and control of false discoveries. The review of the literature, in this work aims to contribute to the understanding and advancement of multiple hypotheses testing, fostering the development of practical, feasible, and sensible methods in this field.

Multiple hypotheses testing has gained significant importance in various scientific disciplines, including chemistry (metabolomics), biology (genomics, proteomics), medicine (fMRI), and social sciences. As the scale of testing has expanded, controlling for multiplicity has become a critical concern due to the inflation of Type I error rates resulting from simultaneous testing. Traditional approaches like the Bonferroni Family-wise Error Rate (FWER), which has been used for a long time, are conservative and

Corresponding Author: Swarnita Chakraborty
Email: s.chakraborty4@wsu.edu

hence, impractical when dealing with millions of hypotheses being tested. This research area starting with Holm (1979), Simes (1986), Hochberg (1988), Hommel (1988), Sarkar (1998) has persevered for the "gold standard" and has see-sawed between techniques that are very conservative or too liberal. The introduction of the FDR (False Discovery Rate) by Benjamini and Hochberg (1995) has marked a significant advancement, providing a flexible and powerful framework for controlling false discoveries. The False Discovery Rate (FDR) is a measure that estimates the expected proportion of false discoveries among all the hypotheses that are rejected. Since the introduction of the FDR, numerous studies have enhanced and refined its methodology. Benjamini and Yekutieli (2001) proposed a modified procedure to accommodate dependence structures, ensuring valid control of the FDR. Efron *et al.* (2001) introduced empirical Bayes methods that borrow strength across hypotheses to improve FDR estimation. Building on these ideas, Efron (2004) developed the "local FDR" approach, allowing for more precise estimation of the FDR. Storey (2002) introduced the concept of q-values, which provide an intuitive interpretation of the FDR, enabling researchers to control the proportion of false discoveries at various thresholds. Furthermore, Storey (2003), Storey (2007), Heller *et al.* (2006), Dudoit *et al.* (2002), Dudoit *et al.* (2003), Pan (2002), Nichols and Holmes (2002), Nichols and Hayasaka (2003), Worsley (2003), Ge *et al.* (2003), Storey (2011) provides a general discussion of further developments related to FDR. These studies have significantly enhanced our understanding of the False Discovery Rate (FDR) framework, shedding light on its practical applications and uncovering its potential limitations. Storey (2011) also provided a comprehensive review of these advancements, offering a valuable resource for researchers in the field. While the FDR has been widely adopted and has greatly influenced the field of multiple hypotheses testing, researchers continue to search for an optimal method that balances statistical power and control of false discoveries.

Looking from a different perspective, practitioners often rely on available software and commonly used packages in R, which incorporate a ranked "top-table" approach following multiplicity corrections. Smyth *et al.* (2003) and Smyth and Speed (2003) highlight this practice and emphasize the importance of revisiting the top-table approach through the lens of multiplicity control. The fundamental question posed by practitioners is how to design studies that allow the identification of features of interest without being overwhelmed by multiplicity corrections and rigid notions of statistical significance. In response to this question, researchers have explored selection-based-on-ranking approaches within the multiplicity framework. Notable contributions in this line of research include works by Smyth *et al.* (2003), Smyth (2005), Kuo and Zaykin (2011), Kuo and Zaykin (2013), Knecht *et al.* (2003), Abbott *et al.* (2010).

Continuing in the same vein of research, Dasgupta *et al.* (2016) introduced the notion of "r-power" to provide a mathematical framework for the top-table approach. r-power is defined as the probability that no false positives exist among the test candidates included in the top-table. However, their analysis assumes independence among the hypothesis testing units, which is often an unreal assumption to implement in practice. Our study aims to relax the assumption of independence and re-formulate r-power under dependence making it applicable to real-life scenarios. It begins by considering the simplest case of equicorrelation and subsequently extends the analysis to more realistic scenarios involving block diagonal correlation structures.

While numerous approaches have been proposed to assess dependence among test candidates in multiple hypotheses problems, our method based on r-power offers a fresh perspective on this issue. Recent methods, such as the one proposed by Leek and Storey (2008), construct a dependence kernel to ensure independence of test statistics. Kim

and van de Wiel (2008) propose a method that assesses dependence using a constrained random correlation matrix. Sun and Tony Cai (2009) introduce a data-driven approach to minimize the false non-discovery rate, assuming a two-state hidden Markov model for the observed data. Additionally, Friguet *et al.* (2009) propose a conditional false discovery rate (FDR) based on a factor model. Furthermore, Liu *et al.* (2016) develop a method to assess dependence in multiple hypotheses testing using graphical models, where latent binary Markov random fields represent the underlying true states of hypotheses, and the observed test statistics appear as coupled mixture variables.

In contrast, our method takes a different perspective. We focus specifically on estimating the probability of false positives within the selected list of hypotheses, rather than considering the entire dataset and we incorporate block diagonal correlation structure to assess dependence among the test candidates. By adopting this approach, our computational framework becomes efficient and easily understandable from a practitioner's point of view.

2. Introducing r-power

In the following, we reintroduce r-power, dropping the assumption of independence. We only present the one-sided case: one-sided hypotheses, for one-sample problems, as it is the foundation of our main results in the sections to follow. Further details on the formulation of r-power in the two-sided case are available in Dasgupta *et al.* (2016). A practical approach to large scale testing, r-power focuses on selection-based ranking and answer the question: can one merely rank a test-statistic and identify the top r candidates from a set of hypotheses generated? By determining r-power, we measure the reliability of this "top table", with a focus on prioritizing features of interest over multiplicity corrections. We now present the underlying multiple testing problem in its canonical form.

2.1. Testing for normal means

Let $\vec{X}$ be a random vector following a multivariate normal distribution such that $\vec{X} \sim N(\boldsymbol{\mu}, \vec{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ is the mean vector and $\vec{\Sigma}$ is the covariance matrix, which can be one of the following: (i) an identity matrix (ii) an equicorrelated matrix (iii) a block diagonal matrix with each block being equi-correlated.

Focusing on t-tests as the statistic of interest, we assume that the number of observations is large, and that the t statistics can be approximated by normal z statistics. Without loss of generality, we define our alternative hypothesis to be in the greater than direction. Comparing to a known mean μ_0, our hypotheses of interest are given by $H_{0i} : \mu_i \leq \mu_0$ and $H_{Ai} : \mu_i > \mu_0$. We assume that for K of them are from $\mu_i = \mu_0$ and $N - K$ of them are from $\mu_i = \mu_1$, where $\mu_1 > \mu_0$. Letting $\bar{y}_i, s_i$ denote the sample mean and sample standard deviation for the i^{th} hypothesis of interest, in the one-sided case, we define our test statistic $t_i = \sqrt{n}(\bar{y}_i - \mu_0)/s_i$. We assume that the number of observations is large, and thus, t_i are approximately $N(0, 1)$ for K of the hypotheses and $N(\delta, 1)$ for $N - K$ of the hypotheses, with corresponding effect size $\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$.

2.2. Determining r-power

To determine r-power in the testing problem above, we consider the top r hypotheses among the test statistics $t_1, \ldots, t_N$. Let G_0, G_A denote the groups of hypotheses

supporting the null and alternative hypotheses respectively, with respective test statistics t_i^0 and t_i^A. In the case of independence, it is assumed that the N hypotheses are independent, with equal variances, reducing the covariance structure to the identity matrix, and t_i^0 and t_i^A are i.i.d. $N(0,1)$ and $N(\delta,1)$, respectively. We denote the respective null and alternative order statistics by $Z_{(i)}, U_{(j)}, i = 1,\ldots,K, j = 1,\ldots,N-K$. Misclassification occurs if the largest member of G_0 is greater than or equal to the $(N-K-r)^{th}$ order statistic from G_A, or equivalently, $Z_{(K)} \geq U_{(N-K-r)}$. We define r-power as the probability of correct classification, that is,

$$r_P = P(Z_{(K)} < U_{(N-K-r)}). \tag{1}$$

Assuming independence, we can write $\vec{\Sigma} = \sigma^2 \vec{I}_N$, resulting in an r-power of

$$r_P{}^{(1)}(N,K,r,\delta) = \int_0^1 \Phi(\Phi^{-1}(t)+\delta)^K \beta(N-K-r, r+1, t)\, dt, \tag{2}$$

where Φ, ϕ denote the respective standard normal distribution and density. In practice, $r \leq N-K$ is chosen by the researcher. Ranking selection based methods such as r-power require some knowledge on the true number of null hypotheses. There have been various methods proposed for estimating the null proportion K/N, such as Jin (2008), Chen (2018), Sijuwade *et al.* (2023).

3. Incorporating dependence

With growing dimension and complexity comes an increased risk of Type I error inflation and thus the assumption of independence becomes less realistic. We consider a more general but practical option: a block diagonal correlation structure. This approach is inspired by the success or similar methods from omics studies, in which genes, lipids and metabolites tend to be related based on common chemical or biological properties. Some compelling examples include the following. Perrot-Dockès *et al.* (2019) estimated block diagonal covariance matrices to study seed quality based on omics information. Pacini *et al.* (2017) established a method to reduce false discoveries in gene expression studies using block diagonal correlation structures. To reduce computation complexity in a sensitivity analysis problem, Broto *et al.* (2020) developed a method to estimate high dimensional block-diagonal covariance matrices for Gaussian data. In practice, unstructured dependence is most realistic to consider for multiple testing, however, we show that our proposed method is general enough to approximate it, but simple enough to obtain reasonable estimates of r-power for implementation.

4. Motivating example: A GWAS study

Genome-Wide Association Studies (GWAS) aim to identify associations between genetic variants, specifically Single-nucleotide polymorphisms (SNPs), and observed traits or phenotypes. This study focuses on the association between SNPs and human cholesterol levels. The dataset used in this study is based on 323 individuals from India, China, and Malaysia, with 2,527,458 SNPs and cholesterol level measurements based on the Singapore Integrative Omics Study Saw *et al.* (2017). The purpose of this study was to use this data as an example and understand the performance of r-power when the test candidates (here, SNPs) are dependent. We focused our analysis on a subset consisting of 316 individuals and 32,010 SNPs from Chromosome 1.

Data description:

We downloaded the dataset from a public github repository basing their GWAS analysis and tutorial on data from the Singapore Integrative Omics Study `https://github.com/monogenea/GWAStutorial/tree/master/public`. Along with their methods in this tutorial, we also followed the GWAS methods for data pre-processing from Reed *et al.* (2015). The dataset includes three sub-parts:

- Genotype: A SNP matrix with columns representing SNPs and rows representing sample IDs. Genotype values range from 0 to 2, indicating different allele combinations.
- Mapping File: Contains sample IDs, SNP IDs, chromosome numbers, SNP positions, and allele types.
- Phenotype: Includes sample IDs and continuous-scale cholesterol level measurements.

Data pre-processing

SNPs with high missingness, low variability and genotyping errors were filtered out. We conducted our entire analysis in R and utilized the libraries **SNPRelate** and **snpStats** from the **BiocManager** package in R alongside commonly used packages for data handling, visualization and parallel processing - **tidyverse, doParallel, foreach** and wrote our own function for conducting the GWAS, based on the following

- **Call Rate:** The percentage of individuals in the study with available SNP information. SNPs with a call rate below 1 were discarded, removing missing information.
- **Minor Allele Frequency (MAF):** MAF denotes the proportion of least common alleles for each SNP. SNPs with MAF below 0.1 were discarded, focusing on those with a higher frequency of less common alleles.
- **Heterozygosity & Hardy Weinberg Equilibrium (HWE):** Heterozygosity occurs when each of the two alleles are present at a given SNP within an individual. HWE is a condition where the population does not evolve over generations. More specifically, this means that the alleles and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.

  A measure of HWE is given by the Inbreeding Coefficient: $|F| = |1 - H/H_{\text{exp}}|$, where H is the observed heterozygosity, $H_{\text{exp}} = 2pq$ is the expected heterozygosity and p, q are the frequencies of the respective dominant and recessive alleles 'A' and 'a'. We retain samples that are not too heterozygous (affecting sample quality) or too homozygous (indicating inbreeding), discarding those with $|F| > 0.1$.
- **Linkage Disequilibrium (LD):** The presence of a statistical association between allelic variants within a population due to the history of recombination, mutation, and selection in a genomic region.
- **Kinship Coefficient:** A measure of relatedness among the individuals. It denotes the probability that a pair of randomly sampled homologous alleles is identical by descent. SNPs with a kinship coefficient above 0.2 were discarded, reducing relatedness bias.

After filtering based on call rate and MAF, 795668 SNPs remained. Following preprocessing, 316 individuals and 32010 SNPs were retained for analysis.

Analysis and results

We fitted a generalized linear model for each of the 32010 SNPs using the top 20 principal components and the Origin variable (dichotomized) as the covariates with our model structure. Our approach was motivated by Reed *et al.* (2015), Lipka *et al.* (2012); Price *et al.* (2006), and Wang and Zhang (2021). We conducted principal component analysis on a LD pruned dataset with an LD cut-off of 0.2. To understand the population structure, we have conducted a principal component analysis on the SNPs. We have pruned the SNPs with a linkage disequilibrium value higher than 0.2. We did so to understand the underlying population substructure, if any, through principal components. In our analysis, the top 21 principal components explained approximately 70% of the variability and we included these PCs as covariates in our model. The first two principal

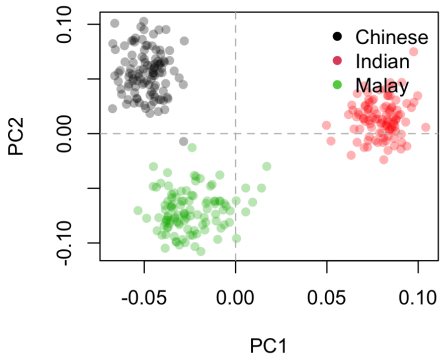


Figure 1: PCA Plot and Difference by Origin

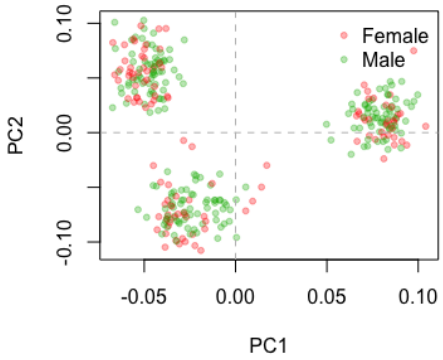


Figure 2: PCA Plot and Difference by Gender

components account for 2.42% of the variability in the data. When we plot PC1 versus PC2, we observe that SNPs originating from the same population tend to cluster together (see Figure 1). Interestingly, when we examine the distribution of gender, we find it to be fairly uniform across the different origins (see Figure 2). To ensure the reliability of our findings and minimize the occurrence of false positives, we employed a robust set of statistical procedures, including the Benjamini-Hochberg, Holmes, Sidak, and Benjamini-

Yukuteli methods. SNPs identified by these methods are shown in Table 1.

Table 1: Comparison of Methods to Control False Discovery

Method	SNPs Identified	SNP ID
No Correction	1502	rs7527051, rs12406924, rs4844688,...
Bonferroni	1	rs7527051
Benjamini Hochberg	2	rs7527051, rs12140539
Benjamini & Yekutieli	0	-
Holm	1	rs7527051
Sidak Single-Step	1	rs7527051

The effect of the multiplicity correction methods on the number of identified SNPs can been seen in 1, with 1502 SNP selections in the absence of any correction. Bonferroni's method, known for its stringent control of family-wise error rate (FWER), only selected 1 SNP. On the other hand, Benjamini Hochberg's method, aimed at controlling the false discovery rate (FDR), chose 2 SNPs. Interestingly, the Benjamini-Yekutieli method, which considers positive regression dependency among SNPs instead of assuming independence, did not select any SNPs. Additionally, Holm's and Sidak's procedures each reported 1 SNP. The discrepancy between the number of significant SNPs obtained from the various correction methods raises an important question: How many SNPs should we follow up on? Whilst no multiplicity correction resulted in a large number of significant SNPs, FDR & FWER corrections, yielded one or two significant SNPs. Striking the right balance is essential to ensure the accuracy and reliability of our findings. Hence, in addition to these established methods, we introduce our formulation of r-power allowing dependence in our formulation. In the following section, we generalize the idea of r-power under dependence.

5. Equicorrelation

Returning to the testing problem 2.1, we consider the equicorrelated case corresponding to a joint distribution $N_N(\boldsymbol{\mu}, \vec{\Sigma})$, where $\vec{\Sigma} = \sigma^2[\rho \vec{1}_N \vec{1}_N^T + (1-\rho)\vec{I}_N], 0 \leq \rho < 1$. We present the following result, from which the probability of correct classification can readily be determined.

Theorem 1: r-power under Compound Symmetry

Under the equicorrelated testing scenario, we have

1. $\vec{\Sigma}^{-1/2} = a\vec{1}_N\vec{1}_N^T + (b-a)\vec{I}_n$, where

$$a = \frac{1}{\sigma N}\left(\frac{1}{\sqrt{1+(N-1)\rho}} - \frac{1}{\sqrt{1-\rho}}\right),\ b = a + \frac{1}{\sigma\sqrt{1-\rho}}.$$

2. $\vec{\Sigma}^{-1/2}\vec{Y} \sim N_N(\vec{\Sigma}^{-1/2}\boldsymbol{\mu}, \vec{I}_N)$.

3. For $1 - \sigma^{-2} < \rho < 1$, the probability of misclassification is always less than the equivalent probability under the independent testing scenario 2.1 with equivalent dimensions, with corresponding classification probability $r_P{}^{(1)}(N, K, r, (b-a)\delta)$, following (2).

Proof:

1. Let $\vec{e}_k$ denote the k^{th} standard basis vector of $\mathbb{R}^N$, $\vec{S}_k = \sum_{i=1}^{k} \vec{e}_i$. Since $\rho \neq (N-1)^{-1}$, the Sherman-Morrison formula $(\vec{A} + \vec{u}\vec{v}^T)^{-1} = \vec{A}^{-1} - (\vec{A}^{-1}\vec{u}\vec{v}^T\vec{A}^{-1})/(1 + \vec{v}^T\vec{A}^{-1}\vec{u})$ with $\vec{u} = \vec{v} = \vec{1}_N, \vec{A} = (1-\rho)\vec{I}_N$, implies that $\vec{\Sigma}$ has symmetric positive definite inverse $\vec{\Sigma}^{-1} = C_1(\vec{I}_N - C_2\rho\vec{1}_N\vec{1}_N^T)$, $C_1 = \sigma^{-2}(1-\rho)^{-1}, C_2 = (1+(N-1)\rho)^{-1}$, since $\vec{1}_N\vec{1}_N^T$ has spectrum $\lambda_1 = N, \lambda_2 = \cdots = \lambda_N = 0$, and eigenvectors $\vec{u}_1 = \vec{1}_N, \vec{u}_j = \vec{e}_1 - \vec{e}_j, j = 2, \ldots, N$.

   Let $\vec{D} = \text{diag}(l_1, \ldots, l_N)$ where $l_j = 1/\sqrt{C_1(1 - C_2\rho\lambda_j)}$, and let $\vec{U}$ denote the matrix with columns $\vec{u}_j$. Letting $\vec{H} = (\vec{1}_N\vec{1}_N^T + \text{diag}(0, -N, \ldots, -N))/N$, $\vec{H}\vec{U} = [\vec{1}_N, \vec{0}_N, \ldots, \vec{0}_N]^T + [(\vec{e}_1 - \vec{S}_N), \vec{e}_2, \ldots, \vec{e}_N]^T = \vec{I}_N$. Since $\vec{H}, \vec{U}$ and their product are symmetric, they commute, and we write $\vec{H} = \vec{U}^{-1}$.

   Let $d_1 = \vec{D}_{11} = 1/(\sigma\sqrt{1+(N-1)\rho}), d_2 = \vec{D}_{22} = 1/(\sigma\sqrt{1-\rho})$ and determine the inverse square root $\vec{\Sigma}^{-1/2} = \vec{U}\vec{D}\vec{U}^{-1}$, as indeed by the Spectral Theorem, $(\vec{\Sigma}^{-1/2})^2 = \vec{U}\vec{D}^2\vec{U}^{-1} = \vec{\Sigma}^{-1}$. Finally, we have the inverse square root $\vec{\Sigma}^{-1/2} = \vec{U}[d_1\vec{1}_N, d_2(\vec{S}_N - N\vec{e}_2), \ldots, d_2(\vec{S}_N - N\vec{e}_N)]^T/N$, so set $a = (d_1 - d_2)/N, b = (d_1 + (N-1)d_2)/N = a + d_2$.

2. Using characteristic functions, let $i = \sqrt{-1}$ denote the imaginary unit and $\vec{r} = (r_1, \ldots, r_N)$ denote an arbitrary deterministic vector. Setting $\vec{s} = (\vec{\Sigma}^{-1/2})^T\vec{r}$, we have $\mathbb{E}(e^{i\vec{r}^T\vec{Y}^*}) = \mathbb{E}(e^{i\vec{r}^T\vec{\Sigma}^{-1/2}\vec{Y}}) = \mathbb{E}(e^{i\vec{s}^T\vec{Y}})$ since $\vec{Y} \sim N_N(\boldsymbol{\mu}, \vec{\Sigma})$. Observing that $\vec{\Sigma}^{-1/2}\vec{\Sigma}(\vec{\Sigma}^{-1/2})^T = \vec{I}_N$ and $\vec{s}^T\vec{s} = \vec{r}^T\vec{r}$ due to the symmetry of $\vec{\Sigma}^{-1/2}$, we have $e^{i\vec{s}^T\boldsymbol{\mu} - \vec{s}^T\vec{\Sigma}\vec{s}/2} = e^{i\vec{r}^T(\vec{\Sigma}^{-1/2}\boldsymbol{\mu}) - \vec{r}^T\vec{\Sigma}\vec{r}/2}$, as desired.

3. Without loss of generality, the mean vector can be written as $\boldsymbol{\mu} = \mu_0(\vec{S}_N - \vec{S}_k) + \mu_1(\vec{S}_N - \vec{S}_{N-k})$. Then, $\vec{\Sigma}^{-1/2}\boldsymbol{\mu} = (\mu_0{}^*)(\vec{S}_N - \vec{S}_k) + (\mu_1{}^*)(\vec{S}_N - \vec{S}_{N-k})$, where $\mu_0{}^* = \mu_0(a + (k-1)b) + \mu_1 b(N-k)$, $\mu_1{}^* = \mu_0 bk + \mu_1(a + b(N-k-1))$. The result follows from the monotonicity of Φ and the observation that in the two-sided case, with effect size $\delta^* = \sqrt{b}(\mu_1{}^* - \mu_0{}^*)/\sigma$, we have $\delta^*/|\mu_1 - \mu_0| = (b-a) = (\sigma\sqrt{1-\rho})^{-1} > 1$ when $1 - \sigma^{-2} < \rho < 1$. We then compare with equation (2).

□

6. Block diagonal approach

We now move to a more general scenario based on the equicorrelated testing problem in our theorem, as equicorrelation is still too restrictive for practical applications. As we mentioned on in Section 3, we extend the equicorrelated case to obtain an analytic form for the probability of correct classification. Based on this recent research, and its adjacency to r-power in application, we consider a block-diagonal correlation structure based on the equicorrelated case which tends toward unstructured as the block size increases. The probability of misclassification can be determined based on the distribution of the within-block order statistics. We assume independent blocks in which each block corresponds to test candidates belonging to either the null or alternative hypotheses or a mix of both.

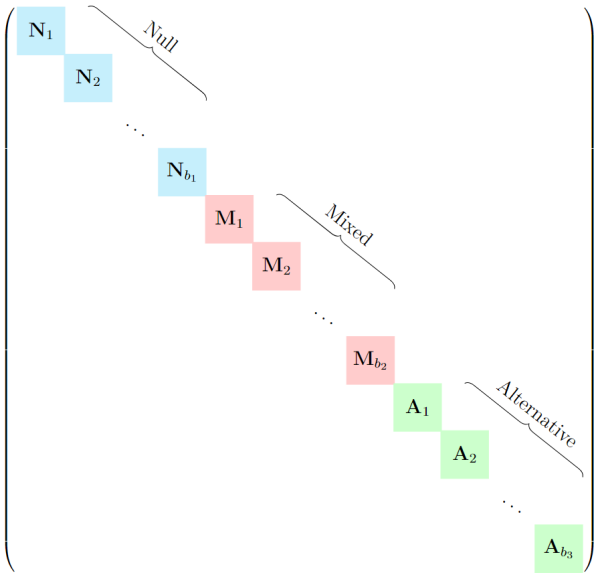
We now define the $N \times N$ block diagonal correlation matrix $\vec{B}$. For $j = 1, 2, 3$, we assume respective null, mixed and alternative test candidate counts $b_j : \sum_j b_j = N$

and define the respective null, mixed and alternative major blocks and indices by $\vec{B}^{(j)} = diag(\vec{B}_1^{(j)}, \ldots, \vec{B}_{b_j}^{(j)})$, defined as follows: each minor block $\vec{B}_i^{(j)}$ is compound symmetric, corresponding to test statistic vectors $\vec{t}_i^{(j)}$ as in the normal means problem, with assumed joint distribution $N_{N_i^{(j)}}(\boldsymbol{\mu}_i^{(j)}, \vec{B}_i^{(j)}), i = 1, \ldots, b_j$ for $j = 1, 2, 3$, based on a common effect size $\delta > 0$:

$$\boldsymbol{\mu}_i^{(j)} = \delta \begin{cases} \vec{0}_{K_i^{(j)}} & j = 1 \\ \vec{S}_{N_i^{(j)}} - \vec{S}_{K_i^{(j)}} & j = 2 \\ \vec{1}_{N_i^{(j)}} & j = 3 \end{cases} \tag{3}$$

$$\vec{B}_i^{(j)} = \sigma_{ij}^2[\rho_{ij}\vec{1}_{N_i^{(j)}}\vec{1}_{N_i^{(j)}}^T + (1 - \rho_{ij})\vec{I}_{N_i^{(j)}}].$$

Here, $\vec{S}_k$ is defined as in our theorem, $0 < \rho_{ij} < 1, \sigma_{ij}^2 > 0$ denote the respective correlation and variance of $\vec{B}_i^{(j)}$. Without loss of generality, we order the minor blocks such that $\vec{B} = diag(\vec{B}^{(1)}, \vec{B}^{(2)}, \vec{B}^{(3)})$. We assume that each block corresponds to $N_i^{(j)}$ test candidates, null candidates $K_i^{(j)} : \sum_{i,j} K_i^{(j)} = K$ and research hypotheses $r_i^{(j)} : \sum_{i,j} r_i^{(j)} = r$, noting that $N_i^{(1)} = K_i^{(1)}, K_i^{(3)} = 0$, since $j = 1, 3$ correspond to only null or alternative candidates respectively. We denote the null and alternative order statistics corresponding to each sub-block $\vec{B}_i^{(j)}$ by $(Z_i^{(j)})_{(k)}, (U_i^{(j)})_{(l)}, k = 1, \ldots, K_i^{(j)}, l = 1, \ldots, N_i^{(j)} - K_i^{(j)} - r_i^{(j)}$ and for convenience, denote the null, mixed and alternative major blocks by $\vec{N}, \vec{M}, \vec{A}$ corresponding to the null, mixed and alternative major blocks $\vec{B}^{(j)}$: $\vec{N} = \vec{B}^{(1)}, \vec{M} = \vec{B}^{(2)}, \vec{A} = \vec{B}^{(3)}$. Write $\vec{B} = diag(\vec{N}, \vec{M}, \vec{A})$, where $\vec{N} = diag(\vec{N}_1, \ldots \vec{N}_{b_1})$, and likewise for $\vec{M}, \vec{A}$. We visualize the block diagonal $\vec{B}$ below,



adopting the notation $\prod_{\vec{N}} = \prod_{i,j:\vec{B}_i^{(j)} \in \vec{N}}$ and likewise for $\vec{M}, \vec{A}$. Under the above assumptions, misclassification occurs when any members of the top r table come from the null, which occurs when any members the $r_i^{(j)}$-th hypotheses come from the corresponding null group of any sub-block $\vec{B}_i^{(j)}$: the largest of the $K_i^{(j)}$-th null order statistics is at least

the smallest of the $N_i^{(j)} - K_i^{(j)} - r_i^{(j)}$-th alternative test statistics. We compute r-power corresponding to the block diagonal $\vec{B}$ as the classification probability

$$r_P = P\left(\max_{\vec{N},\vec{M}}(Z_i^{(j)})_{(K_i^{(j)})} < \min_{\vec{A}}(U_i^{(j)})_{(N_i^{(j)}-K_i^{(j)}-r_i^{(j)})}\right). \tag{4}$$

We note that due to the structure of the means $\boldsymbol{\mu}_i^{(j)}$, the null and alternative test statistics corresponding to the *mixed* block $\vec{M}$ are not exchangeable. However, the test statistics corresponding to the null block $\vec{N}$ and alternative block $\vec{A}$ are (respectively), and their distributions can be readily determined based on the normality in (3), via Theorem 5.3.1 of Tong (1990b) to obtain the r-power analytically. We show that if $b_2 \leq 1$, r_P is completely determined by the distribution

$$F_{i,j,k,\mu}(x) = \int_{\mathbb{R}} H_{ijk}\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)\phi(z)\,dz, \tag{5}$$

where $H_{ijk}(z) = \sum_{m=k}^{N_i^{(j)}} \binom{N_i^{(j)}}{m}\Phi(z)^m\Phi(-z)^{N_i^{(j)}-m}$.

6.1. One-sided case, no mixed candidates

We use the shorthand $Z_{ij} = (Z_i^{(j)})_{(K_i^{(j)})}, U_{ij} = (U_i^{(j)})_{(N_i^{(j)}-K_i^{(j)}-r_i^{(j)})}, K_{ij} = (K_i^{(j)})$ and likewise for $N_i^{(j)}, r_i^{(j)}$. Define $M_{ij} = N_{ij} - K_{ij} - r_i^{(j)} > 0$ and let G_{ij} denote the distribution function of U_{ij} with corresponding density g_{ij}. Starting with the probability of misclassification, due to our assumptions (3), we use the exchangeability of the $t_i^{(j)}$ and proceed as in Tong (1990a). The density corresponding to the distribution (5) is given by

$$f_{i,j,k,\mu} = \sigma_{ij}^{-1}(1-\rho_{ij})^{-1/2}\int_{\mathbb{R}} h_{ijk}\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)\phi(z)\,dz, \tag{6}$$

where $h_{ijk}(z) = k\binom{N_{ij}}{k}\Phi^{k-1}(z)\Phi^{N_{ij}-k}(-z)\phi(z)$. Since $b_2 = 0$, integrating by parts, we obtain

$$\begin{aligned}
r_P &= P\left(\max_{\vec{N}} Z_{ij} \leq \min_{\vec{A}} U_{ij}\right) = \prod_{\vec{N}} P\left(Z_{i1} \leq \min_{\vec{A}}(U_{k3})\right) = \prod_{\vec{N}}\int_{\mathbb{R}} P\left(Z_{i1} \leq u|u\right) g_{k3}(u)du \\
&= \prod_{\vec{N}}\int_{\mathbb{R}}\left(\int_{-\infty}^{u} P\left(\sigma_{i1}(Z_{i1}\sqrt{1-\rho_{i1}} + Z\sqrt{\rho_{i1}}) \leq x\right)dx\right) g_{k3}(u)du \\
&= \prod_{\vec{B}_i\in\vec{N}}\int_{\mathbb{R}}\left(\int_{-\infty}^{u} f_{i,1,K_{i1},0}(x)dx\right)\cdot\frac{\partial}{\partial u}\left(1 - \prod_{\vec{B}_k\in\vec{A}}(1 - F_{k,3,M_{k3},\delta}(u))\right)du \\
&= \prod_{\vec{B}_i\in\vec{N}}\int_{\mathbb{R}} -\left(\int_{-\infty}^{u} f_{i,1,K_{i1},0}(x)dx\right)\left(\frac{\partial}{\partial u}\prod_{\vec{B}_k\in\vec{A}}(1 - F_{k,3,M_{k3},\delta}(u))\right)du \\
&= \prod_{\vec{N}}\int_{\mathbb{R}} f_{i,1,N_{i1},0}(u)\prod_{\vec{A}}(1 - F_{k,3,N_{k3}-r_{k3},\delta}(u))\,du.
\end{aligned}$$

6.2. One-sided case, one mixed candidate

Let p denote the probability that the minimum of the top r table corresponds to the mixed block $\vec{M}$. Since the test candidates corresponding to the mixed block $\vec{M}$ are not exchangeable, to streamline our analytical formulation, we assume the existence of at most one mixed block. Since $b_2 = 1$, using the density (6) and performing the change of variables $t = \Phi\left(\frac{(x-\mu)/\sigma_{ij} + z\sqrt{\rho_{ij}}}{\sqrt{1-\rho_{ij}}}\right)$, we obtain

$$
\begin{aligned}
p &= P(\min_{\vec{A}} U_{ij} = U_{12}) = \prod_{\vec{A}} P(U_{i3} \geq U_{12}) \\
&= \prod_{\vec{A}} \int_{\mathbb{R}} P(U_{i3} \geq u|u) g_{12}(u)\, du \\
&= \frac{1}{\sigma_{12}\sqrt{\rho_{12}}} \prod_{\vec{A}} \int_{\mathbb{R}} P(U_{i3} \geq u|u) \int_0^1 \phi\left(\frac{\Phi^{-1}(t) + (\delta_{12} - u)\sigma_{12}^{-1}}{\rho_{12}^{1/2}(1-\rho_{12})^{-1/2}}\right) \\
&\beta(M_{12}, K_{12} + r_{12} + 1, t)\, dt du \\
&= \prod_{\vec{A}} \int_{\mathbb{R}} f_{i,3,N_{i3}-r_{i3},\delta}(u) F_{1,2,M_{12},\delta_{12}}(u)\, du,
\end{aligned}
$$

where $\delta_{ij} = \delta 1_{\{i,j:\vec{B}_i^{(j)} \in \vec{M}, i > K_{ij}\}}$. We apply our theorem to obtain

$$
\begin{aligned}
P_1 &= P(\max_{\vec{N},\vec{M}} Z_{ij} \geq U_{12}) \\
&= P\left(Z_{i2} \geq U_{12}\right) \prod_{\vec{N}} \int_{\mathbb{R}} \left(\int_u^\infty f_{i,1,K_{i1},0}(x) dx\right) f_{1,2,M_{12},\delta_{12}}(u)\, du \\
&= \left[1 - r_P^{(1)}\left(N_{12}, K_{12}, r_{12}, \delta/\left(\sigma_{12}\sqrt{1-\rho_{12}}\right)\right)\right] \prod_{\vec{N}} \int_{\mathbb{R}} \left(1 - F_{i,1,K_{i1},0}(u)\right) f_{1,2,M_{12},\delta_{12}}(u)\, du, \\
P_2 &= P(\max_{\vec{N}} Z_{ij} \geq \min_{\vec{A}} U_{ij}) \\
&= \prod_{i,\vec{N}} \int_{\mathbb{R}} - \left(\int_s^\infty f_{i,1,K_{i1},0}(x) dx\right) \left(\frac{\partial}{\partial s} \prod_{k,\vec{A}} (1 - F_{k,3,M_{k3},\delta}(s))\right) ds, \\
&= \prod_{\vec{N}} \left(1 - \int_{\mathbb{R}} f_{i,1,N_{i1},0}(s) \prod_{\vec{A}} (1 - F_{k,3,N_{k3}-r_{k3},\delta}(s))\, ds\right)
\end{aligned}
$$

and finally, we have $r_P = p(1 - P_1) + (1 - p)(1 - P_2)$.

6.3. Limiting behavior and the two-sided case

We examine 6.1 to determine the limiting behavior of r_P due to the structural similarity in each case. As the null proportion K/N tends to 1, since $r \leq N - K$, across $\vec{A}, F_{i,j,M_{ij},\mu}$ tends to $F_{i,j,0,\mu} = \int_{\mathbb{R}} \phi(z) dz = 1$ (using the binomial theorem), resulting in vanishing products over $\vec{A}$ and an r-power of zero. The situation in which $r \to N - K$ is similar. Likewise, as $\delta \to \infty$, $F_{i,j,k,\delta} \to 0$, following the limiting behavior of the terms $\Phi^m(z)$ as $z \to -\infty$, r_P tends to $\prod_{\vec{N}} \int_{\mathbb{R}} f_{i,1,N_{i1},0}(u)\, du = 1$. This aligns with our intuition

from the independent case that as the effect size increases, it is easier to distinguish the alternative from the null, and vice versa with increasing null proportion.

In the two-sided case, the test statistics $t_i^{(j)} = \sqrt{N_{ij}}|X_{ij} - \bar{X}|/\sigma$ are assumed to be jointly distributed according to the folded normal, with mean vector entries 0 or δ depending on whether or not they correspond to the null or alternative groups for their respective blocks. We assume a Gaussian copula $C(U_1, \ldots, U_N) = \Phi_{\vec{B}}(\Phi^{-1}(F_1(X_1)), \ldots, \Phi^{-1}(F_N(X_N)))$ with covariance matrix $\vec{B}$ as in (3) and $\sigma_{ij} \equiv 1$. For $1 \leq l \leq N$, the distributions F_l are given by $2\Phi(z) - 1$ and $\Phi(z+\delta) + \Phi(z-\delta) - 1$ respectively. We then determine the r-power as $r_P = P(\max_{N,\vec{M}} \Phi^{-1}(F_k(Z_{ij})) < \max_{\vec{A}} \Phi^{-1}(F_l(U_{ij})), 1 \leq k \leq b_1 + b_2 \leq l \leq N$ and proceed as in 6.1 and 6.2, replacing Φ, ϕ with F_l and its derivative in $H_{ijk}(z)$ from (5). Since r_P has no closed form expression in the block diagonal scenarios, we approximate it numerically. One approach is to reexamine $F_{i,j,k,\mu}(x)$ in (5):

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_{m=k}^{N} \binom{N}{k} \Phi(Az+B)^m (1 - \Phi(Az+B))^{N-m} \phi(z) dz \\
&= \sum_{m=k}^{N} \sum_{j=0}^{N-k} \binom{N}{k} \binom{N-k}{j} (-1)^j \int_{\mathbb{R}} \Phi^{m+j}(Az+B)\phi(z) dz.
\end{aligned}$$

As in Owen (1980), Hartmann (2017), an application of the Fubini-Tonelli theorem and a change of variables $z_k = y_k + x - B/A, k = 1, \ldots, m, \vec{s} = (x, y_1, \ldots, y_m)$ yields

$$\begin{aligned}
&\int_{\mathbb{R}} \Phi^m(Az+B)\phi(z)dz \\
&= \int_{\mathbb{R}} \prod_{k=0}^{m} \Phi(Az+B)\phi(z)dz \\
&= \int_{\mathbb{R}} \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \prod_{k=0}^{m} \phi(Az_k+B)\phi(x)\, dz_1 \ldots dz_k dx. \\
&\frac{1}{\sqrt{(2\pi)^{m+1}|\vec{V}|}} \int_{\mathbb{R}} \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \exp\left(\frac{-1}{2} \vec{s}^T \vec{V}^{-1/2} \vec{s}\right) d\vec{y} dx, \\
&= \frac{1}{\sqrt{(2\pi)^m |\vec{\Sigma}_A|}} \int_{-\infty}^{-B/A} \cdots \int_{-\infty}^{-B/A} \exp\left(\frac{-1}{2} \vec{s}^T \vec{\Sigma}_A^{-1} \vec{s}\right) d\vec{y}, \\
&\vec{V} = \vec{e}_1^T \vec{e}_1 - 2(\vec{e}_1^T \vec{1}_m + \vec{1}_m^T \vec{e}_1) + 4(\vec{1}_m^T \vec{1}_m) + \text{diag}(0, 1, \ldots, 1)/A^2, \vec{\Sigma}_A = \vec{1}_m \vec{1}_m^T + A^{-2} \vec{I}_m.
\end{aligned}$$

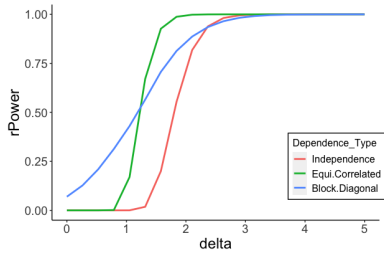
We obtain the multivariate normal distribution function $\vec{F}_m(\vec{0}_m, -(B/A)\vec{1}_m, \vec{\Sigma}_A)$, which can be accurately approximated, as in Genz (1992). The two-sided case can be handled similarly using $F_l(z) = 2\Phi(z) - 1$ or $\Phi(z+\delta) + \Phi(z-\delta) - 1$ respectively, depending on correspondence with the null or alternative.

7. Simulation study

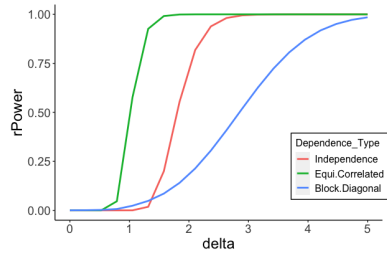
In Section 6, we provided analytical formulae and examined the structure of r-power under our block diagonal assumptions. To support our results, we conducted

an empirical simulation study. We simulated r_P under different scenarios for the one and two-sided cases, letting $b_1 = b_3 = 2, b_2 = 1, N_{ij} \equiv 10^4$. We also varied the number of nulls ($K_{ij} = (1, 3, 5, 7, 9) \cdot 10^3$), null proportions $10^{-4}K_{ij}$, top-table size ($r = 50, 100, 250, 500, 750$) and effect size ($\delta = 0.1, 0.5, 1, 2, 3$). We compare the results for the block diagonal against the independence and equicorrelated scenarios.

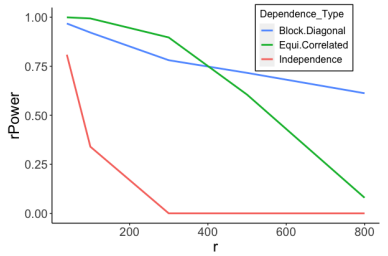
Our results are plotted in Figure 3 for the one-sided case and Figure 4 for the two-sided case. In each figure, we look at r-power as a function of effect size, δ and top-table/list size r. For illustration, we have provided the case with $N = 5$ blocks, assuming $b_1 = b_3 = 2$ and writing $\rho_k, k = 1, \ldots, N$ as the correlation corresponding to block $\vec{B}_k$, *i.e.*, $\rho_1 = \rho_{11}, \rho_2 = \rho_{21}, \rho_3 = \rho_{12}, \rho_4 = \rho_{13}, \rho_5 = \rho_{23}$. To highlight the impact of changing correlation, we vary ρ across the null and alternative blocks respectively, starting with ($\rho_1 = \rho_2 = .7$ and $\rho_4 = \rho_5 = .6$ with $\rho_3 = .5$). Our findings support our expectations from part 3 of our theorem and the tendency of the misclassification probability to increase with r, mentioned in section (6.3). Since $1 - \sigma_{ij}^{-2} = 0 < \rho < 1$, equicorrelation overtakes independence, given equal top table size and effect size with the block diagonal case generally falling somewhere in between the two. The situation changes depending on how the null and alternative correlations compare to each other. Additional results are available upon request from the authors.



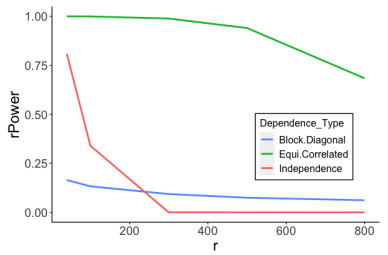
$\rho_1 = 0.7, \rho_2 = 0.7, \rho_3 = 0.4, \rho_4 = 0.5, \rho_5 = 0.5$



$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.8, \rho_4 = 0.9, \rho_5 = 0.9$



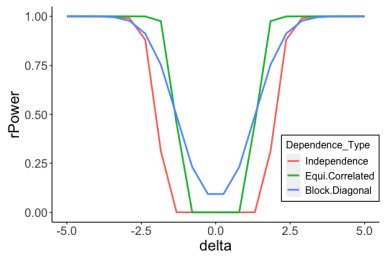
$\rho_1 = 0.9, \rho_2 = 0.9, \rho_3 = 0.4, \rho_4 = 0.4, \rho_5 = 0.4$



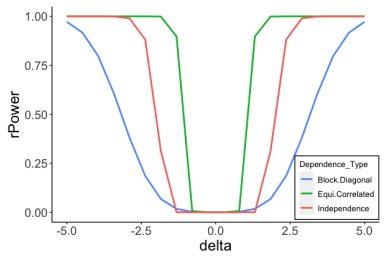
$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.9, \rho_4 = 0.9, \rho_5 = 0.9$

Figure 3: One-sided case: r-power vs effect size δ (top left, top right) and hypothesis selection size r for (bottom left, bottom right)

Since r-power depends upon the size of the top-table r and the number of nulls k, if $r > N - k$, the top-table becomes unreliable as its length exceeds that of the number of alternatives, thus containing members from the null. This supports our findings in Figure 3 (bottom left): as we increase r, we run the risk of this scenario occurring regardless of the dependence structure. However, as the dependence among the alternatives increases, dimensionality is impacted, reducing $N-K$, and when the true $N-K$ tends to be smaller than a given estimate, we are more likely to undershoot for a given choice of top-table size r, reducing the r-power. On the other hand, if the dependence among the alternatives is much smaller than that of the nulls ($\max_{\vec{A}} \rho_{ij} \ll \min_{\vec{N}} \rho_{kl}$), we see a reversal and expect



$\rho_1 = 0.7, \rho_2 = 0.7, \rho_3 = 0.4, \rho_4 = 0.5, \rho_5 = 0.5$



$\rho_1 = 0.4, \rho_2 = 0.4, \rho_3 = 0.9, \rho_4 = 0.9, \rho_5 = 0.9$

Figure 4: r-power vs effect size

larger r-power than the compound symmetry case. This is supported by Figure 4, also where the null correlations are lower than that of the alternatives (*e.g.* .4 vs .9, as shown), r-power for the block diagonal performs worse than the independent case.

8. Revisiting GWAS results

We have seen in 4 that there was a clear discrepancy between the number of significant SNPs obtained from the various correction methods which raised an important question: How many SNPs should we follow up on? Whilst no multiplicity correction resulted in a large number of significant SNPs, FDR & FWER corrections yielded one or two significant SNPs. Hence, in addition to these established methods, we introduced r-power for the block diagonal testing setup as in (3), comparing the reliability of selecting top-5 SNPs. The Manhattan Plots denote the position of top-1 SNP with Bonferroni's cut-off and with Top-5 SNPs based on the ranked test statistic value, respectively.

The Manhattan Plots in Figure 5 show the SNP that was selected from the existing methods (Bonferroni,Holm,Sidak,Benjamini-Hochberg) (above) and the position of the top-5 SNPs (below). From these, we determine the confidence of these selected lists based on r-power. First, we need to estimate the proportion of null hypotheses before using r-power. For this study, we employed the Laplace-transform-based estimator from Sijuwade *et al.* (2023) for its low mean square error in comparison to other estimators. The resulting estimate yielded a null proportion of $\pi_0 = 0.9017$, indicating that there are 28,864 null hypotheses and 3,146 alternative hypotheses. Before evaluating r-power, we also need to determine the block diagonal correlation approximation from the SNP correlation matrix. We construct this by dividing the null and alternative groups and performing variable clustering. The steps involved in determining the parameters of r-power are as follows:

1. Conduct a cluster analysis on the SNPs in both the alternative and null groups based on their mean values, assuming that the null and the alternative groups are well separated.
2. Perform a clustering analysis using CLARA, an extension of the k-medoids algorithm, which is suitable for handling large-scale data Kaufman and Rousseeuw (2008).
3. Assess cluster quality using the widely adopted silhouette method to determine the optimal number of clusters.

We analyzed the alternative group, identifying the top $r = 5$ SNPs, as depicted in Figure 6 and observed two clusters displaying a wide range of correlation values.

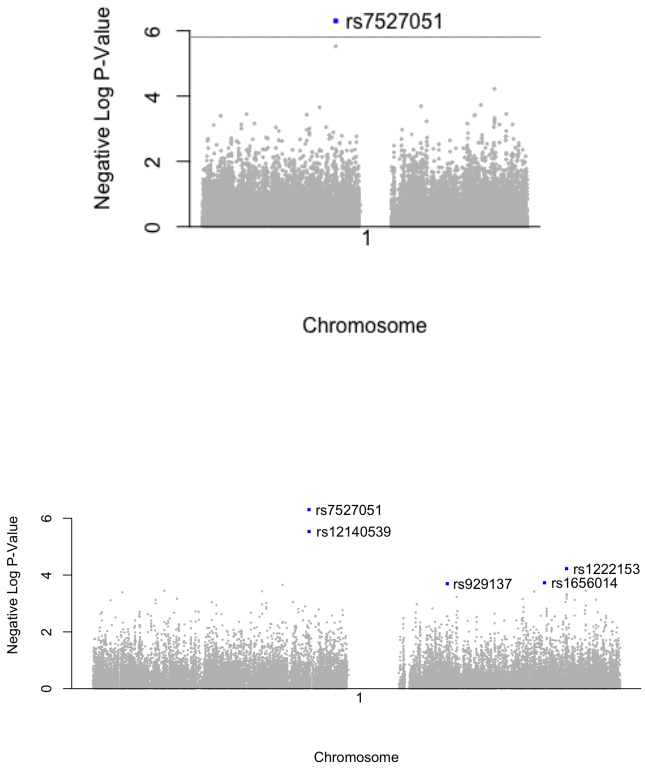


Figure 5: Bonferroni SNPs (top), Top 5 SNPs(bottom)

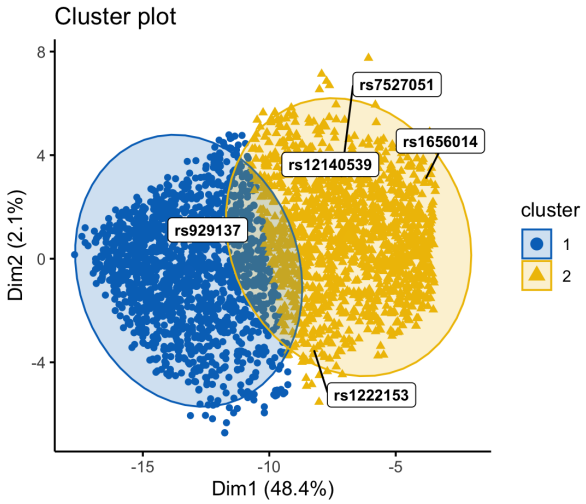


Figure 6: Cluster Plot of the Test Candidates

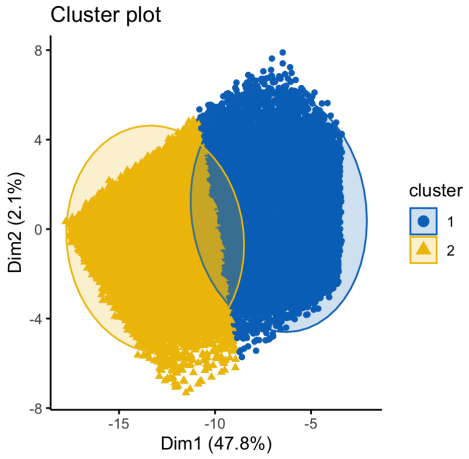


Figure 7: Cluster Plot of the Null Candidates

To enhance the analysis, we subdivided each of these clusters into different blocks of approximately equicorrelated variables using LD-pruning, resulting in an approximated block diagonal correlation matrix. LD is calculated based on R^2 values, and we considered the absolute correlation values of the SNPs since our formulation on r-power is based on positive correlation. Thus, for evaluating r-power, we consider 8 blocks, as illustrated in Table 2.

In the analysis of the null group, we also identified two clusters. However, due to the large number of candidates in the null hypothesis, calculating the correlation matrix for all SNPs was not feasible. Instead, we focused on the number of null candidates with a linkage disequilibrium (measured as R-squared) less than or equal to 0.4 from each cluster. Since our goal in calculating r-power is to ensure that none of the selected $top-r$ candidates are from the null group, we want to avoid over-penalizing the probability by considering all SNPs in the null group, regardless of their correlation. To do so, we choose SNPs with low correlation within the null group to calculate r-power. Cluster 1 in the null group which originally consisted of 14639 SNPS, has 5675 markers, with a linkage disequilibrium threshold value of 0.4. Cluster 2 from the null group, which originally consisted of 14225 SNPs and also has 5682 markers with linkage disequilibrium threshold value 0.4.

We calculated r_P for a block diagonal correlation with 10 blocks - 8 blocks from alternative and 2 from the null as illustrated in Table 2. Although we have 10 blocks, we needed to find the block allocation of the top-5 SNPs to calculate r-power. In our study, the top-5 selected SNPs are "rs7527051" , "rs12140539" ,"rs1222153" , "rs1656014" and "rs929137". The cluster allocation is described in Table 2.

Under the assumption of block diagonal correlation, the r-power for selecting the top 5 significant SNPs with an effect size of 4 was reported to be 91%, indicating a high probability of correctly identifying the relevant SNPs. At an effect size of 3, the r-power was reported to be 60%. Thus, the r-power method not only provides a powerful tool for confidently selecting relevant SNPs but also offers valuable insights into the relationship between effect size and r-power. By visualizing the r-power as a type of power curve, researchers can gain a better understanding of how to choose the optimal value for r in their r-power analysis.

Table 2: Cluster Information and Correlation Among SNPs

Cluster	Hypothesis Group	Block Size	Cluster Correlation	No of SNPs
Cluster 1	Null	5675	0.4	0
Cluster 2	Null	5682	0.4	0
Cluster 1	Alternative	56	0.1	0
Cluster 1	Alternative	921	0.5	0
Cluster 1	Alternative	232	0.7	1
Cluster 1	Alternative	437	1	0
Cluster 2	Alternative	45	0.1	0
Cluster 2	Alternative	855	0.5	2
Cluster 2	Alternative	142	0.7	1
Cluster 2	Alternative	466	1	1

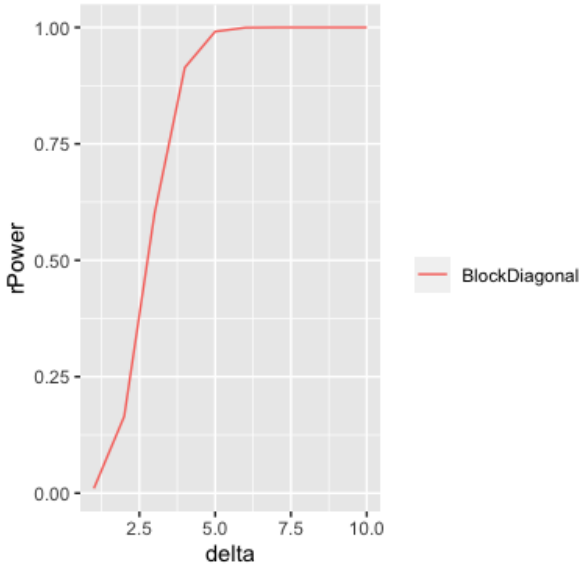


Figure 8: r-power of the top 5 selected SNPs

9. Conclusion

In this article, we addressed a fundamental issue concerning dependence with respect to the normal means problem, making positive steps towards addressing the complexity of the unstructured scenario by investigating dependence patterns, deriving analytical formulae and offering practical solutions to multiplicity issues in large-scale multiple-hypothesis testing problems. Our comprehensive simulation experiments serve to support our findings and demonstrate robustness. Our simulation results consistently show that a positive equicorrelated structure yields higher r-power compared to independence among hypotheses and that the correlation structure within blocks significantly effects the classification probability calculation.

Focusing on top tables, r-power offers insights into the robustness of the systematic selection of candidates based on combinatorial methods. We find that high within-group correlation reduces the effective dimensionality of the top-r table, in which case testing becomes more conservative and in this way, r-power provides insight into test reliability. From our findings, the correlation within the null group surpasses that of the research group, r-power under the block diagonal setup tends to outperform the equicorrelated scenario. Our formulation is built to address scenarios in which sources of variation are

difficult to identify and various features are clustered. Examples of relevant domains for future consideration and applications include but are not limited to large-scale testing within genomics, metabolomics, proteomics and fMRI studies. Our GWAS results in particular, highlight the advantage of our approach in determining test reliability compared to traditional methods, especially in in SNP detection and we are developing an R library for its implementation.

References

Abbott, D. F., Waites, A. B., Lillywhite, L. M., and Jackson, G. D. (2010). fMRI assessment of language lateralization: An objective approach. *NeuroImage*, **50**, 1446–1455.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Broto, B., Bachoc, F., Clouvel, L., and Martinez, J. M. (2020). Block-diagonal covariance estimation and application to the Shapley effects in sensitivity analysis. *arXiv:1907.12780 [math, stat]*, . arXiv: 1907.12780.

Chen, X. (2018). Estimators of the proportion of false null hypotheses: I "universal construction via Lebesgue-Stieltjes integral equations and uniform consistency under independence". *arXiv preprint arXiv:1807.03889*, .

Dasgupta, N., Lazar, N. A., and Genz, A. (2016). A look at multiplicity through misclassification. *Sankhya B*, **78**, 96–118. Publisher: Springer.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

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

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, **104**, 1406–1415.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Hartmann, M. (2017). Extending Owen's integral table and a new multivariate bernoulli distribution. *arXiv preprint arXiv:1704.04736*, .

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage*, **33**, 599–608.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, **75**, 383–386.

Jin, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**, 461–493.

Kaufman, L. and Rousseeuw, P. J. (2008). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc, New Jersey.

Kim, K. I. and van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, **9**, 1–12.

Knecht, S., Jansen, A., Frank, A., Van Randenborgh, J., Sommer, J., Kanowski, M., and Heinze, H. (2003). How atypical is atypical language dominance? *NeuroImage*, **18**, 917–927.

Kuo, C.-L. and Zaykin, D. (2013). The ranking probability approach and its usage in design and analysis of large-scale studies. *Plos One*, **8**, e83079.

Kuo, C.-L. and Zaykin, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics*, **189**, 329–340.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, **105**, 18718–18723.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.

Liu, J., Zhang, C., and Page, D. (2016). Multiple testing under dependence via graphical models. *Annals of Applied Statistics*, **1**, 1699–1724.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, **12**, 419–446.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, **15**, 1–25.

Owen, D. B. (1980). A table of normal integrals: A table. *Communications in Statistics-Simulation and Computation*, **9**, 389–419.

Pacini, C., Ajioka, J. W., and Micklem, G. (2017). Empirical Bayes method for reducing false discovery rates of correlation matrices with block diagonal structure. *BMC Bioinformatics*, **18**, 213.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Perrot-Dockès, M., Lévy-Leduc, C., and Rajjou, L. (2019). Estimation of large block structured covariance matrices: Application to "multi-omic" approaches to study seed quality. *arXiv:1806.10093 [stat]*, . arXiv: 1806.10093.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., and Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, **34**, 3769–3792.

Sarkar, S. K. (1998). Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. *Annals of Statistics*, , 494–504.

Saw, W.-Y., Tantoso, E., Begum, H., Zhou, L., Zou, R., He, C., Chan, S. L., Tan, L. W.-L., Wong, L.-P., Xu, W., et al. (2017). Establishing multiple omics baselines for three southeast asian populations in the singapore integrative omics study. *Nature Communications*, **8**, 653.

Sijuwade, A. J., Chakraborty, S., and Dasgupta, N. (2023). An inverse Laplace transform oracle estimator for the normal means problem. *Metrika*, **1**, 1–18.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer.

Smyth, G. K. and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical issues in cDNA microarray data analysis. In *Functional Genomics*, pages 111–136. Springer.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**, 2013–2035.

Storey, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 347–368.

Storey, J. D. (2011). FDR. In *International Encyclopedia of Statistical Science*, pages 504–508. Springer.

Sun, W. and Tony Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71**, 393–424.

Tong, Y. L. (1990a). Order statistics of normal variables. In Tong, Y. L., editor, *The Multivariate Normal Distribution*, Springer Series in Statistics, pages 123–149. Springer, New York, NY.

Tong, Y. L. (1990b). Positively dependent and exchangeable normal variables. In Tong, Y. L., editor, *The Multivariate Normal Distribution*, Springer Series in Statistics, pages 91–122. Springer, New York, NY.

Wang, J. and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, **19**, 629–640.

Worsley, K. (2003). Detecting activation in fMRI data. *Statistical Methods in Medical Research*, **12**, 401–418.