Computational Statistics and Data Analysis 167 (2022) 107379

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

The general goodness-of-fit tests for correlated data ☆

Hong Zhang [a], Zheyang Wu [b,*]

[a] Biostatistics and Research Decision Sciences, Merck Research Laboratories, Rahway, NJ 07065, USA
[b] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609, USA

ARTICLE INFO

Article history:
Received 30 March 2021
Received in revised form 15 September 2021
Accepted 21 October 2021
Available online 29 October 2021

Keywords:
Global hypothesis testing
Goodness-of-fit
Correlated data
Data-adaptive test
Genetic association studies

ABSTRACT

Analyzing correlated data by goodness-of-fit type tests is a critical statistical problem in many applications. A unified framework is provided through a general family of goodness-of-fit tests (GGOF) to address this problem. The GGOF family covers many classic and newly developed tests, such as the minimal p-value test, Simes test, the GATES, one-sided Kolmogorov-Smirnov type tests, one-sided phi-divergence tests, the generalized Higher Criticism, the generalized Berk-Jones, etc. It is shown that the omnibus test that automatically adapts among GGOF statistics for given data, i.e., the GGOF-O, is still contained in the GGOF family and is computationally efficient. For analytically controlling the type I error rate of any GGOF tests, exact calculation is deduced under the Gaussian model with positive equal correlations. Based on that, the effective correlation coefficient (ECC) algorithm is proposed to address arbitrary correlations. Simulations are used to explore how signal and correlation patterns jointly influence typical GGOF tests' statistical power. The GGOF-O is shown robustly powerful across various signal and correlation patterns. As demonstrated by a study of bone mineral density, the GGOF framework has good potential for detecting novel disease genes in genetic summary data analysis. Computational tools are available in the R package *SetTest* on the CRAN.

1. Introduction

As a fundamental statistical approach for the global hypothesis testing problem, the goodness-of-fit (GOF) tests have been developed and widely applied for about a century (Kolmogorov, 1933; Kotz and Johnson, 2012). They are broadly used in enormous applications for signal detection, meta-analysis, and data integration. For example, in genetic association studies, the p-values of single-nucleotide polymorphisms (SNPs) have been combined by various GOF type tests for detecting novel disease genes (He and Wu, 2011; Li et al., 2011; Yang et al., 2014). In recent years, association studies based on summary statistics have become an increasingly important strategy for dissecting the genetics of complex traits (Pasaniuc and Price, 2017). The GOF type tests are a flexible and powerful strategy for this purpose. The popularity and the potential of these tests are due to both their convenience to apply and their excellent performance. For example, a collection of GOF tests are proven asymptotically optimal for detecting weak and rare signals, e.g., the Higher Criticism (HC) (Donoho and Jin, 2004), the Berk-Jones (BJ) test (Berk and Jones, 1979), and a spectrum of ϕ-divergence tests (Jager and Wellner, 2007).

☆ Partial funding support from the National Science Foundation (NSF grants DMS-1309960 and DMS-1812082).
* Corresponding author.
E-mail addresses: hong.zhang8@merck.com (H. Zhang), zheyangwu@wpi.edu (Z. Wu).

https://doi.org/10.1016/j.csda.2021.107379
0167-9473/

The GOF tests have been broadly studied under the assumption of independence. However, correlation is a ubiquitous phenomenon in real data analyses. For example, adjacent SNPs are often correlated due to linkage disequilibrium (LD) (Reich et al., 2001), and therefore it is vital to account for the SNP correlations in such a testing process. A common strategy for addressing correlated data is to extend an existing test statistic by explicitly incorporating correlations into the new statistic. The motivations have been argued from the convenience of analytical p-value calculation or an interpretation perspective. For example, the GATES statistic (Li et al., 2011) extended Simes statistic (Simes, 1986) so that the p-value can be obtained easily under correlations. The GHC and the GBJ extended the original HC and BJ statistics, respectively, motivated by the statistic formulas' interpretations under correlations (Barnett et al., 2017; Sun and Lin, 2019). These extended test versions provided reasonable practical solutions to address correlated data analysis. At the same time, compared with the original versions, the extended statistic formulas are much more complicated and the computations are much slower. The statistical power is either very similar or has pros and cons depending on correlation patterns.

This paper presents an alternative solution for correlated data analysis based on two considerations. First, it considers the problem of p-value calculation separately from the necessity of extending original statistics. In principle, the original statistics remain legitimate to analyze correlated data. It is a computational problem of controlling type I errors for properly utilizing them under correlations. The computational problem is highly desired to resolve because many traditional GOF statistics have excellent performances for analyzing correlated data (Chicheportiche and Bouchaud, 2011). Second, instead of addressing statistics individually, the paper considers all statistics of a similar type as a whole family and provides a general framework to deal with correlations. Different GOF statistics have relative advantages under different signal and correlation patterns; therefore, an omnibus procedure that automatically adapts to statistics of complementary advantages would robustly retain high statistical power.

This paper has three main contributions. First, it establishes a framework for applying a general family of goodness-of-fit tests (the GGOF (Zhang et al., 2020)) to analyze correlated data. The GGOF family is well interpretable under correlations, following the essential idea of determining whether the input p-values have a good "fit" to their "null behavior." It is well defined based on a general statistic formula, or equivalently, an expression of rejection boundary. It broadly includes many existing tests, such as the traditional GOF tests, the minimal p-value test (minP), Simes test, the GATES, the GHC, the GBJ, the one-sided ϕ-divergence tests, etc. The testing framework includes the p-value calculation for GGOF tests. The exact calculation is deduced under positive equal-correlation. Based on that, the effective correlation coefficient (ECC) method is proposed for addressing arbitrary correlations. Besides its generality, the ECC is much faster than existing statistic-specific calculations (e.g., the effective number of independent test (ENIT) method for the GATES, and the *GBJ*-package for the HC/BJ/GHC/GBJ), and is more accurate in many situations.

As the second contribution, the paper shows that the GGOF-O, the omnibus test that automatically adapts over multiple GGOF statistics for a given data, is still in the GGOF family. Therefore, the same p-value calculation can be applied for efficient computation. Moreover, by adapting to different GGOF statistics that possess relative advantages under different signal and correlation patterns, the GGOF-O is robustly powerful in broad scenarios. For example, the GGOF-O that adapt to the minP, the HC, and the BJ is shown a faster and often more powerful test when comparing with the extended tests (such as GATES / GHC / GBJ) and some other omnibus tests (such as SKAT-O (Lee et al., 2012)).

This paper's third contribution lies in a systematic study on the influence of correlation patterns to both p-value calculation and the statistical power of typical GGOF tests. The accuracy and the limitations of related p-value calculation methods are revealed under a wide variety of correlation patterns. Similarly, the power study demonstrates how signal and correlation patterns together influence the absolute and the relative power of the GGOF tests. For example, some tests powerful only for sparse signals under independence (such as the minP and Simes) become dominating for both sparse and dense signals under reasonably strong positive correlations.

The remainder of the paper is organized as follows. In Section 2 we define the GGOF family and reveal their performances by the rejection boundaries. A universal p-value calculation approach is developed and assessed in Section 3. Section 4 presents the statistical power study. An applicational study of detecting novel genes associated with bone mineral density is given in Section 5. Section 6 concludes the work and discusses the limitations and future plans.

2. The GGOF family

The GGOF is defined following the essential idea of the goodness-of-fit testing procedure. Based on a group of *input p-values* $P_1, \cdots, P_n$, as indicated by its name, the test determines whether these input p-values has a good "fit" to a given global null hypothesis. For that, a summary *test statistic* is formulated, and the *test p-value* is obtained to decide whether or not the null is rejected. This paper considers that P_i's could be dependent, but they have an identical marginal distribution Uniform$(0, 1)$ under the null hypothesis. Under the alternative, some P_i's are stochastically smaller than Uniform$(0, 1)$ due to signals or effects, which fits the testing problems in signal detection or meta-analysis. The hypotheses can be further specified, e.g., under the Gaussian mean model to be defined in (8), where the dependences among P_i's are well characterized based on the correlation matrix and the analytical calculation for the test p-value can be deduced accordingly. Meanwhile, in principle, the GGOF remains legitimate for global testing problems in general, even without the Gaussian assumption.

Specifically, a GGOF statistic is defined by the supremum of monotone functions of ordered input p-values $P_{(1)} \leq \cdots \leq P_{(n)}$ (Zhang et al., 2020):

$$S_{n,f,\mathcal{R}} = \sup_{i\in\mathcal{R}} f(\frac{i}{n}, P_{(i)}), \tag{1}$$

where $f(\frac{i}{n}, x)$ denotes any function that is decreasing in x at each fixed i/n. The GGOF statistic is for the one-sided test, which makes sense when smaller input p-values (rather than larger ones) indicate the alternative. The smaller the input p-values (in their ordered form), the larger the statistic, and the more evidence against the global null hypothesis. $\mathcal{R} \subset \{1, ..., n\}$ represents any truncation scheme for the $P_{(i)}$'s based on the index i and/or the magnitude of $P_{(i)}$. For example,

$$\mathcal{R} = \{i : k_0 \le i \le k_1 \text{ and } \alpha_0 \le P_{(i)} \le \alpha_1\},$$

for some $k_0 \le k_1 \in \{1, \dots, n\}$ and $\alpha_0 \le \alpha_1 \in [0, 1]$. f and $\mathcal{R}$ determine a specific statistic, and thus a test when the size α is given.

Under independence, the GGOF statistics can be interpreted as a measure of the maximum departure of the $P_{(i)}$'s from their null expectations $\mathbb{E}(P_{(i)}|H_0) = i/(n+1)$. Here i/n is applied following the tradition of the GOF. The equation does not hold under dependence. However, the essential idea of the "goodness-of-fit" still follows. Because n is fixed for a given data, we can write $f(\frac{i}{n}, x) = f_i(x)$, which is decreasing in x at each i. Since smaller input p-values always indicate more departure from the null, a larger GGOF statistic provides more evidence against the null. Therefore, a GGOF statistic can still be interpreted as a legitimate measure of how well the $P_{(i)}$'s fit their null behavior.

The GGOF can be equivalently defined based on the rejection region's expression, following the monotonicity of f_i and the supremum over $\mathcal{R}$. Specifically, at each fixed i, denote the inverse function $u_i(x) = f_i^{-1}(x)$. The distribution function is

$$\mathbb{P}(S_{n,f,\mathcal{R}} \le b) = \mathbb{P}(\sup_{i\in\mathcal{R}} f_i(P_{(i)}) \le b) = \mathbb{P}(P_{(i)} > u_i(b), \text{for all } i \in \mathcal{R}). \tag{2}$$

Let b_α denote the threshold that controls the type I error rate at α, i.e., $\mathbb{P}(S_{n,f,\mathcal{R}} > b_\alpha|H_0) = \alpha$. The test will reject the null whenever the observed statistic $s_{n,f,\mathcal{R}} > b_\alpha$. Therefore, the *size-$\alpha$ rejection region* of the test is

$$\mathbf{R}_{S,\alpha} = \{P_{(i)} : P_{(i)} \le u_i(b_\alpha), \text{for at least one } i \in \mathcal{R}\}. \tag{3}$$

On the other hand, for any test whose rejection region can be written in (3), we can always find a proper function $f_i(x) = f(\frac{i}{n}, x)$ in the form of (1). Therefore, the GGOF test can be equivalently defined by the rejection region in (3). We call the series $\{u_1(b_\alpha), \dots, u_n(b_\alpha)\}$ the *rejection boundary* (RB) of this GGOF test, as rejection happens whenever a $P_{(i)}$ crosses it.

2.1. Examples of GGOF statistics

The GGOF family broadly covers many classic and newly developed tests. For example, the initial Kolmogorov-Smirnov test statistic was written based on the cumulative distribution function (CDF) and the empirical CDF (Kolmogorov, 1933). It can also be written in the form that measures the difference between i/n and $P_{(i)}$ (cf. (Shao, 2010) Section 6.5.2 and (Moscovich et al., 2016)). Specifically, a one-sided statistic is $\max_i\{i/n - P_{(i)}\}$, which corresponds to the GGOF family in (1) with the f function

$$f_i(P_{(i)}) = f_{KS^+}\left(\frac{i}{n}, P_{(i)}\right) = \frac{i}{n} - P_{(i)}.$$

Its inverse function $u_i(x) = f_i^{-1}(x) = i/n - x$. Because smaller input p-values more likely indicate signals, it is better to reweigh the absolute difference $i/n - P_{(i)}$ with regard to $P_{(i)}$. Such a scaled Kolmogorov-Smirnov statistic (also called the HC, interpreted from the perspective of multiple-testing (Donoho and Jin, 2004)) corresponds to the f function

$$f_i(P_{(i)}) = f_{HC}\left(\frac{i}{n}, P_{(i)}\right) = \frac{\sqrt{n}(\frac{i}{n} - P_{(i)})}{\sqrt{P_{(i)}(1 - P_{(i)})}}. \tag{4}$$

Its inverse function $u_i(x) = f_i^{-1}(x) = \frac{1}{2(x^2+n)}\left(x^2 - x\sqrt{x^2 + 4i(1 - i/n)} + 2i\right)$. Another well-known statistic is the one-sided Berk-Jones (BJ) statistic, where

$$f_i(P_{(i)}) = f_{BJ}\left(\frac{i}{n}, P_{(i)}\right) = \left[\frac{i}{n}\log\left(\frac{\frac{i}{n}}{P_{(i)}}\right) + \left(1 - \frac{i}{n}\right)\log\left(\frac{1 - \frac{i}{n}}{1 - P_{(i)}}\right)\right]\{P_{(i)} < \frac{i}{n}\}.$$

Here, even though $u_i(x)$ does not have a simple closed-form, it can be numerically obtained because $f_i(x)$ is strictly decreasing at each i.

Furthermore, the family of ϕ-divergence statistics introduced by (Jager and Wellner, 2007) follow the format in (1), where the f function is indexed by a statistic-defining parameter s:

$$f_s^{\phi}(\frac{i}{n}, P_{(i)}) = \frac{1}{s(1-s)}\left(1 - (\frac{i}{n})^s P_{(i)}^{1-s} - (1-\frac{i}{n})^s (1-P_{(i)})^{1-s}\right), s \neq 0, 1,$$

$$f_1^{\phi}(\frac{i}{n}, P_{(i)}) = \frac{i}{n}\log\left(\frac{\frac{i}{n}}{P_{(i)}}\right) + \left(1-\frac{i}{n}\right)\log\left(\frac{1-\frac{i}{n}}{1-P_{(i)}}\right), \text{and}$$

$$f_0^{\phi}(\frac{i}{n}, P_{(i)}) = P_{(i)}\log\left(\frac{P_{(i)}}{\frac{i}{n}}\right) + (1-P_{(i)})\log\left(\frac{1-P_{(i)}}{1-\frac{i}{n}}\right).$$

Since only small p-values $P_{(i)} < \frac{i}{n}$ (instead of large p-values) indicate signals or effects, it is appropriate to consider the one-sided version ϕ-divergence statistics, which are covered in the GGOF family. A version of the one-sided statistics is (Zhang et al., 2020)

$$f_s(\frac{i}{n}, P_{(i)}) = \begin{cases} \sqrt{2n f_s^{\phi}(\frac{i}{n}, P_{(i)})} & P_{(i)} \leq \frac{i}{n}, \\ -\sqrt{2n f_s^{\phi}(\frac{i}{n}, P_{(i)})} & P_{(i)} > \frac{i}{n}, \end{cases}$$

which exactly covers the HC and the reverse HC statistics (Donoho and Jin, 2008) for $s = 2$ and -1, respectively: It also covers the Berk-Jones (BJ) and the reverse Berk-Jones statistics (Berk and Jones, 1979) for $s = 1$ and 0, respectively.

Based on the HC and the BJ, the GHC statistic and the GBJ statistic were created to explicitly account for correlations (Barnett et al., 2017; Sun and Lin, 2019). Their designs were primarily motivated from the interpretation perspective. For example, GHC considers a special case that two-sided input p-values come from normally distributed input statistics: $T_i \sim N(0,1)$ and $P_i = 2P(N(0,1) > |T_i|) = 2\bar{\Phi}(|T_i|)$, $i = 1, \cdots, n$. In this case, the HC statistic can be equivalently written as

$$HC = \sup_{t \in \mathcal{R}^*} \frac{S(t) - 2n\bar{\Phi}(t)}{\sqrt{2n\bar{\Phi}(t)(1-2\bar{\Phi}(t))}},$$

where $S(t) = \sum_i \{|T_i| > t\}$. The denominator $\sqrt{2n\bar{\Phi}(t)(1-2\bar{\Phi}(t))}$ is the standard deviation of the binomial random variable $S(t)$ under independence. Therefore, the idea of the GHC is to replace the denominator with the estimated standard deviation of $S(t)$ under correlation (with no truncation, i.e., $\mathcal{R} = \{1 \leq i \leq n\}$). Similarly, the BJ statistic can be viewed as a function of probabilities regarding $S(t)$. The GBJ statistic is to replace these probabilities with the versions that are conditional on the correlated input statistics (with fixed $\mathcal{R} = \{1 \leq i \leq n/2\}$). The GHC and GBJ statistics' formulas are rather complicated, but their rejection regions can be written in the form of (3). Therefore, they are included in the GGOF family.

The GGOF family also covers some tests that may not traditionally be considered goodness-of-fit tests. For example, the minimal p-value test (minP) takes $P_{(1)}$ as the statistic. Simes test takes $\min_i\{P_{(i)} \cdot n/i\}$ as the statistic (Simes, 1986). Moreover, an extended Simes procedure (called the GATES) was proposed to accommodate correlations (Li et al., 2011). The GATES statistic is $\min_i\{P_{(i)} \cdot n_e/i_e\}$, where i_e is the "effective number of independent p-values" among $P_{(1)}, \cdots, P_{(i)}$, $i = 1, ..., n$. It is straightforward to show that all three statistics can be equivalently expressed in the format of (1), and their rejection regions follow the form of (3). Therefore, they belong to the GGOF family.

2.2. The GGOF-O

The GGOF-O is a data-adaptive omnibus test that automatically selects function f and truncation scheme $\mathcal{R}$ for a given data. This procedure is important to practice because different f and $\mathcal{R}$ often possess advantages for different signal patterns. For example, it has been shown that the minP, Simes, and the HC are more powerful for sparser signals, while the BJ is more powerful for denser signals (Li and Siegmund, 2015; Zhang et al., 2020). The p-value truncation scheme $\mathcal{R}$ is also quite relevant. For example, the "modified HC" with $\mathcal{R} = \{1 < i \leq n/2, P_{(i)} \geq 1/n\}$ and the original HC that does not impose the truncation $\{P_{(i)} \geq 1/n\}$ have different merits under different data and signal patterns (Donoho and Jin, 2004; Li and Siegmund, 2015; Zhang et al., 2020). Ideally, by allowing general f and $\mathcal{R}$, the GGOF family will retain all merits to gain robustly high power in analyzing various data.

Define function $G_{n,f,\mathcal{R}}(x) = \mathbb{P}(S_{n,f,\mathcal{R}} > x|H_0)$. The test p-value at an observed statistic $s_{n,f,\mathcal{R}}$ is $G_{n,f,\mathcal{R}}(s_{n,f,\mathcal{R}})$. For a given data of fixed n, the GGOF-O statistic is the smallest test p-value (representing the highest significance) over a set of GGOF statistics indexed by f and $\mathcal{R}$:

$$S_{n,o} = \inf_{f,\mathcal{R}} G_{n,f,\mathcal{R}}(S_{n,f,\mathcal{R}}). \tag{5}$$

Deduction shows that for any constant $s_0 \in [0,1]$,

$$\mathbb{P}(S_{n,o} > s_o) = \mathbb{P}(P_{(1)} > u_1^{\star}(s_o), \ldots, P_{(n)} > u_n^{\star}(s_o)), \tag{6}$$

where for each $i = 1, \ldots, n$,

$$u_i^\star(s_o) = \sup_{f,\mathcal{R}}\{u_{f,\mathcal{R},i}(s_o)\}, \text{ with } u_{f,\mathcal{R},i}(s_o) = f_i^{-1}(G_{n,f,\mathcal{R}}^{-1}(s_o)).$$

Let $s_{o,\alpha}$ denote the threshold that controls the type I error rate of the omnibus test at α, i.e., $\mathbb{P}(S_o < s_{o,\alpha}|H_0) = \alpha$ (note that the smaller the $S_{n,o}$ the more significant the test is). The size-α rejection region of the GGOF-O is

$$\mathbf{R}_{S_o,\alpha} = \{P_{(i)} : P_{(i)} \le u_i^\star(s_{o,\alpha}), \text{for at least one } i \in \{1,\dots,n\}\}. \tag{7}$$

The RB of the GGOF-O is $\{u_1^\star(s_{o,\alpha}),\dots,u_n^\star(s_{o,\alpha})\}$. Following the definition of the GGOF by (3), the above equation indicates that the GGOF-O also belongs to the GGOF family. This self-containing property is attractive in both theory and practice. The GGOF related analytical and computational techniques can be directly applied to the GGOF-O.

For practical convenience we consider the GGOF-O over a finite number of m GGOF statistics. Specifically, m GGOF statistics S_j, $j = 1,\cdots,m$, are defined by the statistical functions $f_{ji}, i = 1,\cdots,n$, and the truncation schemes $\mathcal{R}_j$. Accordingly, the GGOF-O statistic expression in (5) is simplified to

$$S_o = \min_j G_j(S_j),$$

where $G_j(x) = \mathbb{P}(S_j > x|H_0)$. Regarding the survival function of S_o in (6), the boundaries are now

$$u_i^\star(s_o) = \max_j u_{ji}(s_o),$$

where

$$u_{ji}(s_o) = \begin{cases} f_{ji}^{-1}(G_j^{-1}(s_o)) & \text{if } i \in \mathcal{R}_j \\ 0 & \text{if } i \notin \mathcal{R}_j. \end{cases}$$

2.3. Rejection boundaries

Through the rejection boundaries in (3) and (7), we can see why the GGOF tests could work well for various signal patterns at given n, α, and correlations. Because H_0 is rejected whenever at least one $P_{(i)}$ goes below the RB, a higher RB curve indicates a higher statistical power at a given setting. At the same time, signal patterns relate to how the input p-values will likely be associated with true signals. For example, strong and sparse signals are likely associated with the top-ranked input p-values, i.e., the $P_{(i)}$'s at small i values, while weak or dense signals are often associated with $P_{(i)}$'s at large i values.

Fig. 1 illustrates the RBs at the logarithm scale for four GGOF statistics: the HC, the BJ, the minP, and the GGOF-O that adapts to the previous three statistics. The minP and the HC have higher RBs at the top-ranked input p-values. Therefore, these statistics are more sensitive (and thus have higher statistical power) to strong and sparse signals. On the other hand, the BJ is more sensitive to the lower-ranked p-values, indicating its advantage to detect weaker or denser signals. The GGOF-O provides a robust solution because over $i = 1,\cdots,n$ its RB is point-wisely close to (but slightly lower than) the highest RB among the adapted statistics. This phenomenon can be explained by formulas. Specifically, at a given significance level α, the RB of the GGOF-O is

$$u_i^\star(s_{o,\alpha}) = \max_j f_{ji}^{-1}(G_j^{-1}(s_{o,\alpha})),$$

while the RB of the j's involved GGOF statistic is $f_{ji}^{-1}(G_j^{-1}(\alpha))$. The maximum makes $u_i^\star(s_{o,\alpha})$ close to the highest RB $\max_j f_{ji}^{-1}(G_j^{-1}(\alpha))$ point-wisely. Meanwhile, $u_i^\star(s_{o,\alpha})$ is still slightly lower because $s_{o,\alpha} \le \alpha$, which indicates a "trade-off" due to the stochastic selection process among multiple involved GGOF statistics. Another interesting observation is that when data correlation increases (comparing the three panels from left to right), the RB curves of the HC and the BJ are dropping while the minP's curve is rising. This pattern means that the minP gains more advantage when data are more strongly correlated.

Fig. 2 gives the RBs of the ϕ-divergence statistics with $s \in \{-2,-1,0,1,2,3\}$. The RB curves show an interesting pattern: statistics with larger s have higher RBs at the top-ranked p-values. If signals are strong and sparse, they are among the smallest input p-values. Therefore, these statistics are more sensitive to such signals and have higher statistical power. On the other hand, statistics with small s are more sensitive to the lower-ranked p-values, indicating their advantages to detect weaker or denser signals. Such observation also indicates that a proper p-value truncation domain $\mathcal{R}$ could benefit the statistic power by focusing on certain ordered input p-values. As shown in the right panel, the GGOF-O test over various f_s functions provides a more balanced and robust solution. At each position i, the RB curve of the GGOF-O test is close to the highest RB at that point. Furthermore, the GGOF-O adapting over a dense grid of s values (e.g., $s = -1, 0, 1, 2, 3$) is similar as that adapting over a sparse grid (e.g., $s = 1, 2$). This observation indicates that the omnibus procedure does not have to include many similar statistics. For both easy computation and high power, it is better to include fewer statistics that have complementary sensitiveness to broader signal patterns.

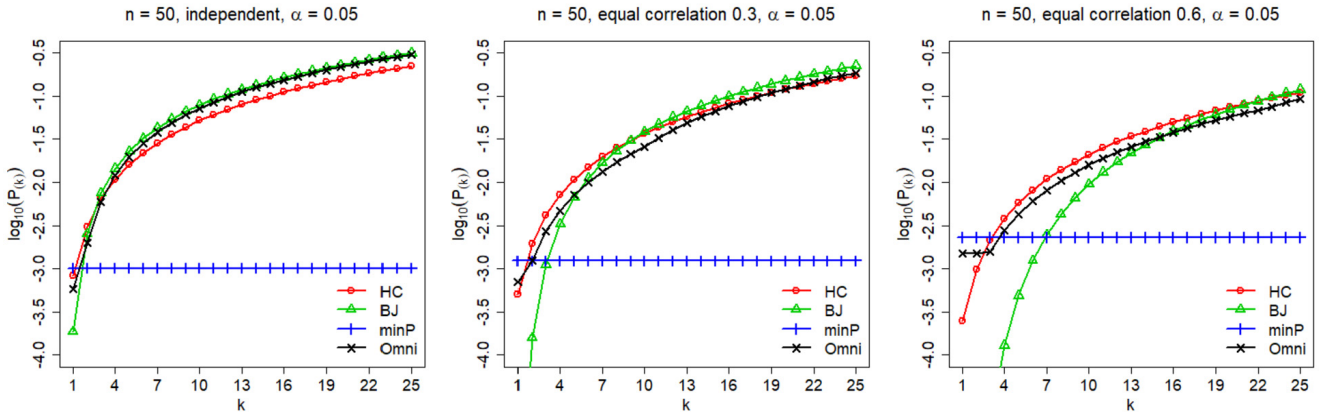

Fig. 1. The log-scaled rejection boundary (RB) of the minP, the HC, the BJ, and the GGOF-O (Omni) adapting to them. The null hypothesis is rejected as long as one value of $\log_{10}(P_{(i)})$ is below the boundary. Panels from left to right: equal-correlations among input statistics with values 0 (i.e., independence), 0.3 and 0.6, respectively. $\alpha = 0.05$.

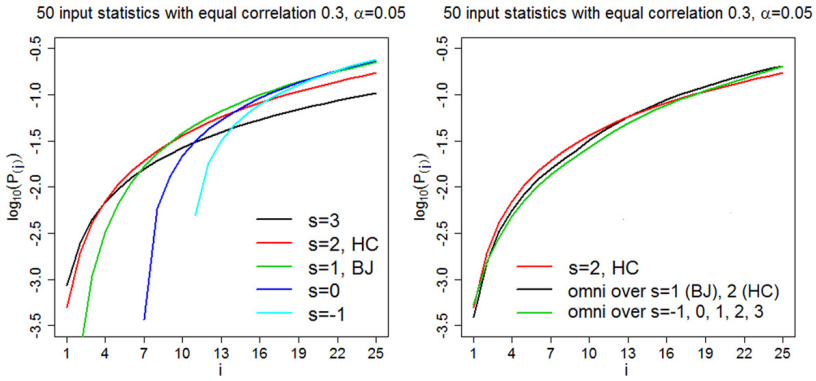

Fig. 2. The log-scaled rejection boundaries (RB) of the ϕ-divergence statistics and the corresponding GGOF-O tests are compared. The null hypothesis is rejected as long as one value of $\log_{10}(P_{(i)})$ is below the boundary. Left panel: RBs for ϕ-divergence statistics with $s = 3$ (black), 2 (i.e., the HC, red), 1 (i.e., the BJ, green), 0 (i.e., reverse BJ, blue) and -1 (i.e., reverse HC, cyan). Right panel: RBs for the HC (red), the GGOF-O over the HC and the BJ (black), and the GGOF-O over ϕ-divergence statistics with $s \in \{-1, 0, 1, 2, 3\}$ (green). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

3. The *p*-value calculations

This section proposes an analytical approach for calculating the test p-value of any GGOF test under the Gaussian mean model (GMM). The calculation is exact under positive equal-correlation. Based on that, an approximation method, the effective correlation coefficient (ECC), is proposed for arbitrary correlation patterns.

The GMM is a typical setting for characterizing correlated data (Hall and Jin, 2010). It assumes that n *input test statistics* $T_1, \ldots, T_n$ are multivariate Gaussian, i.e.,

$$T = (T_1, \ldots, T_n) \sim N(\boldsymbol{\mu}, \Sigma), \tag{8}$$

where Σ is a correlation matrix (assuming T_i's are standardized with unit variance). Σ is known or can be reliably estimated (e.g., in the context of linear models with large sample size). The global null hypothesis is about the unknown vector $\boldsymbol{\mu}$, i.e., $H_0 : \boldsymbol{\mu} = 0$. The input p-values could be one- or two-sided depending on specific scientific problem.

$$\text{One-sided: } P_i = \bar{\Phi}(T_i); \quad \text{Two-sided: } P_i = 2\bar{\Phi}(|T_i|), i = 1, \ldots, n, \tag{9}$$

where $\Phi(x)$ denotes the CDF of $N(0, 1)$ and $\bar{\Phi}(x) = 1 - \Phi(x)$. We always have marginal $P_i \sim \text{Uniform}(0, 1)$ under H_0. However, the type of p-values in (9) and Σ together determine the dependence among P_i's.

3.1. Exact calculation under positive equal correlation

The distribution of the GGOF statistic in (2) and that of the GGOF-O statistic in (6) are both about a cross-boundary probability of $P_{(i)}$'s, $i = 1, \cdots, n$. We give its calculation under positive equal correlation. Specifically, consider the input statistics T in (8) with $\mu = 0$, $\Sigma_{ij} = \rho$, $0 \le \rho < 1$, for all $i \ne j$. Denote $U_{(1)} \le \cdots \le U_{(n)}$ the order statistics of $U_i \overset{iid}{\sim} \text{Uniform}(0, 1)$, $i = 1, \cdots, n$. Denote $\phi(z)$ the density function of $N(0, 1)$. We have

$$\mathbb{P}(P_{(i)} > u_i, \text{for all } i \in \mathcal{R}) = \int_{-\infty}^{\infty} \mathbb{P}(U_{(i)} > c_i(z), \text{ for all } i \in \mathcal{R})\phi(z)dz, \tag{10}$$

where

$$c_i(z) = \bar{\Phi}\left(\frac{\bar{\Phi}^{-1}(u_i) - \sqrt{\rho}z}{\sqrt{1-\rho}}\right)$$

for the one-sided P_i's, or

$$c_i(z) = \bar{\Phi}\left(\frac{\bar{\Phi}^{-1}(u_i/2) - \sqrt{\rho}z}{\sqrt{1-\rho}}\right) + \bar{\Phi}\left(\frac{\bar{\Phi}^{-1}(u_i/2) + \sqrt{\rho}z}{\sqrt{1-\rho}}\right)$$

for the two-sided P_i's in (9). The rejection boundary $u_i = u_i(b)$ for the GGOF in (2) or $u_i = u_i^*(s_o)$ for the GGOF-O in (6). The proof of (10) is given in Appendix A.1.

For the cross-boundary probability $\mathbb{P}(U_{(i)} > c_i, \text{ for all } i \in \mathcal{R})$ in (10), we have developed both exact and approximated calculations (Zhang et al., 2020). For example, for $\mathcal{R} = \{i : k_0 \le i \le k_1\}$, $k_0 \le k_1 \in \{1, \ldots, n\}$, the exact calculation is given in the following. Denote $F_{B(\alpha,\beta)}(x)$ the CDF of Beta(α, β) distribution and $\bar{F}_{B(\alpha,\beta)}(x) = 1 - F_{B(\alpha,\beta)}(x)$. Let $m = n - k_1 + 1$. Define

$$a_{k_1} = \frac{n!}{(n-k_1+1)!}\bar{F}_{B(1,m)}(c_{k_1}) \text{ and}$$
$$a_k = \frac{n!}{(n-k+1)!}\bar{F}_{B(k_1-k+1,m)}(c_{k_1}) - \sum_{j=1}^{k_1-k}\frac{c_{k+j-1}^j}{j!}a_{k+j} \text{ for } k = k_1 - 1, \cdots, 1.$$

We have

$$\mathbb{P}(U_{(i)} > c_i, \text{ for all } i = k_0, \cdots, k_1) = \bar{F}_{B(k_1,m)}(c_{k_1}) - \sum_{i=k_0}^{k_1-1}\frac{c_i^i}{i!}a_{i+1}.$$

3.2. ECC approximation under arbitrary correlations

In real data analysis, the structure of Σ is often more complicated than equal correlation. We propose a method to approximate the distribution of any GGOF statistic under arbitrary Σ. The method takes advantage of the calculation under the equal correlation in (10), so we call it the *effective correlation coefficient* (ECC) method. The ECC is somewhat analogous to the method of the effective number of independent tests (ENIT), which was developed initially for the correlated multiple-hypothesis testing problem (Bailey and Grundy, 1999), and was also utilized for global testing problem (Li et al., 2011).

The idea of the ENIT is that n dependent p-values could be approximated by n_e independent p-values (i.e., the effective number of independent tests). Obviously, $n_e = n$ under independence, and $n_e = 1$ under perfect correlation because all p-values are equal. For the correlations in-between, one could find $n_e \in (1, n)$ by interpolation. Many interpolation methods have been developed based on the ordered eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ of the correlation matrix Σ. For example, the classic Cheverud's ENIT consider n_e as a linear function of λ_i's sample variance: $s^2 = \sum_i(\lambda_i - 1)^2/(n-1)$ (Cheverud, 2001). Following the fact that $n_e = n$ and 1 correspond to $s^2 = 0$ and n, respectively, a linear interpolation gives

$$n_e = n - (n-1)s^2/n. \tag{11}$$

Other ENIT methods have been proposed (Galwey, 2009; Li and Ji, 2005; Moskvina and Schmidt, 2008). In the GATES test, a calculation is recommended (with λ_i obtained from the correlation matrix of p-values):

$$n_e = n - \sum_{i=1}^{n}(\lambda_i - 1)I(\lambda_i > 1). \tag{12}$$

The ECC method is to find a proper positive equal-correlation matrix Σ_ρ as the surrogate of Σ, then use the calculation in (10) to approximate the null distribution of any GGOF statistic. As an analogy to n_e, we call ρ the effective correlation coefficient. Σ is mapped to Σ_ρ by matching proper features of their eigenvalues, which guarantees two restraints: $\rho = 0$ under independence and $\rho = 1$ under perfect correlation. In this paper we consider a flexible framework that matches the L_r-norm of the magnitudes of centralized eigenvalues: $\sum_{i=1}^n |\lambda_i - 1|^r$, $r > 0$. Because the largest eigenvalue of Σ_ρ equals $1 + (n-1)\rho$ and the rest eigenvalues equal $1 - \rho$, we have

$$(n-1)^r\rho^r+(n-1)\rho^r=\sum_{i=1}^{n}|\lambda_i-1|^r \Rightarrow \rho(r)=\left(\frac{\sum_{i=1}^{n}|\lambda_i-1|^r}{(n-1)^r+(n-1)}\right)^{1/r}. \tag{13}$$

This matching has several meaningful cases. When $r=1$, the ECC is consistent with the ENIT of the GATES in (12) because both measure the dependence based on the eigenvalues that are larger than 1. To see this, denote i^* the largest index such that $\lambda_i>1$. Since $\sum_i\lambda_i=n$, we have

$$\sum_{i=1}^{n}|\lambda_i-1|=\sum_{i=1}^{i^*}(\lambda_i-1)+\sum_{i=i^*+1}^{n}(1-\lambda_i)=2\sum_{i=1}^{n}(\lambda_i-1)I(\lambda_i>1).$$

So,

$$\rho=\frac{1}{n-1}\sum_{i=1}^{n}(\lambda_i-1)I(\lambda_i>1).$$

When $r=2$, we have $\rho^2=s^2/n$, which is consistent with Cheverud's ENIT in (11) in the sense that both methods capture the character of the correlations by s^2. Moreover, a large r value captures the correlations by the largest eigenvalue λ_1 because $\lim_{r\to+\infty}\rho(r)=\frac{\lambda_1-1}{n-1}$. This choice is also reasonable since λ_1 is often considered as a main characteristic of the correlation matrix (Friedman and Weisberg, 1981).

In principle, by choosing different r values, we can control the correlation strength $\rho(r)$, and thus the inflation or conservativeness of a GGOF's p-value. Particularly, when $r>1$ and n is reasonable large, we have

$$\rho(r)=\frac{\left(\sum_{i=1}^{n}|\lambda_i-1|^r\right)^{1/r}}{n-1}\left(\frac{(n-1)^r}{(n-1)^r+(n-1)}\right)^{1/r}\approx\frac{1}{n-1}\left(\sum_{i=1}^{n}|\lambda_i-1|^r\right)^{1/r}.$$

In this case $\rho(r)$ is decreasing in r. Accordingly, an r value larger than 2 could be interpreted as an option that balances between $r=2$ (which captures correlations by s^2) and $r=\infty$ (which captures correlations by λ_1). Through extensive simulations (e.g., Supplementary Figure S1), we found that $r=3$ is a robust choice over different test statistics and correlation structures.

3.3. Computation time

Computation time for a GGOF procedure involves two components: the time of getting the statistic (or equivalently, getting the u_i's in (2) or (6)) and the time of calculating the test p-value (i.e., getting the cross-boundary probability itself). Fig. 3 shows the total computation time of these two components for several typical GGOF tests. All computations were conducted using R 3.5.0 on an Intel Core i5-8350U CPU at 1.70 GHz with 16 GB RAM.

Comparing with the HC and the BJ statistics, their extended versions, the GHC and the GBJ statistics, are significantly slower to compute (see HC_GP vs. GHC_GP and BJ_GP vs. GBJ_GP). Moreover, the p-value calculation methods designed for GHC and GBJ (implemented in *GBJ*-package (GP) (Barnett et al., 2017; Sun and Lin, 2019)) can be applied to the HC and the BJ. They are much slower than the ECC when n is large (see HC_GP vs. HC_ECC and BJ_GP vs. BJ_ECC). It is worth mentioning that the real differences are much more significant than those demonstrated in the figure because the ECC was implemented in R, while the cross-boundary probability calculation in the *GBJ*-package was implemented in a much faster language C++. Similarly, the GATES is much slower than Simes_ECC. The GATES needs to calculate i_e, which is computationally expensive, and it needs to repeat for all $i=1,\cdots,n$. In principle, the ENIT method could be applied to GGOF tests with a necessary "extension", such as extending Simes to the GATES. However, in principle, it will be computationally intensive due to the need to get i_e's. Overall, keeping neat and simple test statistics, such as the classic ones like the HC, the BJ, and Simes, has a clear computational benefit.

Under the GMM, the minP's test p-value can be calculated based on the multivariate normal distribution (minP_MVN, implemented by R package *mvtnorm*). However, that is also much slower than the ECC (minP_ECC). In general, the ECC is efficient. It retains the same computation for any given cross-boundary probability in (2), and it is relatively less sensitive to the increase of n (as shown by the flatter slope of the time curves). Based on the ECC, the GGOF-O test is faster than the GHC, the GBJ, the minP_MVN, and the GATES when n is large. In principle, the GGOF-O is efficient because the boundaries $u_i^\star$ of the GGOF-O in (6) are directly obtained based on the u_i functions of the individual statistics to be adapted. There is no cost for estimating the correlations among these individual statistics (and also no loss of accuracy), which is often needed in other omnibus testing procedures (Sun and Lin, 2019).

3.4. Accuracy of p-value calculations

In this section, we apply extensive simulations to comprehensively evaluate the influence of relevant factors under the GMM and the linear models.

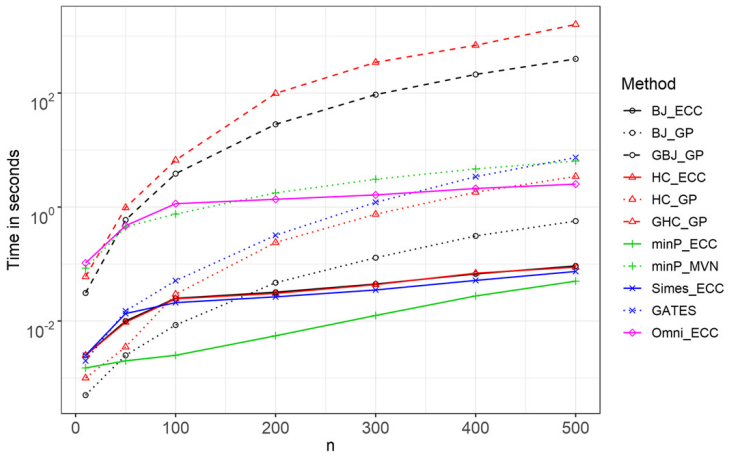

Fig. 3. The computation time for combinations of statistics (HC, BJ, GHC, GBJ, minP, Simes, GATES) and p-value calculation methods (GP: *GBJ*-package; ECC (solid curves); MVN: multivariate normal distribution implemented by R package *mvtnorm*). Correlation matrix $\Sigma = D_n(1)$ as specified in Table 1. Y-axis: real computation time; x-axis: dimension n.

Table 1
Correlation matrix Σ involved in GMM simulations. Three organizational types based on (14) and (15): Whole matrix, upper-block matrix (type I), two-block matrix (type II).

Type	Whole matrix	Upper-block (I)	Two-block (II)
Equal(ρ)	$\Sigma = E_n(\rho)$	$\Sigma_{11} = E_{n/2}(\rho)$; $\Sigma_{22} = I$	$\Sigma_{11} = \Sigma_{22} = E_{n/2}(\rho)$
Poly(κ)	$\Sigma = D_n(\kappa)$	$\Sigma_{11} = D_{n/2}(\kappa)$; $\Sigma_{22} = I$	$\Sigma_{11} = \Sigma_{22} = D_{n/2}(\kappa)$
Inv-Equal(ρ)*	$\Sigma = (E_n(\rho))^{-1}$	$\Sigma_{11} = (E_{n/2}(\rho))^{-1}$; $\Sigma_{22} = I$	$\Sigma_{11} = \Sigma_{22} = (E_{n/2}(\rho))^{-1}$
Inv-Poly(κ)*	$\Sigma = (D_n(\kappa))^{-1}$	$\Sigma_{11} = (D_{n/2}(\kappa))^{-1}$; $\Sigma_{22} = I$	$\Sigma_{11} = \Sigma_{22} = (D_{n/2}(\kappa))^{-1}$

* Σ is standardized to become a correlation matrix with unit diagonal.

3.4.1. Under the GMM

We assess the influences of correlations Σ, the significance level α, the number of input p-values n, and indeed the test statistics themselves. Before going to the details of many settings and results, let us briefly summarize the findings. First, overall the ECC is an adequate universal approach for GGOF tests under low and medium correlations. It is often more accurate than the *GBJ*-package for HC, GHC, BJ, and GBJ, and the ENIT for GATES. Second, under strong block-wise correlations, all these methods face the challenge of controlling stringent α (< 0.001). In this case, one practical solution is to test block by block. For example, in genetic studies, it is common to estimate independent LD blocks and study them separately (Berisa and Pickrell, 2016). Third, different tests have different sensitiveness to data correlations. For example, the HC/GHC, Simes, and minP are less sensitive to correlations. In contrast, the BJ/GBJ are much more sensitive (notably, the GBJ is challenging to control under negative correlations).

We considered two basic correlation patterns: equal and polynomial-decaying correlations, which represent relatively dense and sparse correlations in the matrix, respectively:

$$\text{Equal-correlation } E_k(\rho) : E_k(i,j) = \rho, 1 \le i \ne j \le k \text{ and } 0 \le \rho < 1, \tag{14}$$

$$\text{Polynomial-decay } D_k(\kappa) : D_k(i,j) = 1/|i-j|^{\kappa}, 1 \le i \ne j \le k \text{ and } \kappa > 0. \tag{15}$$

Two parameters ρ and κ are set to control correlation strength. We further considered negative correlations by matrix inversion. Specifically, $(E_k(\rho))^{-1}$ is an equal correlation matrix with negative entries, $(D_k(\kappa))^{-1}$ has negative correlations with the magnitude decaying roughly at a polynomial speed (Hall and Jin, 2010; Sun, 2005). We also organized Σ in three ways. First, the whole matrix Σ follows the correlation patterns described above. Second, we considered that Σ follows certain block-wise structures, which are often interested in practice (e.g., the LD blocks in genetics). Specifically, consider Σ a 2×2 block matrix

$$\Sigma_{n\times n} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix},$$

where each block is a $(n/2) \times (n/2)$ matrix. Two blocking types were considered: (I) The upper block matrix Σ_{11} follows the above correlation patterns and the lower block $\Sigma_{22} = I$ for independence; (II) $\Sigma_{11} = \Sigma_{22}$, both follow above correlation patterns. In total, 12 structures of Σ are summarized in Table 1. The dimension $n = 10$, 50 or 100. Two-sided input p-values were simulated for comparison because the GATES and the GHC/GBJ are for two-sided p-values only.

We used the ratio between the empirical type I error rate and the nominal α to measure the accuracy of p-value calculation. The empirical type I error rate is the proportion of the calculated p-values that are smaller than the nominal

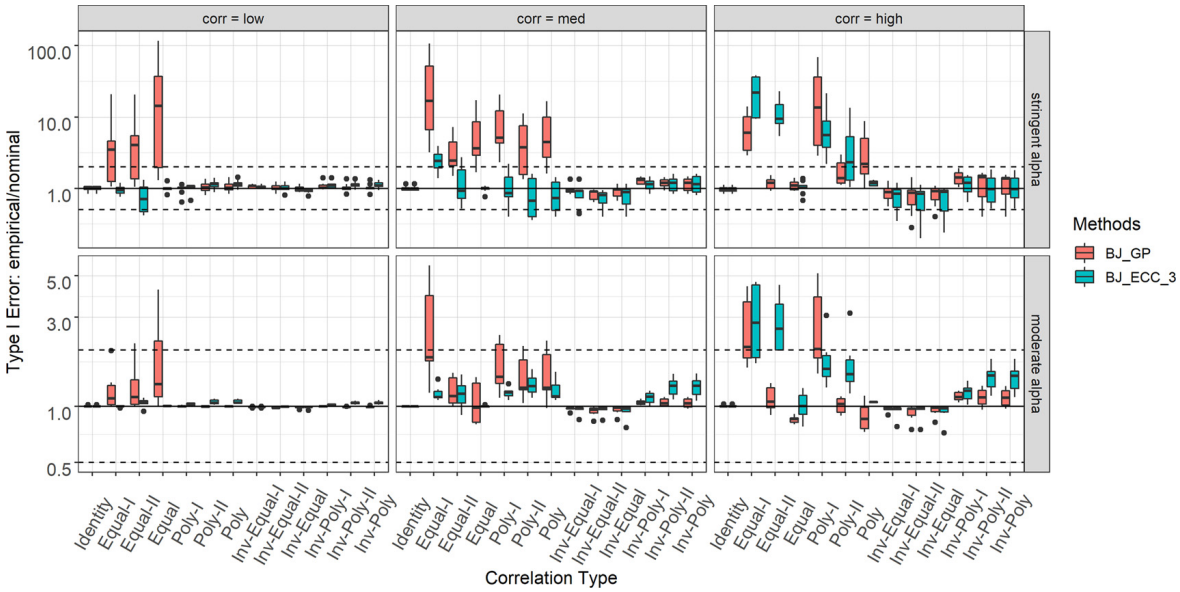

Fig. 4. The ratio between empirical type I error rate and nominal α for BJ. In each panel, the y-axis is the ratio between the empirical type I error rate and the nominal α. The two dash lines mark a range of "adequate" ratios between 0.5 and 2. The x-axis lists the independence case ($\Sigma = I$) and the 12 correlation patterns in Table 1. The rows of panels are arranged according to significance levels: "moderate α" combines results of $\alpha \geq 0.001$, and "significant α" combines results of $\alpha \leq 10^{-4}$. The columns of panels are arranged by correlation strength: low, medium, and high, corresponding to $\rho = 0.1, 0.5, 0.9$ (or $\kappa = 3, 1, 0.2$), respectively. GP: *GBJ*-package; ECC_3: the ECC with $r = 3$.

α. A ratio around 1 indicates accuracy; a ratio significantly larger (or smaller) than 1 indicates an inflated (or conservative) test because the rejections are more (or fewer) than expected. Here 1×10^7 simulations were carried out and $\alpha = 0.05$, 0.01, 0.001, 10^{-4}, 10^{-5} and 2.5×10^{-6} were considered.

The accuracy of the ECC versus the *GBJ*-package (GP) for the BJ test is demonstrated in Fig. 4. When correlations are low or medium, the ECC is primarily adequate, and its improvements over the *GBJ*-package are significant. Meanwhile, the block-wise positive high correlations (e.g., Equal-I and Equal-II with $\rho = 0.9$) remain challenging. Comparisons for other statistics are given in Supplementary Materials: the HC, the GHC, the GBJ, the GATES, Simes, minP, and the GGOF-O (Supplementary Figures S2–S7). Similarly, the ECC is mostly adequate and often improves the accuracy under various settings, unless under block-wise positive high correlations. Further Supplementary figures show the comparisons from different angles too. In particular, Figures S8 – S12 illustrate the results at each α level, indicating that smaller α is often harder to control. Figures S13 – S17 make the comparisons across different n values, indicating that the comparison patterns are consistent over different n values. Figures S18 – S22 make comparisons after combining all correlation patterns to represent a comprehensive scenario of overall applications. The ECC is shown reasonably acceptable in general.

3.4.2. Under linear models

To better understand the performance of relevant p-value calculations in practice, we considered the linear-model based analysis, especially in genetic association studies. Specifically, we simulated 1,290 haplotypes for a genome region of 250k base-pairs by the Cosi2 package following a coalescent model with the typical LD pattern of the European population (Shlyakhter et al., 2014). Two haplotypes were randomly drawn with replacement from the haplotypes to form each subject's genotypes. Genes to be tested were defined at random locations containing n consecutive rare SNPs (with the minor allele frequency (MAF) $< 5\%$) (McClellan and King, 2010).

Under the null of no genetic associations, quantitative and binary traits of the ith individual, $i = 1, \cdots, N$, were generated from regression and logit model:

$$\text{Quantitative trait: } Y_i = 0.5Z_{1i} + 0.1Z_{2i} + \epsilon_i, \text{ where } \epsilon_i \sim N(0,1), \tag{16}$$

$$\text{Binary trait: } \text{logit}(P(Y_i = 1)) = -1.25 + 0.5Z_{1i} + Z_{2i}, \tag{17}$$

where two independent environmental factors were assumed: $Z_1 \sim$Bernoulli(0.5) is a binary factor and $Z_2 \sim N(0,1)$ is a continuous factor. Two-sided input p-values were obtained by the marginal score statistics with the correlation matrix being estimated by the maximum likelihood method (Sun and Lin, 2019; Zhang et al., 2020).

The results for quantitative trait are summarized in Fig. 5. For the GATES, the ECC is more accurate than the original p-value calculation. For the BJ, when $n = 50$ and 100, the *GBJ*-package appears to inflate the type I errors at a small α level. In these cases, the ECC works better. For the HC, *GBJ*-package and the ECC both control the type I error well. Finally, the ECC performs well for the GGOF-O (adapting to HC, BJ, and minP) in all cases considered. The results for binary trait show very similar comparisons (Supplementary Figure S23).

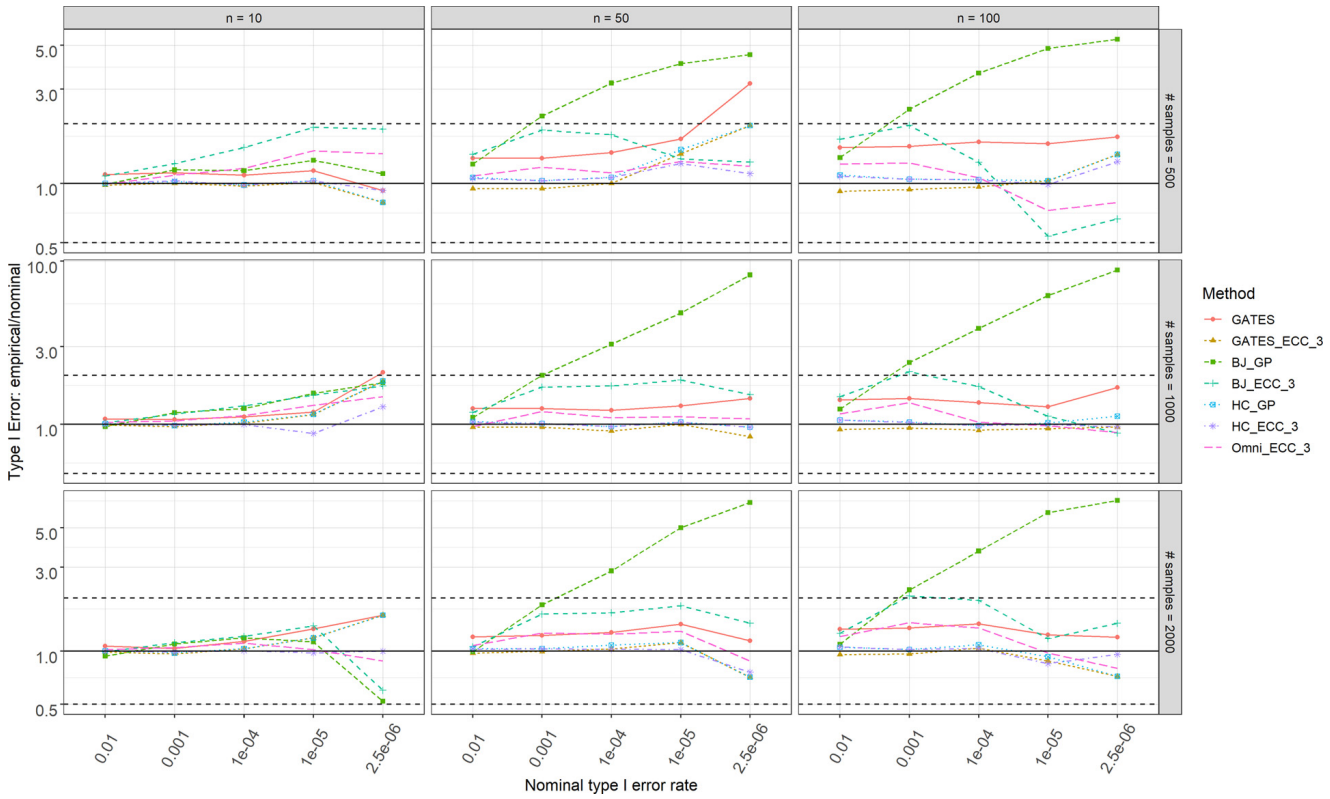

Fig. 5. The ratios between empirical type I error rates and the nominal α levels for quantitative traits simulated by the regression model. For each panel, the y-axis is the ratio between the empirical type I error rate and nominal α; the x-axis is α levels. Panels are organized by the sample sizes $N = 500, 1000, 2000$ (in rows) and the gene sizes $n = 10, 50, 100$ (in columns). Tests: GATES, BJ, HC, and GGOF-O; Calculations: *GBJ*-package (GP), the ECC with $r = 3$ (ECC_3).

4. Statistical power

To equitably reveal true statistical power, we empirically controlled the type I error rate at $\alpha = 0.01$ based on 10^5 simulations of the null. Potential inaccuracy in the p-value calculation was avoided in this way. Settings of the GMM and the linear models were similar to those in Section 3.4 except signals were added.

Under the GMM, signals were configured by the mean vector with $\mu_i = \mu$, for $i \in M^* = \{i_1, \ldots, i_K\}$ and $\mu_i = 0$ otherwise. M^* was uniformly distributed on $\{1, \ldots, n\}$. Various signal strength $\mu = 0, 0.25, 0.5, \cdots, 3$, and number of true signals $K = 1, 2, 3, 5, 10$ and $0.5n$ were considered. Figs. 6 shows the results under two correlations: (a) a positive polynomial decaying Σ and (b) its inverse, i.e., negative correlations with decaying magnitude. First, it is interesting to see the power comparisons between the original tests and their extended versions. The HC has almost identical power as the GHC in almost all settings considered in this paper. Similarly, Simes and the GATES have very similar power, which has also been pointed out in the original paper of GATES (Li et al., 2011). Meanwhile, the difference between the BJ and the GBJ could be significant, and the relative merits highly depend on the correlation patterns. Under high positive correlations, the GBJ could be much more powerful than the BJ, whereas, under high negative correlations, the GBJ could become much less powerful (in which we observed that the power of the GBJ could even drop over increasing signal strength).

Second, it is interesting to observe how correlation and signal patterns influence the statistical power across different tests. Under positive correlations (shown in Fig. 6 part (a)), the influence has two aspects: overall power and relative merits. Regarding the overall power, at each given signal number and strength, all tests show power increasing when correlation becomes lower. This trend holds no matter signals are sparse or dense. In the meanwhile, the amount of power change varies across different tests. Therefore, the relative merits of different tests can change significantly depending on signal sparsity. Specifically, under a low correlation, minP, Simes/GATES, and HC/GHC are more powerful for sparse signals, while BJ/GBJ is more powerful for dense signals. This observation is consistent with literature results under independence (Li and Siegmund, 2015; Zhang et al., 2020). However, when correlation becomes higher, minP and Simes/GATES gain more advantages for sparse and dense signals. On the other hand, under negative correlations (Fig. 6 part (b)) the power patterns remain similar except for the GBJ (low power).

Third, we observe that the GGOF-O is a robustly powerful test under all settings. As discussed above, different tests have different merits under different signal and correlation patterns. For example, when BJ has low power, minP and HC often have high power, and vice versa. Therefore, by including tests with complementary performances, the GGOF-O test can automatically adapt to a proper test and robustly provide high power across various scenarios. Moreover, because the

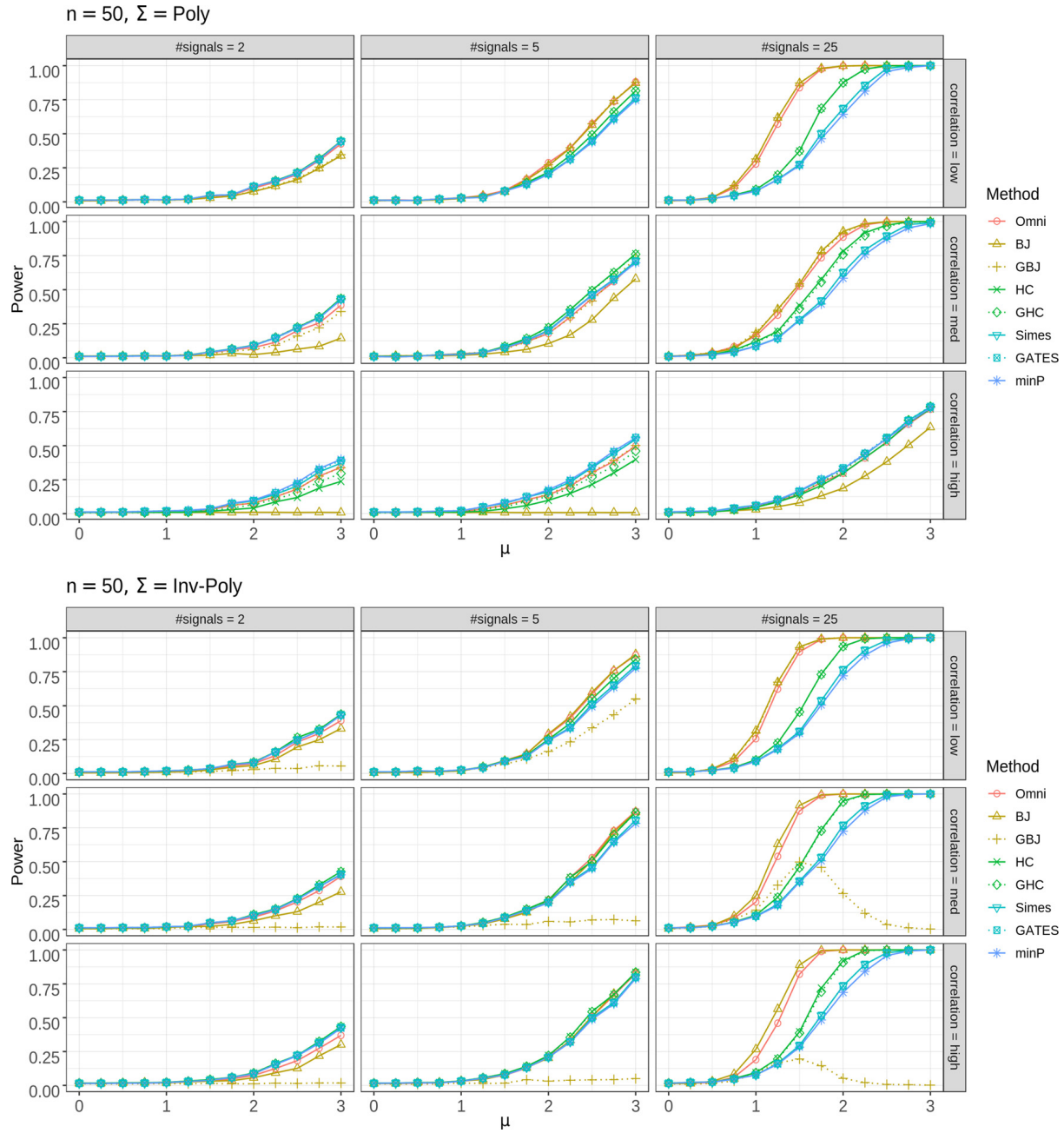

Fig. 6. Statistical power across signal number and strength. The dimension is $n = 50$. (a) Poly: Polynomial decaying positive correlations; (b) Inv-Poly: The inverse of (a) gives negative correlations with decaying magnitude. Each panel shows the power curves of different tests over signal strength μ. The panels' rows correspond to low, medium, and high correlations with $\kappa = 3, 1, 0.2$ in (15). The columns of the panels correspond to signal numbers $K = 2, 5, 25$. Omni: GGOF-O adapting to minP, HC, and BJ. X-axis: Signal strength $\mu_i = \mu \in (0, 3)$.

excellent performance of traditional GGOF tests has already covered broad scenarios of data patterns, the GGOF-O does not need to incorporate the extended versions of GGOF tests, which are often computationally expensive. For example, even in their best scenarios, GHC/GBJ is not much better than GGOF-O, whereas the latter could be much better in other scenarios. Overall, when signal and correlation patterns are unknown, it is good to apply the GGOF-O.

We also compared the power under the regression model of a quantitative trait. The genotype data were generated by the same procedure described before. The quantitative trait model followed (16) plus SNP genotype covariates with coefficients $\beta_i = \beta$, for $i \in M^* = \{i_1, \ldots, i_K\}$. M^* is the set of true SNPs uniformly distributed on $\{1, \ldots, n\}$. Various effects $\beta = 0.2, 0.4, 0.6$ and numbers of true signals $K = 1, 2, \cdots, 10$ were considered. We included the likelihood-ratio test (LRT), Wald test (Fahrmeir, 1987), the SKAT test, and its omnibus test SKAT-O because of their popularity in association studies and the rare-variant analysis.

Figs. 7 illustrates the power comparisons. The relative advantages of various tests are similar to those under the GMM with moderate positive correlations (expected under typical LD). The minP, the GATES, and the HC/GHC are more powerful

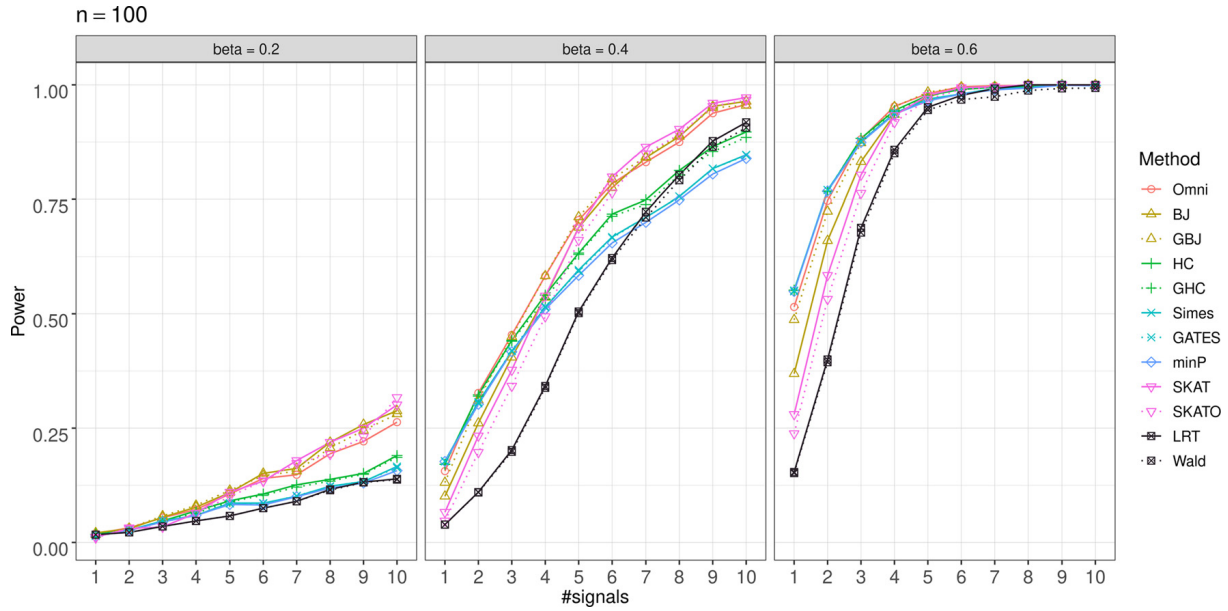

Fig. 7. The statistical power is compared under the quantitative-trait model. Cosi2 simulated genotypes of $n = 100$ SNPs. X-axis: The number of effective SNPs. Panels from left to right: Signal strength $\beta = 0.2$, 0.4 and 0.6. Omni: GGOF-O adapting to minP, HC, and BJ.

than the rest when genetic effects are sparser and stronger, while the BJ/GBJ and the SKAT are better for denser and weaker effects. Regarding the comparisons between the classic and the extended versions, again, the HC and the GHC, as well as Simes and the GATES, are almost identical. The GBJ is similar to the BJ but could be better for sparse and strong signals (e.g., at $\beta = 0.6$ and K=1, where the HC is even better than the GBJ). The LRT and Wald tests have similar power and do not show an advantage here. In all cases, the GGOF-O is robustly powerful; it is similar to the SKAT-O for dense signals but has significantly higher power for sparse signals. More simulation based power comparisons are given in Supplementary Figures S24 – S30, for other parameter settings and from different angles of comparisons (e.g., by correlation patterns).

5. Application to genetic summary data analysis

For dissecting the genetics of complex traits, analyzing genetic summary data plays an increasingly critical role. In contrast to traditional individual-level data, the summary data features more availability because of less privacy concern, easier data cumulation for larger sample size, and the efficiency in conveying information (Lin and Zeng, 2010; Pasaniuc and Price, 2017). The GGOF tests are naturally a flexible and powerful framework for this type of data analysis. Here, we demonstrate the application of the GGOF in a gene-based analysis of bone mineral density (BMD) based on SNP p-values. The results show that the GGOF tests are promising for detecting gene-trait associations and that the ECC algorithm can well control the genome-wide type I errors in such studies.

The SNP p-values were taken from the summary statistics of the GEFOS consortium on femoral neck BMD (FN-BMD) (Estrada et al., 2012). In total, 1,346,818 SNPs were grouped into 26,790 genes based on their genomic locations. Because highly correlated SNPs deliver similar genetic association information, if two SNPs were in high LD (correlation coefficient $r > 0.9$), the SNP with a larger p-value was trimmed. Based on the data source, it is reasonable to assume that the summary statistics were asymptotically Gaussian, reasonably satisfying the GMM assumption in (8). It fits the practice of estimating the correlation matrix Σ by the LD of an external reference panel, such as the 1000 Genomes Project (Lin and Zeng, 2010; Siva, 2008). Preferring efficient computation in practice, we applied the minP, the HC, the BJ, and the GGOF-O (adapting to the minP, the HC, and the BJ) with the ECC ($r = 3$) being used to calculate gene-level association p-values. For comparisons, we also applied the *GBJ*-package and the *mvtnorm*-package to calculate the gene p-values of the corresponding tests: minP_MVN, BJ_GP, and HC_GP.

The results show that the ECC method is adequate for controlling the genome-wide error rate. It is evidenced by both the QQ-plot (Supplementary Figure S31) and the genomic inflation factors. The QQ-plot shows that all tests are reasonably controlled overall. Meanwhile, the genomic inflation factors are 0.94, 1.01, and 1.00 for the ECC-based GGOF-O, BJ, and HC, respectively. They are close to the ideal value 1 (Yang et al., 2011). The genomic inflation factors of the BJ_GP (0.77) and the HC_GP (0.79) are slightly smaller than 1.

Moreover, the results show that the GGOF-O has higher power than the individual tests. Specifically, we got 65 gene-hits based on a genome-wide significance threshold 2.50×10^{-6} by at least one of these seven tests. Among these, 50 genes are considered known because they were reported in the literature as being associated with BMD or osteoporosis (the disease characterized by low BMD), or they are adjacent ($\pm$ 250kb) to a known disease marker. Among these 50 known genes,

the GGOF-O successfully replicated 40 of them, while the individual tests replicated fewer: 30 (minP_ECC), 34 (BJ_ECC), 33 (HC_ECC), 30 (minP_MVN), 37 (BJ_GP) and 34 (HC_GP). The list of top hits are given in Supplementary Table S1.

Among the 65 top gene-hits, 15 genes are considered putative novel genes for FN-BMD. Eleven of them are reasonably associated with other bone related phenotypes. In particular, the loci of three genes *COBLL1*, *AC019181.2* and *SLC38A11* have close locations (chr 2: 165510134 – 165812035); two of them (*COBLL1* and *SLC38A11*) are associated with body mass index (BMI) and waist-hip ratio (Pulit et al., 2019). Gene *CDH12* (chr 5: 21750782 – 22853731) is associated with BMI (Graff et al., 2017). Gene *SLC44A1* (chr 9: 108006903 – 108201452) is associated with waist-hip ratio (Lotta et al., 2018; Kichaev et al., 2019). Gene *MRPS35* (chr 12: 27863706 – 27909228) is associated with Osteoarthritis (Styrkarsdottir et al., 2018; Zhou et al., 2018). With close loci, three genes *RP11-793H13.10*, *ATF7*, and *ATP5G2* (chr 12: 53900472 – 54071192) are reported in association with height (He et al., 2015; Kichaev et al., 2019). Another two closely located genes *PRR15L* and *CDK5RAP3* (chr 17: 46029333 – 46059140) are reportedly associated with BMI and height (See UK BioBank report: http://www.nealelab.is/uk-biobank/). For the rest genes *CDH18* (chr 5: 19473060 – 20575982), *DCAF13P3* (chr 15: 51236860 – 51238193), *CTD-2377D24.8* (chr 17: 46760729 – 46781844), and *CTD-2582D11.1* (chr 17: 71890888 – 71908017), we did not find direct associations with BMD related phenotypes. Further functional analysis and validation would be helpful to decide their relevance.

6. Discussion

This paper proposes a unified framework for applying the GGOF tests, a generic family of supremum-based statistics based on ordered input p-values, into analyzing correlated data. The GGOF-O, i.e., the omnibus test that adapts to GGOF statistics, is still in the GGOF family, and its test p-value can be efficiently computed by the same method. Due to the simplicity and the complementary performance of minP, HC, and BJ, the corresponding GGOF-O is shown to possess advantages in computation and statistical power over GATES, GHC, and GBJ under various patterns of signals and correlations. Therefore, instead of extending individual statistics to incorporate correlations explicitly, the GGOF framework provides an alternative solution to address correlated data.

A few limitations of this work are to be addressed in our future studies. First, as a universal approach, the ECC moves forward the status quo in terms of generality, computational speed, and accuracy. However, it remains challenging to accurately calculate a very small p-value under specific strong correlations, such as block-wise positive high correlations. No satisfactory solutions are available yet. Practical strategies could avoid the issue, such as testing block by block or obtaining empirical p-value by resampling-based methods. Still, improving analytical computation remains an exciting topic. For that, further development based on the ECC idea is possible. For example, we could fine-tune the r parameter in (13) and recommend specific values for specific tests and correlation patterns. We could also develop more sophisticated algorithms based on the general idea of matching the features of correlation matrices. Second, we will carry out a theoretical study on the optimality of GGOF-O under various signal and correlation patterns. For example, it is of interest to determine whether GGOF-O inherits the optimalities of the individual tests that were adapted. For that, studying the rejection boundaries in (7) could be a critical strategy. Moreover, novel GGOF procedures could also be designed directly based on such rejection boundaries to address specific signal patterns.

Appendix A

A.1. Proof of Equation (10)

The elements of T can be written as $T_i = \sqrt{1-\rho}Z_i + \sqrt{\rho}Z_0$, $i = 1, \ldots, n$, where $Z_0, Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$. First, consider the one-sided input p-values. Following (2) and noting $P_{(i)} = \bar{\Phi}(T_{(n-i+1)})$ and $U_{(i)} \overset{D}{=} \bar{\Phi}(Z_{(n-i+1)})$, $i = 1, \cdots, n$, we have

$$\begin{aligned}
\mathbb{P}(S_{n,f,\mathcal{R}} < b) &= \mathbb{P}(P_{(i)} > u_i, \text{ for all } i \in \mathcal{R}) \\
&= \mathbb{P}(T_{(n-i+1)} < \bar{\Phi}^{-1}(u_i), \text{ for all } i \in \mathcal{R}) \\
&= \mathbb{P}(\sqrt{1-\rho}Z_{(n-i+1)} + \sqrt{\rho}Z_0 < \bar{\Phi}^{-1}(u_i), \text{ for all } i \in \mathcal{R}) \\
&= \int_{-\infty}^{\infty} \phi(z)\mathbb{P}(Z_{(n-i+1)} < \frac{\bar{\Phi}^{-1}(u_i) - \sqrt{\rho}z}{\sqrt{1-\rho}}, \text{ for all } i \in \mathcal{R})dz \\
&= \int_{-\infty}^{\infty} \phi(z)\mathbb{P}(U_{(i)} > \bar{\Phi}\left(\frac{\bar{\Phi}^{-1}(u_i) - \sqrt{\rho}z}{\sqrt{1-\rho}}\right), \text{ for all } i \in \mathcal{R})dz.
\end{aligned}$$

When the input p-values are two-sided, the calculation is adjusted accordingly.

$$
\begin{aligned}
\mathbb{P}(S_{n,f,\mathcal{R}} < b) &= \mathbb{P}(P_{(i)} > u_i, \text{ for all } i \in \mathcal{R}) \\
&= \mathbb{P}(2\bar{\Phi}(|T|_{(n-i+1)}) > u_i, \text{ for all } i \in \mathcal{R}) \\
&= \mathbb{P}(|\sqrt{1-\rho}Z + \sqrt{\rho}Z_0|_{(n-i+1)} < \bar{\Phi}^{-1}(u_i/2), \text{ for all } i \in \mathcal{R}) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(|\sqrt{1-\rho}Z + \sqrt{\rho}z|_{(n-i+1)} < \bar{\Phi}^{-1}(u_i/2), \text{ for all } i \in \mathcal{R})\phi(z)dz \\
&= \int_{-\infty}^{\infty} \mathbb{P}(U_{(i)} > \bar{F}_z\left(\bar{\Phi}^{-1}(u_i/2)\right), \text{ for all } i \in \mathcal{R})\phi(z)dz,
\end{aligned}
$$

where $\bar{F}_z(x) = 1 - F_z(x)$ with $F_z(x)$ being the CDF of $|\sqrt{1-\rho}Z + \sqrt{\rho}z|$:

$$
F_z(x) = \Phi\left(\frac{x - \sqrt{\rho}z}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{-x - \sqrt{\rho}z}{\sqrt{1-\rho}}\right).
$$

In summary, define $c_{1i} = \bar{\Phi}\left(\frac{\bar{\Phi}^{-1}(u_i) - \sqrt{\rho}z}{\sqrt{1-\rho}}\right)$, $c_{2i} = \bar{F}_z\left(\bar{\Phi}^{-1}(u_i/2)\right)$, then

$$
\mathbb{P}(S_{n,f,\mathcal{R}} < b) = \int_{-\infty}^{\infty} \mathbb{P}(U_{(i)} > c_{t,i}, \text{ for all } i \in \mathcal{R})\phi(z)dz, \quad t = 1, 2.
$$

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107379.

References

Bailey, T.L., Grundy, W.N., 1999. Classifying proteins by family using the product of correlated p-values. In: Proceedings of the Third Annual International Conference on Computational Molecular Biology, pp. 10–14.

Barnett, I., Mukherjee, R., Lin, X., 2017. The generalized higher criticism for testing SNP-set effects in genetic association studies. J. Am. Stat. Assoc. 112 (517), 64–76.

Berisa, T., Pickrell, J.K., 2016. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32 (2), 283.

Berk, R.H., Jones, D.H., 1979. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. Probab. Theory Relat. Fields 47 (1), 47–59.

Cheverud, J.M., 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87 (1), 52–58.

Chicheportiche, R., Bouchaud, J.-P., 2011. Goodness-of-fit tests with dependent observations. J. Stat. Mech. Theory Exp. 2011 (09), P09003.

Donoho, D.L., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat. 32 (3), 962–994.

Donoho, D.L., Jin, J., 2008. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. Proc. Natl. Acad. Sci. USA 105 (39), 14790–14795.

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y.-H., Duncan, E.L., Ntzani, E.E., Oei, L., Albagha, O.M., Amin, N., Kemp, J.P., et al., 2012. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat. Genet. 44 (5), 491–501.

Fahrmeir, L., 1987. Asymptotic testing theory for generalized linear models. Statistics 18 (1), 65–76.

Friedman, S., Weisberg, H.F., 1981. Interpreting the first eigenvalue of a correlation matrix. Educ. Psychol. Meas. 41 (1), 11–21.

Galwey, N.W., 2009. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet. Epidemiol. 33 (7), 559–568.

Graff, M., Scott, R.A., Justice, A.E., Young, K.L., Feitosa, M.F., Barata, L., Winkler, T.W., Chu, A.Y., Mahajan, A., Hadley, D., et al., 2017. Genome-wide physical activity interactions in adiposity—a meta-analysis of 200,452 adults. PLoS Genet. 13 (4), e1006528.

Hall, P., Jin, J., 2010. Innovated higher criticism for detecting sparse signals in correlated noise. Ann. Stat. 38 (3), 1686–1732.

He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J.-Y., Johnson, T.A., Li, H., Yang, X., Dai, J., et al., 2015. Meta-analysis of genome-wide association studies of adult height in East asians identifies 17 novel loci. Hum. Mol. Genet. 24 (6), 1791–1800.

He, S., Wu, Z., 2011. Gene-based higher criticism methods for large-scale exonic single-nucleotide polymorphism data. In: BMC Proceedings, vol. 5. Springer, p. S65.

Jager, L., Wellner, J.A., 2007. Goodness-of-fit tests via phi-divergences. Ann. Stat. 35 (5), 2018–2053.

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., Price, A.L., 2019. Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. 104 (1), 65–75.

Kolmogorov, A., 1933. Sulla determinazione empirica di una leggi di distribuzione (on the empirical determination of a distribution function). G. Ist. Ital. Attuari 4, 83–91.

Kotz, S., Johnson, N.L., 2012. Breakthroughs in Statistics: Foundations and Basic Theory. Springer Science & Business Media.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., Lin, X., 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91 (2), 224–237.

Li, J., Ji, L., 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95 (3), 221–227.

Li, J., Siegmund, D., 2015. Higher criticism: p-values and criticism. Ann. Stat. 43 (3), 1323–1350.

Li, M.-X., Gui, H.-S., Kwan, J.S., Sham, P.C., 2011. Gates: a rapid and powerful gene-based association test using extended Simes procedure. Am. J. Hum. Genet. 88 (3), 283–293.

Lin, D.-Y., Zeng, D., 2010. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika 97 (2), 321–332.

Lotta, L.A., Wittemans, L.B., Zuber, V., Stewart, I.D., Sharp, S.J., Luan, J., Day, F.R., Li, C., Bowker, N., Cai, L., et al., 2018. Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors. JAMA 320 (24), 2553–2563.

McClellan, J., King, M.C., 2010. Genetic heterogeneity in human disease. Cell 141 (2), 210–217.
Moscovich, A., Nadler, B., Spiegelman, C., 2016. On the exact Berk-Jones statistics and their p-value calculation. Electron. J. Stat. 10 (2), 2329–2354.
Moskvina, V., Schmidt, K.M., 2008. On multiple-testing correction in genome-wide association studies. Genet. Epidemiol. 32 (6), 567–573.
Pasaniuc, B., Price, A.L., 2017. Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. 18 (2), 117–127.
Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al., 2019. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum. Mol. Genet. 28 (1), 166–174.
Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al., 2001. Linkage disequilibrium in the human genome. Nature 411 (6834), 199.
Shao, J., 2010. Mathematical Statistics. Springer Verlag.
Shlyakhter, I., Sabeti, P.C., Schaffner, S.F., 2014. Cosi2: an efficient simulator of exact and approximate coalescent with selection. Bioinformatics 30 (23), 3427–3429.
Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73 (3), 751–754.
Siva, N., 2008. 1000 genomes project. Nat. Biotechnol. 26 (3), 256.
Styrkarsdottir, U., Lund, S.H., Thorleifsson, G., Zink, F., Stefansson, O.A., Sigurdsson, J.K., Juliusson, K., Bjarnadottir, K., Sigurbjornsdottir, S., Jonsson, S., et al., 2018. Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. Nat. Genet. 50 (12), 1681–1687.
Sun, Q., 2005. Wiener's lemma for infinite matrices with polynomial off-diagonal decay. C. R. Math. 340 (8), 567–570.
Sun, R., Lin, X., 2019. Genetic variant set-based tests using the generalized Berk–Jones statistic with application to a genome-wide association study of breast cancer. J. Am. Stat. Assoc., 1–13.
Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'connell, J.R., Mangino, M., et al., 2011. Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. 19 (7), 807–812.
Yang, L., Xuan, J., Wu, Z., 2014. A goodness-of-fit association test for whole genome sequencing data. In: BMC Proceedings, vol. 8. Springer, p. S51.
Zhang, H., Jin, J., Wu, Z., 2020. Distributions and power of optimal signal-detection statistics in finite case. IEEE Trans. Signal Process. 68, 1021–1033.
Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al., 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50 (9), 1335–1341.