Genome analysis

The null hypothesis of GSEA, and a novel statistical model for competitive gene set analysis

Birgit Debrabant

Department of Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, 5000 Odense C, Denmark

Associate Editor: John Hancock

Received on January 21, 2016; revised on December 7, 2016; editorial decision on December 8, 2016; accepted on December 21, 2016

Abstract

Motivation: Competitive gene set analysis intends to assess whether a specific set of genes is more associated with a trait than the remaining genes. However, the statistical models assumed to date do not enable a clear cut formulation of the competitive null hypothesis. This is a major handicap to the interpretation of results obtained from a gene set analysis.

Results: This work presents a hierarchical statistical model based on the notion of dependence measures, which overcomes this problem. The two levels of the model naturally reflect the modular structure of many gene set analysis methods. We apply the model to show that the popular GSEA method, which recently has been claimed to test the self-contained null hypothesis, actually tests the competitive null if the weight parameter is zero. However, for this result to hold strictly, the choice of the dependence measures underlying GSEA and the estimators used for it is crucial.

Contact: bdebrabant@health.sdu.dk

Supplementary information: Supplementary material is available at Bioinformatics online.

1 Introduction

Genome-wide association studies (GWAS) search for association between a trait and a large number of single nucleotide polymorphisms (SNPs) across the genome. They have successfully identified many loci contributing to complex traits. However, the polymorphisms identified so far only account for a small portion of the heritability, see Manolio et al. (2009). One possible explanation is given by the quasi-infinitesimal hypothesis stating that the number of loci contributing to a trait might be high with few large effects and many effects of modest or small size. Especially the latter ones are likely to be missed to great extent by GWAS which after simultaneous testing of a very large number of hypotheses and the corresponding adjustment for multiple testing suffer from low power.

In order to overcome these drawbacks, gene set analysis methods have been proposed, which aim at combining statistical evidence from a set of SNPs or genes into one test. For an overview over current methods we refer to Wang et al. (2010, 2011); Fridley and Biernacka (2011). It is hoped for these methods to potentially increase the power. By combining SNPs, the dimensionality of the problem is reduced, and bounds on P-values are thus less stringent. Furthermore, by aggregating markers which individually are weakly associated, stronger evidence for association might emerge at the level of the combined set. From a biological point of view, the rationale for such an analysis is that it is the collective function of the gene products in a pathway which are playing a role in the cellular context, and, therefore, the combined effect is very reasonable to study.

One currently very popular approach to gene set analysis is the GSEA method, which has been put forward in the context of GWAS by Wang et al. (2007) as an adaption of gene set enrichment analysis Subramanian et al. (2005) to SNPs. In this gene-based approach, genes are ranked according to their association with the trait and a running sum is calculated for investigated gene sets based on the ranks of the genes within the set relative to those of the genes outside. The maximum value is called ‘enrichment score’ and is intended to measure the overrepresentation of genes with low ranks within the set relative to its outside. Significance is assessed permuting sample labels.
For this article we consider the following SNP-based version of GSEA: For a set \( G \) of genes and a subset \( G' \subset G \) of genes of interest, let \( S_1, \ldots, S_N \) be the SNPs mapped to \( G \) and \( J, \subset \{1, \ldots, N\} \) the indices corresponding to \( G' \). The enrichment score \( ES(G') \) is given by

\[
ES(G') = \max \left\{ \frac{\sum_{i \in G'} \left| r_i \gamma \right|^p}{\sum_{i \in G' \cup \neg G'} \frac{1}{N} \right\},
\]

where \( r_i \) are statistics obtained from association tests between the SNPs' and the investigated trait, ordered according to their values reflecting the degree of association, and where \( \gamma = \sum_{i \in J} \left| r_i \gamma \right|^p \) and \( p \geq 0 \) is a weight parameter to enhance the impact of outstanding SNPs.

There exist several refinements of the original GSEA method, cp. e.g. GSEA-SNP (Holden et al., 2008), SEA (Weng et al., 2011), i-GSEA4GWAS (Zhang et al., 2010), Guo et al. (2009) or SeqGSEA (Wang and Cairns, 2013). However until today, GSEA remains the most widely used method for competitive gene set analysis.

While a lot of attention has been paid to incorporating and exploiting the structure of SNPs and genes in a right way, the basic statistical setup such as the underlying model, null and alternative hypotheses are still being discussed in order to be fully understood. This incertitude hinders correct interpretation of results issuing from GSEA. Especially, the null hypothesis was for a long time considered to be the competitive hypothesis \( H_{\text{comp}} \): The genes in the set \( G' \) show the same or lower magnitude of association with the trait compared with the remaining genes in the complementary set \( \neg G' \). Thus, one is contrasting genes within the set relative to those outside, cp. e.g. Goeman and Bühlmann (2007). Only recently, Maciejewski (2013) pointed out that the sample permutation scheme determines the null hypothesis, which therefore amounts to \( H_0 \) : ‘No gene, neither in \( G' \) nor in \( \neg G' \), is associated with the trait.’, not being of any practical interest. However, it was shown by Debrabant and Soerensen (2014) that Maciejewski’s definition of \( H_0 \) is too narrow, and that GSEA actually tests the broader self-contained null \( H_{\text{comp}}^\text{null} \) : ‘No gene in \( G' \), is associated with the trait.’ This is essentially due to the fact that the test statistic is tailored to competitive alternatives in which \( G' \) is more associated than \( \neg G' \), and a situation where only genes in \( G' \) are associated does not give rise to any power and can be considered part of the null. Unfortunately, amongst self-contained methods for gene set analysis, GSEA is expected to be outperformed by other approaches designed to test \( H_{\text{comp}}^\text{null} \), which impairs its practical relevance.

The present article is concerned with the special case where the weight parameter \( P \) of GSEA is chosen to be 0. This case differs in that the enrichment score (1) reduces to the common Kolmogorov-Smirnov statistic, which is based on rank transformed gene statistics only. Outstanding genes are not given a greater weight, and high values of \( ES(G') \) are due to differences in ranks between genes in \( G' \) and genes in \( \neg G' \). In this case, it seems plausible, that, if \( G' \) and \( \neg G' \) are equally associated with the trait, \( ES(G') \) does not stand out, but has a distribution similar to the one under \( H_0 \), which is obtained by permutations. Consequently, the null hypothesis of GSEA would also encompass the competitive null, which GSEA originally intended to assess, and the practical worth of GSEA would be confirmed. However, it has so far not been possible to formally prove such a statement due to deficiencies in the currently considered statistical models underlying competitive gene set analysis when it comes to pinpointing what is being understood as ‘more associated’.

Maciejewski (2013) presented four statistical models, which are currently believed to underlie the different types of gene set analysis approaches. According to this work, GSEA tests \( H_0 \) and relies on his Model 4, which consists of independent individual observations, each consisting of a trait value, and genotypes from \( G' \) and \( \neg G' \). An alternative to Model 4 for competitive methods is Model 3, assuming iid observations given by \( r = (r_i), i \in J \), and \( r' = (r_i), i \in \{1, \ldots, N\} \setminus J \), to be loosely seen as describing the association of genes in \( G' \) and \( \neg G' \). Based on Model 3 competitive hypotheses could now be formulated with the help of the marginal distributions of \( r \) and \( r' \). Unlike Model 4, Model 3 directly addresses the association between the observations and genes from \( G' \), respectively \( \neg G' \), and enables comparison of the strength of association. However, it ignores that our original sample consists of individual trait values and genotypes, hence it does not directly correspond to the observed data. Especially, when gene randomization is used to derive the null distribution of a corresponding test statistic, this hinders interpretation of the results.

The present article overcomes the above mentioned problems by introducing a hierarchical statistical model for competitive gene set analysis, which combines features of Maciejewski’s models 3 and 4. It founds on the general notion of dependence measures to model the (degree of) association with the trait for each gene. Thereby, the gene specific dependence measures are themselves viewed as a random sample of two distributions \( F \) resp. \( F_c \) reflecting the association of the genes in \( G' \) resp. \( \neg G' \) with the trait. Within this kind of model, testing \( H_{\text{comp}}^\text{null} \) can be formulated as \( H_{F_c}^\text{null} : F_c = F \) or \( H_{\text{comp}}^\text{null} : F_c \geq F \), where \( \geq \) denotes stochastic dominance. Our model is suitable for GSEA and other competitive methods beyond.

The article is organized as follows: We present our new statistical model in Section 2 and give an example to illustrate how our model can be applied. This is the only example currently known to the authors. In Section 3 we show under suitable conditions in Proposition 1, that the distributions of GSEA given \( H_0 \) and given \( H_{\text{comp}}^\text{null} \) equal. We further apply the proposition to our example. Finally, in Section 4, we simulate data and apply GSEA to it in order to illustrate our proposition. Since it is generally difficult to find examples in which the conditions of Prop. 1, which crucially depend on the dependency measures and corresponding estimates underlying GSEA, are satisfied, we also include a setup, where the assumptions of Prop. 1 are not satisfied, to investigate the consequences of this.

In the remainder of this article, we restrict the weight parameter to \( P = 0 \).

2 A statistical model for competitive gene set analysis

We consider data \( Z_1, \ldots, Z_n \) from \( n \) individuals, where \( Z_i = (X_i, S_i) \), further \( Y_i \) is the trait of interest (with e.g. \( Y_i \in \mathbb{R} \) for continuous or \( Y_i \in \{0, 1\} \) for binary traits), and \( S_i = (S_{i1}, \ldots, S_{iN}) \in \mathbb{E}^N \) (with e.g. \( E = \{0, 1, 2\} \)) are the SNP-genotypes. Let \( J, \subset \{1, \ldots, N\} \) denote the indices for SNPs corresponding to \( G' \). Denote by \( F_Z \) the univariate distribution of the individual observations \( Z_1, \ldots, Z_n \).

Apart from not assuming independency of the observations \( Z_1, \ldots, Z_n \), up to now, the above corresponds to Model 4 of Maciejewski (2013), stated to be underlying competitive methods with sample randomization as e.g. GSEA and SAFE. Within this model, it is well possible to infer the self-contained hypothesis \( H_0^\text{null} \) i.e. whether there exist SNPs in \( G' \), associated with the trait. However, the competitive null \( H_{\text{comp}}^\text{null} \) not only infers the presence of association, but seeks to compare its magnitude between \( G' \) and \( \neg G' \). Unfortunately, the above model is not
rich enough as it leaves unspecified how the magnitude of association can be measured and how it exactly relates to the random variables introduced. This lack of concreteness hinders interpretation of the tested results and study of the distribution of the relevant test statistic under the null.

To overcome this drawback we refer to the notion of measures of dependence for two random variates, e.g. Spearman’s ρ, Kendall’s τ or Hoeffding’s D. Concretely, we extend the present model in the following hierarchical way: Assume the distribution of $Z_1, \ldots, Z_n$ is parameterized by $\rho = (\rho_1, \ldots, \rho_N)$ being measures of dependence between the trait $Y$ and each of the SNPs. Further, assume the dependence measures $\rho_i$ are themselves sampled from two underlying populations corresponding to SNPs mapped to $G_i$, resp. $G_j^c$ and there exist two univariate distribution functions $F_i, F_j$ such that $\rho_i$ has distribution function $F_i$ for $j \in J_i$, resp. $F_j$ for $j \in J_j$. Assume that $F_i$ and $F_j$ are either continuous or equal to $1_{[0, \infty)}$, the distribution function of a point measure concentrated in 0 reflecting no association between trait and corresponding SNPs. In the given hierarchical model, independence is not required, neither for the random vectors $Z_i$ nor the dependence measures $\rho_i$. Especially, if the SNPs are in linkage disequilibrium (LD), that is the components of $S_i$ are dependent, the corresponding dependence measures cannot be assumed independent. We do however in the sequel assume that the basic dependency structure, which is related to the LD structure of the SNPs, is fixed. This will be formulated using the concept of a copula, (Nelsen, 2006), which describes the dependence structure of a random vector autonomously from its marginal distributions, by assuming the copula of $\rho$ to be fixed, i.e. to be the same for every possible pair $F_i$ and $F_j$.

Altogether, the model can be formulated as follows:

### Statistical Model

Complete individual observations: $Z_i = (Y_i, S_i), i = 1, \ldots, n$

Genotypes: $S_i = (S_{i1}, \ldots, S_{in}), i = 1, \ldots, n$

Index set corresponding to $G_i$, $J_i \subset \{1, \ldots, N\}$

Dependence measures: $\rho = (\rho_1, \ldots, \rho_N)^T$

with

\[ Z_i | \rho \sim F_i^Z, \quad i = 1, \ldots, n, \quad (2) \]

i.e. the distribution of $Z_i$ is parametrized by $\rho_i$, and $\rho_i \sim F_i, \quad j = 1, \ldots, N$

\[ F_i = \begin{cases} F_i, & j \in J_i, \\ F_j, & \text{else} \end{cases} \]

whereby the copula of $\rho$ is given by $C$

Thereby, the notation ‘$X \sim F$’ stands for ‘$X$ has distribution function $F$’. Moreover, $F_i^Z$ denotes the $(N + d)$-dimensional distribution of a single individual’s observations, i.e. its phenotype and genotypes, which is parameterized by the $N$-dimensional parameter $\rho$, and where $d$ denotes the dimension of the phenotypes $Y$. Further, $F_i$ denotes the (univariate) distribution function of $\rho$, and $F_i$ and $F_j$ are univariate continuous distribution functions.

Note, that the dependency between the genotypes is not modelled directly, but implicitly reflected by the copula of $\rho$. This is due to the fact, that the former is not addressed specifically by GSEA and thus does not need to be formulated directly during our analysis. However, this assumes by no means independent genotypes. Instead, the copula $C$ implicitly needs to be consistent with the LD structure of the genotypes in a way that a corresponding distributional model exists. We refer to our Example 1 for an illustration of this aspect.

Based on the above model, the null hypothesis $H_0$ of Maciejewski and the competitive null $H_{0\text{comp}}$ can be expressed as

\[ H_0 \text{ (no association)} : F_i = F_j = 1_{[0, \infty)} \]

\[ H_{0\text{comp}} \text{ (higher association in } G_i^c) : F_i(x) \geq F_j(x) \quad \forall x, \]

where $F_i$ and $F_j$ are to be interpreted as distribution functions in the latter case. Inferring the above $H_{0\text{comp}}$ can be interpreted as inferring whether the SNPs/genes in the set $G_i$ show the same or lower magnitude of association with the trait compared to the SNPs/genes in the complementary set $G_j^c$.

Note that the described two level model naturally reflects the modular structure of GSEA and several other gene set methods (Ackermann and Strimmer, 2009), which include an initial marginal analysis resulting in SNP level statistics, that are afterwards summarized on the level of the gene set.

Subsequently, we present an example illustrating model (2).

**Example 1. (latent variable model)** Consider a dataset with a continuous trait $Y \in \mathbb{R}$ and $N$ genotyped SNPs $S_1, S_2, \ldots, S_N \in \{0, 1\}$. Assume the genotypes are generated by underlying normal variables $S_j \sim N(0, 1)$, such that $S_j = 0$, $S_j = 1$ or $S_j = -1$. Assume further that the genotypes are independent of the trait $Y$.

\[ Y | S_j \sim \mathcal{N}(\mu_j, \sigma_j) \quad j = 1, \ldots, N \]

where $\mu_j$ and $\sigma_j$ are the mean and standard deviation of the $j$th genotype. Further, assume that the genotypes are independent of each other.

The introduced statistical model (2) is a combination of Maciejewski’s Models 3 and 4. Thereby, Model 3 considers data $r = (r_j), j \in J_i$, and $r' = (r_j), j \in \{1, \ldots, N\} \setminus J_i$, sampled from two unknown distributions, which intend to describe the association between the trait and the genes in $G_i$ resp. $G_j^c$. The loose interpretation connected to Model 3 appears vague especially because of
the fact, that the actually observed individual trait values and genotypes are not referred to. This drawback is eliminated with the help of the hierarchical structure introduced in Model (2), and the reference to the concept of dependence measures is now clear-cut.

3 GSEA and the competitive null hypothesis

Model (2) provides a suitable statistical framework for an investigation of the hypotheses connected with Maciejewski’s $H_0$ and the competitive null $H_{0 \text{ comp}}$. Especially, we can now show, that for GSEA, the distributions of the test statistic under $H_0$ and $H_{0 \text{ comp}}$ coincide, provided the test statistic applied at the SNP-level is consistent for the underlying dependence parameter, and converging at a comparable speed for different SNPs.

**Proposition 1.** Let $F_r$ and $F^*_r$ be two univariate continuous distributions, and $C$ an $N$-dimensional copula. Let $Z_1, \ldots, Z_n$, $n \in \mathbb{N}$, be distributed according to model (2).

Let the parameter $P$ for the calculation of the enrichment score $ES(G_r)$ be 0. Further, given genotypes $S_{ij}, \ldots, S_n$ for a SNP $j \in \{1, \ldots, N\}$, let the SNP statistics $R^{(n)} = \{R^{(n)}_1, \ldots, R^{(n)}_{S_n} : S_n \} \subset \mathbb{R}$, underlying the calculation of $ES(G_r)$ be strongly consistent estimators for the corresponding association parameter $\rho_j$, i.e.

$$R^{(n)}(Y_1, \ldots, Y_n, S_{ij}, \ldots, S_n) \xrightarrow{n \to \infty} \rho_j \quad \text{almost surely.} \quad (4)$$

Further, assume that for $F = 1_{\{G_r\}}$, there exist a stricly increasing transformation $T_a: \mathbb{R} \to \mathbb{R}$ and a continuous distribution $G$ such that for all $i = 1, \ldots, N$

$$T_a(R^{(n)}(Y_1, \ldots, Y_n, S_{ij}, \ldots, S_n)) \xrightarrow{D} G \quad \text{for } n \to \infty. \quad (5)$$

Then, the following holds:

1. For $F_r = F^*_r$ continuous and $n \to \infty$, the distribution of $ES(G_r)$ coincides with the distribution obtained for $F_r = F^*_r = 1_{\{G_r\}}$.

2. Moreover, if $F_r$ and $F^*_r$ are continuous and $F_r$ is stochastically dominated by $F^*_r$, i.e. $F_r(x) \geq F^*_r(x)$ for all $x$, then, for $n \to \infty$, the distribution of $ES(G_r)$ is stochastically dominated by the distribution obtained when $F_r = F^*_r$.

The proof is contained in the Appendix A1 As a consequence of Proposition 1 we find:

**Corollary 3.1.** For $P = 0$, methods based on an enrichment score can still be considered to test a competitive null $H_{0 \text{ comp}}$ against a competitive alternative $H_{0 \text{ comp}}^*$. □

**Example 1** (Continued). Estimating the correlation $\rho_j$ between $Y$ and $S_j$ based on the thresholded observations $S_j$ is a known problem. In this connection $\rho_j$ is called the polyserial correlation coefficient, cp. Olsson et al. (1982), and a possible estimator is given by

$$\hat{\rho}_j = \frac{r_{S_j Y}}{\hat{\phi}(\hat{\mu}_1) + \hat{\phi}(\hat{\mu}_2)} \quad (6)$$

where $r_{S_j Y}$ is the sample product moment correlation of $Y$ and $S_j$, $s_j$ is the sample standard deviation of $S_j$. Moreover, $\hat{\mu}_j = \Phi^{-1}(P_j)$ with $P_j$ being the cumulative marginal proportions of $S_j$, and finally $\hat{\phi}$ resp. $\Phi^{-1}$ are density resp. inverse distribution function of the standard normal distribution.

It is shown in (Bedrick and Breslin, 1996, Prop. 1), that the estimator (6) is consistent, whereby $\sqrt{\lambda}(\hat{\rho}_j - \rho_j) \xrightarrow{D} N(0, \sigma^2)$, where $\sigma^2$ is given by formula (6) of the referenced article. Especially, for $\rho_j = 0$, we obtain in our example

$$\sigma^2 = \frac{\text{Var}(S_j)}{E(S_j S_j)} = \frac{\left(\{x_j - x_{2j}\} \Phi_{1j} + (x_{2j} - x_{3j}) \Phi_{2j}\right)^2}{\left((x_{1j} - x_{2j})\Phi_{1j} + (x_{2j} - x_{3j})\Phi_{2j} + x_{3j}\right)^2}$$

where $\Phi_{1j} = \Phi(\hat{\mu}_j)$ and $\Phi_{2j} = \Phi(\hat{\mu}_2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{w_j^2}{2}\right\}$. The transformation $T_a$ required for our Prop. 1, can therefore be chosen to be $T_a(r) = \sqrt{nr}$. However, the limit variance of $T_a(\hat{\rho}_j)$ still depends on the SNP $j$. In order to find a unique variance, it is necessary to choose the genotypic values $x_{kj}$ depending on the allele frequency of $S_j$. One can show, that this is always possible, if all investigated SNPs are common, i.e. such that minor allele frequencies are bounded from below e.g. by MAF > 0.01, and in Hardy–Weinberg equilibrium (HWE). Then, one can find a $C > 0$ such that for all $j = 1, \ldots, N$ there exist $x_{kj} < x_{kj} < x_{kj}$ with $\sigma^2 = C$, cp. Appendix A2 in the Supplementary Material.

Note, that the above example of a multivariate latent variable model based on underlying correlation measures is the only explicit example of a model admitting the form (2), which is presently known to the authors. Moreover, only in the case of continuous traits $Y$, we are aware of estimators satisfying the conditions of Prop. 1, cp. Ex. 1 (cont.). We do however believe, that the asymptotic equality stated in Prop. 1 is usually only corrupted slightly to moderately in other situations. The next section illustrates the conclusions of Prop. 1 by simulations also addressing this aspect. Especially, Setup II discusses a case (standard values), where the conditions of Prop. 1 do not hold.

4 Simulations

In this section, we illustrate the conclusions of the last section with the help of simulations. Therefore, we consider both binary outcomes as well as continuous outcomes.

4.1 Setup I: binary outcome

Here, we simulate 1000 realizations of the enrichment score, each based on case-control data with 1000 independent samples, half of which are cases, and with 100 genes, 40 out of which constitute $G_r$. We assume all genes to have three possible genotypes and simulate gene-level statistics directly, that is, we omit the construction of gene-level statistics based on SNP-level statistics as this is not the focus of the present article. Therefore, we focus on the following hypotheses:

- $H_0$: ‘No gene in $G_r$ is associated with the trait.’ This is the null hypothesis according to Maciejewski (2013). We simulate the gene statistics of $G_r$ and $G_r^*$ as independent $\chi^2$-random variables, e.g. issuing from a $\chi^2$-test based on a $2 \times 3$ contingency table in a situation where genotypes and phenotypes are independent. The obtained distribution of $ES(G_r)$ is asymptotically equal to the distribution obtained by permuting sample labels.

- $H_0$, $H_0$, $H_0$: ‘Both, genes in $G_r$ as well as genes in $G_r^*$ are associated with the trait.’ These are competitive hypotheses, where either $G_r$ and $G_r^*$ are associated equally or with different magnitudes. We simulate genotypes based on the binomial distribution as follows: Independently for each gene in $G_r$ resp. $G_r^*$, we draw random
The null hypothesis of GSEA

Fig. 1. (A) Setup I: Kernel density estimates for ES. (B) Setup I: Quantile–quantile plot for ES. (C) Setup I: Quantile–quantile plot for ES (using adapted values). (D) Setup I: Quantile–quantile plot for ES (using standard values). (E) Setup II: Kernel density estimates for ES. (F) Setup II: Quantile–quantile plot for ES (using adapted values). (G) Setup II: Quantile–quantile plot for ES (using standard values)
numbers \( pr_c \) resp. \( pr_c' \) from \([0,0.5]\) and simulate genotypes based on binomial probabilities from the distribution \( B(2, |y – pr_c|) \) resp. \( B(2, |y – pr_c'|) \) for all samples independently. Hereby, \( y \) denotes the respective sample’s phenotype set to either 0 for controls or 1 for cases. The closer \( pr \) is to the value of 0.5, the weaker is the association of the corresponding gene, whereby \( pr = 0.5 \) corresponds to no association. Especially, we distinguish between the following cases:

- \( H_2: \) \( pr, pr_c \sim U[0,0.5] \), that is uniformly distributed on the interval \([0,0.5]\) (equal magnitude of association)
- \( H_2: \) \( pr, pr_c \sim U[0,0.45] \) and \( pr_c' \sim U[0,0.5] \) (\( G_c \) is more associated than \( G_c' \))
- \( H_2: \) \( pr, pr_c \sim U[0,0.5] \) and \( pr_c' \sim U[0,0.45] \) (\( G_c' \) is more associated than \( G_c \))

We use the \( \chi^2 \)-statistic issuing from the contingency tables for the calculation of the enrichment score. These are closely linked to Cramér’s coefficient given here by \( \varphi_j = \sqrt{\frac{\chi^2(S,Y)}{N}} \), where \( \chi^2(S,Y) \) is the value of the \( \chi^2 \)-statistic for the underlying contingency table, cp. Conover (1999). Cramér’s coefficient can be considered an estimator of the population dependency parameter \( \rho_j = \frac{\sum_{i=0}^1 \sum_{l=0}^2 \left( (Y=S_i)-0.5 \right)^2}{\sum_{i=0}^1 \sum_{l=0}^2 (Y=S_i)} \).

The resulting Figures 1(A-D) confirm the results from Proposition 1 and Corollary 3.1. The distributions of \( ES(G_c') \) under \( H_0 \) and \( H_2 \) coincide approximately, whereas the distribution under \( H_0 \) dominates the one under \( H_2 \) since \( G_c \) was more associated than \( G_c' \). Oppositely, the distribution of \( ES(G_c') \) under \( H_2 \) is dominated by the one under \( H_0 \).

4.2 Setup II: continuous outcome

We simulate data according to the latent variable model presented in Example 1. We again generate 1000 realizations of the enrichment score, each based on 1000 independent samples with 100 genes, whereby 40 constitui \( G_c \). The latent structures \( S_j \) are assumed to be standard normal, and thresholds \( w_{1j}, w_{2j} \) chosen such that the resulting minor allele frequency is \( \text{MAF} = 0.4 \) for \( j \in G_c \) and \( \text{MAF} = 0.1 \) for \( j \in G_c' \) and all genes are in HWE. The covariance matrix \( \Sigma \) is chosen block diagonal such that correlations within either of \( G_c \) or \( G_c' \) are constant 0.7.

In the sequel, we compare the following null hypotheses, both modelling situations where the association between outcome and \( G \) equals that between outcome and \( G_c' \):

- \( H_2: \) \( \rho = (0,\ldots,0)^T \) (no association between genes and outcome)
- \( H_2: \) \( \rho_i \sim U[0.56, 0.58] \) for \( i = 1,\ldots,N \) and the copula of \( \rho \) corresponds to the uniform distribution on \( D \)

The calculated enrichment scores are based on the estimate given by (6) for the polyserial correlation coefficient. In order for (4) to hold, and thus Proposition 1 to be applicable, we chose the genotypes values to equal \( x_1 = -200, x_2 = -150 \) and \( x_3 = -149 \) for \( j \in G_c \) resp. \( -265, -150 \) and \( -100 \) for \( j \in G_c' \), whereby the values correspond to homozygous (minor allele), heterozygous and homozygous (major allele) genotypes. We call these genotype values ‘adapted values’. Choosing these adapted values assures that, if \( \rho = 0 \) and \( n \to \infty, \sqrt{np_j} \to N(0, \sigma^2) \) with \( \sigma^2 \approx 2.2 \) for every \( j \), cp. Example 1 (continued) (7). To investigate the impact of deviations from the adapted values we also consider ‘standard values’ (\( x_{1j} = 0, x_{2j} = 1, x_{3j} = 2 \) for every \( j \)).

The results of these simulations are displayed in Figures 1(E–G): Figure 1E shows that the distributions of the enrichment score under \( H_0 \) and \( H_2 \) overall are very similar using either adapted or standard values. Figures 1F and G compare \( H_0 \) with \( H_2 \) using adapted (Fig. 1F) or standard (Fig. 1G) values. Especially, for adapted values, the distribution of \( ES \) under \( H_0 \) roughly coincides with the one obtained under \( H_2 \) which confirms Prop. 1. For standard values, although being similar, the two distributions deviate somewhat more. The type-1-error rate achieved using thresholds based on \( H_0 \) instead of \( H_2 \), which is what is done, when the distribution of \( ES \) is estimated by permuting sample labels, we obtain 0.097, 0.042 and 0.007 for significance levels of 0.1, 0.05 and 0.01 if adapted values are used. For standard values, we find 0.064, 0.019 and 0.003, which corresponds to a conservative test.

5 Conclusions

In this article we presented a novel two-level model (2) for competitive gene set analysis methods, which relies on the notion of dependence measures. By this, we overcome the deficiencies from former models and the definition of the competitive hypothesis becomes clear cut. The two levels of the model reflect the modular structure of many gene set analysis methods.

Especially, we applied the model to the popular GSEA method to understand its null hypothesis for \( P = 0 \), that is when the enrichment score reduces to an ordinary Kolmogorov-Smirnov statistic. Under suitable conditions, we showed, that GSEA indeed tests the competitive null hypothesis. This result helps to reestablish the practical relevance of GSEA, which has recently suffered from being considered a self-contained and not very powerful method.

However, the choice of the dependence measure and corresponding estimates underlying GSEA is crucial. Especially, it seems difficult to find explicit multivariate models admitting the form (2), additionally satisfying the conditions from Prop. 1, and the latent variable model together with the ad hoc estimate for the polyserial correlation coefficient is the only example known to the authors at the moment. We do however believe, that deviations from the mentioned conditions usually only result in slight to moderate deviations in the asymptotic distribution of the enrichment score. More research is needed in order to investigate the consequences of deviations from the necessary conditions in more detail.

In practice, GSEA is often used together with \( P > 0 \), when it tests the self-contained null. Although the statistical model (2) is still applicable, the notion of independency would suffice in order to formulate \( H_0^{\text{ad}} \) and referring to dependency measures can be considered overly complex. Proposition 1 does not hold for \( P > 0 \), since the enrichment score no longer exclusively relies on the ranks of the gene-statistics as is the case for \( P = 0 \). The presented statistical model is applicable to other versions of GSEA applied in different contexts, such as gene expression analysis, GWAS or the analysis of sequencing data. It can help to understand competitive gene set analysis methods beyond GSEA.

Conflict of Interest: none declared.

Appendix A1

Proof of Proposition 1

Proof. 1. Let \( R_j = R_j^{(s)}(Y_1,\ldots,Y_n,S_{i1},\ldots,S_{in}) \) for \( j = 1,\ldots,N \), and \( F := F_s = F_p \). Note that the distribution of \( ES(G_c') \) is
asymptotically independent of \( F \), if the probabilities \( P(R_{n1}^{(n)} \leq \cdots \leq R_{nk}^{(n)}) \) under a given \( F \) are independent of \( F \) as \( n \to \infty \) for all \( k \in \mathbb{N} \) and \( \{1, \ldots, k\} \subseteq \{1, \ldots, N\} \).

Consider first continuous \( F \). The strong convergence (4) of the estimators \( R_{n1}^{(n)}, \ldots, R_{nk}^{(n)} \) to \( \rho \) implies that all marginals converge in distribution to \( F \) for \( n \to \infty \). Moreover, the copula of \( R_{n1}^{(n)}, \ldots, R_{nk}^{(n)} \) converges to the copula of \( \rho \). Therefore we find

\[
P(R_{n1}^{(n)} \leq \cdots \leq R_{nk}^{(n)}) = n^{-1} \int_{[0,1]^N} I(u_1 \leq \cdots \leq u_k) C(\text{d}u_1, \ldots, \text{d}u_N),
\]

which does not depend on \( F \).

Now let \( F = \mathbb{1}_{[0, \infty)} \) and hence \( \rho_i = 0, j = 1, \ldots, N \). Moreover, note that since \( T_n \) is strictly increasing, the copula of \( T_n(R_{n1}^{(n)}), \ldots, T_n(R_{nk}^{(n)}) \) coincides with the copula of \( R_{n1}^{(n)}, \ldots, R_{nk}^{(n)} \) and converges in distribution to \( C \) for \( n \to \infty \). Further, the marginals converge to \( G \). Therefore,

\[
P(R_{n1}^{(n)} \leq \cdots \leq R_{nk}^{(n)}) = P(T_n(R_{n1}^{(n)}), \ldots, T_n(R_{nk}^{(n)}))
\]

\[
\lim_{n \to \infty} \int_{[0,1]^N} I(u_1 \leq \cdots \leq u_k) C(\text{d}u_1, \ldots, \text{d}u_N)
\]

\[
= \int_{[0,1]^N} I(G(u_1) \leq \cdots \leq G(u_k)) C(\text{d}u_1, \ldots, \text{d}u_N)
\]

\[
= \frac{1}{M} \sum_{i=1}^M I(u_i \leq \cdots \leq u_k) C(\text{d}u_1, \ldots, \text{d}u_N),
\]

which coincides with (8). This proves the first part.

2. This follows immediately from formula (1).

References


