Abstract—Epistasis detection studies focus on finding interactions between Single Nucleotide Polymorphisms (SNPs) that may be linked with susceptibility to and development of complex disease states. Since existing search and score methods for detecting significant SNP combinations focus heavily on the search algorithm, the question of how to best evaluate the epistatic contribution of these interactions still lacks a satisfactory answer. This dissertation proposes a novel methodology for evaluating the performance of six widely used objective functions for epistasis detection based on genotype distribution in the dataset. This analysis reveals a correlation between high scoring power, extreme frequency table values and a simplified form of heritability, based on which a threshold is defined for each objective function in order to pinpoint the cases where SNP interactions can be reasonably assumed to carry epistatic significance. Below this threshold, the combination of two and three complementary objective functions in a multi-objective approach demonstrates an increase in scoring power. This frequency table based approach is innovative in the sense that there is not currently a defined methodology for evaluating and comparing the performance of objective functions. The defined parameters can be applied to real datasets, representing a first step in validating the results of existing epistasis detection methods and promoting the choice of the least complex scoring method possible for specific datasets.

Index Terms – Single Nucleotide Polymorphism; epistasis; frequency table; objective function; multi-objective optimization; case-control studies

1. Introduction

The identification of genes associated with complex human diseases is currently one of the most challenging problems in the field of epidemiology. Epistasis detection studies, which search for SNP-SNP interactions contributing to complex disease states, are of particular interest [1]. However, the problem of epistasis detection is a complex one, representing a heavy computational burden [2] (given the large size of Genome Wide Association Studies (GWAS) datasets and the high complexity of scoring functions) and lacking a suitable methodology for evaluating the contribution of gene interactions to the phenotype [3].

In this dissertation a novel methodology is proposed for evaluating the performance of several of the most used objective functions for epistasis detection – the Chi-Squared test, the G-test, the Mutual Information Score, the AIC/BIC scores, the K2-score and the Gini score – based on dataset characteristics. This frequency table based analysis constitutes a new approach which aims to ultimately validate the performance of objective functions on real datasets for epistasis detection. A connection is established between the presence of extreme values in the frequency table of true associated SNP combinations and high scoring power. Two parameters $d_{h1} + d_{h2}$ and $h$ are defined based on this principle. These parameters allow for the definition of a threshold above which, for the simulated datasets analysed, a single objective scoring approach demonstrates satisfactory scoring power. Below it, this work further demonstrates that a multi-objective Pareto optimized approach combining two or three scoring functions significantly increases scoring power for all functions. Specifically, it is shown that objectives which demonstrate better performance in datasets with different characteristics are more effective in boosting each other’s scoring power.

2. Background

In order to perceive the true complexity of the epistasis detection problem, we must first have an understanding of its underlying biological and mathematical concepts. Given the variety of methods available which deal with SNP-SNP interactions, a broad classification of 50 epistasis detection methods is attempted, selected both for being heavily cited and their employ of the more common scoring methods for interactions. From these, the more used objective functions are singled out for comparison: the Chi-Squared Test, the G-Test, Mutual Information, the AIC and BIC scores, the K2-Score and the Gini Score.

2.1. Gene Expression and Single Nucleotide Polymorphisms

Genes are the basic unit of heredity, involved in the transmission of information to an individual’s offspring. At each genetic locus, humans have two alleles, one inherited from each parent. Alleles are variant forms of genes, with each pair representing what is called the genotype for that specific gene. While humans share the same genes, each human does not have identical patterns of alleles. Assuming that a certain gene has two possible alleles $A$ (dominant) and $a$ (recessive), we call a locus homozygous if its alleles are the same – $AA$ or $aa$ – and heterozygous if its alleles
are different – $Aa$ or $aA$. These variations in alleles will affect the observable traits of an organism, like physical characteristics or the presence of diseases, which we call the phenotype.

The most common type of genetic variation in humans is the SNP [4]. A SNP represents a difference in a single DNA nucleotide that is shared by at least 1% of the population [5]. The frequency of the second most common allele for a given SNP in a certain population is labeled the Minor Allele Frequency (MAF), and is especially useful in order to determine the rarity of a specific variant. Although the presence of a SNP may not manifest itself in any way, SNPs are linked with susceptibility to and development of a wide range of complex diseases like Alzheimer’s [6].

Genome Wide Association Studies (GWAS) find common DNA sequence variants in individuals with or without a common trait like a disease using genome-wide SNP arrays. Variants that are associated with a disease will be found in greater frequency in cases (individuals with the trait) than in controls (individuals without the disease). The most common approach to GWAS deals with the association between SNPs at single loci and a disease state. Studies like these have contributed significantly to the research of single gene diseases, but have provided limited explanations regarding more complex ones, due to problems such as “missing heritability” [7], lack of reproducibility and the existence of a large number of factors that could potentially be related to disease status [8].

2.2. Epistasis

An alternative approach to GWAS appears with the concept of epistasis. Biological epistasis [9] occurs when the effect of an allele at a certain genetic variant is dependant on the existence or absence of a second genetic variant [1]. Statistical epistasis, which is what we refer to when we use computational methods for epistasis detection, is a population phenomenon, made possible by the variability of genotypes in individuals [10], and evaluates the global contribution of genetic variants at different loci to the phenotype [11].

Epistatic interactions between two (pairwise) or more (high-order) SNPs have been found and proven in several complex diseases, such as breast cancer [12] and Alzheimer’s Disease [6]. However, the study of epistasis lends itself to several difficulties, due both to the nature of genomic data as well as the limits of computational capacity in today’s technology.

2.2.1. Evaluating SNP-SNP Interactions. Epistasis detection methods apply dedicated search algorithms in order to check datasets of SNPs for gene–gene interactions. The contribution of candidate SNP combinations to the phenotype is then scored through one or multiple objective functions. The type of search algorithm used to find epistatic combinations is the main focus of a majority of methods for epistasis detection, but it is only part of the problem at hand. The contribution to the phenotype of any SNP combinations found by epistasis detection algorithms must be evaluated in order to detect any meaningful results, and so scoring functions play an important role on deciding which SNP combinations will be chosen as optimal solutions. However, a consensus has not been found regarding which objective function or functions are better indicated to tackle this problem. Given the great variety of scoring methods which can be used in problems like this one, with some authors even introducing their own new tests for use with a particular method [13], finding coherence in the way epistasis detection algorithms are tested may facilitate comparison of the power of different methods.

2.2.2. Single and Multi-Objective Methods. The definition of single objective methods is intuitive. A single test function is applied to the combinations found by the search algorithm, from which a score is obtained that will identify the most statistically significant interactions. However, given the complexity of disease models, the use of a single objective function may not be enough to provide accurate results [14], [15], [16]. In order to reduce the rate of false negative or false positive cases in the final set of solutions as well as increase expressive power [17], multi-objective methods have seen a rise in popularity. These methods leverage the benefits of two or more objective functions in order to obtain a more accurate final solution set, ideally combining them in a way that will balance the score of each test.

2.2.3. Classification of methods for epistasis detection. As the basis for this dissertation, 50 studies introducing epistasis detection methods were chosen for analysis based on several criteria: the type of search algorithm used, the order of epistasis detection (pairwise versus high order), whether one or multiple objective functions are used to evaluate the results of the epistatic search (single versus multi-objective methods), which evaluation scores were used and finally how these tests were combined to create a final score for each SNP combination.

Relative to the objective function used both in the SNP combination search stage and during possible post processing steps (Evaluation stage), each of the 50 epistasis methods is classified according to which objective or objectives are applied, singling out the Chi-Squared test, the G-test, the K2-Score, Information Theory scores AIC and BIC and the Gini score as the most widely applied scoring functions. In total, 18 of the methods studied apply the Chi-Squared test in some stage (either the search stage or post processing of solutions) of the epistasis detection process, followed by 16 for Mutual Information, 9 for the K2-Score, 8 for the G-test, 7 for the Gini score and 5 for Information Theoretic scoring. The most utilized objective function is the $\chi^2$ test when considering all stages of the epistasis detection process. Considering only the search stage, however, the use of Mutual Information (applied in 16 methods) surpasses that of the Chi-squared, which follows it.

This breakdown of epistasis detection methods reveals that no observable pattern is apparent in the application of objective functions. The most widely used tests are
applied by themselves or alongside others, for a wide variety of different search algorithms. In earlier works especially, methods tended to use a wider variety of objectives (or the Chi-Squared test).

2.3. Objective Functions

A brief mathematical description is now provided of the singled out objective functions– the Chi-Squared test, the G-test, the K2 score, Information Theoretic scoring functions AIC and BIC, the Gini Index and Mutual Information.

For all objectives, \( I = 3^k \) and represents the number of possible genotype combinations, which in the context of two-locus epistasis (\( k = 2 \)) corresponds to \( I = 9 \). \( J \) corresponds to the total number of possible phenotypes, which for a case-control study is 2.

2.3.1. Chi-Squared Test. The \( \chi^2 \) Chi-Squared Test is a statistical goodness-of-fit test that is used to determine how well the proportions for the samples at hand fit the population proportions in the null hypothesis [18]. For the problem of epistasis detection, we consider the null hypothesis as there being no epistatic interaction in a given dataset. Generally speaking, when talking about the standard Chi-squared test, authors refer to Pearson’s Chi-squared test, which for the problem of two locus epistasis takes the form

\[
\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},
\]

where \( O_{ij} \) corresponds to \( n_{ij} \), that is, the number of observations where the SNPs take the \( i \)th genotype and the \( j \)th phenotype. Taking \( n_i \) as the number of observations taking the \( i \)th genotype and \( n_j \) as the number of observations taking the \( j \)th phenotype, then \( E_{ij} = \frac{n_i \times n_j}{n} \).

Pearson’s Chi-squared test has been widely used in epistasis detection methods like EpiMode [19] and AntiEpiSeeker [20]. However, the presence of strong marginal effects on individual SNPs in real datasets is likely to cause false positive reports (type I errors) [13].

2.3.2. G-test. The G-test is being increasingly recommended as a replacement for the \( \chi^2 \) test [21]. For a large enough sample size, the G-test and the Chi-squared test will lead to the same conclusions, and for some cases, the G-test is always better than \( \chi^2 \) [22]. For the problem of two-locus epistasis detection, the value of \( G^2 \) is given by

\[
G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} \ln \frac{O_{ij}}{E_{ij}},
\]

with \( O_{ij} \) and \( E_{ij} \) representing the observed and expected frequencies for a particular SNP combination with a particular phenotype expression, respectively.

Recently, methods like FHSA-SED [23] have used the G-test in order to isolate false positives in the solution set found during the search stage, and a recent method SEE [16] employs the G-test in combination with other objective functions.

2.3.3. Gini Score. The Gini index is used to measure the impurity and inequality of a dataset [24], measuring the probability of misclassification of data [25]. The formulation for this function is

\[
Gini = \sum_{i=1}^{I} P_i \left( 1 - \sum_{j=1}^{J} p_{ij}^2 \right),
\]

where \( p_{ij} = \frac{n_{ij}}{n_i} \) is the estimated probability that the \( i \)-th genotype combination is actually associated with the phenotype \( j \). \( P_i = \frac{n_i}{n} \) is the percentage of the \( i \)-th genotype combination in the sample set.

The lower the Gini score, the stronger the association between a specific SNP combination and the phenotype [24].

In addition to being used as one of the objectives during the search stage of several epistasis detection methods like SEE [16], the Gini Index is used by the CART [26] (classification and regression tree), the base classifier used in decision tree based methods like epiForest [27].

2.3.4. Bayesian Networks - The K2-Score. Bayesian Networks are probabilistic graphic models represented by Directed Acyclic Graphs (DAG) and based on the Bayes Theorem. DAGs consist of nodes representing a random variable set, connected by edges which determine the structure of the graph and represent direct dependencies between the variables [28]. In the case of a GWAS, the nodes in a DAG represent genotypes and the phenotype, and their conditional dependencies are represented by the edges. That is, a directed edge going from SNP nodes to the phenotype node implies a correlation between the SNP combination and the phenotype. Thus, a Bayesian network score can measure these dependencies in order to calculate the causality between SNPs and diseases [17]. The K2 score is defined as

\[
K2 = \prod_{i=1}^{I} \left( \prod_{j=1}^{J} \frac{(J - 1)!}{(n_i + J - 1)!} \prod_{j=1}^{J} n_{ij}! \right)
\]

In addition, several methods like MACOED [14] and EIMOABC/D [17] refer to the K2 score in its logarithmic form,

\[
K2^2 = \sum_{i=1}^{I} \left( \sum_{b=1}^{J} \log(b) - \sum_{j=1}^{J} \sum_{d=1}^{J} \log(d) \right).
\]

2.3.5. Information-Theoretic Scoring - AIC and BIC. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used to measure quality of dataset statistical models which deal with the trade-off between the goodness of fit and complexity of the model [29].

Methods like MACOED [14] and FAACOSE [29] utilize logistic regression in order to model the relationship between genotypes and the phenotype, computing the maximized log-likelihood \( \log \text{lik} \) of the model and then applying obtaining the AIC score of the model [14] through the formula

\[
AIC = -2 \log \text{lik} + 2x,
\]
where \( x \) represents the number of estimated parameters in the model. Note that by simply changing the penalty \( 2k \) in the AIC score to \( \ln(n)k \), the formula for the BIC score [30] can be obtained as

\[
BIC = -2 \log \hat{lik} + \ln(n)x. \tag{7}
\]

Calculating the log-likelihood required by these objectives is done through binary logistic regression (since disease state can only take values 1 and 0, for positive and negative respectively); more specifically, using the Iteratively Reweighted Least Squares (IRLS) method [31].

### 2.3.6. Mutual Information

Based on information theory, mutual information quantifies the amount of information obtained about a random variable through another random variable [32]. Given the entropy \( H(Y) \) of a disease state (phenotype) \( y \) considering the combination of SNPs \( X \) with 3 possible genotypes, the joint uncertainty \( H(X) \) of \( k \) SNPs and of the SNP combinations and the phenotype \( H(X,Y) \); the mutual information score [33] of \( X \) and \( Y \) is given by

\[
I(X,Y) = H(X) + H(Y) - H(X, Y) \tag{8}
\]

where \( p_{xy} = p(x_1, \ldots, x_k, y_i) \), \( p_x = p(x_1, \ldots, x_k) \) and \( p_y = p(y_i) \).

Variations of mutual information have been widely used to measure genotype/phenotype associations. ClusterMI [34] screens potentially significant SNP pairs using conditional mutual information that will be used to search for high-order interactions, and ESMO [35] also applies it to quantify the information contribution of a \( k \)-order epistatic combination to disease state (or vice versa).

### 2.4. Multi-objective Analysis

In addition to the identification of suitable objective functions, there is also the issue of how to combine them to obtain a final score. Ideally, a certain SNP combination would be scored at the same level by all objective functions applied. This, however, is often not the case. The score for an interaction may be the best according to one objective, and worse according to another, giving rise to the problem of finding a compromise between scores that will lead to a practical optimal solution.

The Pareto Optimality approach is a comprehensive and broadly applicable solution to this problem, used by several methods like MACOED [14], EIMOABC/D [17] and FAACOSE [29].

Considering a pair of solutions for a multi-objective problem, with \( f_e \) (with \( e = 1, \ldots, M \)) an objective function applied to the problem with \( M \) objectives, and assuming that the operator \( \prec \) denotes worse and \( \succ \) denotes better, a solution \( S_1 \) is said to dominate over another solution \( S_2 \) if [36]:

- \( S_1 \) is no worse than \( S_2 \) for all objectives, i.e. \( \forall e \in \{1, \ldots, M\}, f_e(S_1) \not\preceq f_e(S_2) \);
- \( S_1 \) is strictly better than \( S_2 \) for at least one objective, i.e. \( \exists e \in \{1, \ldots, M\}, f_e(S_1) \succ f_e(S_2) \).

If these conditions are satisfied, we can then say that \( S_2 \) is dominated by \( S_1 \). After applying these rules to all pairs of solutions found by the search algorithm, two sets of solutions are said to be obtained, a dominated set and a non-dominated set. Overall, and despite all the different ways objective functions have been combined in different studies, there is not one approach which we can say is the ideal one for multi-objective evaluation. Most methods in the field of epistasis detection tend to focus most of their attention on the search algorithm being used and its computational efficiency, both to solve the problem of high-order epistasis and to allow for better analysis of large GWAS datasets. The problem of multi-objective evaluation of SNP interactions is one which has not yet found a satisfactory answer, and even several newer methods still apply only a single objective. This issue is one that deserves attention, both for the apparent benefit of leveraging advantages from functions which work in different ways as well as the possibility of finding a multi-objective evaluation strategy which is accurate, widely applicable and independent of the search algorithm.

### 3. Implementation

Based on the issues presented in the Background section, an approach is now introduced for characterizing the performance of the Chi-Squared, G-test, Mutual Information, AIC/BIC, K2-Score and Gini Score objectives based on extreme frequency table values, as well as improving scoring power through the use of multi-objective optimization.

### 3.1. Datasets

A SNP dataset \( D \) for epistasis detection contains \( X + 1 \) variables, which correspond to the \( X \) SNPs to be evaluated plus the disease state (phenotype) \( y \) for each observation. Variables representing each SNP take as possible values 0, 1 or 2, corresponding respectively to the homozygous dominant genotype \( AA \), the heterozygous genotype \( Aa \) and the homozygous recessive genotype \( aa \). The phenotype variable may take as values 1, for a positive disease state (case); or 0, for a negative disease state (control).

The evaluation in this dissertation is done over simulated datasets comprised of one pair of true associated SNPs following a specific epistatic model or penetrance function and \( X - 2 \) noise SNPs following random penetrance values.

#### 3.1.1. Frequency Table

The dataset study and characterization employed in this dissertation relies on the use of frequency tables. For each pairwise SNP combination in a dataset, a \( 2 \times 9 \) frequency table is built, which will then be used as the basis for the calculation of the objective function score. In this table, rows represent the disease state (0 or 1) and columns represent each possible combination
of genotypes between the SNPs. Each cell corresponds to the total number of observations in a dataset where a SNP combination takes the $i$th genotype and the $j$th disease state.

The algorithm and source code for both the processing of datasets and the construction of frequency tables used in the implementation of this dissertation was adapted from the MACOED algorithm [14], which is available at [http://www.csbio.sjtu.edu.cn/bioinf/MACOED/](http://www.csbio.sjtu.edu.cn/bioinf/MACOED/).

### 3.2. Implementation of Objective Functions

After the frequency table for a particular combination of SNPs is calculated, it will then be scored according to each of the objectives studied in this dissertation. In order to provide an analysis of objective functions that is independent from the search algorithm, an exhaustive search algorithm is employed, in which a frequency table is created and scored for each possible SNP combination in each dataset. After each interaction is scored the top SNP combination is determined.

The objective functions Chi-Squared, G-test, Mutual Information, K2-score, AIC, BIC and Gini score were implemented in C++, and no optimization strategies other than simplifications to the algorithms themselves were applied. The scoring for all functions is based on values taken from the frequency table for each SNP combination.

The algorithms for each objective function were, as much as possible and with the exception of some optimizations which did not affect the final result, based on the existing computations of these functions by other methods. This was done in order to maintain some cohesiveness between these implementations and the literature, allowing for possible later comparison with applications in specific epistasis detection methods.

### 3.3. Objective Function Complexity and Run Times

Comparison of the objective functions studied, in addition to being based on the ability of the objective to identify the true associated SNP combination, may rely as well on the different complexity of the algorithms in question. For datasets following parameters in which a less computationally complex objective function demonstrates reasonable success in identifying the true associated SNP combination, the application of this objective may be preferable to a more computationally intensive one with a similar degree of success.

Table 1 displays both the complexities for each objective function, as well as the average run time of scoring the same randomly chosen dataset by each function, with 10 repetitions. In this table, $I$ represents the total number of possible genotype combinations for a SNP pair, $n$ to the number of observations in a dataset, and $iter$ to the number of iterations of the Newton Method applied in IRLS. $\theta_{size}$ is a vector of size 4 in this context. The Chi-Squared, G-test, Mutual Information and Gini scores indicate similar run times, as is the case for the AIC and BIC scores. The run time for AIC/BIC is nearly three times that of the Chi-squared test, and approximately 2.1 times that of the K2-Score.

<table>
<thead>
<tr>
<th>Objective Function</th>
<th>Time Complexity</th>
<th>Mean Run Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>$O(1)$</td>
<td>341.1137</td>
</tr>
<tr>
<td>$G^2$</td>
<td>$O(1)$</td>
<td>347.9209</td>
</tr>
<tr>
<td>MI</td>
<td>$O(1)$</td>
<td>339.3945</td>
</tr>
<tr>
<td>K2</td>
<td>$O(n^2 \times I)$</td>
<td>763.055</td>
</tr>
<tr>
<td>AIC</td>
<td>$O(\theta_{size}^3 \times iter)$</td>
<td>927.1619</td>
</tr>
<tr>
<td>BIC</td>
<td>$O(\theta_{size}^3 \times iter)$</td>
<td>956.712</td>
</tr>
<tr>
<td>Gini</td>
<td>$O(1)$</td>
<td>354.3187</td>
</tr>
</tbody>
</table>

### 3.4. Correlating single objective performance with frequency table values

From cross-comparison between the scoring results for each dataset and the frequency tables of both top scored and true associated SNP combinations, a relationship can be established between the distribution of genotypes in these tables and the scoring performance of each objective function.

Two new variables $d_{91} + d_{82}$ and $h$ are defined in this work in order to condense a frequency table into a single value. These variables demonstrate a correlation to the success or failure of an objective function in identifying the true associated SNP combination in a simulated dataset; and are successful in partitioning the datasets analysed in order to define a subset for which each objective function demonstrates near maximum accuracy identifying this interaction. This partitioning is done through calculation of a threshold for these parameters applied to the frequency table of the top scored SNP combination by an objective function in a particular dataset. Above this threshold, the validity of the top scored SNP combination’s epistatic significance can reasonably be assured.

#### 3.4.1. Extreme Frequency Values

Let us consider, for a specific genotype in a frequency table, the difference between the number of cases and controls, defined as $d_i$, and sort the frequency table according to these values, from lowest $d_1$ to highest $d_9$. The presence of extreme negative or positive values for $d_i$ – representing either a large number of controls over cases or a large number of cases over controls – in a frequency table has been found to positively influence the score of a SNP interaction by all objective functions studied.

Specifically, the distance between the highest and lowest values of $d_i$ in a frequency table, defined as the difference between them, has shown a significant degree of correlation to the outcome of scoring (that is, if the true associated SNP combination was found) by each objective function.

The parameter $d_{91} + d_{82}$,

$$d_{91} + d_{82} = (d_9 - d_1) + (d_8 - d_2).$$

(9)
corresponds to the sum of the distance between the values of the highest and lowest values of \( d_i \), \( d_{91} \); and the distance between the values of the second highest and second lowest values of \( d_i \), \( d_{82} \).

### 3.4.2. Heritability

The \( \hat{h} \) parameter is a variation on the heritability as defined by GAMETES 2.0 for dataset generation. Given the specific context of the analysis in this dissertation, such as basis on the frequency table and datasets with the same number of observations, a simplified version of heritability is introduced in this work which takes the form

\[
\hat{h} = \sum_{i=1}^{I} \frac{n_{i}^{2}}{n_{i}}.
\] (10)

This simplified heritability parameter, unlike \( d_{91} + d_{82} \), considers all the genotypes in a frequency table. This distinction between the two parameters allows for different characterizations of the frequency table. When applied to the group of datasets below the threshold established for \( d_{91} + d_{82} \), it is able to further partition the datasets, identifying an even wider group of datasets for which the objective function demonstrates reasonable accuracy.

### 3.5. Single Objective Analysis and Partitioning

The parameters established in Section 3.4, applied to both the top scored and true associated SNP combinations for each dataset, provide a way to characterize datasets depending on whether each single objective scoring function identified the true associated SNP combination.

First, for each objective, a value 1 or 0 is attributed to each dataset depending on whether the true associated SNP combination received the top score for that objective.

In order to analyse the performance of the individual objective functions, the power metric was applied,

\[
\text{power} = \frac{\text{number of correct datasets}}{\text{total number of datasets}},
\] (11)

where number of correct datasets refers to the number of datasets where the objective function attributed the top score to the true associated SNP combination in the simulated dataset, and total number of datasets refers to the total number of datasets scored.

The values of \( d_{91} + d_{82} \) and \( \hat{h} \) are then calculated for the frequency table of the top scored SNP combination by each objective. Two partitioning stages are applied in order to identify a threshold for these parameters above which the true associated SNP combination is identified in at least 95% of the datasets. In the first stage, this threshold is calculated based on the value of \( d_{91} + d_{82} \) for all datasets. In the second stage, only datasets below this threshold are considered, and a new one is calculated based on \( \hat{h} \). After the application of these two partitioning stages, a group of datasets remains for which the performance of the objective functions could not be characterized by either \( d_{91} + d_{82} \) or \( \hat{h} \). This group contains a combination of datasets where the true associated SNP was found and of those in which it was not. In other words, datasets where the performance of the objective functions under study is not reliable. Increasing the scoring accuracy in these datasets serves as motivation for the application of multi-objective optimization.

The metrics chosen for the purpose of establishing the threshold for \( d_{91} + d_{82} \) or \( \hat{h} \) are the Positive Predictive Value (PPV), and the Negative Predictive Value (NPV), which correspond to the proportions of positive and negative results that are true positives and true negatives, respectively.

In order to determine where the threshold lies for each objective and each of the two parameters, cut points were calculated for dividing the datasets using the R package cutpointsr [37]. This package allows for calculating “optimal” cut points for a variable based on an input constraint that is maximized or minimized.

### 3.6. Multi-objective Algorithm

In several groups of datasets following different Models and MAFs, a single objective scoring approach reveals low rates of success in identifying the true associated SNP combination. Increasing the scoring power in these groups is the main motivation for the implementation of Multi-objective or Pareto Optimization.

The goal of this optimization is to find objectives whose scoring methods are complementary. That is, that demonstrate greater rates of success for different distributions of frequency tables, and as such may broaden the variety of datasets for which there is a reasonable degree of confidence in the scoring accuracy.

The implementation of Pareto optimization in this dissertation is a simplified version, specific to the goal of obtaining the first front of non-dominated SNP pairs for each SNP dataset and each combination of objectives. By analysing which SNP combinations make up this first front, and specifically if the correct SNP interaction (the one containing epistatic information) is included in the first front, we can measure the power of each multi-objective combination.

In order to measure the performance of the different combinations of objective functions in a multi-objective implementation, the same power metric used for evaluating single objective performance is utilized, as described in Eq. (11). In this context, however, # of correct datasets refers to the number of datasets which included the true associated SNP combination in the reported Pareto first front.

### 4. Results

Abridged results are now presented which validate the approach defined in this work. This includes an analysis of single and multi-objective performance, frequency table analysis which supports the defined parameters and a demonstration of threshold generation for the AIC score objective.
4.1. Datasets

The datasets used in this work were generated using the GAMETES_2.0 [38] software, an epistatic model generator which allows for the specification of parameters for the true associated SNP combination in a generated dataset.

Three 2-locus DME (displaying marginal effects) models – one additive, one multiplicative and one threshold – with four different sets of parameters were used to generate these datasets. These models, defined by Jing et al. [27] are the same ones utilized in simulated dataset testing for the MACOED method [14]. For each model, four different values of MAF were considered, 0.05, 0.1, 0.2 and 0.5.

4.2. Single Objective Analysis

4.2.1. Power. Despite the scope of this dissertation including three models, for the purposes of this summarized analysis we will focus on datasets following Model 1, for which the objectives exhibit the lowest performance of all three models and the most variation in scoring power between functions.

The power (as defined in Eq. (11)) of each objective function in Model 1 datasets for each MAF considered is represented in Figure 1a. Since the performance of the AIC and BIC scores is identical, only the AIC score is represented. This similarity can be attributed to the fact that the difference in these scoring methods lies simply in the penalty term added to the result of the calculated log-likelihood, as can be seen in Eqs. (6) and (7). Regardless, this objective yields the best results overall, followed by the K2-score, Mutual Information score, the G-test, the Chi-Squared test and lastly the Gini score.

It is clear that different objectives perform better or worse depending on the the MAF attributed to the true associated SNP combination in the dataset. The most extreme example of this is the Gini score, which presents the lowest power in most cases, but claims the highest power when the MAF of the true associated SNP is 0.5. On the other hand, the power for all AIC, K2, Mutual Information and $G^2$ scores drops when the MAF rises to 0.5, while $\chi^2$ also fares better than for other MAFs.

The performances of the G-test and Mutual Information score also present nearly identical results This similarity, much like the one between AIC and BIC, can be attributed to the also similar computation of these two scores. There is also some correspondence between the performance of these two objectives and that of the $\chi^2$ test, which is to be expected given that the G-test is a variation on the $\chi^2$ test, and both test statistics follow a Chi-squared distribution.

4.2.2. Relationship between frequency distribution and scoring power. Given the power reported above, the relationship between performance and frequency table values can now be explored. Figure 2 represents, for datasets following Model 1 and MAF 0.05, the single objective power for each objective, the median distribution of genotypes across the frequency table of the true associated SNP combination, the distribution of $d_i$ for each genotype $i$ in the frequency table of the true associated SNP combination and the distribution of $d_i$ for each genotype $i$ in the frequency table of the top scored SNP pair by the AIC score. The values of $d_i$ in both plots are sorted from the lowest to the highest. Figure 3 corresponds to the same plots, but for datasets following Model 3 and MAF=0.5.

From comparison of the two Figures, it is obvious that the different models and MAFs considered for dataset generation will result in differences in the frequency tables for the true associated SNP combination in each dataset. Datasets with MAF=0.05 for both true associated SNPs present, as expected, very high counts of the AA/BB combination of genotypes, since the set frequency for the minor allele is so low. For the same reason, the next highest genotype combinations are AA/Bb and Aa/BB. The genotype values tend toward equalization with the increase of MAF, and datasets with MAF=0.5 in general present much more balanced frequency tables across all the genotypes, since there is an approximately equal number of major and minor alleles for these SNPs across all observations.

We can also see that Model 1 datasets present a more balanced number of cases versus controls than Model 3. For MAF=0.05 the values of $d_1$ approach 0 for all genotypes $d_2$ through $d_7$. Given the low counts for most genotypes in the frequency tables for this MAF, it stands to reason that only 3 genotypes, $d_1$, $d_4$ and $d_5$ diverge from 0. The similarity between the $d_i$ values for the top scored SNP combination and the true associated one in Figure 2 suggests that the reason for the low power of the AIC score in this group of datasets is due to the similarity in the values of $d_i$ between the true associated SNP and the randomly assigned noise SNPs that constitute the rest of the dataset.

Model 3 presents both power very close to 1 for all objectives with the exception of the Gini score and the most extreme values for $d_1$ when compared to other Models. Given that the noise SNPs for this group of datasets can be assumed to follow the same distribution as those for Model 1, we can infer that the values of $d_i$ for the true associated SNP combination in these datasets present much more extreme values than those of the noise SNPs. This further supports the hypothesis that extreme $d_i$ values boost the score of the true associated SNP combination.

4.2.3. Partitioning of datasets by $d_{g1} + d_{g2}$. The value of $d_{g1} + d_{g2}$ for the frequency table of the top SNP combination for each simulated dataset according to the AIC score is plotted in Figure 4. In this plot, data points in blue represent datasets where the true associated SNP combination was correctly identified as the top score, while data points in red represent datasets where the true associated SNP combination did not obtain the top score. Datasets are numbered from 1 to 12000, with the lowest numbered datasets corresponding to Model 1 and MAF 0.05, and follow the Model/MAF increase, with the highest numbered datasets corresponding to Model 3 and MAF=0.5.

A value for $d_{g1} + d_{g2}$ can be determined for the datasets analysed above which the correct result is obtained for
the majority of datasets. Below this threshold, this metric fails to separate datasets for which the true associated SNP combination was identified from those where it was not.

Similar distributions can be plotted for the remaining objective functions, preserving the general “shape” of the plot and varying slightly in the values that can be attributed to the threshold.

The concrete threshold for separation of datasets which identifies the true associated SNP combination was established through the use of cut points, and for the AIC score is at $d_{91} + d_{82} = 143$.

For $\chi^2, G^2, \text{Mutual Information, the K2-score and AIC}$, roughly half of the total number of datasets remain on each subset above and below the threshold, while for the Gini score, this ratio highly favors the subset below the threshold. This is to be expected, given the low power of this objective.

For the values of $d_{91} + d_{82}$ above the threshold and at least up to the maximum value of $d_{91} + d_{82}$ recorded in this study, it may be reasonably stated that if a SNP combination with $d_{91} + d_{82}$ value in this range is determined by one of these objectives to possess epistatic significance, this classification is accurate.

After excluding the datasets above the cut point for each objective, a number of datasets remain for which $d_{91} + d_{82}$ cannot accurately separate datasets in which the true associated SNP combination was identified from those where it was not. A further exploration of parameters which can better characterize the frequency tables of the top SNP combinations in these remaining datasets is then called for, particularly one which makes use of all genotype counts.

4.2.4. Partitioning of datasets by $\hat{h}$. By plotting the value of $\hat{h}$ for the frequency tables of the top scored SNP combination by the AIC score in the datasets below the thresholds determined for $d_{91} + d_{82}$, the results in Figure 5 are obtained.

In this plot, a further subset of datasets can be identified for which the true associated SNP combination is found reliably, and a new threshold can be established which separates these datasets. For a SNP combination with a value of $d_{91} + d_{82}$ below the threshold previously established for this parameter, a degree of certainty in correct classification may still be attributed if the value of $\hat{h}$ for this combination falls above the new threshold determined to correspond to $\hat{h} = 409.03$ for the AIC score.

Establishing the threshold for $\hat{h}$ for is done the same way as for the calculation of the $d_{91} + d_{82}$ threshold.
However, from observation of Figure 5 it is obvious that a number of datasets remain below the threshold that can be defined for $h$. This suggests that additional parameters to the ones defined in this study exist which play a part in defining the relationship between the frequency table and objective performance. In other words, given a dataset where the objective function attributed the top score to a SNP combination with $d_{31} + d_{82}$ and $h$ between these ranges, these parameters cannot provide reasonable confidence that this combination would have true epistatic power.

5. Multi-Objective Power

A potential solution for increasing the scoring power of these functions, especially in datasets with values of $d_{31} + d_{82}$ and $h$ falling within the determined non-characterizable ranges, is the use of multi-objective optimization for scoring SNP combinations.

A graphical representation of the power for two objective and three objective implementations for Model 1 is represented in Figures 1b and 1c. For Models 2 and 3, most objectives achieve near maximum power for all groups of datasets when applied in a single objective approach. Therefore, the implementation of multi-objective scoring is more relevant in the context of Model 1.

The most significant increase in power comes from a two objective versus single objective implementation, with only slight increases achieved through the addition of a third objective function. This suggests that for these datasets a possible four-objective implementation with these functions may show little improvement in performance over a three objective one. An increase to near maximum power in
Model 1 and MAFs 0.05, 0.1 and 0.2 is unlikely even if all six objective functions are combined in multi-objective optimization. The heavy computational cost of calculating scores according to a large number of objective functions must also be considered versus the possible improvement in power.

Overall, the combination of the AIC, K2 and Gini scores demonstrates the best performance over all groups of datasets. Out of the objective functions studied in the context of this dissertation, the AIC and K2 scores are the most computationally complex, with AIC requiring the computation of a logistic regression and the K2-score the calculation of several triangular numbers. This more complex computation, combined with the different approach of the Gini score when compared to all other objectives may be the explanation for the success of this combination.

The goal of multi-objective optimization in the context of epistasis detection is increasing both power and applicability over a wide range of models and dataset characteristics. The outcome of this study indicates that when considering a dataset with unknown attributes, introducing a multi-objective scoring approach effectively increases the changes of identifying SNP combinations with epistatic significance, making headway towards the achievement of this goal.

6. Conclusion

This work successfully identified genotype distribution based parameters which translate to correct identification of true associated SNP pairs by each objective. These parameters establish a connection between extreme frequency table values and high reported power by an objective function. By focusing on direct examination of the datasets and comparison to objective scoring power, this approach sets the stage for validation of real GWAS dataset scoring, since it does not rely on the simulated dataset model.

The implementation of objective functions developed in this dissertation takes advantage of frequency table data for calculating epistasis scores. This has the potential to lower computational complexity for multi-objective scoring, since only one analysis of the dataset itself needs to be done independently of the number of objectives.

This work demonstrates that while for some datasets a single objective approach may be sufficient in identifying SNP interactions, multi-objective approaches significantly increase the scoring power. A Pareto optimization algorithm, reporting only the first front of non-dominated solutions, revealed different levels of performance depending on the combination of objectives used and the Model and MAF of the true associated SNP interaction. A two objective approach was found to improve scoring performance significantly over a single objective one, with lesser improvements when a third objective is added.

6.1. Future Work

This dissertation represents a first step in correlating frequency table data with objective function performance. However, further analysis into the parameters defined in this work is necessary in order to validate their applicability. Exploring datasets generated using different models than the ones already adopted has the potential to either expand the applicability of this study or highlight other factors which favorably or negatively influence the score of an SNP combination. Relaxing the definition of correctly identified SNP interactions” for a single objective approach may also reveal further information when compared to the more strict metric applied in this dissertation.

References


