mathematics

Article

HSIC$_{CR}$: A Lightweight Scoring Criterion Based on Measuring the Degree of Causality for the Detection of SNP Interactions

Junxi Zheng [1], Juan Zeng [1], Xinyang Wang [2,*], Gang Li [3], Jiaxian Zhu [4], Fanghong Wang [5] and Deyu Tang [1,*]

[1] School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China
[2] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China
[3] School of Information Technology Engineering, Guangzhou College of Commerce, Guangzhou 511363, China
[4] School of Computer Science, Zhaoqing University, Zhaoqing 526061, China
[5] School of Bussiness, Zhijiang College of Zhengjiang University of Technology, Shaoxin 310024, China
* Correspondence: wxyyuppie@bjfu.edu.cn (X.W.); tangdeyu@gdpu.edu.cn (D.T.)

Abstract: Recently, research on detecting SNP interactions has attracted considerable attention, which is of great significance for exploring complex diseases. The formulation of effective swarm intelligence optimization algorithms is a primary resolution to this issue. To achieve this goal, an important problem needs to be solved in advance; that is, designing and selecting lightweight scoring criteria that can be calculated in $O(m)$ time and can accurately estimate the degree of association between SNP combinations and disease status. In this study, we propose a high-accuracy scoring criterion (HSIC$_{CR}$) by measuring the degree of causality dedicated to assessing the degree. First, we approximate two kinds of dependencies according to the structural equation of the causal relationship between epistasis SNP combination and disease status. Then, inspired by these dependencies, we put forward this scoring criterion that integrates a widely used method of measuring statistical dependencies based on kernel functions (HSIC). However, the computing time complexity of HSIC is $O(m^2)$, which is too costly to be an integral part of the scoring criterion. Since the sizes of the sample space of the disease status, SNP loci and SNP combination are small enough, we propose an efficient method of computing HSIC for variables with a small sample in $O(m)$ time. Eventually, HSIC$_{CR}$ can be computed in $O(m)$ time in practice. Finally, we compared HSIC$_{CR}$ with five representative high-accuracy scoring criteria that detect SNP interactions for 49 simulation disease models. The experimental results show that the accuracy of our proposed scoring criterion is, overall, state-of-the-art.

Keywords: lightweight scoring criterion; causality; SNP interactions; measuring statistical dependencies

MSC: 62H20

check for updates

Citation: Zheng, J.; Zeng, J.; Wang, X.; Li, G.; Zhu, J.; Wang, F.; Tang, D. HSIC$_{CR}$: A Lightweight Scoring Criterion Based on Measuring the Degree of Causality for the Detection of SNP Interactions. *Mathematics* **2022**, *10*, 4134. https://doi.org/10.3390/math10214134

Academic Editor: Konstantin Kozlov

Received: 24 September 2022
Accepted: 3 November 2022
Published: 5 November 2022

1. Introduction

Since many complex diseases are usually caused by multiple genes and multiple factors, in recent years, with the emergence of high-throughput genotypic technology, genome-wide association analysis (GWAS) has been one of the main methods used to study complex diseases. Furthermore, the identification of single-nucleotide polymorphism (SNP) interactions from GWAS data is of great importance for exploring the explanation, prevention and treatment of complex diseases [1,2]. Therefore, over the past decade, this research topic has attracted considerable attention [3–9].

It is well known that SNP interactions represent combinations of multiple SNPs that affect complex diseases in a linear or non-linear manner, also known as k-order epistasis SNPs. The research topic of detecting k-order epistasis SNPs is a typical case of combinatorial optimization problems in k-dimensional discrete space ($k \in \{2, 3, 4, 5\}$ in practice), and swarm intelligence optimization (SIO) algorithms are one of the main methods used to solve the problems [9–11]. For this study to be successful, an important problem needs to be solved in advance; that is, designing and selecting lightweight scoring criteria that can be calculated in $O(m)$ time and can accurately estimate the degree of association between SNP combinations and disease status.

To date, few lightweight scoring criteria can accurately estimate the degree of association of SNP combinations with disease status in most disease models due to the widely varying characteristics of different disease models. As one of the primary methods used to work on this combinatorial optimization problem, SIO algorithms mostly tackle this issue by combining multiple criteria [9–13]; however, using too many objective functions will often make the proposed algorithm difficult to converge effectively. Therefore, picking a few high-accuracy objective functions instead of using too many objective functions can dramatically improve the performance of the used algorithms [14,15].

This paper's goal is not to contribute toward fixing the issue entirely but to use a different methodology to propose a scoring criterion that can accurately estimate the associations in most disease models. The contributions of this paper are:

1. We propose a high-accuracy scoring criterion based on measuring the degree of causality that integrates a widely used method of measuring statistical dependencies (HSIC);
2. We put forward an efficient algorithm of computing HSIC on two variables with a small sample in $O(m)$ time, thus enabling us to compute HSIC_{CR} in $O(m)$ time in practice.

2. Related Works

So far, the proposed lightweight scoring criteria can be roughly divided into two categories.

The first category covers various approaches, which are so-called Bayesian scoring criteria. The Bayesian scoring criteria calculate the posterior probability distribution, proceeding from a prior belief on the possible DAG models, conditional on the data [16]. The K2-Score is an efficient Bayesian scoring criterion that obtains priors under the assumption that all DAG models are equally likely. Other such scoring criteria representatives include the Bayesian Dirichlet equivalent (BDe) scoring criterion and the Bayesian Dirichlet equivalent uniform (BDeu) scoring criterion [4,17–19].

The second category is usually known as the information-theoretic scoring criteria. Mutual information (Mi) is a lightweight method but has preferences for certain disease models [6,20]. The JS divergence is a symmetrized divergence measure, derived from the Kullback–Leibler (KL) divergence, which is an asymmetric divergence measure of two probability distributions [21]. This approach can be utilized to evaluate the SNP genotype deviation between control samples and case samples. Lately, joint entropy (JE) and normalized distance with joint entropy (ND-JE) have been proposed as criteria for guiding harmony search algorithms to discover clues for exploring the epistasis of SNP combinations [9].

There are also a few approaches that do not fall into any of the above two main categories. For example, the LR is a composite indicator that reflects both sensitivity and specificity and can be used for a related measure to find the likelihood difference between a disease-causing SNP combination and an SNP combination that is not involved in the disease process [22,23].

In statistics, G-test is a significant test method of natural ratio or maximum likelihood. In recent years, scholars have tended to use the G-test independence test instead of the chi-square independence test recommended in the past. In genome association analysis, G-test has been extensively used. Different from other scoring criteria, G-test will provide its p-value when measuring the relationship between an SNP combination and sample state, which can indicate whether the SNP combination has a significant relationship with the sample state [24].

Published research has found that the results were different when employing different scoring criteria, The K2-Score has been widely used to evaluate the association. This measure has a high capacity for detecting SNP interactions and is superior in discriminating certain disease models with low marginal effects. However, for the interaction model with low minor allele frequencies (MAFs) and low genetic heritability (h), the K2-Score has a low performance in detecting high-order SNP interactions. The ND-JE is proposed based on the properties of the disease-causing SNP combination models without marginal effects, so this metric is more suitable for evaluating diseases with this type of mode. The LR-score aims to discover the relationship between likelihood differences in functional SNP combinations and non-functional SNP combinations. The method is well adapted to unbalanced datasets of cases and controls. In practice, the use of the G-test as a single evaluation criterion for detection is found to be inadequate, as there are often many SNP combinations with G-test values close to 0 [6,9,11].

The scoring criterion proposed in this work is based on the theory of causality. It is distinct from the theoretical approach taken by the current existing criteria. From the perspective of a comparison with correlation, causality strictly distinguishes "cause" variables and "result" variables, and plays an irreplaceable role in revealing the mechanism of the occurrence of things and guiding intervention behaviors [25]. Thus, the proposed criterion is a useful for and complementary to the current existing criteria.

3. Methodology

3.1. Concepts and Terms

In this work, $x = \{x_1, x_2, \ldots, x_n\}$ represents a set of n SNP loci, and $X = \begin{bmatrix} x_{11} & x_{21} & . & x_{1n} \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & . & x_{mn} \end{bmatrix}$ is a set of m samples of x; $Y = \{y_1, y_2, \ldots, y_m\}^t$ denotes a set of m samples of disease status y. For $\forall 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m$, $i, j \in \mathbb{Z}$, $x_{ij} \in \{0, 1, 2\}$, x_{ij} is equal to 0, 1 and 2, which implies that it is the homozygous major allele (AA), heterozygous allele (AT), and homozygous minor allele (TT), respectively; $y_j = 0$ for control and $y_j = 1$ for case; $D = (X,Y)$ is a dataset with m samples.

Definition 1 (k-order epistasis SNP combination)**.** *Let S_k = {{$x_{i_1}, x_{i_2}, \ldots, x_{i_k}$}} be a collection of a set with k SNP loci ($1 < k < n$). $f(D)$: $S_k \rightarrow \mathbb{R}^+$ is a score function used for measuring the association between any k SNP loci and disease status y based on a dataset D. If x has the k-order epistasis SNP combination on y (denoted as s_k^*, $s_k^* \in S_k$), and $f(D)$ is a correct score function (or scoring criterion), then, for $\forall s_k \in S_k$, $s_k \neq s_k^*$, $f(D)(s_k^*) < f(D)(s_k)$ or $f(D)(s_k^*) > f(D)(s_k)$.*

3.2. Causal Relationship

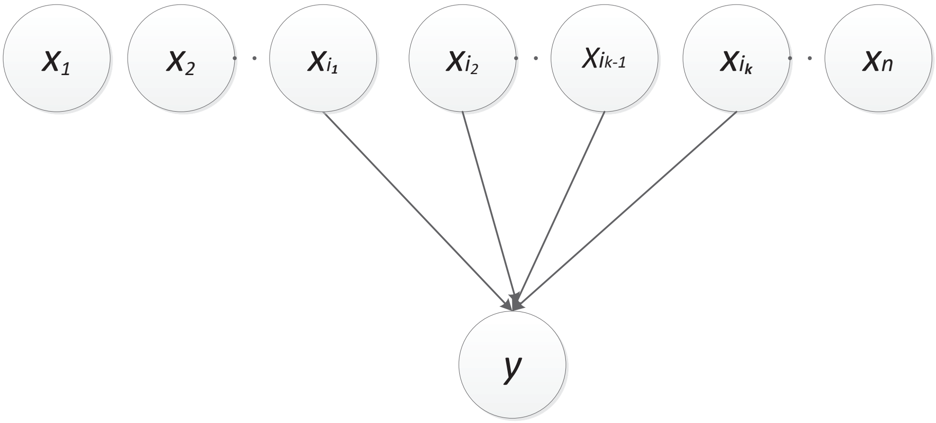
According to how the data are generated, the structural equation of the causal relationship between epistasis SNP combination and disease status can be modeled as [26]:

$$y = f(s_k^*) + e_y, \tag{1}$$

where e_y is the noise variable.

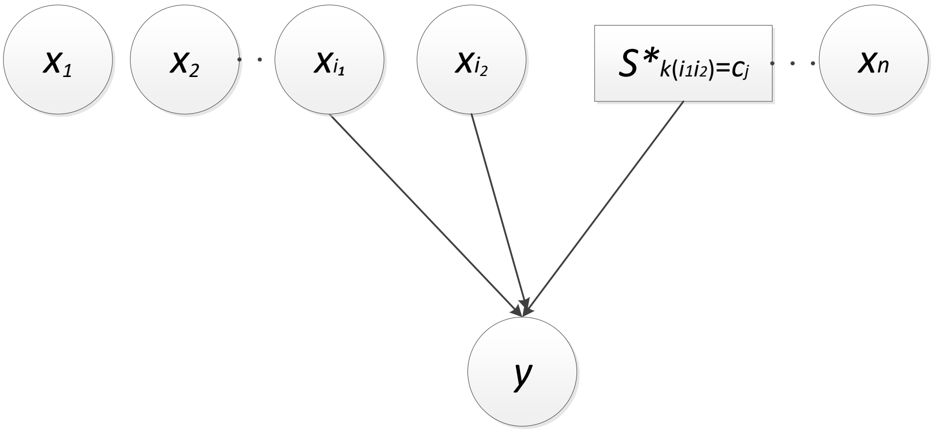
From the above equation, we can find that s_k^* and e_y are independent (denoted as $s_k^* \perp e_y$). In other words, among all $s_k \in S_k$, s_k^* and e_y have the lowest degree of dependency. However, it is unrealistic to measure the dependence degree of any s_k and e_y as the evaluation criterion for epistasis detection, because it requires too high a computational cost to obtain data generated by e_y based on the regression method.

Thus, this paper herein let e_y be the constant 0, i.e., $y \approx f(s_k^*)$, which approximately introduces two kinds of dependencies as described in Figures 1 and 2, respectively [25]. Obviously, the dependence between s^* and y is direct (denoted as $s_k^* \not\perp y$); and the other is derived from the v-structure, i.e., x_{i_1} and x_{i_2} are dependent given a value of y and $s_{k(i_1 i_2)}^*$ (denoted as $x_{i_1} \not\perp x_{i_2} | y = y_i, s_{k(i_1 i_2)}^* = c_j$). Let $s_k^* = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{k-1}}, x_{i_k}\}$, $s_{k(i_1 i_2)}^*$ represent $s_k^* \setminus \{x_{i_1}, x_{i_2}\}$. In particular, the set $s_{k(i_1 i_2)}^*$ is empty when k is equal to 2.



$$S_k^* \not\perp y, S_k^* = \{X_{i_1},\ X_{i_2} ..., X_{i_{k-1}},\ X_{i_k}\}$$

Figure 1. Direct dependence.



$$S^*_{k(i_1 i_2)} = \{X_{i_3},\ X_{i_4} ..., X_{i_{k-1}},\ X_{i_k}\}, \forall S_{k(i_1 i_2)} \in \{c_1, c_2, ..c_{3^{k-2}}\}$$

$$X_{i_1} \perp X_{i_2}, X_{i_1} \not\perp X_{i_2} | (y = y_i, S^*_{k(i_1 i_2)} = c_j), \forall y_i \in \{0,1\}$$

Figure 2. V-structure-related dependence.

3.3. Scoring Criterion

These two kinds of dependencies described above inspire us to raise this scoring criterion, which integrates a widely used method of measuring statistical dependencies based on kernel functions (HSIC).

For $\forall s_k \in S_k$, then, for $\forall x_{i_1}, x_{i_2} \in s_k$, given $y = p$ ($p \in \{0,1\}$) and $s_{k(i_1 i_2)} = q$ ($q \in \{c_1, c_2, \ldots, c_{3^{k-2}}\}$), let $D^{pq}_{i_1 i_2} = (X^{pq}_{i_1 i_2}, Y^{pq}_{i_1 i_2})$ be a slice of D on x_{i_1}, x_{i_2} under the constraint; $m^{pq}_{i_1 i_2}$ is the number of rows of the data slice ($m^q_{i_1 i_2} = m^{0q}_{i_1 i_2} + m^{1q}_{i_1 i_2}$); $\mathrm{HSIC}(X, Y)$ is used to measure the degree of statistical dependence of two random variables (x and y) based on dataset (X, Y); the scoring criterion can be computed by the following Equations (2)–(6).

$$b^q_{i_1 i_2} = \frac{m^{0q}_{i_1 i_2}}{m^q_{i_1 i_2}} \mathrm{HSIC}(X^{0q}_{i_1}, X^{0q}_{i_2}) + \frac{m^{1q}_{i_1 i_2}}{m^q_{i_1 i_2}} \mathrm{HSIC}(X^{1q}_{i_1}, X^{1q}_{i_2}) \tag{2}$$

$$b^q = \sum_{i_1 \neq i_2, x_{i_1}, x_{i_2} \in s_k} \left(\frac{m^q_{i_1 i_2}}{\sum_{i_a \neq i_b, x_{i_a}, x_{i_b} \in s_k} m^q_{i_a i_b}} b^q_{i_1 i_2} \right) \tag{3}$$

$$\overline{m}^q = (\sum_{i_a \neq i_b, x_{i_a}, x_{i_b} \in s_k} m^q_{i_a i_b})/c_k^2 \quad (4)$$

$$\text{HSIC}^q_{CR} = \frac{m}{m+\overline{m}^q}\text{HSIC}(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k}) + \frac{\overline{m}^q}{m+\overline{m}^q} b^q \quad (5)$$

$$\text{HSIC}_{CR}(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k} : data) = \sum_{q=1}^{q=3^{k-2}} \text{HSIC}^q_{CR} \quad (6)$$

To facilitate the reader's understanding, we now define following notations:

1. The value of $b^q_{i_1 i_2}$ is a linear weighted sum of HSIC$(X^{0q}_{i_1}, X^{0q}_{i_2})$ and HSIC$(X^{1q}_{i_1}, X^{1q}_{i_2})$ based on respective sample sizes;
2. For all $x_{i_1}, x_{i_2} \in s_k$, the value of b^q is a linear weighted sum of all $b^q_{i_1 i_2}$ since there are c_k^2 v-structures given $s_{k(i_1 i_2)} = q$;
3. The value of HSIC$^q_{CR}$ is a linear weighted sum of HSIC$(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k})$ and b^q based on sample size of $D_{i_1 i_2 \ldots, i_k}$ and average sample size of all $D^q_{i_a i_b}$, which is a component and basis of HSIC$_{CR}$;
4. In particular, $b^q_{i_1 i_2} = b^q$ as is HSIC$^q_{CR}$ = HSIC$_{CR}$ when $k = 2$;
5. For robustness purposes, let $b^q_{i_1 i_2}$ = 0 if and only if the denominator of the weighted factor term is 0, like b^q;
6. The effort to calculate scoring criterion is reduced to calculate HSIC$(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k})$ once, and is reduced up to $C_k^2 * 3^{k-2}$ times to calculate the type of problem $b^q_{i_1 i_2}$ (fortunately, $k \in \{2,3,4,5\}$ in practice).

Thus, our estimate of s_k^* can eventually be obtained by solving the problem

$$\underset{s_k \in S_k}{maxf}(D)(s_k) = \text{HSIC}_{CR}(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k}) \quad (7)$$

3.4. Method for Measuring Statistical Dependence

3.4.1. HSIC

HSIC is a measuring statistical dependence criterion proposed by other authors [27,28] based on the eigenspectrum of covariance operators in reproducing kernel Hilbert spaces (RKHSs), denoted by HSIC(P_{xy}) as follows:

$$\begin{aligned} \text{HSIC}(P_{xy}) = {} & E_{x,x',y,y'}[k(x,x')l(y,y')] + \\ & E_{x,x'}[k(x,x')]E_{y,y'}[l(y,y')] - \\ & 2E_{x,y}[E_{x'}[k(x,x')]E_{y'}[l(y,y')]], \end{aligned} \quad (8)$$

where $k(,)$ and $l(,)$ are two kernel functions.

Let $\mathbb{X}$ and $\mathbb{Y}$ be the separable sample spaces of random variables x and y, respectively, assuming that $(\mathbb{X}, \Gamma)$ and $(\mathbb{Y}, \Lambda)$ are furnished with probability measures p_x, p_y, respectively (Γ being the Borel sets on $\mathbb{X}$, and Λ the Borel sets on $\mathbb{Y}$); p_{xy} is a joint measure over $(\mathbb{X} \times \mathbb{Y}, \Lambda \times \Gamma)$; HSIC$(P_{xy}) \geq 0$ (the higher the degree of dependence of x and y, the greater the value) and HSIC(P_{xy}) is zero if and only if x and y are independent.

In order to show that HSIC is a practical criterion for measuring independence or the degree of dependence given a finite number of observations, it consists of an empirical estimator with $O(m^{-1})$ expectation bias, denoted by HSIC(D), formulated as follows:

$$\text{HSIC}(D) = (m-1)^{-2} trace(KHLH), \quad (9)$$

where D:=$\{(x_1, y_1), \ldots, (x_m, y_m)\} \sqsubseteq \mathbb{X} \times \mathbb{Y}, K, H, L \in \mathbb{R}^{m \times m}$,$K_{i,j} := k(x_i, x_j), L_{i,j} := l(y_i, y_j)$, $H_{i,j} := \delta_{ij} - m^{-1}$.

An advantage of HSIC compared with other kernel-based independence criteria is that it can be computed in $O(m^2)$ time. However, such computational costs are too high as an integral part of the scoring criterion. Fortunately, the sample space of the disease status, SNP loci and k-order SNP combination is finite discrete. Thus, immediately below, we put forward an efficient HSIC calculation method for variables with a small sample, which can be approximately calculated in $O(m)$ time. Thus, as $k \in \{2,3,4,5\}$, we can compute HSIC$_{CR}$ in $O(c_k^2 \times m) \approx O(m)$ time in practice.

3.4.2. Efficient Computation

Proposition 1 (Efficient computation)**.** *Let x and y be two random discrete variables with p and q states, respectively, where $p^2 \times q^2 < m$, or $p^2 \times q^2 \approx m$. Then, we can compute $trace(KHLH)$ in $O(m)$ time.*

Proof. Let e be a column vector with a length of m, $e = \begin{bmatrix} 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$, $L = \begin{bmatrix} {l_1}^t \\ \cdot \\ \cdot \\ {l_m}^t \end{bmatrix}$, and I be an identity matrix with a size of $m * m$; we have $H = I - \frac{1}{m}ee^t$ and $LH = L - \frac{1}{m}Lee^t$.

As $\frac{1}{m}Le = \begin{bmatrix} \frac{1}{m}{l_1}^t e \\ \cdot \\ \cdot \\ \frac{1}{m}{l_m}^t e \end{bmatrix}$, which implies that each i-th element of $\frac{1}{m}Le$ is the mean of the corresponding row elements of L, we have $L - \frac{1}{m}Lee^t = \begin{bmatrix} {l_1}^t \\ \cdot \\ \cdot \\ {l_m}^t \end{bmatrix} - \begin{bmatrix} \frac{1}{m}\overline{l_1} \\ \cdot \\ \cdot \\ \frac{1}{m}\overline{l_m} \end{bmatrix} e^t$, where each $\overline{l_i}$ is the sum of the corresponding row elements of L.

Let $\overline{L} = \begin{bmatrix} \frac{1}{m}\overline{l_1} \\ \cdot \\ \cdot \\ \frac{1}{m}\overline{l_m} \end{bmatrix} e^t$ be an $m \times m$ matrix ($\overline{L}_{ij} = \frac{1}{m}\overline{l_i}$) and $\overline{K} = \begin{bmatrix} \frac{1}{m}\overline{k_1} \\ \cdot \\ \cdot \\ \frac{1}{m}\overline{k_m} \end{bmatrix} e^t$ be an $m \times m$ matrix ($\overline{K}_{ij} = \frac{1}{m}\overline{k_i}$); we have $trace(KHLH) = trace((K - \overline{K})(L - \overline{L}))$.

Let P be an $m \times m$ row transformation matrix; we have $trace(PKHPLH) = trace(P(K - \overline{K})P(L - \overline{L})) = trace((K - \overline{K})(L - \overline{L}))$=$trace(KHLH)$.

Thus, without a loss of generality, we can assume that: D=$\{(c_1, y_{i_1}), \ldots (c_1, y_{i_2 - 1}), (c_2, y_{i_2}), \ldots (c_2, y_{i_3 - 1}), \ldots (c_p, y_{i_p}), \ldots (c_p, y_{i_m})\}$, i.e., the number of observed instances of x with the value of c_j (denoted as $xt(c_j)$) is $i_{j+1} - i_j (i_{p+1} = i_m + 1)$; K can be viewed as a $p \times p$ partitioned matrix. Let K^{jl} be the $j \times l$th block having $xt(c_j) \times xt(c_l)$ elements, all having the same value ($k(c_j, c_l)$). Let $\hat{K} = K - \overline{K}$, where all elements in $\hat{K}^{jl}$ have the same value equal to $k(c_j, c_l) - \frac{\sum_{n=1}^{p} xt(c_n) \times k(c_j, c_n)}{m}$ (denoted as $\hat{K}(j,l)$).

Let $yx(j,i)$ be the number of observed instances with the value (c_i, d_j) ($1 \leq j \leq q$, d_j is the j-th state of y); the definition of $\hat{L}$ and $\overline{L}$ is similar to that of $\hat{K}$ and $\overline{K}$. We also view $\hat{L}$ as a $p \times p$ partitioned matrix, where each L^{jl} has the same number of rows and columns as K^{jl}; for $\forall p, 1 \leq p \leq m$, $\frac{1}{m}\bar{l}_p = \frac{\sum_{v=1}^{q} yt(d_v) \times k(d_h, d_v)}{m}$ (denoted as $\bar{l}(d_h)$), where $y_p = d_h$.

As $trace(\hat{K}\hat{L}) = \sum_{i=1}^{p} \sum_{j=1}^{p} < \hat{K}^{ij}, (\hat{L}^{ji})^T >$ (<.,.> denoted as inner product operator) and $< \hat{K}^{ij}, (\hat{L}^{ji})^T >= \hat{K}(i,j) \times \sum_{u=1}^{q} yx(u,j) \times \sum_{v=1}^{q} (yx(v,i) \times (l(d_u, d_v) - \bar{l}(d_u))$, we can obtain that the computational complexity of $trace(\hat{K}\hat{L})$ is $O(p^2 q^2)$.

As described above, we can know that the total computational complexity of xt, yt and yx is $O(m)$, and that those of $\overline{K}$ and $\overline{L}$ are $O(p^2)$ and $O(q^2)$, respectively.

Hence, we have that the total computational complexity of HSIC is $O(m)$, where $p^2 \times q^2 < m$, or $p^2 \times q^2 \approx m$.

The proof is complete. □

In fact, the proof above gives the simplified process of efficient computation to HSIC. The detailed processes are shown in Algorithms 1–5. Algorithm 1 is the main process of the method, consisting of three functions:

1. (X, Y) is m observations of a tuple of x and y with p and q states, respectively;
2. $kernels$ includes two kernel functions used to calculate $K_{i,j}$ and $L_{i,j}$, the parameters of which are $delta(1)$ and $delta(2)$, respectively;
3. $GetInfo$ (see Algorithm 2) is used to calculate xt, yt and yx ;
4. For all $1 \leq j, l \leq p$, KH (see Algorithm 3) is used to calculate $\hat{K}(j, l)$;
5. $Trace$ (see Algorithm 4) is used to calculate $trace(\hat{K}\hat{L})$;
6. $RowAverage$ (see Algorithm 5) is used to calculate $\overline{K}$ and $\overline{L}$.

Algorithm 1 Calculate $value = \text{HSIC}(X, Y, p, q, m, kernels, deltas)$

```
Require: $|X| = |Y| = m$
$[xt, yt, yx] \Leftarrow GetInfo(X, Y, p, q, m)$
$\hat{K} \Leftarrow KH(xt, p, kernels(1), deltas(1), m)$
$value \Leftarrow Trace(\hat{K}, yt, yx, p, q, kernels(2), deltas(2), m)$
```

Algorithm 2 Calculate $[xt, yt, yx]$= $GetInfo(X, Y, p, q, m)$

```
$xt \Leftarrow zeros(1, p)$
$yt \Leftarrow zeros(1, q)$
$yx \Leftarrow zeros(q, p)$
$col \Leftarrow 1$
while $col \leq m$ do
    $stx \Leftarrow X(col)$
    $sty \Leftarrow Y(col)$
    $xt(stx) \Leftarrow x(stx) + 1$
    $yt(sty) \Leftarrow y(sty) + 1$
    $yx(sty, stx) \Leftarrow yx(sty, stx) + 1$
    $col \Leftarrow col + 1$
end while
```

Algorithm 3 Calculate $\hat{K} = KH(xt, p, kernel, delta, m)$

```
$\overline{K} \Leftarrow RowAverage(p, k, delta, m, xt)$
$\hat{K} \Leftarrow zeros(p, p)$
$i \Leftarrow 1$
while $i \leq p$ do
    $j \Leftarrow 1$
    while $j \leq p$ do
        $\hat{K}(i, j) \Leftarrow kernel(i, j, delta) - \overline{K}(i)$
        $j \Leftarrow j + 1$
    end while
    $i \Leftarrow i + 1$
end while
```

Algorithm 4 Calculate $value = Trace(\hat{K}, yt, yx, p, q, kernel, delta, m)$

```
$\overline{L} \Leftarrow RowAverage(q, kernel, delta, m, yt)$
$value \Leftarrow 0$
$i \Leftarrow 1$
while $i \leq p$ do
    $j \Leftarrow 1$
    while $j \leq p$ do
        $u \Leftarrow 1$
        $t \Leftarrow 0$
        while $u \leq q$ do
            $v \Leftarrow 1$
            $s \Leftarrow 0$
            while $v \leq q$ do
                $s \Leftarrow s + yx(v, i) \times (kernel(u, v, delta) - \overline{L}(u))$
                $v \Leftarrow v + 1$
            end while
            $t \Leftarrow t + yx(u, j) \times s$
            $u \Leftarrow u + 1$
        end while
        $value \leftarrow value + \hat{K}(i, j) \times t$
        $j \Leftarrow j + 1$
    end while
    $i \Leftarrow i + 1$
end while
$value \Leftarrow value/((m-1)*(m-1))$
```

Algorithm 5 Calculate $\overline{M} = RowAverage(p, kernel, delta, m, y)$

```
$\overline{M} \Leftarrow zeros(1, p)$
$i \Leftarrow 1$
while $i \leq p$ do
    $j \Leftarrow 1$
    while $j \leq p$ do
        $\overline{M}(i) \Leftarrow \overline{M}(i) + y(j) * kerenl(i, j, delta)$
        $j \Leftarrow j + 1$
    end while
    $\overline{M}(i) \Leftarrow \overline{M}(i)/m$
    $i \Leftarrow i + 1$
end while
```

4. Experiments

We employed representation of the data in a matrix $X_{i_j}^{pq} \in \{0,1,2\}^{m_{i_j}^{pq} \times 1}$ to calculate HSIC$(X_{i_1}^{pq}, X_{i_2}^{pq})$ by using a Gaussian kernel ($\sigma^2 = 0.1$). In addition, we mapped $X_{i_1 i_2 \ldots, i_k} \in \{0,1,2\}^{m \times k}$ onto $X_{i_1 i_2 \ldots, i_k} \in \{0,1\}^{m \times 3k}$ to compute HSIC$(X_{i_1 i_2 \ldots, i_k}, Y_{i_1 i_2 \ldots, i_k})$ by also using a Gaussian kernel ($\sigma^2 = 1$), i.e., $0 \mapsto (1,0,0)$, $1 \mapsto (0,1,0)$ and $2 \mapsto (0,0,1)$. The advantages and disadvantages of the two representations have been explained by the other authors [29].

4.1. Evaluation Criterion

The evaluation criterion that we adopted in the experiments is by [9]:

$$Power = \frac{\sharp S}{\sharp T}, \tag{10}$$

where $\sharp S$ is the number of found disease-causing SNP combinations (the epistasis SNPs score the highest) and $\sharp T$ is the number of datasets. Each dataset includes one disease-causing SNP combination. *Power* is a measure of the accuracies of scoring criteria from genome data.

4.2. Simulated Datasets

For any data set, the worst-case scenario for checking the correctness of the scoring criteria is extensive testing of all SNP combinations. It is too computationally expensive for $k = 4$ and $k = 5$ cases. Therefore, tests were only conducted for $k = 2$ and $k = 3$.

4.2.1. Disease Models with $k = 2$

For $k = 2$, we used thirty-five disease models without marginal effects (DNME1–35) and six disease models with marginal effects (DME1–6). The models were designed based on interaction structures with different diseases, MAFs, prevalence (p) and h (the parameter settings are described in the supplementary files). Each data set contains 1000 SNPs and includes pairs of interacting SNPs (M0P0 and M1P1) generated according to the disease model setting, while other SNPs are generated using MAFs uniformly selected in [0.05, 0.5). For each model, we generated two simulated 100 data sets using the software GAMETES2.1 [30] with sample sizes of 400 (200 controls and 200 cases) and with sample sizes of 800 (400 control and 400 cases) [31].

Disease Models without Marginal Effects

We divided all DNMEs into seven subgroups for analysis according to the different combined values of h and MAF (DNME1–5 MAF = 0.2, $h = 0.2$; DNME6–10 MAF = 0.4, $h = 0.2$; DNME11–15 MAF = 0.2, $h = 0.1$; DNME16–20 MAF = 0.4, $h = 0.1$; DNME21–25 MAF = 0.2, $h = 0.05$; DNME25–30 MAF = 0.4, $h = 0.05$; DNME1–5 MAF = 0.2, $h = 0.025$).

The analysis results of subgroups of DNME1–35, each of which has 400 samples, are shown in Figure 3:

1. Except for Mi, using tests on DNME1–10, the accuracy of all scoring criteria is close to 100%;
2. All criteria are not very accurate using tests on DNME21–25 and DNME31–35;
3. Mi has an extremely poor accuracy on all subgroup tests;
4. LR has the highest accuracy using tests on DNME11–15 and DNME16–20, close to 100%, but is only a little more accurate than Mi on DNME21–25, DNME26–30 and DNME31–35 tests;
5. The accuracy rates of both ND-JE and G-test rank in the middle overall, but G-test has the highest accuracy on the DNME26–30 test;
6. The accuracy rate of K2-Score ranks second on DNME11–15 and DNME21–25 tests, third on the DNME26–30 test and slightly worse than ND-JE, HS_{CR} and G-test on the DNME16–20 test, but is only a little more accurate than Mi on the most difficult model (DNME31–35) test;
7. HSIC_{CR} has the highest accuracy on the two most difficult model subgroup (DNME21–25 and DNME31–35) tests, especially on DNME31–35, where the accuracy is much higher than other criteria, and the overall accuracy on other model subgroups tests is similar to the other four criteria.

When the size of samples increased from 400 to 800, the accuracy of all criteria was greatly improved. The analysis results of subgroups of DNME1–35, each of which has 800 samples, are shown in Figure 4:

1. Except for Mi, the accuracy of all criteria is close to 100% excluding tests on the two most difficult model subgroups (DNME21–25 and DNME31–35);
2. Although the accuracy rate of Mi can be significantly improved with the increase in the size of samples, it is still relatively poor overall;
3. With the number of samples increasing, there is still no change in the overall ranking, but the accuracy of the K2-Score on the DNME31–35 test rises to second;
4. HSIC_{CR} has the highest accuracy on the two most difficult model subgroup tests.

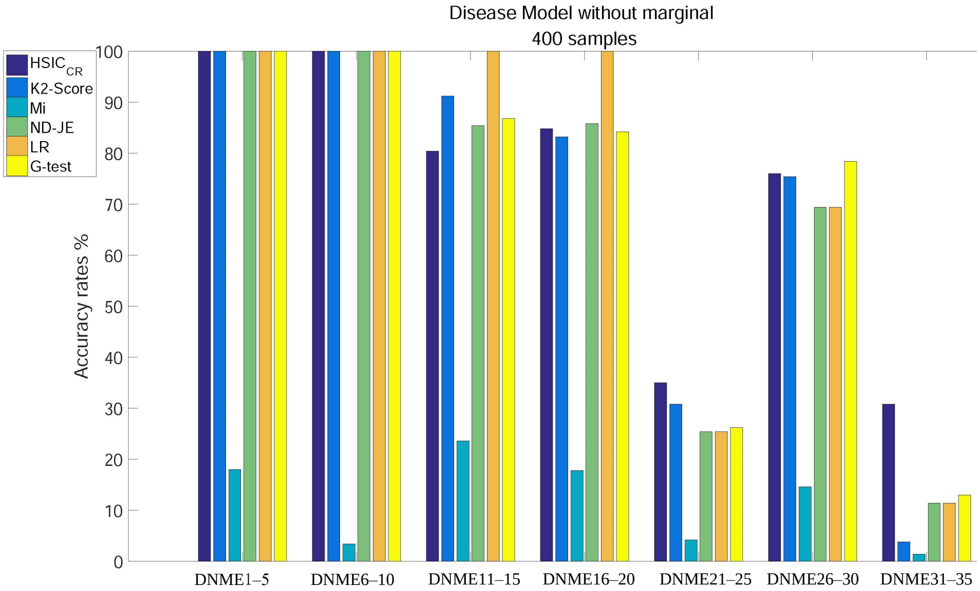


Figure 3. Disease model without marginal effects, with 400 samples and with $k = 2$.

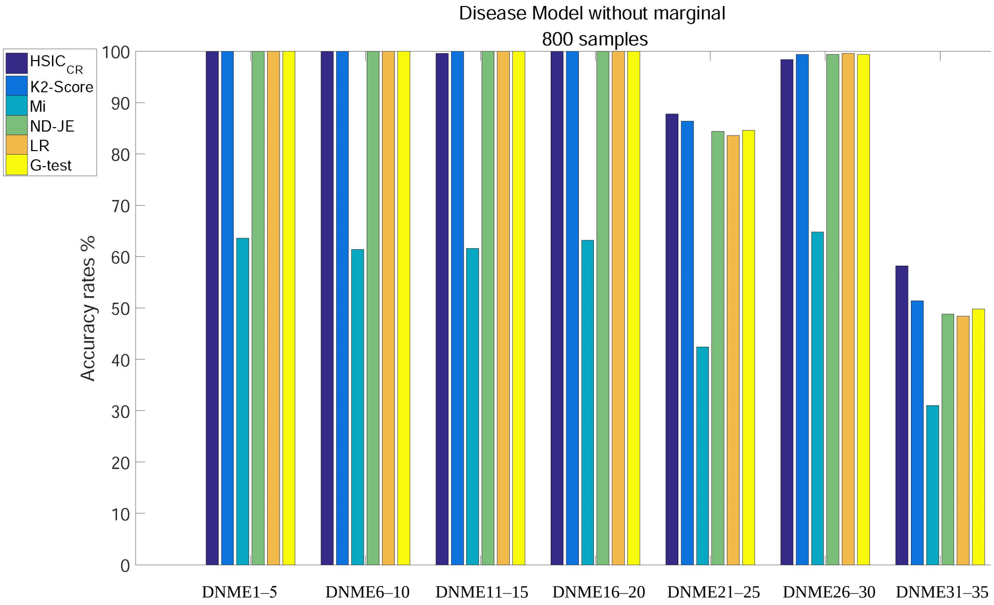


Figure 4. Disease model without marginal effects, with 800 samples and with $k = 2$.

Table 1 reveals the total average accuracy. From Table 1, we can find that Mi has a poor average accuracy; HSIC$_{CR}$ has the best average accuracy regardless of the model's sample scale of 400 or 800; although HSIC$_{CR}$ is only slightly higher than the other four criteria on the total average accuracy, and the average accuracy on the most difficult model subgroup test is much better than other criteria.

Table 1. The number of times, out of 3500 data sets generated by 35 models without marginal effects, where $k = 2$, that each scoring criterion identified epistasis SNPs of snp1000 for sample sizes of 400 and 800. The fourth column gives the total accuracy over all sample sizes. The last column gives the accuracy over all sample sizes in the most difficult subgroup models. The scoring criteria are listed in descending order of total accuracy.

Scoring Criterion	400 Samples	800 Samples	Total (%)	DNME31–35 (%)
HSIC$_{CR}$	2535	3220	5755 (82.2%)	445 (44.5%)
K2-Score	2479	3186	5665 (80.9%)	336 (33.6%)
G-test	2443	3169	5612 (80.2%)	314 (31.4%)
ND-JE	2440	3163	5603 (80.0%)	301 (30.1%)
LR	2437	3158	5595 (79.9%)	297 (29.7%)
Mi	494	1971	2465 (35.2%)	192 (19.2%)

Disease Models with Marginal Effects

We tested six DMEs for analysis according to MAF = 0.1 and the different combined values of heritability and prevalence (DME1 h = 0.031 and p = 0.050; DME2 h = 0.014 and p = 0.050; DME3 h = 0.01 and p = 0.050; DME4 h = 0.016 and p = 0.046; DME5 h = 0.009 and p = 0.026; DME6 h = 0.008 and p = 0.017).

The analysis results of DME1–6, each of which has 400 samples, are shown in Figure 5:

1. The accuracy of all scoring criteria is close to 100% tested on DME1, except for Mi;
2. Mi has extremely poor accuracy on all six models tests;
3. Except for Mi, the accuracy rate of LR is worse than the other four criteria on DME2–6 tests, except that the accuracy on the DME3 test is nearly the same as that of HSIC$_{CR}$;
4. The accuracy rate of ND-JE ranks third on DME1 and DME3 tests, and fourth on the other four models tests;
5. The accuracy rate of G-test ranks first on the DME1 test, second on DME2 and DME3 tests and third on the other four models tests;
6. HSIC$_{CR}$ has the highest accuracy rate on DME4 and DME6 tests, its accuracy rate on the DME5 test is slightly worse than LR and the accuracy rate on DME1–2 tests ranks third, whereas the accuracy rate on the DME3 test is a little better than Mi;
7. K2-Score has the highest accuracy rate on DME1–3 and DME5 tests, its accuracy rate on DME4 and DME6 tests ranks second and it significantly outperforms the others on the most difficult model (DME3) test (although its accuracy rate in DME3 is below 50%).

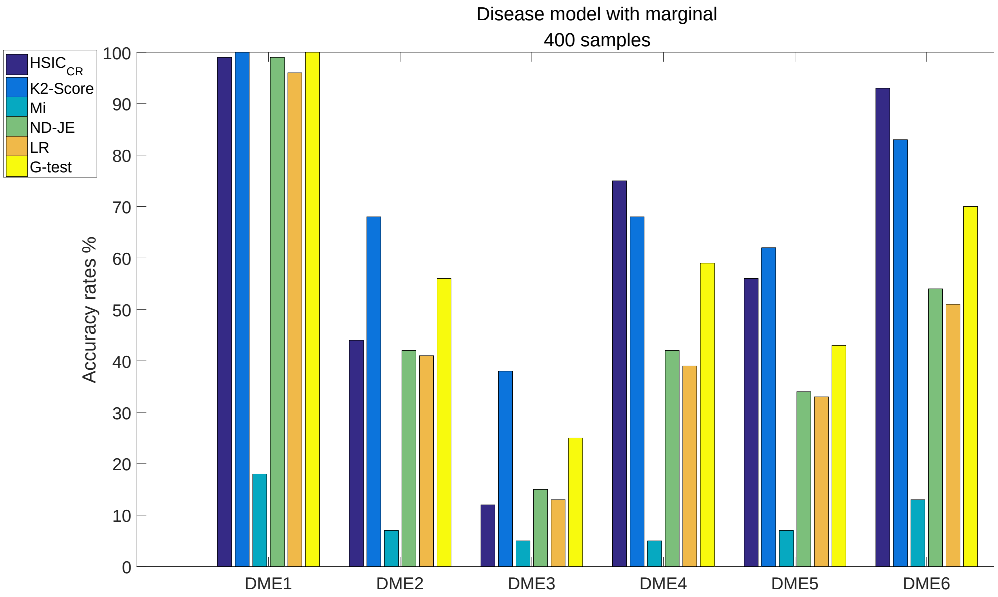


Figure 5. Disease model with marginal effects, with 400 samples and $k = 2$.

When the size of samples increased from 400 to 800, the accuracy of all criteria was greatly improved. The analysis results of DME1–6, each of which has 800 samples, are shown in Figure 6:

1. Although the accuracy of Mi can be significantly improved with the increase in the size of samples, it is still relatively poor overall;
2. The accuracy rates of the other five scoring criteria all exceed 95% tested by the models, except on DME3;
3. K2-Score has the highest accuracy rate on the most difficult model test (over 90%), the accuracy rate of G-test ranks second (over 80 %) and the accuracy rates of ND-JE, HSIC_{CR} and LR are not good enough, at just over 70%.

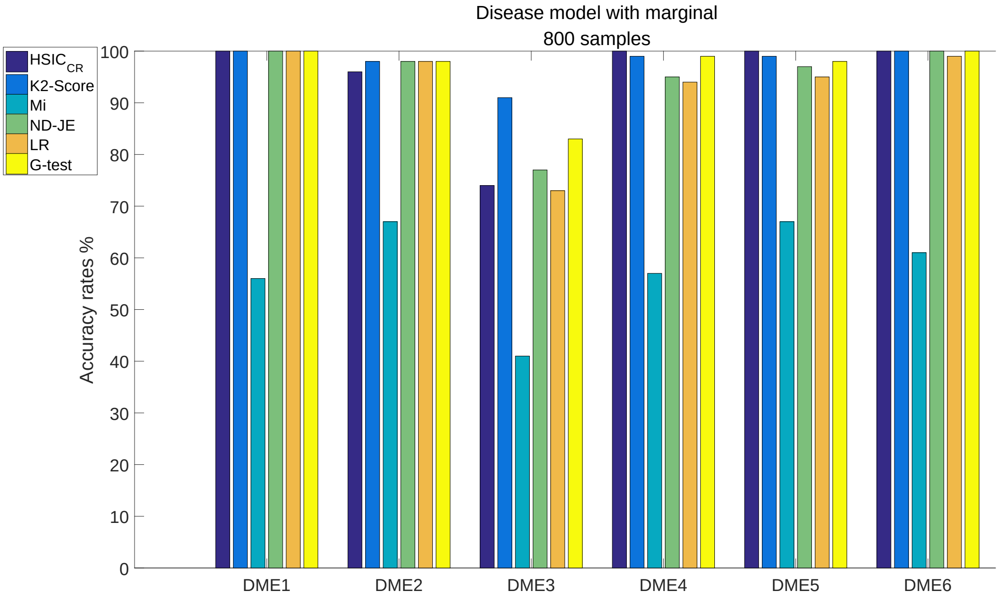


Figure 6. Disease model with marginal effects, with 800 samples and $k = 2$.

Table 2 reveals the total average accuracy. From Table 2, we can find that Mi has a poor average accuracy; the K2-Score has the best average accuracy rate regardless of the model's sample scale of 400 or 800, where the main reason is that its accuracy rate on the DME3 test is much better than the other five scoring criteria; although the average accuracy of HSIC_{CR} tested on DME3 is not good enough, it ranks second in the overall average accuracy rate.

Table 2. The number of times, out of 600 data sets generated by six models with marginal effects, where $k = 2$, that each scoring criterion identified epistasis SNPs of snp1000 for sample sizes of 400 and 800. The fourth column gives the total accuracy over all sample sizes. The last column gives the accuracy over all sample sizes in the most difficult model. The scoring criteria are listed in descending order of total accuracy.

Scoring Criterion	400 Samples	800 Samples	Total (%)	DME3 (%)
K2-Score	419	586	1005 (83.8%)	159 (79.5%)
HSIC_{CR}	379	570	949 (79.1%)	86 (43%)
G-test	353	578	931 (77.6%)	108 (54%)
ND-JE	286	567	853 (71.1%)	92 (46%)
LR	273	559	832 (69.3%)	86 (43%)
Mi	55	349	404 (33.7%)	48 (24%)

4.2.2. Disease Models with $k = 3$

For $k = 3$, the data sets are generated by eight third-order epistasis pathogenic models (DM1–8), which are modeled by GAMETES2.1 according to the combinations of different MAFs ([0.2, 0.4]) and different heritability ([0.025, 0.05, 0.1, 0.2]) (DM1 MAF = 0.2, h = 0.025;

DM2 MAF = 0.2, h = 0.05; DM3 MAF = 0.2, h = 0.1; DM4 MAF = 0.2, h = 0.2; DM5 MAF = 0.4, h = 0.025; DM6 MAF = 0.4, h = 0.05; DM7 MAF = 0.4, h = 0.1; DM8 MAF = 0.4 h = 0.2). The ♯quantiles of each combination is five. Every quantile of each pathogenic model corresponds to 100 simulated data files. Each file contains 100 SNPs and 1600 samples (800 normal, 800 diseased), and includes three interacting SNPs (M0P0, M1P1 and M2P2) generated according to the disease model settings, while other SNPs were generated using MAFs uniformly selected in [0.05, 0.5]. Therefore, the total number of the data sets is 4000 [6]. Detailed parameter settings are described in the supplementary file.

The analysis results of DM1–8, each of which has 1600 samples, are shown in Figure 7:

1. The accuracy of all scoring criteria is close to 100% tested on DM2, DM3–4 and DM7–8;
2. The accuracy of all scoring criteria is close to 80% on the DM1 test;
3. The K2-Score has a poor accuracy rate on the most difficult model (DM5) test, whose accuracy is just close to 10%;
4. The accuracy rates of the scoring criteria on the DM6 test are good enough, and the accuracy rates of the scoring criteria are close to 100%, except the K2-Score;
5. HSIC$_{CR}$ has the highest accuracy rate on the DM5 test, and its accuracy rate is the only one that exceeds 60% among all scoring criteria.

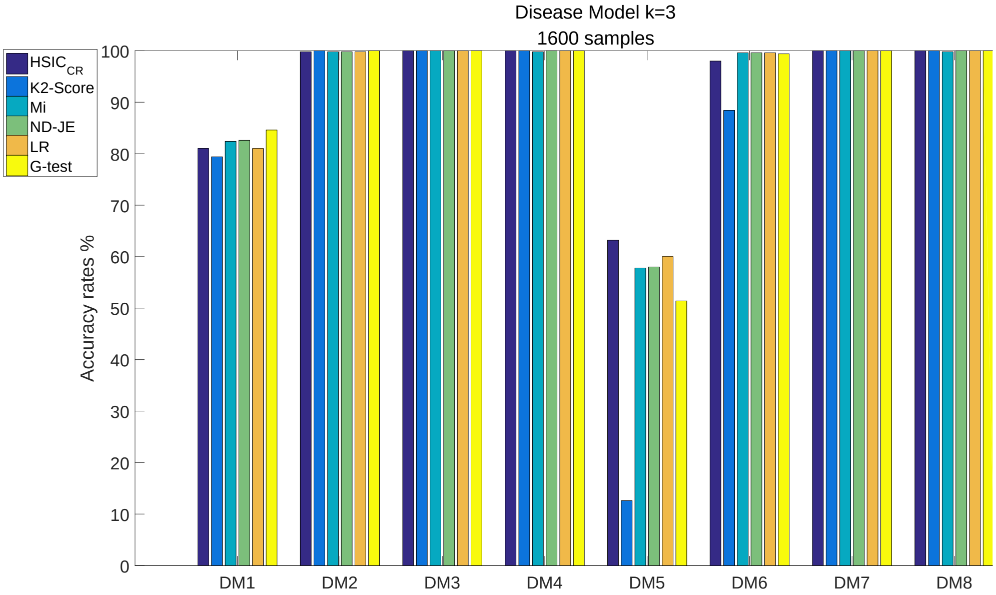


Figure 7. Disease model with 1600 samples and $k = 3$.

Table 3 reveals the total average accuracy. From Table 3, we can find that the K2-Score has a poor average accuracy on the most difficult model test; the total average accuracy rates for all scoring criteria are good enough, with all criteria except the K2-Score achieving over 90% accuracy. Although HSIC has the best total average accuracy, its accuracy is not significantly better than other four criteria; however, tested on the most difficult model, the accuracy rate of HSIC$_{CR}$ outperforms LR by 3.2%, and significantly outperforms the other four criteria, especially the K2-Score.

Table 3. The number of times, out of 4000 data sets generated by eight models, where $k = 3$, that each scoring criterion identified epistasis SNPs of snp100 for 1600 samples. The second column gives the total accuracy over a sample size of 1600. The last column gives the accuracy over a sample size of 1600 in the most difficult model. The scoring criteria are listed in descending order of total accuracy.

Scoring Criterion	Total (%)	DM5 (%)
HSIC$_{CR}$	3710 (92.8%)	316 (63.2%)
LR	3702 (92.6%)	300 (60%)
ND-JE	3700 (92.5%)	290 (58%)
Mi	3696 (92.4%)	289 (57.8%)
G-test	3677 (91.9%)	257 (51.4%)
K2-Score	3402 (85.1%)	63 (12.6%)

4.2.3. The Running Time Analysis

To demonstrate that our proposed method can be used as a lightweight scoring criterion, we proved in the previous section that its time complexity is $O(m)$. Furthermore, we calculated the average running time per dataset (unit in seconds) for the two-order with a sample size of 800 and the three-order in the simulation experiments, and we found that the average running time per dataset of our proposed method is between the other five lightweight methods (see Table 4). This further demonstrates the applicability of our proposed method as a lightweight scoring criterion.

In the experiment, all scoring criteria were implemented based on Matlab, and all tests were run on the environment of Windows 10 64 desktop computer with 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80 GHz, and 16.0 GB memory.

Table 4. Average running time (s) for the six scoring criteria per dataset for both the two-order tests and the three-order tests in the simulation experiments.

Scoring Criterion	2-Order (s)	3-Order (s)
HSIC$_{CR}$	120.2426	80.3045
LR	77.3055	55.4438
ND-JE	134.922	76.5377
Mi	78.3	55.6138
G-test	77.6589	56.5221
K2-Score	125.0655	81.4237

4.3. Case Study: A Real Chronic Dialysis Data

A real data set of 193 cases and 704 controls was selected from the mitochondrial D-loop region of chronic dialysis patients who were observed in a study by other authors [32]. The genotypes and locations of 77 SNPs are presented in Table 5 [33].

The 77 SNPs contained in the subset of the chronic dialysis data set were used in the case study, which aims to give our readers more specificity regarding our proposed scoring criterion. First, for this dataset, we performed a full-space two-order SNP combination detection, meaning that the HSIC$_{CR}$ values were evaluated for 2926 (C_{77}^{2}) possible combinations. Then, we selected the top ten HISC$_{CR}$-valued combinations as candidate two-order epistasis SNP combinations to be raised for medical researchers, and the 10 candidate combinations are presented in Table 6.

Table 5. Positions of chronic dialysis-associated 77 SNPS in mitochondrial d-loop region. [a] Left and right letters are major and minor genotypes, respectively. The number is the SNP position in the mitochondrial D-loop region.

SNP	D-Loop Position				
1∼5	A16051G [a]	T16086C	T16092M	T16093C	C16108T
6∼10	C16111T	T16126C	G16129A	T16136C	T16140C
11∼15	G16145A	C16148T	T16157C	A16162G	A16164G
16∼20	C16167T	T16172C	T16209C	T16217C	C16218T
21∼25	T16223C	A16227G	C16234T	A16235G	T16243M
26∼30	C16248T	T16249C	C16256T	C16257W	C16260T
31∼35	C16261T	C16266D	A16272G	G16274A	C16278T
36∼40	C16290T	C16291T	C16295T	C16297T	C16298T
41∼45	C16304T	A16309G	T16311C	A16316G	G16319A
46∼50	T16324C	C16327T	A16335G	C16355T	T16356C
51∼55	T16357C	T16362C	G16390A	A16399G	A16463G
56∼60	C16519T	A93G	G103A	T146M	C150T
61∼65	C151T	T152C	A153G	G185A	A189G
66∼70	C194T	T195C	T199C	A200G	T204C
71∼75	G207A	A210G	T217C	A234G	A235G
76∼77	T317C	C461T			

Table 6. The top ten highest HISC$_{CR}$-valued two-order SNPs were used as ten candidate combinations.

Rank	Combination	HSIC$_{CR}$
1	41, 21	0.035922
2	52, 21	0.033105
3	41, 17	0.019069
4	56, 21	0.018961
5	68, 39	0.018545
6	21, 19	0.017254
7	60, 21	0.014506
8	17, 8	0.011645
9	17, 14	0.0097405
10	75, 36	0.0095467

5. Conclusions

In this paper, we verified with rigorous mathematical proof that HSIC$_{CR}$ can be computed in $O(m)$ time. Moreover, we compared HSIC$_{CR}$ with five representative scoring criteria for 49 simulation disease models. The experimental results show that: Mi has a poor accuracy on two-order disease models; the K2-Score has a poor accuracy on three-order difficult disease models; the accuracy rates of LR are not good enough on two-order disease models tests; HSIC$_{CR}$, G-test and ND-JE have a high accuracy on all three classes of disease models tests; the accuracy rates of HSIC$_{CR}$ rank first on two-order disease models without marginal effects tests and three-order disease model tests, and rank second on two-order disease models with marginal effects tests, although its advantage is not significant.

The advantages of HSIC$_{CR}$ are: the methodology used is different from other scoring criteria, which makes it more complementary to other scoring criteria; it has a high accuracy on most disease models.

In the future, we will further investigate proposing efficient SIO algorithms to solve this problem by combining HSIC$_{CR}$ and other effective lightweight criteria that already exist as weighted single or multi-objective functions. In addition, we will work with several local medical research institutions to use their real disease case-control study data to mine for disease-related SNP combinations by using our proposed approach. This will ultimately provide new guidance for drug development in complex diseases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10214134/s1, Table S1: Model with marginal effects when $k = 2$; Table S2: Models 1 to 10 without marginal effects when $k = 2$; Table S3: Models 11 to 20 without marginal effects when $k = 2$; Table S4: Models 21 to 30 without marginal effects when $k = 2$; Table S5: Models 31 to 35 without marginal effects when $k = 2$.

Author Contributions: Conceptualization, J.Z. (Junxi Zheng); data curation, J.Z. (Junxi Zheng) and J.Z. (Jiaxian Zhu); formal analysis, J.Z. (Juan Zeng), J.Z. (Jiaxian Zhu) and F.W.; funding acquisition, J.Z. (Junxi Zheng); investigation, G.L. and F.W.; methodology, J.Z. (Junxi Zheng); project administration, D.T. and X.W.; supervision, D.T. and X.W.; validation, J.Z. (Junxi Zheng) and J.Z. (Juan Zeng); visualization, G.L.; writing—original draft, J.Z. (Junxi Zheng) and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangdong provincial medical research foundation of China (No. A2022531), the national natural science foundation of China (No. 61976239), and the natural science foundation of Guangdong province, China (No. 2020A1515010783).

Data Availability Statement: The data that support the findings of this study can be acquired from the corresponding author.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

Abbreviations

The following abbreviations are used in this manuscript:

SNP	single-nucleotide polymorphism
GWAS	genome-wide association analysis

References

1. Carlson, C.S.; Eberle, M.A.; Kruglyak, L.; Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* **2004**, *429*, 446–452. [CrossRef] [PubMed]
2. Wei, W.H.; Hemani, G.; Haley, C.S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **2014**, *15*, 722–733. [CrossRef]
3. Guo, X.; Meng, Y.; Yu, N.; Pan, Y. Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. *BMC Bioinform.* **2014**, *15*, 102. [CrossRef]
4. Guo, X.; Zhang, J.; Cai, Z.; Du, D.Z.; Pan, Y. Searching genome-wide multi-locus associations for multiple diseases based on bayesian inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *14*, 600–610. [CrossRef]
5. Gyenesei, A.; Moody, J.; Semple, C.A.; Haley, C.S.; Wei, W.H. High-throughput analysis of epistasis in genome-wide association studies with BiForce. *Bioinformatics* **2012**, *28*, 1957–1964. [CrossRef]
6. Liyan, S. The Research on Epistasis Detection Algorithm in Genome-wide Association Study. Ph.D. Thesis, Jilin University, Changchun, China, 2020.
7. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147. [CrossRef] [PubMed]
8. Wang, X.; Cao, X.; Feng, Y.; Guo, M.; Yu, G.; Wang, J. ELSSI: Parallel SNP–SNP interactions detection by ensemble multi-type detectors. *Brief. Bioinform.* **2022**, *23*, bbac213. [CrossRef] [PubMed]
9. Tuo, S.; Liu, H.; Chen, H. Multipopulation harmony search algorithm for the detection of high-order SNP interactions. *Bioinformatics* **2020**, *36*, 4389–4398. [CrossRef]
10. Sun, Y.; Shang, J.; Liu, J.X.; Li, S.; Zheng, C.H. epiACO—A method for identifying epistasis based on ant Colony optimization algorithm. *BioData Min.* **2017**, *10*, 23. [CrossRef]
11. Tuo, S.; Zhang, J.; Yuan, X.; He, Z.; Liu, Y.; Liu, Z. Niche harmony search algorithm for detecting complex disease associated high-order SNP combinations. *Sci. Rep.* **2017**, *7*, 11529. [CrossRef]
12. Aflakparast, M.; Salimi, H.; Gerami, A.; Dubé, M.; Visweswaran, S.; Masoudi-Nejad, A. Cuckoo search epistasis: A new method for exploring significant genetic interactions. *Heredity* **2014**, *112*, 666–674. [CrossRef] [PubMed]
13. Jing, P.J.; Shen, H.B. MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics* **2015**, *31*, 634–641. [CrossRef] [PubMed]
14. Cheng, R.; Jin, Y.; Olhofer, M.; Sendhoff, B. A reference vector guided evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **2016**, *20*, 773–791. [CrossRef]
15. Shouheng, T.; Hong, H. DEaf-MOPS/D: An improved differential evolution algorithm for solving complex multi-objective portfolio selection problems based on decomposition. *Econ. Comput. Econ. Cybernet. Stud. Res.* **2019**, *53*, 151–167.

16. Verzilli, C.J.; Stallard, N.; Whittaker, J.C. Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.* **2006**, *79*, 100–112. [CrossRef]
17. Cooper, G.F.; Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **1992**, *9*, 309–347. [CrossRef]
18. Jiang, X.; Neapolitan, R.E.; Barmada, M.M.; Visweswaran, S. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinform.* **2011**, *12*, 89. [CrossRef]
19. Zhang, Y.; Liu, J.S. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **2007**, *39*, 1167–1173. [CrossRef]
20. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
21. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
22. Bush, W.S.; Edwards, T.L.; Dudek, S.M.; McKinney, B.A.; Ritchie, M.D. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinform.* **2008**, *9*, 238. [CrossRef] [PubMed]
23. Neyman, J.; Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **1928**, *20A*, 175–240.
24. Stamatis, D.H. *Essential Statistical Concepts for the Quality Professional*; CRC Press: Boca Raton, FL, USA, 2012.
25. Pearl, J. *Models, Reasoning and Inference*; Cambridge University Press: Cambridge, UK, 2000; Volume 19.
26. Schaid, D.J. Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Hum. Hered.* **2010**, *70*, 109–131. [CrossRef]
27. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings of the International Conference on Algorithmic Learning Theory, Singapore, 8–11 October 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 63–77.
28. Gretton, A.; Fukumizu, K.; Teo, C.; Song, L.; Schölkopf, B.; Smola, A. A kernel statistical test of independence. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 585–592.
29. Kodama, K.; Saigo, H. KDSNP: A kernel-based approach to detecting high-order SNP interactions. *J. Bioinform. Comput. Biol.* **2016**, *14*, 1644003. [CrossRef]
30. Urbanowicz, R.J.; Kiralis, J.; Sinnott-Armstrong, N.A.; Heberling, T.; Fisher, J.M.; Moore, J.H. GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min.* **2012**, *5*, 16. [CrossRef] [PubMed]
31. Yang, C.H.; Chuang, L.Y.; Lin, Y.D. Multiobjective multifactor dimensionality reduction to detect SNP–SNP interactions. *Bioinformatics* **2018**, *34*, 2228–2236. [CrossRef]
32. Chen, J.B.; Yang, Y.H.; Lee, W.C.; Liou, C.W.; Lin, T.K.; Chung, Y.H.; Chuang, L.Y.; Yang, C.H.; Chang, H.W. Sequence-based polymorphisms in the mitochondrial D-loop and potential SNP predictors for chronic dialysis. *PLoS ONE* **2012**, *7*, e41125. [CrossRef]
33. Yang, C.H.; Kao, Y.K.; Chuang, L.Y.; Lin, Y.D. Catfish Taguchi-based binary differential evolution algorithm for analyzing single nucleotide polymorphism interactions in chronic dialysis. *IEEE Trans. Nanobiosci.* **2018**, *17*, 291–299. [CrossRef]