University of Reading

The optimal parameters of spline regression for SNP-set analysis in genome-wide association study

Article

Published Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Open Access

Sookkhee, S., Kirdwichai, P. and Baksh, F. ORCID: https://orcid.org/0000-0003-3107-8815 (2021) The optimal parameters of spline regression for SNP-set analysis in genome-wide association study. Science & Technology Asia, 26 (1). pp. 39-52. ISSN 2586-9027 doi: 10.14456/scitechasia.2021.5 Available at https://centaur.reading.ac.uk/97535/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

Identification Number/DOI: 10.14456/scitechasia.2021.5
<https://doi.org/10.14456/scitechasia.2021.5>

Publisher: Thammasat University

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia
Science & Technology Asia
Vol. 26 No.1 January - March 2021
Page: [39-52]
Original research article

The Optimal Parameters of Spline Regression for SNP-Set Analysis in Genome-Wide Association Study

Sirikanlaya Sookkhee[1], Pianpool Kirdwichai[1,*], Fazil Baksh[2]

[1]*Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology of North Bangkok, Bangkok 10800, Thailand*
[2]*Department of Mathematics and Statistics, School of Mathematics, University of Reading RG6 6AH, UK*

Received 27 March 2020; Received in revised form 9 November 2020
Accepted 16 November 2020; Available online 16 March 2021

ABSTRACT

This research aims to develop a method that is capable and reliable for identifying significant regions in Genome-Wide Association Study based on Spline regression. We evaluate the optimal parameters in the Splines by smoothing and tuning p-values obtained from two methods, Sequence Kernel Association Test using normal weight (SKAT normal weight) and Generalized Higher Criticism (GHC) for testing SNP-set. False positive (FP) and True positive (TP) rates were evaluated under different genetic models for disease with significant thresholds adjusted for multiple hypothesis testing based on the permutation method. The simulated data used in this research are constructed from a control data set in a study of Crohn's disease which is repeated 1,500 replicates for studies of size 3,000 cases and 3,000 controls. The simulation result shows that the optimal parameter in the Splines on the p-value of SKAT normal weight and GHC under the one disease SNP model simulation are at the degree of freedom 1,000. GHC is shown to be preferable in terms of comparing FP and TP rates but it is disadvantageous compared to SKAT in terms of computational burden time. Finally, the optimal parameter of both methods was applied to real data on Crohn's disease. Both methods found the important regions of genes NOD2 which are strongly associated with the development and the importance of gene NOD2 which causes Crohn's disease.

Keywords: Sequence kernel association test; Generalized higher criticism; Permutation test; Spline regression; B-spline; GWAS

***Corresponding author:** pianpool.k@sci.kmutnb.ac.th **doi:** 10.14456/scitechasia.2021.5

1. Introduction

A Genome-wide association study (GWAS) is the study of the association of different individuals in the genetic variant with a phenotype. GWAS typically focuses on the association between the Single Nucleotide Polymorphism (SNPs) with a phenotype which is the data that plays a very important role in identifying the location of DNA which causes the disease [1]. SNPs do not directly cause the disease, but they can indicate the risk of disease. Many diseases are still unable to determine the exact gene locations that cause the disease. There are also hundreds of thousands of SNPs that need to analyzed, so the computational burden is another issue. Therefore, identifying significant regions of disease-gene association in high dimensional genomewide studies is developed and evaluated. A computationally efficient method for obtaining the optimal tuning parameter is also evaluated using simulation.

The simplest and most commonly used method for analyzing SNPs and the disease is a single SNP analysis that can identify SNPs that cause the disease by analyzing only one location at a time. It is the traditional testing method that has been considered as the approach for genetic association analysis [2-3]. However, the results of the single-SNP analysis are not sufficiently informative to interpret without explicit references to linkage disequilibrium (LD) patterns of candidate variants [4]. Many studies have discovered that SNPs that cause the disease can be located at the same chromosome [5-7]. The Sequence Kernel Association Test (SKAT) was proposed to analyze the association between SNP-set with disease outcome under a logistic kernel machine model. This method aggregates individual SNP score statistics in the SNP set and efficiently compute SNP-set level p-values [8]. SKAT is a supervised and flexible computationally efficient regression method to test for the association between common or rare variants and disease [9-12]. SKAT performance depends on weight configuration. Therefore, varying weights of SKAT testing models are used in this research to find the appropriate weight. They are 1) default weight 2) Madsen and Browning weight 3) inverse means weight, and 4) normal weight. We then compare and select the best method to study in the next step.

Generalized Higher Criticism (GHC) has been recently proposed for testing multiple SNPs in genome-wide association studies. The technique uses a correlation matrix to construct a new test statistic. The GHC method is flexible to the correlation structure and is computationally efficient, providing a p-value without the need for the simulation of the null distribution [13]. Finally, the efficiency SKAT and GHC were selected to find the region where the SNP-set together affects the disease by using Spline Regression [14].

In this research, a novel method that is capable of reducing the FP rate using spline regression to identify the significant regions in the genome-wide association study is developed and evaluated. The adjusting p-value using b-spline with the cubic function that gives optimal smoothing and tuning parameters is considered.

Another important aspect of GWAS is the testing of many hypotheses simultaneously, resulting in high false positives and incorrectly ascribing scientific significance to a statistical test [15]. Bonferroni adjustment is a popular method for controlling the probability of the type I error. However this approach tends to be highly constringent and conservative [16-17]. Therefore, the permutation method was selected as the alternative way of controlling the type I error rate which is based on the nonparametric method. It is a good choice for the hypothesis test of unknown distribution. This approach will be estimating the sampling distribution of a test statistic under the null hypothesis that a set of genetic variants does not affect the outcome. In this research, 10,000 replicates were used to compute the multivariate sampling distribution under the null hypothesis with no gene effect. This

approach provides a highly reliable distribution and gives the exact p-value of the test statistics [18-19]. The genotype data [20] was used in this simulation study from 1,504 individuals in the 1958 British Birth Cohort on Chromosome 16 which is being associated with Crohn's disease. But It is not clear that which SNPs or genes are responsible for this disease.

Therefore, the objective of this research is to find the optimal parameters of b-spline to identify the significant SNP-set from the SKAT and GHC methods. True positive (TP) and false positive (FP) rates use the permutation threshold as considered to find the efficiency method. Finally, the optimal parameters of both methods were selected to apply to the real data from the WTCCC study on Crohn's disease [20].

2. Methods

This section will present the Spline Regression which was selected to identify the gene regions. The permutation method was selected for finding the appropriate thresholds of the theoretical SKAT and GHC methods.

2.1 Spline regression

A spline [21] is a piece-wise polynomial with the piece defined by a sequence of knots in the range of X

$$\xi_1 < \xi_2 < \cdots < \xi_k,$$

such that the pieces join smoothly at the knots. For a spline of degree m, one usually requires the polynomial and their first m-1 derivatives to agree at the knots, so that m-1 derivatives are continuous. A spline of degree m can be represented as:

$$A(X) = \sum_{j=0}^{m} b_j X^j + \sum_{j=1}^{k} \lambda_j (X - \xi_j)^m, \qquad (1)$$

where the notation

$$(X - \xi_j) = \begin{cases} X - \xi_j & , X > \xi_j \\ 0 & , \text{otherwise} \end{cases}.$$

A(X) is the spline function of degree m. b_j is the associated spline coefficient with k knots and there are k+1 polynomial of degree m. X^j is the set of basis function and $\boldsymbol{\lambda}_j$ is the coefficients for the polynomial [22].

The most popular spline is the cubic spline

$$A(X) = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \sum_{j=1}^{k} \lambda_j (X - \xi_j)^3. \quad (2)$$

A much better representation of splines for computation is as linear combinations of a set of basis splines called Basis splines (B-spline) which are non-zero over a limited range of knots. A B-spline is a curve created from sub-curves in each range that can change the coefficient of the control point, which will affect the shape of the curve only near the control point. This makes the B-spline curves easy to shape and does not affect the overall curve. These sub-curves are created using the n polynomial. The value of n affects the smoothness of the curves. The basic properties of B-spline throughout this work**,** let $r_1, r_2, ..., r_k$ be the order of knots included in a real interval $[a, b]$. A spline of order $p \geq 0$ is a piecewise polynomial function of order p such that its derivatives up to order p-1 are continuous at every knot $r_1, r_2, ..., r_k$. The set of splines order p over the knots $r = r_1, r_2, ..., r_k$ is a vector space of dimension $p + k + 1$.

A spline basis is the truncated power basis: $\{x^0, x^1, ..., x^p, (x - r_1)^p, ..., (x - r_k)^p\}$. Reference [23] introduced B-spline as more adapted to computational implementation of spline regression.

B-spline is a spline with a non-zero over $[x_k, x_{p+k+1}]$ for some k. For

$i = 1, 2, ..., p + k + 1$, the i - th B-spline of order p is noted $N_{i,p}(x)$ and defined by

$$N_{i,p}(x) = \frac{x - r_i}{r_{i+p} - r_i} N_{i,p-1}(x) + \frac{r_{i+p+1} - x}{r_{i+p+1} - r_{i+1}} N_{i+1,p-1}(x). \quad (3)$$

In this research, a B-spline was selected fitting the p-value which uses a function in R for calculating the basis. Therefore, the main problem was finding the optimal parameter of bs function in R. The parameter of bs() function is bs(x, df = null, knots = null, degree = 3) where x is a predictor, df is degree of freedom, knots are the internal breakpoints to set the default which based on quantile of x and degree of 3 which is the cubic spline. We compute B-spline coefficients for regression quantile B-spline with fixed knots and specify the df. The optimal df can be obtained from tuning which will specify the df value in one hundred increments in the range of 100 - 900.

2.2 Permutation method

A permutation method (also called a randomization test, re-randomization test, or an exact test) is a nonparametric method for estimating the sampling distribution of a test statistic under the null hypothesis that a set of genetic variants has no effect on the outcome. This approach provides a highly reliable distribution of the test statistic but requires many samples generated under the null model. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels [18]. A permutation test is a good choice for hypothesis test of unknown distribution. It works regardless of the shape and size of the population to give the exact p-value [24]. In this research, we use 10,000 replicates for computing the multivariate sampling distribution under the null hypothesis with no gene effect and to establish significance thresholds giving a Type I error close to 0.05. We use linear interpolation for finding the thresholds

2.3 SNP-set methods

The Sequence Kernel Association Test or SKAT [17] is the test for the joint effects of multiple variants in a region of the genome on the disease which the regions were defined by genes. P-values were calculated for an association underlying the test procedure that can be viewed within the kernel machine regression framework [9, 25]. The semiparametric logistic regression model was used in the feature of SKAT including the parametric and nonparametric functions component effect the conditional probability of a dichotomous outcome [26]. The semiparametric logistic regression model for the i^{th} individual is

$$\text{logit}P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{C}_i + h(\mathbf{T}_i), \quad (4)$$

where y_i is a binary disease outcome taking values 0 (no disease) or 1 (disease) where $i = 1, 2, \ldots, n$, $\boldsymbol{\alpha}_0$ is an intercept term, $\boldsymbol{\alpha}'$ is the vector of regression coefficients for the covariates $\mathbf{C}_i$ and $\mathbf{T}_i$ are the observed variants and are related to disease through a nonparametric function $\mathbf{h}(\cdot)$, which is assumed to lie in a functional space generated by a positive semidefinite kernel function $K(\cdot,\cdot)$ [17]. $H_0 : h(\mathbf{T}) = 0$ is the null hypothesis of no association between the disease and gene region which were tested by assuming that the $n \times 1$ vector $\mathbf{h} = [h(\mathbf{T}_1), ..., h(\mathbf{T}_n)]'$ for the genetic effects of the n subjects follows a distribution with mean 0 and covariance $\tau\mathbf{K}$, where τ is a variance component. The semiparametric logistic regression model [10] is equivalent to

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{C}_i + \boldsymbol{\beta}'\mathbf{T}_i, \quad (5)$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients for the p observed variants in the gene region with each b_j following an arbitrary distribution with mean of 0 and a variance of $w_j\tau$, where w_j is a pre-specified weight for variant j and τ is the variance component. The null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to the hypothesis $H_0 : \tau = 0$, which may be tested with a variance-component score test statistic

$$S = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{Y} - \hat{\boldsymbol{\mu}}), \qquad (6)$$

where $\mathbf{K} = \mathbf{TWT}'$, $\hat{\boldsymbol{\mu}}$ is the predicted mean of $\mathbf{Y} = (y_1, \ldots, y_n)$ under H_0 that is $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\hat{\alpha}_0 + \mathbf{C}\hat{\boldsymbol{\alpha}})$, $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ are estimated under the null model by regressing $\mathbf{y}$ on the covariate $\mathbf{C}$ and $\mathbf{T} = [\mathbf{T}_1, \ldots, \mathbf{T}_n]$ is an $n \times p$ matrix with elements variant j of individual i, and $\mathbf{W} = \text{diag}(w_1, \ldots, w_p)$ contains the weights of the p variants. SKAT uses the variance-component score test statistic to test the null of no genetic effect but exploits the semiparametric regression approach in computing $\mathbf{K}$. The form for $\mathbf{K}$ used by SKAT is an $n \times n$ symmetric matrix with elements $\mathbf{K}(\mathbf{T}_i, \mathbf{T}_{i'})$ that measures genetic similarity between the $i-th$ and $i'-th$ subjects in the study. The weighted linear kernel $\mathbf{K}(\mathbf{T}_i, \mathbf{T}_{i'}) = \sum_{j=1}^{p} w_{ij} T_i T_{i'}$ was selected in this paper. Weight functions can be specified in the SKAT package in R which are based on the Beta density function $\text{Beta}(x_j : a, b)$

$$\sqrt{w_j} = \frac{x_j^{a-1}(1-x_j)^{b-1}}{B(a,b)}, 0 < x_j < 1;\ a, b > 0, \qquad (7)$$

where B denotes the beta function, a and b are prespecified scale and shape parameters and x_j is the estimated minor allele frequency (MAF) for SNP j using all cases and controls. Four weights were considered which are default, Madsen and Browning, inverse mean and normal weight.

The default weight chooses a small a and large b as $\text{Beta}(x_j : 1, 25)$ which substantially up-regulates rare variants and down-regulates common variants [10]. The Madsen and Browning weight was defined as $\text{Beta}(x_j : 0.5, 0.5)$, which corresponds to $\sqrt{w_j} = 1/\sqrt{\text{MAF}_j(1-\text{MAF}_j)}$; that is w_j is the inverse of the variance of the genotype marker j. The inverse mean is equivalent to $\text{Beta}(x_j : 0.5, 1)$the weight based on function $w_j = 1/\sqrt{x_j}$. The normal weight is $\text{Beta}(x_j : 10, 10)$ which gives the appearance of a symmetrical distribution similar to the normal distribution [17].

Generalized Higher Criticism or GHC [13,17] is the method for testing associated gene regions by using single variant statistics and their correlation matrix to construct a new test statistic and its distribution. Considering the parametrization of $P(y_i = 1)$ for the $j-th$ variant in a set of p variants,

$$\text{logit}\, P(y_i = 1) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}'\mathbf{C}_i + \boldsymbol{\beta}_j T_{i,j}, \qquad (8)$$

where b_j is the effect of the $j-th$ variant and $T_{i,j}$ is the observed $j-th$ variant with the $i-th$ subject, and the other terms are as in the previous section. The GHC approach exploits the fact that while p might be large in the test of the global null $H_0 : \boldsymbol{\beta} = \mathbf{0}$, in a genetic construct variants are likely to be correlated and generally only a small subset

of variants are signals for association. In other words, a sparse set of the $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is not zero. GHC aims to account for both sparse signals and correlation among SNPs when combining individual marker test statistics.

Let $\mathbf{T}'_j = (T_{1,j}, ..., T_{n,j})$ be the vector of observed variants at the j-th marker, $\mathbf{Y} = (y_1, \ldots, y_n)'$ be the observed disease status and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, ..., \hat{\mu}_n)'$ be the predicted mean for **Y** under the assumption of no genetic effect. The statistical the score for β_j, is under the global null

$$Z_j = \frac{\mathbf{T}_j'(\mathbf{Y} - \hat{\mu})}{\sqrt{\mathbf{T}_j' \mathbf{P} \mathbf{T}_j}}, \qquad (9)$$

of $P = W - WC(C'WC)^{-1}C'w$ and $W = \text{diag}\{\hat{\mu}_1(1-\hat{\mu}_1), \ldots, \hat{\mu}_n(1-\hat{\mu}_n)\}$. These individual variants test the statistics of asymptotically jointly distributed as $\mathbf{Z} \sim \text{MVN}(0, \boldsymbol{\Sigma})$, where the $(i,k)^{th}$ component of $\boldsymbol{\Sigma}$ is estimated by

$$\hat{\boldsymbol{\sigma}}_{jk} = \frac{\mathbf{T}_j' \mathbf{P} \mathbf{T}_k}{\sqrt{\mathbf{T}_j' \mathbf{P} \mathbf{T}_j} \sqrt{\mathbf{T}_k' \mathbf{P} \mathbf{T}_k}}. \qquad (10)$$

Define $S(t)$ by

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|z_j| \geq t\}}. \qquad (11)$$

The generalized higher criticism test statistic, defined as

$$T_{GHC} = \sup_{t \geq t_0} \left\{ \frac{S(t) - 2p(1-\phi(t))}{\sqrt{\widehat{\text{var}}(S(t))}} \right\}, \qquad (12)$$

where $\phi(t)$ is the standard normal distribution function and $\widehat{\text{var}}(S(t))$, calculated by accounting for the correlation between the $Z_j's$. The p-value

$$P(GHC) \geq T_{GHC} \qquad (13)$$

is also calculated accounting the correlation. The algorithm behind both methods is implemented in the R packages SKAT [8] and GHC [27], respectively.

3. The Data and Model Simulation

The genotype data used in this simulation are the SNPs on Chromosome 16 from 1,504 unaffected individuals in the WTCCC study of Crohn's disease. The 13, 479 SNPs from each individual are used to construct two haplotypes to give a total of 3,008 haplotypes for use in the simulation study. The simulation is repeated 1,500 times for a study size of 3,000 cases and 3,000 controls.

The new genotype data were generated and assigned disease status based on two disease SNPs. The first SNP rs3789038 is located at position 50711672bp in gene HMOX2 and has MAF equal to 0.31. The second, SNP rs3785142 has MAF equal 0.48 and is located at position 50753236bp in gene CYLD. There is a total of 7 SNPs in the data on gene HMOX2 with pairwise correlation ranging between 0.93 and 0.99, with median 0.99, while there are 8 SNPs in the data on CYLD having pairwise correlations between 0.51 and 0.99, with the median equal to 0.93. [17].

The model for one disease SNP used to generate disease status is

$$P(\text{diseased} \mid T) = \frac{e^{\alpha_0 + \beta_1 T}}{1 + e^{\alpha_0 + \beta_1 T}}, \qquad (14)$$

where T is the number of copies of the rare allele of the disease SNP, α_0 is a pre-specified baseline relative risk of disease and β_1 is the gene effect. In this study, 0.1, 0.2, 0.4 and 0.7 are arbitrarily chosen for gene effect of β_1 and we found that small gene effects are able to classify rare variants

better than large gene effects. Therefore, we have chosen the gene effect equal to 0.2.

The disease model for two disease SNPs is

$$P(\text{diseased} \mid T_1, T_2) = \frac{e^{\alpha_0 + \beta_1 T_1 + \beta_2 T_2}}{1 + e^{\alpha_0 + \beta_1 T_1 + \beta_2 T_2}}. \quad (15)$$

This model assumes the two disease SNPs act linearly on the logit scale and two situations are investigated. The first is for gene effect $\beta_1 = 0.1$ and $\beta_2 = 0.2$ while in the second case the gene effects are fixed at $\beta_1 = 0.2$ and $\beta_2 = 0.1$. Once again small gene effects for β_1 and β_2 give a better result in classifying rare variants. The example genotype data in simulation were shown in Table 1. The haplotypes are next coded to 0 (major allele) and 1 (minor allele) and used to construct pairs of parental genotypes using randomly selected haplotypes. Each pair of parental genotypes are then used to construct the genotype of an individual in the study. The disease status (y = 0,1) is determined using the probability from the logistic regression in Equation 5 and 6, respectively. The cases and controls are generated under the randomly selected assumption and disease SNP are rs3789038 and rs3785142.

Determining the status of an individual is performed by considering random numbers from the uniform distribution to compare with the probability. If the probability value is more than the random numbers, we define as case (disease = 1) and control otherwise (no disease = 0). The SNPs are coded in three fashions, with 0, 1, and 2 corresponding to homozygotes for the major allele, heterozygotes, and homozygotes for the minor allele, respectively. The genotype data from generating simulated data is shown in Table 1.

Table 1. The genotype data from generating simulated data 3,000 cases and 3,000 controls.

Individual	Disease status	SNP_1	SNP_2	...	$SNP_{13,479}$
1	1	0	1	2	2
2	1	2	1	1	1
⋮	1	⋮	⋮	⋮	⋮
3000	1	0	0	1	1
3001	0	2	1	1	1
3002	0	0	1	1	0
⋮	0	⋮	⋮	⋮	⋮
6000	0	0	0	0	1

4. Simulation Study

In this research we considered the efficiency of the model by using the false positive (FP) and true positive (TP) rates. The FP is the number of disease SNPs or SNP-set that are not identified, whereas the TP is the number of times the disease SNPs were detected. The result in this research are based on 1,500 replicates. The equation for calculating FP and TP rates in the study are as follows:

$$FP = \frac{\#\text{significant SNP except disease SNP}}{\#\text{A total SNPs} \times \#\text{A total replicates}}. \quad (16)$$

$$TP = \frac{\#\text{significant SNP in disease SNP}}{\#\text{A total replicates}}. \quad (17)$$

In this research, we present the simulation result of SKAT and GHC with B-spline for two cases, one disease SNP case and two disease SNPs case.

Case I: One disease SNP

The FP and TP rates of SKAT normal and GHC of rs3789038 and rs3785142 as disease SNP with gene effect size $\beta_1 = 0.2$ are provided in Table 2. The result has shown that when the degree of freedom increases FP and TP rates are decreasing. In this case, both methods were confirmed by the simulation results which show that b-spline under the degree of freedom 1,000 which is the optimal parameter with the lowest FP and high TP.

Table 2. The FP and TP rates from SKAT normal and GHC with B-spline based on permutation threshold with rs3789038 and rs3785142 as disease SNP and effect size $b_1 = 0.2$

d.f.	Disease SNP rs3789038				Disease SNP rs3785142			
	SKAT normal		GHC		SKAT normal		GHC	
	FP	TP	FP	TP	FP	TP	FP	TP
400	0.01505	0.95	0.00860	0.91	0.01440	0.50	0.01271	0.58
500	0.01082	0.93	0.00784	0.88	0.01124	0.60	0.01049	0.60
600	0.01152	0.88	0.00703	0.86	0.01003	0.70	0.00861	0.60
700	0.01087	0.86	0.00668	0.83	0.00925	0.74	0.00791	0.61
800	0.01030	0.85	0.00728	0.83	0.00846	0.78	0.00740	0.62
900	0.00963	0.86	0.00616	0.81	0.00782	0.80	0.00684	0.62
1000	0.00953	0.86	0.00616	0.81	0.00771	0.79	0.00679	0.62
1100	0.00953	0.86	0.00616	0.81	0.00771	0.79	0.00679	0.62
1200	0.00953	0.86	0.00615	0.81	0.00771	0.79	0.00679	0.62

Table 3. The TP and FP rates of SKAT normal and GHC method with B-spline based on permutation threshold with rs3789038 and rs3785142 as disease SNP.

d.f.	$\beta_2 = 0.2$ and $\beta_1 = 0.1$				$\beta_1 = 0.1$ and $\beta_2 = 0.2$			
	SKAT normal		GHC		SKAT normal		GHC	
	FP	TP	FP	TP	FP	TP	FP	TP
400	0.07079	0.99	0.03029	0.95	0.08264	0.88	0.03595	0.85
500	0.05783	0.98	0.02557	0.94	0.06972	0.88	0.02970	0.85
600	0.05165	0.97	0.02169	0.93	0.06147	0.91	0.02427	0.85
700	0.04779	0.96	0.02002	0.93	0.05660	0.91	0.02202	0.85
800	0.04483	0.95	0.01909	0.92	0.05273	0.92	0.02053	0.85
900	0.04224	0.96	0.01808	0.91	0.04953	0.94	0.01908	0.85
1000	0.04180	0.95	0.01796	0.91	0.04901	0.93	0.01900	0.85
1100	0.04180	0.95	0.01796	0.91	0.04901	0.93	0.01900	0.85
1200	0.04180	0.95	0.01796	0.91	0.04901	0.93	0.01900	0.85

The FP rate of SKAT normal is seen at 0.00953 to be higher than GHC but this comes at a higher TP rate 0.86. The FP rate of the GHC is 0.00616 and the TP rate is 0.81 show that there is a small degree of difference in the FP and TP.

In the case of rs3785142 as disease SNP, the result has shown that when the degree of freedom increases FP and TP rates are roughly increasing. The simulation results show b-spline under the degree of freedom 1,000 which is the optimal parameter with the lowest FP and high TP. The FP rate of SKAT normal is seen at 0.00771 to be higher than GHC but this comes at a higher TP rate 0.79. The FP rate of the GHC is 0.00679 and the TP rate is 0.62, which show that there is a small difference in FP while much higher TP.

Case II: Two disease SNPs

Comparisons of the FP and TP rates of SKAT normal and GHC of rs3789038 and rs3785142 under the disease model for two disease SNP are shown in Table 3. The FP and TP rates for SKAT normal and GHC with a gene effect sizes of $\beta_2 = 0.2$ and $\beta_1 = 0.1$. This is consistent with the finding in Table 2. The simulation results show that b-spline under the degree of freedom 1,000 is

the optimal parameter with the lowest FP. The FP rate of SKAT normal is seen at 0.04180 to be higher than GHC but this comes at a higher TP rate of 0.95. The FP rate of the GHC is 0.01796 and the TP rate is 0.91.

In the case of the gene effect size β_1 = 0.1 and β_2 = 0.2, the result has shown that when the degree of freedom increases, FP and TP rates are increasing. The simulation results show that b-spline under the degree of freedom 1,000 which is the optimal parameter with the lowest FP and a high TP. The FP rate of SKAT normal is seen at 0.04901 to be higher than GHC but this comes as higher than the TP rate of 0.93. The FP rate of the GHC is 0.01900 and the TP rate is 0.85 showings an above moderate difference in FP and TP.

The finding confirmed that the GHC outperforms SKAT normal. The different disease SNP affects the efficiency of the model which is SNP rs3789038 is driving the disease gene relationship in the simulation model. The ROC curves of the disease model for one disease SNP are shown in Fig. 1 and Fig. 2 and two disease SNP in Fig. 3 and Fig. 4, respectively.

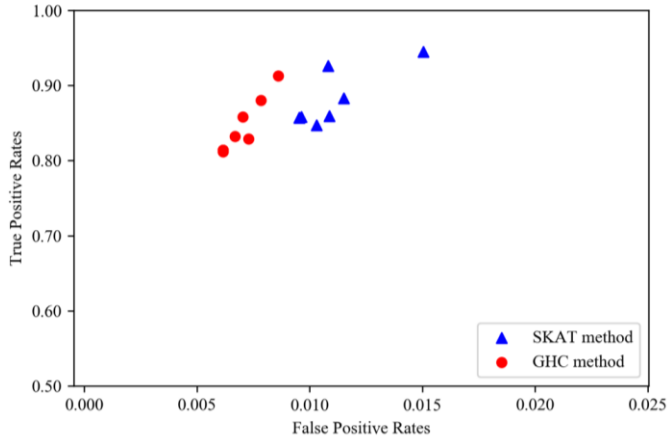

Fig. 1. The ROC curve for FP and TP rates from SKAT and GHC with B-spline with rs3789038 as disease SNP and effect size β_1 = 0.2

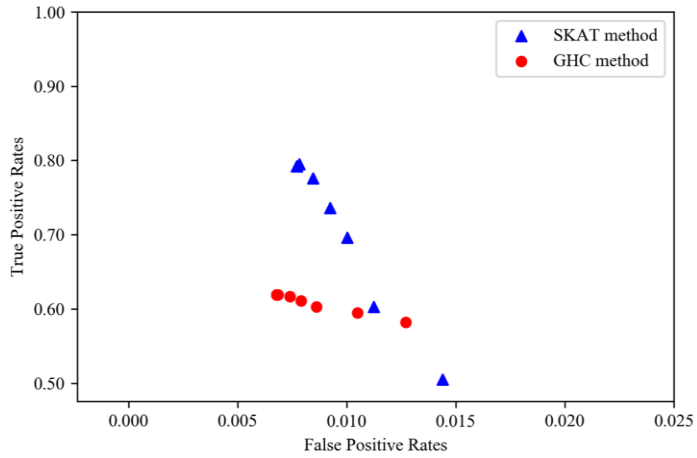

Fig. 2. The ROC curve for FP and TP rates from SKAT and GHC methods using B-spline with rs3785142 as disease SNPs and effect size β_1 = 0.2

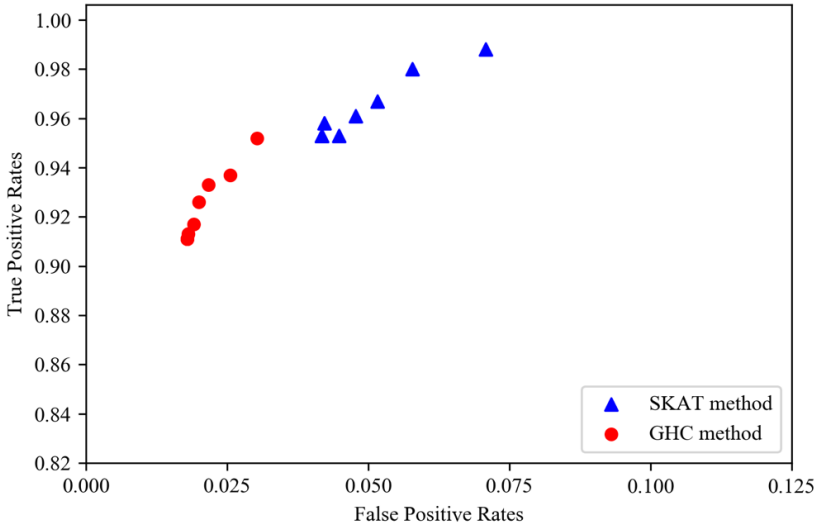

Fig. 3. The ROC curve for FP and TP rates from SKAT and GHC method using B-spline with rs3789038 and rs3785142 as disease SNP and respective effect sizes of $\beta_2 = 0.2$ and $\beta_1 = 0.1$

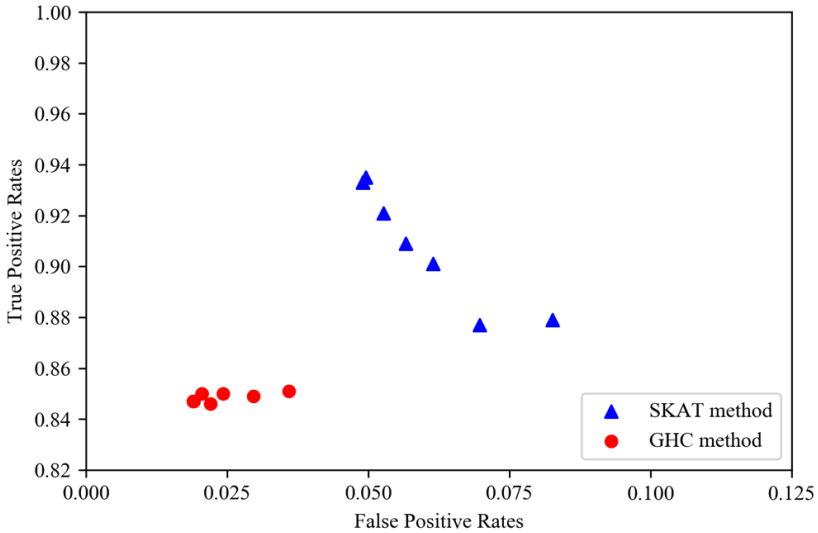

Fig. 4. The ROC curve for FP and TP rates from SKAT and GHC method using B-spline with rs3789038 and rs3785142 as disease SNP and respective effect sizes of $\beta_1 = 0.1$ and $\beta_2 = 0.2$

The result shows that SKAT and GHC give a comparable result. The optimal degree of freedom of both methods is 1,000. The example of the B-spline for declaring the significance of SKAT normal (a) and GHC (b) with degree of freedom 1,000 for one and two disease SNPs genetic models are shown in Fig. 5 and Fig. 6, respectively.

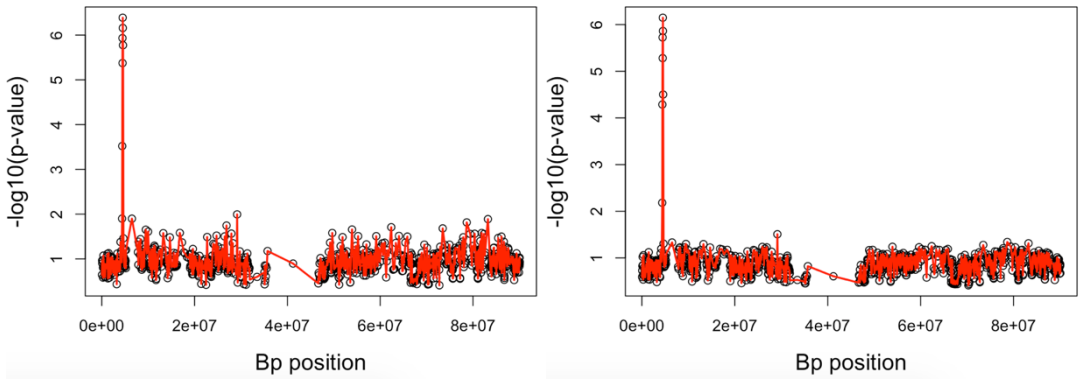

Fig. 5. The horizontal line of permutation threshold for declaring significant of SKAT normal (a) and GHC (b) with B-spline df 1,000 for one disease SNP genetic model.

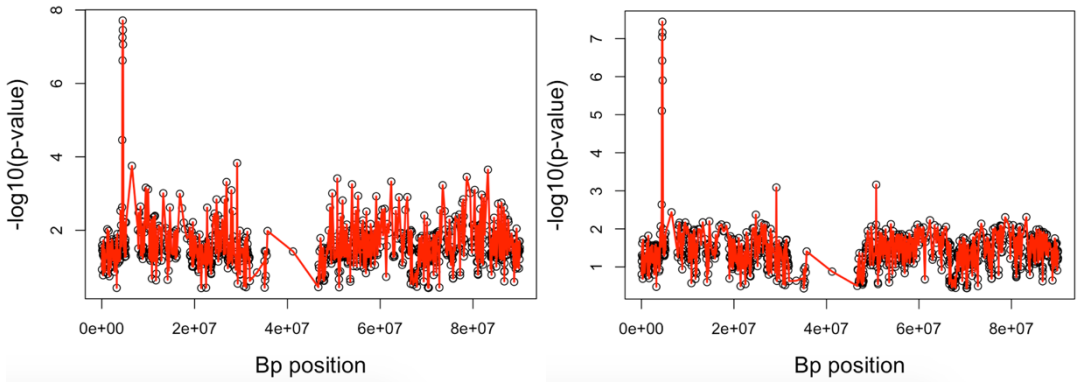

Fig. 6. The horizontal line of permutation threshold for declaring significant of SKAT normal (a) and GHC (b) with B-spline df 1,000 for two disease SNP genetic model.

5. Real Data Application

Both methods that were evaluated in the previous section can be applied to real data to define the location of SNP or the region that causes the disease. The data set used has 13,479 SNPs on Chromosome 16 which comprise 2,005 cases and 1,500 controls of Crohn's disease studies. Table 4 shows the SNPs which are declared as significant by SKAT normal and GHC with B-spline. SKAT normal found four regions. The largest region called region 7 located at 114,090 basepairs and contains a cluster of 27 SNPs. There are 7 SNPs located within an intron gene 174 (a region inside a gene), 12 SNPs were located in NOD2, and 8 SNPs were located in CLDY gene, respectively. Other significant genes were LOC646828, intron gene 89 and SMG1P5. Moreover, the GHC method found four regions, the largest region is called region 7 located at 114,090 basepairs and contains a cluster 27 SNPs. There are 7 SNPs located within an intron (a region inside a gene), 12 SNPs located in NOD2 and 8 SNPs located in the CLDY gene, respectively. Other significant genes were SMG1P5, LINC01566, and intron gene 173.

The result obtained from the real data shows that both methods give the same

regions that cause the disease. Especially, both methods were finding gene NOD2 which declared to concern Crohn's disease [28-30]. However, when comparing the performance, all GHC with B-spline give a lower FP in the simulation studies. In addition, GHC has an advantage in terms of lower computational cost. Therefore, it is clearly seen that GHC with B-spline is preferable.

Table 4. Gene on Chromosome 16 declared as significant SKAT and GHC method without and with B-spline for the WTCCC study of Crohn's disease.

Region	SKAT normal with B-spline	GHC with B-spline
1	LOC646828, Intron Gene 89	-
2	-	-
3	SMG1P5	SMG1P5
4	-	LINC01566
5	Intron Gene 151, KIF18BP1	-
6	-	Intron Gene 173
7	Intron Gene 174, NOD2, CYLD	Intron Gene 174, NOD2, CYLD

6. Discussion and Conclusion

Identifying an optimal parameter and appropriate thresholds for declaring significance are two important and related problems remaining to be solved. It is desirable to have an optimal parameter. The findings showed that b-spline under the degree of freedom 1,000 is the optimal parameter for all conditions. Obviously, when setting the degree of freedom more than 1,000 the FP and TP rates are unvarying. Defining the degree of freedom over the number of variables (SNP-sets) causes the B-spline to overfit with the data set. Moreover, if we are increasing the degree of freedom it will take a lot of time to fit the B-spline. If we compared the efficiency of the model using FP and TP rates, SKAT normal is highly TP and FP while GHC with B-spline is less than TP and gives the lower FP. Both methods have different advantages and disadvantages, SKAT normal gives a high FP while GHC gives the lower FP which we focus on reducing. But it is difficult to specify the efficiency of the model. It can be seen that both models obtain the same region in real data application analysis. There are many factors that should be considered as well, such as the computation. In the process of obtaining the SKAT and GHC p-value, it was found that the GHC takes the most time to analyze, about 750 hours while SKAT normally used only about 100 hours. It is clear that SKAT normal is very advantageous and can reduce computational time while efficiency is the same as GHC.

In the section of real data analysis, it was found that both methods found particularly important regions. Region 4 was involved in genes 174, NOD2 and CYLD. Many researchers found that gene NOD2 is strongly associated with the development and important genetic variant cause Crohn's disease [28-30]. Finally, the researchers expect that this method will be able to apply to other diseases that have not yet been able to identify the SNP-sets that affect the disease.

References

[1] Bush WS, Moore JH. Genome-wide association studies. PLoS Comput. Biol 2012; 8: e1002822.

[2] Lewis CM. Genetic association studies: design, analysis and interpretation. Briefings in Bioinformatics 2002; 3: 146-53.

[3] Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nature Protocols 2011; 6: 121–33.

[4] Lee Y, Luca F, Pique-Regi R, Wen X. Bayesian multi-SNP genetic association analysis: control of FDR and use of summary Statistics. bioRxiv 2018; 316471.

[5] Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. American Journal of Human Genetics 2001; 70: 425-34.

[6] Zhao Y, Chen F, Zhai R, Lin X, Diao N, Chritiani DC. Association Test Based on SNP Set: Logistic Kernel Machine Based Test vs. Principal Component Analysis. PLoS ONE 2012; 7: 1-11.

[7] Kirdwichai P, Baksh MF. The analysis of genomewide SNP data using nonparametric and kernel machine regression. Journal of Applied Science 2019; 18: 20-30.

[8] SKAT package [Internet]. [cited 2019 Nov 10]. Available from: https://cran.r-project.org/web/packages/SKAT/SKAT.pdf

[9] Wu MC, Kraft P, Epstein MP, Taylor D, M, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. American Journal of Human Genetics 2010; 86: 929-42.

[10] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 2011; 89: 82-93.

[11] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American Journal of Human Genetics 2012; 91: 224-37.

[12] Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. American Journal of Human Genetics 2013; 92: 841-53.

[13] Barnett I, Mukherjee R, Lin X. The Generalized higher criticism for testing SNP-set effects in genetic association studies. Journal of the American Statistical Association 2017; 112: 64-76.

[14] Goepp V, Bouaziz O, Nuel Z. Spline regression with automatic knot selection. HAL Archives-Ouvertes.fr 2018: hal-01853459.

[15] Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brein SJ. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 2010; 11: 724.

[16] Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F. Statistical analysis for genome-wide association study. The Journal of Biomedical Research 2015; 29: 285-297.

[17] Sookkhee S, Baksh MF, Kirdwichai P. Efficiency of Single SNP analysis and Sequence Kernel Association Test in Genome-wide Association Analysis. In: Ao SI, Castillo O, Douglas C, Dagan DF, Korsunsky AM, eds. Proceedings of the International MultiConference of Engineers and Computer Scientists; 2018 Mar 14-16; Hong Kong. IAENG International Journal of Applied Mathematics. [cited 2019 Nov 9]. Available form: http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp308-313.pdf

[18] Corcoran CD, Senchaudhuri P, Mehta CR, Patel NR. (2005). Exact Inference for

Categorical Data. In Armitage P, Colton T, eds. Encyclopedia of Biostatistics (2nd ed.) pp.1804-1820. Chichester, UK: John Wiley.

[19] Permutation Test & Monte Carlo Sampling [Internet]. [cited 2019 Jan 10]. Available from : http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Permutation/Permutation-Monte-Carlo-Jianqiang-2009.pdf

[20] The Wellcome Trust Case-Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447: 661-678.

[21] Rodríguez G. Logit Models for Binary Data. [Internet]. [cited January 10, 2019]. Available from : https://data.princeton.edu/wws509/notes/c3.pdf

[22] Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of Spline function procedures in R. BMC Medical Research Methodology 2019; 19:1-16.

[23] De Boore C. A Practical Guide to Splines. New York: Springer-Verlagp; 1978.

[24] Jianqiang MA. Permutation Test & Monte Carlo Sampling. [Internet]. [cited 10 Jan 2019]. Available from : http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/ Permutation-Monte-Carlo-Jianqiang-2009.pdf

[25] Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 2007; 31: 358-362.

[26] Banerjee M, Mukherjee D, Mishra S. Semiparametric binary regression models under shape constraints. Research supported in part by National Science Foundation grants. [Internet]. [cited 4 Nov 2020]. Available from : http://dept.stat.lsa.umich.edu/~moulib/bmmtechrep.pdf

[27] Barnett I. GHC: Computes P-values for the Generalized Higher Criticism. R package version 1.0. [Internet]. [cited 2019 Jan 10]. Available from: https://scholar.harvard.edu/ibarnett/software/generalized-higher-criticism

[28] Michail S, Bultron G, DePaolo RW. Genetic variants associated with Crohn's disease. The Application of Clinical Genetics 2013; 6: 25–32.

[29] Sidiq T, Yoshihama S, Downs I, Kobayashi KS. Nod2: A Critical Regulator of ileal Microbiota and Crohn's Disease. Frontiers in Immunology 2016; 7: 1-11.

[30] Nicholas AK, MBBS F, Christopher AL, MBBS, Susan HB, BS, Alan WW, John M, FRCP, Miles P, Rachel S, BSc, Mark T, Sarah N, Genetics Consortium, Julian P, Chris P, Georgina LH, Charlie WL. The Impact of NOD2 Variants on Fecal Microbiota in Crohn's Disease and Controls Without Gastrointestinal Disease. Inflammatory Bowel Diseases 2018; 24: 583–592.