Thailand Statistician
July 2025; 23(3): 643-656
http://statassoc.or.th
Contributed paper

Tuning Spline Smoothing Parameters in GWAS Using Replication-Based Approach

Chanunya Pailoung [a], Pianpool Kamoljitprapa*[a] and Sirikanlaya Sookkhee [b]

[a] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

[b] Department of Mathematics, Faculty of Education, Sisaket Rajabhat University, Si Sa Ket, Thailand.

*Corresponding author; e-mail: pianpool.k@sci.kmutnb.ac.th

Received: 10 September 2024
Revised: 4 January 2025
Accepted: 11 March 2025

Abstract

Genome-Wide Association Study (GWAS) is an approach for identifying the associations between genetic variants, especially Single Nucleotide Polymorphisms (SNPs), and phenotypes, such as disease risk. GWAS can be conducted either on a single SNP or groups of SNPs. However, analyzing the GWAS data can be challenging due to its high dimensionality, leading to an inflation of type I error rate and computational burdens when conducting multiple hypotheses testing. To address these limitations, this research investigates the association between SNP sets, grouped by gene, and the risk of Crohn's disease. The Sequence Kernel Association Test (SKAT) is employed to assess these associations, while spline regression analysis is used to construct the model and reduce analytical complexity. This research aims to obtain the optimal smoothing parameters, particularly the degree of freedom, for the spline regression model and the optimal number of replications for simulated data, and to apply the optimal model for identifying gene regions associated with Crohn's disease. The results indicate that the degree of freedom of 1,000 is the optimal parameter for the spline regression model, as it provides the lowest false positive rate while maintaining a reasonable true positive rate. Additionally, 1,000 replicates have been identified as the optimal number of replications, as this value ensures the most efficient processing time. Ultimately, the optimized model can effectively identify gene regions associated with Crohn's disease while minimizing the error rate and conserving computational resources during the analysis of extensive data.

Keywords: GWAS, replication, sequence kernel association test, smoothing parameter, spline regression.

1. Introduction

A Genome-Wide Association Study (GWAS) is a research approach used to identify the association between genomic variants and phenotypes, which helps researchers discover DNA locations associated with complex diseases. In GWAS, a primary interest lies in exploring the association between diseases and Single Nucleotide Polymorphisms (SNPs). An SNP is a genetic

variant at a single base position in the DNA. Although SNPs do not harm organisms, their locations are often associated with disease risk or drug response (Sukhumsirichart 2018).

Single SNP Analysis is a straightforward method to identify the association between a disease and one SNP location at a time (Sookkhee et al. 2018, Sookkhee et al. 2021). Though, with approximately 10 million SNP locations in the human genome, conducting simultaneous hypotheses testing can lead to significant challenges, particularly in inflating of the type I error rate. The Bonferroni correction is one approach to mitigate this issue, but it is known for being extremely conservative (Sookkhee et al. 2021, Kamoljitprapa et al. 2023, Kamoljitprapa et al. 2024). Another method, the permutation test, is a nonparametric approach that can estimate the distribution of test statistics under the null hypothesis (Berger 2011). Researchers can select the percentile of test statistics from this distribution that corresponds to or is close to the desired significance level and use it as a threshold for their hypotheses test.

However, an SNP location might not strongly affect the risk of disease, and groups of closely linked SNPs often underlie the development of disease as presented by Sookkhee et al. (2021). Moreover, single SNP analysis can be time-consuming. SNP-set analysis reduces processing time by identifying associations between groups of SNPs and phenotypes. SNPs are commonly grouped by genomic features such as genes and haplotype blocks. The Sequence Kernel Association Test (SKAT) proposed by Wu et al. (2011) is a notable method that identifies associations between SNP sets grouped by gene and phenotype using the logistic kernel machine model.

Since GWAS data is considered high-dimensional, the analysis may face complexity, often involving significant computational burdens due to numerous simulation studies. To address these challenges, spline regression analysis plays an important role in constructing and smoothing the model to reduce both noise and complexity. Moreover, spline regression effectively handles nonparametric data, ensuring reliable outcomes regardless of data assumptions. However, the efficiency of the spline regression model depends on the smoothing parameters, especially the degree of freedom, which can be challenging to optimize (James et al. 2021, Sookkhee et al. 2021, Pailoung et al. 2024).

In this research, we focus on exploring the association between SNP sets and Crohn's disease, by using SKAT for testing the associations and constructing the spline regression model through the R statistical software (R Core Team 2022) with R packages "SKAT" from Lee and Zhao (2023) and "splines" from Bates and Venables (2022). The data used for simulation are sourced from the 1958 British Birth Cohort (Burton et al. 2007). The research objectives include optimizing the smoothing parameter, that is the degree of freedom for spline regression model, and the number of replications for simulated data. Lastly, this research aims to apply the optimal spline regression model to real data in order to identify the gene region associated with Crohn's disease.

2. Method

This section presents the methodology used in this research, including SKAT for identifying the association between SNP sets and the disease, spline regression for constructing and smoothing the models, and the permutation method for adjusting thresholds in multiple hypotheses testing.

2.1. Sequence kernel association test

The Sequence Kernel Association Test (SKAT), proposed by Wu et al. (2011), is a supervised, flexible, and computationally efficient regression method for testing associations between genetic variants in a region, such as a gene or haplotype block, and both continuous and dichotomous phenotypes. SKAT rapidly calculates the p-value for association using a variance-component score test within a mixed-model framework. The logistics model for the i-th individual is defined as

$$\text{log}it\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i + \boldsymbol{\beta}'\boldsymbol{G}_i, \tag{1}$$

where y_i indicates the disease status (1 for case or having disease and 0 otherwise), α_0 is an intercept term, $X_i = (X_{i1}, X_2, \ldots, X_{im})$ is the covariates with $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_m]'$ is the regression coefficient vector for the m covariates, and $\boldsymbol{G}_i = (G_{i1}, G_{i2}, \ldots, G_{ip})$ is the genotypes of variants within a region where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]'$ is the vector of regression coefficient for p variants.

To test whether the variants in a region are associated with the disease, the null hypothesis can be constructed as $\beta_1 = \beta_2 = \ldots = \beta_p$. Each β is assumed to follow an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance-component, and w_j is a pre-specified weight for j^{th} variant in a region where $j = 1, \ldots, p$. The null hypothesis can then alternatively be stated as $\tau = 0$, which can be conveniently tested with a variance-component score test. The test statistic is defined as

$$Q = (y - \hat{\mu})'\boldsymbol{K}(y - \hat{\mu}), \tag{2}$$

where $\hat{\mu}$ is the predicted mean of y under the null hypothesis that $\hat{\mu} = \text{log}it^{-1}(\hat{\alpha}_0 + \boldsymbol{X}\hat{\boldsymbol{\alpha}})$, $\boldsymbol{G}$ is an $n \times p$ genotype matrix, which each element of $\boldsymbol{G}$ is a j^{th} variant of i^{th} individual, and $\boldsymbol{W}$ is a $p \times p$ diagonal matrix containing the w_j weights for p variants. $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'$ is an n×n kernel matrix with the (i, i') element is the kernel function $K(\boldsymbol{G}_i, \boldsymbol{G}_{i'})$ of genotypes in a region from the i^{th} and i'^{th} individual, respectively.

The kernel function measures the similarity between the genetic variants of the sample. Assuming the relationship between SNPs and the disease is linear with no interactions, the weighted linear kernel function $K(\boldsymbol{G}_i, \boldsymbol{G}_{i'}) = \sum_{i=1}^{p} w_j G_{ij} G_{i'j}$ is used for its high power in this research. The weight function is set as Beta distribution, where the random variables are minor allele frequency of the j^{th} SNP or MAF_j in a region.

$$\sqrt{w_j} = Beta(MAF_j; a_1, a_2) = \frac{MAF_j^{a_1 - 1}(1 - MAF_j)^{a_2 - 1}}{B(a_1, a_2)}, \tag{3}$$

where B denotes the beta function, $0 < MAF_j < 1$ and $a_1, a_2 > 0$. The weight used in this research is normal weight $Beta(MAF_j; 10, 10)$. According to Sookkhee et al. (2018), normal weight provides the most efficient result for detecting the association between genetic variants and Crohn's disease.

2.2. Permutation test

The permutation test is a nonparametric method that helps researchers determine the threshold for controlling the error rate in multiple hypothesis tests under the null hypothesis of no effect. The permutation test can estimate the sampling distribution of test statistics, which are highly reliable but require a large number of samples generated under the null model (Berger 2011, Sookkhee et al. 2021). To apply this test, the data labels must be exchangeable under the null hypothesis. One reason the permutation test is commonly used, beyond being a nonparametric test that requires no assumptions about the data distribution, is that the distribution of test statistics is directly obtained from the data itself. This makes the method robust to the shape of the distribution. In this research, the thresholds were adopted from Sookkhee et al. (2021), which involved simulating 10,000 replicates under the null hypothesis that no SNP or SNP sets causes or affects Crohn's disease and

varied according to the degrees of freedom. The simulation was for computing the multivariate distribution that achieved a type I error rate close to 0.05.

2.3. Spline Regression

A spline is a piece-wise polynomial constructed by dividing the x-axis into intervals and fitting a polynomial function of degree d on each interval. These polynomial functions are connected by knots (James et al. 2021). A general model for a spline of degree d with K knots, $\xi_1, \xi_2, \ldots, \xi_K$ are the knot sequence, is given by

$$f(x) = \sum_{i=0}^{d} \beta_i x^i + \sum_{j=d+1}^{K+d} \beta_j (x - \xi_j)_+^d, \tag{4}$$

where $(x-\xi_j)_+ = \begin{cases} x-\xi_j; x > \xi_j \\ 0; \text{otherwise,} \end{cases}$ β_i represent the regression coefficients for the polynomial, x is the independent variable, β_j are the regression coefficients for the spline and $(x-\xi_j)_+^d$ is the basis function of spline. To ensure splines join smoothly and continue at every knot, constraints must be set. Since d is the degree of spline and K is the number of knots, this spline function requires continuity in derivatives up to degree $d-1$ at each knot. Thus, the spline function with $K+1$ distinct polynomial functions of degree d will be tied together smoothly at the K knots.

The most commonly used spline degree is 3, known as the cubic spline, since it provides a reasonably smooth approximation to most non-linear functions (Kirdwichai 2016, Kamoljitprapa and Leelasilapasart 2024). Another reason is that these curves appear perfectly smooth to the human eye (Perperoglou et al. 2019). The spline applied in this research is the B-spline, which basis function can be defined by a recursive function:

$$B_j^d(x) = \frac{x-\xi_j}{\xi_{j+d}-\xi_j} B_j^{d-1}(x) - \frac{\xi_{j+d+1}-x}{\xi_{j+d+1}-\xi_{j-1}} B_{j+1}^{d-1}(x);\ j = 1, \ldots, K+d+1, \tag{5}$$

where $B_j^0(x) = \begin{cases} 1; \xi_j \le x \le \xi_{j+1} \\ 0; x < \xi_j, x > \xi_{j+1}, \end{cases}$ and $B_j^0(x) \equiv 0$ if $\xi_j \equiv \xi_j + 1$. B-spline produces a curve from sub curves created in each interval, known as local estimation. Changing a few knot positions affects only the sub-curves near the knots, making B-spline curves easy to shape without affecting the overall curve (Perperoglou et al. 2019, Sookkhee et al. 2021). Moreover, B-spline uses a recursive function to define its basis, which avoids multicollinearity between basis splines.

However, the efficiency of the spline depends on tuning parameters, such as the number of knots and the degree of freedom, which refers to the number of free parameters in the model. Although directly specifying the number and locations of knots is a straightforward way to optimize the parameters, it might require considerable time and effort for extensive data. In practice, it is common to specify the desired degrees of freedom and then have the R statistical software automatically place the corresponding number of knots at uniform quantiles of the data (James et al. 2021).

3. Data and Model Simulation

The data used for simulation in this research comprises genotype data of 13,479 SNPs on chromosome 16 from 1,504 healthy control individuals obtained from the 1958 British Birth Cohort. These 13,479 SNPs from 1,504 controls will be separated into 3,008 haplotypes. The genotype will be encoded into two values: 0 for the major allele and 1 for the minor allele, with the minor allele

being more likely to be a risk allele, according to Kido et al. (2018). Next, two haplotypes will be randomly selected and combined to create a genotype for each simulated individual. The simulated genotypes for each SNP are represented as 0, 1, and 2, indicating homozygotes for the major alleles, heterozygotes, and homozygotes for the minor alleles, respectively.

A disease SNP will be assigned to the simulated data. Two disease SNPs are arbitrarily chosen: SNP rs3789038 located at 50,711,672bp on the HMOX2 gene and SNP rs3785142 located at 50,753,236bp on the CYLD gene. The HMOX2 gene contains 7 SNPs, while the CYLD gene contains 8 SNPs. Then, the disease status for each simulated individual will be determined via a logistic function:

$$P(y_i = 1 \mid T_i) = \frac{e^{\alpha_0 + \beta T_i}}{1 + e^{\alpha_0 + \beta T_i}}, \tag{6}$$

where α_0 is a pre-specified relative risk set as $\alpha_0 = 0.0$, T_i is the number of copies of the rare allele or the encoded genotype on a disease SNP of an individual, and β is a gene effect that indicates how strongly the disease SNP can affect the i-th individual. In this research, the main interest is in the rare variant that shows a weak effect or is hard to classify. According to Sookkhee et al. (2021), small gene effects can classify rare variants more effectively than large gene effects. Thus, the gene effect is chosen to be $\beta = 0.2$.

To determine the optimal number of replications for simulation, the number will be varied from 1,000, 1,500, 2,000, 2,500, and 3,000. Each replicate comprises the simulated genotype of 3,000 cases and 3,000 controls.

4. Simulation Study

This section presents the simulation results of the models. The model's efficiency is assessed by false positive (FP) and true positive (TP) rates. FP rate occurs when the model incorrectly detects an SNP set not designated as associated with the disease. In contrast, the TP rate occurs when the model correctly identifies the disease-associated SNP set. Since one simulated replication contains only one disease-associated SNP set, the equations for calculating the FP and TP rates across all replications are as follows:

$$FP = \frac{\#\text{significant disease} - \text{unassociated SNP sets}}{\#\text{total unassociated SNP sets} \times \#\text{total replications}}, \tag{7}$$

$$TP = \frac{\#\text{significant disease} - \text{associated SNP sets}}{\#\text{total replications}}. \tag{8}$$

Several steps are taken to identify the optimal model in this research. The optimal number of replications for each degree of freedom is selected first, followed by evaluating the optimal degree of freedom for spline regression.

4.1. SNP rs3789038 as a disease SNP

The ROC curves for FP and TP rates obtained from SKAT with B-spline for rs3789038 as a disease SNP with different degrees of freedom and number of replications are shown in Figure 1.

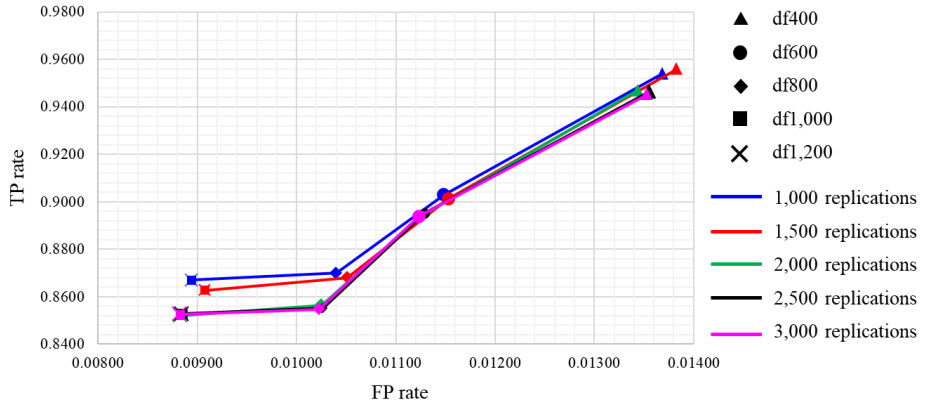

Figure 1 ROC curves for FP and TP rates from SKAT with B-spline with rs3789038 as a disease SNP

Figure 1 illustrates that as the degrees of freedom increase, both FP and TP rates decrease. While increasing numbers of replications slightly lower FP and TP rates. Moreover, FP and TP rates are diverged into two groups: below 2,000 replications and the other above, after passing the degree of freedom of 800 (♦). It shows that FP and TP rates become similar when the degree of freedom is high.

To explore the differences in each degree of freedom and each number of replications, the boxplot of FP and TP rates are shown in Figure 2 and Figure 3:

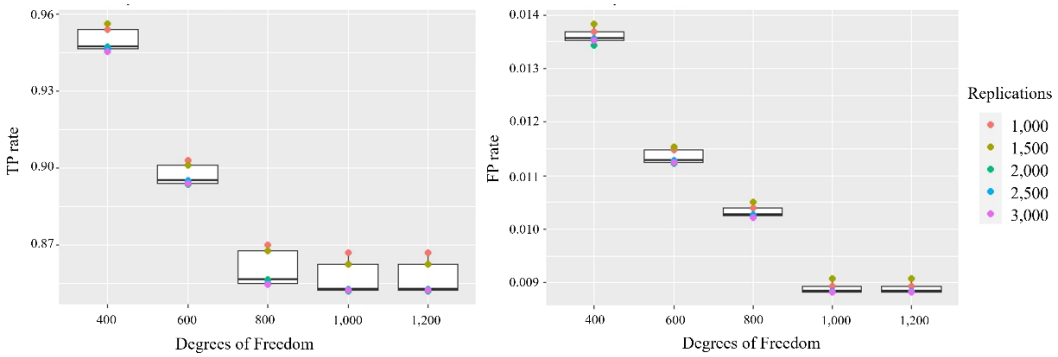

Figure 2 Boxplot for FP rate (right) and TP rate (Burton et al.) from SKAT with B-spline with rs3789038 as a disease SNP, where the x-axis is the degree of freedom

Figure 2 shows the boxplot for FP and TP rate where the x-axis is the degrees of freedom. As the degree of freedom increases, the FP and TP rate boxes gradually shift downward and stabilize after reaching a degree of freedom of 1,000. Additionally, the variability of the FP rate in each replication, indicated by the height of the FP box, decreases with an increase in the degrees of freedom. This observation confirms that increasing the degrees of freedom positively impacts the model's efficiency.

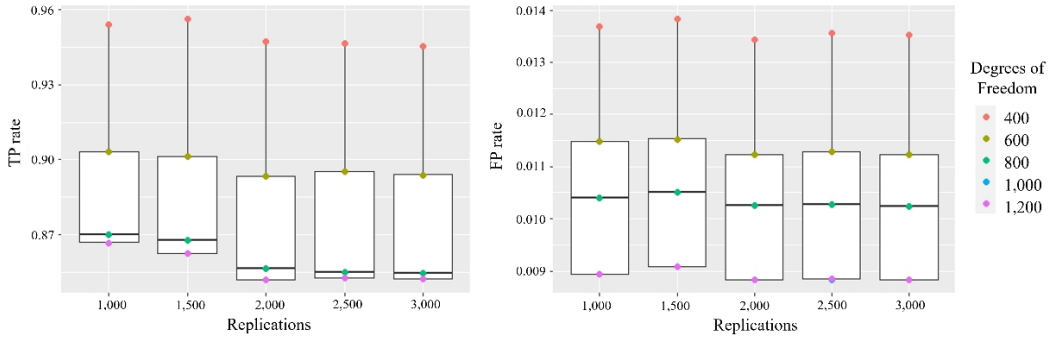

Figure 3 Boxplot for FP rate (right) and TP rate (Burton et al.) from SKAT with B-spline with rs3789038 as a disease SNP, where the x-axis is the replication.

Figure 3 shows the boxplot for FP and TP rate where the x-axis is the number of replications. Conversely, from Figure 2, as the number of replications increases, both FP and TP rate boxes tend to lower slightly and remain consistent in shape, suggesting that increasing the number of replications has a marginal effect on model efficiency. Therefore, evaluating the optimal number of replications for each degree of freedom requires significant processing time. Then, the FP, TP rates and processing time in hours from different degrees of freedom and replications will be presented in Table 1.

Table 1 Achieved FP, TP rates and runtime from each degree of freedom, separated by number of replications from SKAT with B-spline with rs3789038 as a disease SNP

Degree of Freedom	Number of Replications	FP	TP	Time (hours)
400	1,000	0.01369	0.9540	91.72
	1,500	0.01383	0.9560	141.07
	2,000	0.01344	0.9470	190.80
	2,500	0.01356	0.9464	240.07
	3,000	0.01353	0.9453	290.57
600	1,000	0.01148	0.9030	91.77
	1,500	0.01153	0.9013	141.13
	2,000	0.01123	0.8935	190.87
	2,500	0.01129	0.8952	240.15
	3,000	0.01124	0.8940	290.67
800	1,000	0.01040	0.8700	91.90
	1,500	0.01051	0.8680	141.33
	2,000	0.01025	0.8565	191.15
	2,500	0.01027	0.8552	240.50
	3,000	0.01023	0.8547	291.03
1,000	1,000	0.00894	0.8670	92.00
	1,500	0.00908	0.8627	141.48
	2,000	0.00883	0.8520	191.32
	2,500	0.00884	0.8528	240.72
	3,000	0.00882	0.8527	291.35

Table 1 (Continued)

Degree of Freedom	Number of Replications	FP	TP	Time (hours)
	1,000	0.00894	0.8670	92.45
	1,500	0.00908	0.8627	142.13
1,200	2,000	0.00883	0.8520	192.18
	2,500	0.00884	0.8528	241.78
	3,000	0.00882	0.8527	292.63

Table 1 presents FP rates, TP rates, and processing time in hours obtained from SKAT with B-spline with SNP rs3789038 as a disease SNP. As shown in Table 1, the higher the degree of freedom and the number of replications, the lower the FP and TP rates. Overall, the 1,000 replications yield the highest FP and TP rates, while 2,000 and 3,000 replications produce the lowest FP rate, which is the main interest of this study. However, increasing 500 replications results in 50 additional hours.

For the degrees of freedom of 400 and 600, the 2,000 replications return the lowest FP rates, 0.01344 and 0.01135, and considerable TP rates of 0.9470 and 0.8935, with 190.80 and 190.87 runtime hours, respectively. Whereas the 1,000 replications give comparable FP rate of around 0.01369 and 0.01148 for the degrees of freedom of 400 and 600, respectively, which are 1.86% and 2.23% higher than those with 2,000 replications but require only 91.72 and 91.77 hours, representing 51.93% and 51.92% reduction in time.

For the degrees of freedom of 800, 1,000 and 1,200, the lowest FP rates are 0.01023, 0.00882 and 0.00882 obtained from 3,000 replications, requiring an average of 291.67 hours in processing time. The 1,000 replications still give similar FP rates of 0.01040, 0.00894, and 0.00894, respectively, but taking only an average of 92.12 runtime hours, which is 68.42% less time.

4.2. SNP rs3785142 as a disease SNP

The ROC curves for FP and TP rates obtained from SKAT with B-spline for rs378142 as a disease SNP with different degrees of freedom and number of replications are shown in Figure 4.

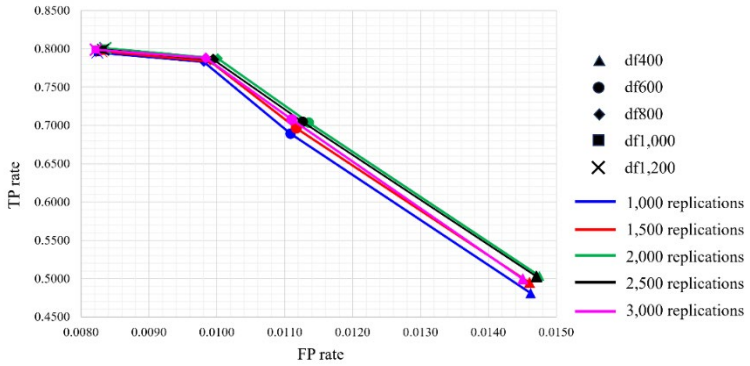

Figure 4 ROC curve for FP and TP rates from SKAT with B-spline with rs3785142 as a disease SNP

Figure 4 indicates that as the degrees of freedom in the model increase, the FP rates decrease, whereas the TP rates increase. Regarding the number of replications, higher replications slightly

increase both FP and TP rates. Additionally, at higher degrees of freedom, the data points from each replication closely align within each degree of freedom, with points nearly overlapping.

To further explore the patterns across degrees of freedom and numbers of replications, boxplots of FP and TP rates are shown in Figures 5 and 6.

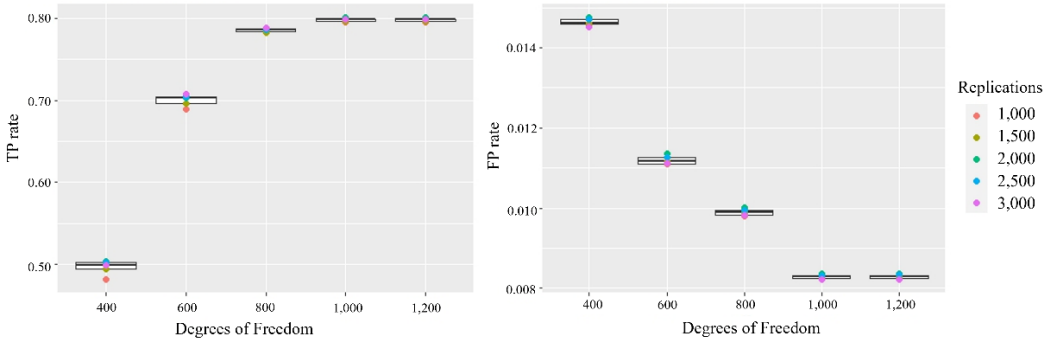

Figure 5 Boxplot for FP rate (right) and TP rate (Burton et al.) from SKAT with B-spline with rs3785142 as the disease-associated SNP, where the x-axis is the degree of freedom

Figure 5 shows the boxplot of FP and TP rates using rs3785142 as the disease-associated SNP, with the x-axis representing the degrees of freedom. When the degree of freedom increases, the FP rate boxes decline, while the TP rate boxes raise, and both stabilize after reaching a degree of freedom of 1,000. Additionally, the boxes become narrower at a higher degree of freedom, indicating that increasing the degree of freedom reduces the variance of FP and TP rates in each replication.

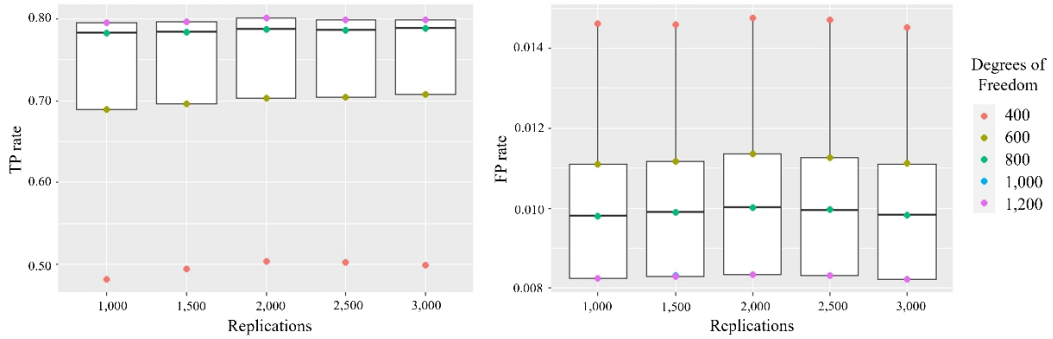

Figure 6 Boxplot for FP rate (right) and TP rate (Burton et al.) from SKAT with B-spline with rs3785142 as the disease-associated SNP, where the x-axis is replications

Figure 6 shows the boxplot of FP and TP rates using rs3785142 as the disease-associated SNP, with the x-axis representing the number of replications. Although TP rate boxes tend to stay higher with more replications, FP rate boxes remain consistent regardless of the number of replications. This suggests that the number of replications rarely affects the model's efficiency, with processing time being a significant factor in determining the optimal number of replications for each degree of freedom. Subsequently, Table 2 will display the FP and TP rates along with runtime (in hours) across various degrees of freedom and numbers of replications.

Table 2 Achieved FP, TP rates and run time from each degree of freedom, separated by number of replications from SKAT with B-spline with rs3785142 as a disease SNP

Degree of Freedom	Number of Replications	FP	TP	Time (hours)
400	1,000	0.01462	0.4810	88.35
	1,500	0.01460	0.4947	133.58
	2,000	0.01475	0.5035	178.85
	2,500	0.01470	0.5028	224.95
	3,000	0.01451	0.4993	270.40
600	1,000	0.01109	0.6890	88.37
	1,500	0.01118	0.6960	133.62
	2,000	0.01136	0.7035	178.92
	2,500	0.01127	0.7044	225.03
	3,000	0.01111	0.7073	270.50
800	1,000	0.00981	0.7830	88.42
	1,500	0.00990	0.7833	133.68
	2,000	0.01001	0.7875	179.00
	2,500	0.00995	0.7860	225.13
	3,000	0.00983	0.7880	270.62
1,000	1,000	0.00824	0.7950	88.60
	1,500	0.00830	0.7967	133.97
	2,000	0.00835	0.8010	179.38
	2,500	0.00832	0.7988	225.60
	3,000	0.00821	0.7987	271.18
1,200	1,000	0.00824	0.7950	89.07
	1,500	0.00830	0.7967	134.65
	2,000	0.00835	0.8010	180.30
	2,500	0.00832	0.7988	226.75
	3,000	0.00821	0.7987	272.55

Table 2 displays the FP rate, TP rate, and processing time obtained from SKAT with B-spline using SNP rs3785142 as the disease-associated SNP. The table indicates that increasing the degree of freedom raises TP rates while reducing FP rates. Conversely, increasing the number of replications slightly increases both FP and TP rates. Among the tested replication settings, 2,000 replications yield the highest TP rate, while the lowest FP rates are observed with 1,000 and 3,000 replications. Additionally, increasing the simulation by 500 replications results in approximately 45 additional runtime hours.

In the degree of freedom of 400, 3,000 replications yield the lowest FP rate at 0.01451, requiring 270.40 processing hours. In contrast, 1,000 replications result in a slightly higher FP rate at 0.01462, which is 0.76% higher than that of 3,000 replications, but with reduced processing time to 88.35 hours, which is 67.38% lower. For the degrees of freedom of 600 and 800, 1,000 replications offer both the lowest FP rates and the shortest processing time which are 0.01109 and 0.00981 with 88.37 and 88.42 hours, respectively.

Finally, 3,000 replications provide identical results degrees of freedom of 1,000 and 1,200, achieving the lowest FP rate of 0.00821 with an average processing time of 271.87 hours. In contrast, 1,000 replications can produce the closely equivalent FP rate to 3,000 replications which are 0.00824, merely 0.36% higher, while requiring only 88.84 processing hours, which is 67.32% less time compared to 3,000 replications.

4.3. Evaluation the optimal spline regression model and the replications

The simulation results, which involve varying the disease SNP, degrees of freedom and the number of replications, confirm that the degree of freedom significantly affects the model's efficiency, while the number of replications has a marginal impact, mainly when the degree of freedom of the model is high. Moreover, the height of each box from the boxplot with the number of replications is an x-axis, is comparable in both simulation cases. Differences in FP rates across distinct numbers of replications are minor, only 1-3%, but the differences in processing time are substantial. Thus, the processing time is a key factor for determining the optimal number of replications. Since the 1,000 replications require the shortest processing time, as shown in Table 1 and Table 2. The optimal number of replications is 1,000.

To determine the optimal degree of freedom of the spline regression model, Figure 7 shows the ROC curves of FP and TP rates obtained from 1,000 replications in two case simulation studies, as shown below:

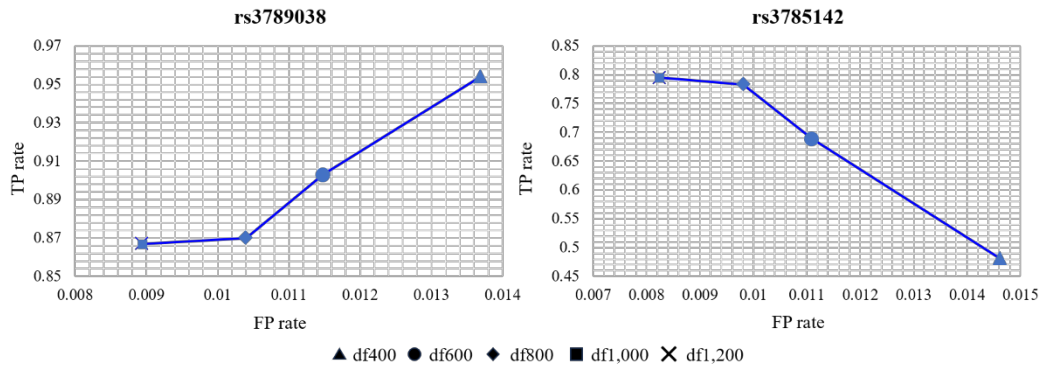

Figure 7 ROC curve for FP and TP rates from SKAT with B-spline obtained from 1,000 replications with disease SNP as rs3789038 (Burton et al.) and rs3785142 (right)

The point of interest on the ROC curve typically lies in the bottom-left region, where the false positive rate is minimized. Figure 7 shows the degrees of freedom of 1,000 (■) and 1,200 (×) provide the identical results: the lowest FP and reasonable TP rates. However, the degree of freedom of 1,000 still surpasses the degree of freedom of 1,200 in terms of processing time, as shown in Table 1 and Table 2. Thus, the optimal degree of freedom for the spline regression model is 1,000.

5. Application to Real Data

The real data used in this part consists of the genotype data for 13,479 SNPs on chromosome 16, including 2,005 Crohn's disease cases and 1,500 healthy controls obtained from the Wellcome Trust Case Control Consortium or WTCCC (Burton et al. 2007). The optimal spline regression model, determined to be at a degree of freedom of 1,000 through simulation studies, is applied to identify

gene regions associated with Crohn's disease. The model successfully detects several significant regions related to Crohn's disease presented in Table 3.

Table 3 Names, positions and achieved p-values of significant genes on chromosome 16 declared by SKAT with B-spline using a degree of freedom of 1,000.

Position	Gene Name	p-value
16p12.2	LOC646828	1.01×10^{-5}
	Intron Gene 89	1.01×10^{-5}
16p12.3	SMG1P5	2.76×10^{-6}
	Intron Gene 151	3.28×10^{-6}
	Intron Gene 174	6.94×10^{-6}
16q12.1	NOD2	1.22×10^{-7}
	CYLD	4.00×10^{-7}

Three specific locations of SNP sets or genes were identified. These locations are as follows: LOC646828 on 16p12.2, containing 2 SNPs; SMG1P5 on 16p12.3, also comprising 2 SNPs; NOD2 and CYLD on 16q12.1, including 12 and 8 SNPs, respectively. Additionally, the model indicates that intron regions—segments within a gene that are not translated into proteins—are associated with Crohn's disease.

The application result shows that the spline regression model with a degree of freedom of 1,000 has efficiency in identifying regions within chromosome 16 that are related to the disease. Especially the NOD2 on 16q12.1, which is the significant gene contributed to Crohn's disease, as suggested in previous studies from Roda et al. (2020) and Ashton et al. (2023).

6. Conclusion and Discussion

Optimizing model parameters remains a challenge due to limited time and computational resources. The results suggest that a degree of freedom of 1,000 is optimal for the spline regression model. It provides the lowest false positive (FP) rate, and a reasonable true positive (TP) rate compared to other degrees of freedom. Additionally, it requires less processing time than a degree of freedom of 1,200, which yields identical results. This finding aligns with the study by Sookkhee et al. (2021), which investigated the optimal parameters for spline regression in the study of the association between SNP sets and Crohn's disease. Furthermore, the degree of freedom reflects the number of free parameters, indicating the model's flexibility. A high degree of freedom may lead to overfitting and increased processing time, whereas a low degree of freedom could result in underfitting (James et al. 2021).

The research outcomes suggest that the number of replications has a minor effect on the model's efficiency, although the processing time varies significantly with different number of replications. Using a large number of replications ensures the most accurate results but requires a substantial amount of time. Conversely, fewer replications can produce comparable outcomes in less time, consistent with the studies by Mundfrom et al. (2011) and Koskan et al. (2023). Therefore, the optimal number of replications is 1,000, as it maintains low FP and reasonable TP rates close to the best results but with reduced runtime.

After obtaining the optimized smoothing parameter and number of replications, the spline regression model with the degree of freedom of 1,000 successfully identifies gene regions associated

with Crohn's disease, particularly NOD2, which is widely recognized as a significant factor in the development of Crohn's disease in established studies (Roda et al. 2020, Ashton et al. 2023).

The differences in the simulation study results are attributable to the Sequence Kernel Association Test (SKAT). Assigning different disease-causing SNPs leads to varied outcomes, demonstrating that the efficiency of SKAT depends on the specific disease-causing SNPs, as suggested by Kirdwichai and Baksh (2019) and Sookkhee et al. (2021).

For future work, many genetic variants are associated with various complex diseases that have not been identified. Testing these associations and constructing an alternative regression model, such as penalized spline regression, would be an interesting approach. While penalized spline regression helps balance smoothing and model fit, determining the optimal tuning parameter is also necessary. Finally, the findings of this research are expected to provide aspects of defining appropriate smoothing parameters for spline regression models and determining the optimal number of replications. This could help identify associations between SNP sets and complex diseases more efficiently by reducing the error rate and computational burden, especially processing time, during an analysis.

References

Ashton JJ, Seaby EG, Beattie RM, Ennis S. NOD2 in Crohn's disease-unfinished business. J Crohns Colitis. 2023; 17(3): 450-458.

Bates DM, Venables WN. An R-package for Regression Spline Functions and Classes version 4.2.1. [monograph on the Internet]. 2020 [cited 2023 Sep 18]. Available from: https://stat.ethz.ch/R-manual/R-devel/library/ splines/html/00 Index.html

Berger D. A Gentle Introduction To Resampling Techniques [monograph on the Internet]. Claremont Graduate University. 2011 [cited 2023 Nov 20]. Available from: https://www.academia.edu /66608980/A_ Gentle_Introduction_to_Resampling_Techniques.

Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145): 661-678.

James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. 2nd ed. New York: Springer; 2021.

Kamoljitprapa P, Baksh FM, De Gaetano A, Polsen O, Leelasilapasart P. Statistical study design for analyzing multiple Gene Loci correlation in DNA sequences. Mathematics. 2023; 11(23): 4710.

Kamoljitprapa P, Leelasilapasart P. Nonlinear models for Influenza patients for different age groups in Thailand. ICIEI2024: Proceedings of the 9th International Conference on Information and Education Innovations; 2024 Apr 12-14; Verbania, Italy: Association for Computing Machinery; 2024. pp. 109-112.

Kamoljitprapa P, Polsen O, Sookkhee S. Statistical analysis for genome data based on multiple SNPs using kernel machine based test. In: Proceedings of the 5th Research, Invention, and Innovation Congress (RI2C2024); 2024 Aug 8-9; Bangkok, Thailand. p. 262-266.

Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, et al. Are minor alleles more likely to be risk alleles? BMC Med Genomics. 2018; 11(1): 3.

Kirdwichai P. An efficient association test for high dimensional data, with application in genetic studies. In: Proceedings of the World Congress on Engineering (WCE2016), Vol II; 2016 Jun 29-Jul 1; London, UK. p. 618-622.

Kirdwichai P. Estimation and use of correlation in multiple hypothesis testing with high dimensional data. In: Proceedings of the 2nd International Conference on Mathematics and Statistics

(ICOMS2019); 2019; Prague, Czech Republic. New York: Association for Computing Machinery; 2019. p. 36-39.

Koskan O, Ergin M, Koknaroglu H. Determination of suitable sample size and number of simulations (resampling) for predicting dry matter intake of feedlot cattle. Int J Nat Eng Sci. 2023; 17: 27-36.

Lee S, Zhao Z. An R-package for SNP-Set (Sequence) Kernel Association Test version 2.2.5 [monograph on the Internet]. 2023 [cited 2023 Sep 27]. Available from: https://cran.r-project.org/web/packages/SKAT /index.html

Mundfrom D, Schaffer J, Kim M-J, Shaw D, Thongteeraparp A, Preecha P, et al. Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. J Mod Appl Stat Methods. 2011; 10: 19-28.

Pailoung C, Kamoljitprapa P, Sookkhee S. Optimization of Smoothing Parameters in Splines in GWAS Using a Replication Strategy. RI2C2024: Proceedings of the 5th Research, Invention, and Innovation Congress; 2024 Aug 8-9; Bangkok, Thailand. pp. 188-192.

Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Med Res Methodol. 2019; 19(1): 46.

Roda G, Chien Ng S, Kotze PG, Argollo M, Panaccione R, Spinelli A, et al. Crohn's disease. Nat Rev Dis Primers. 2020; 6(1): 22.

R Core Team. R: A Language and Environment for Statistical Computing [monograph on the Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022 [cited 2023 Nov 25]. Available from: https://www.R-project.org.

Sookkhee S, Baksh F M, Kirdwichai P. Efficiency of Single SNP analysis and Sequence Kernel Association Test in Genome-wide Association Analysis. IMECS2018: Proceeding of the 18th International MultiConference of Engineers and Computer Scientists; 2018 Mar 14-16; Hong Kong. pp. 308-313.

Sookkhee S, Kirdwichai P, Baksh F. The optimal parameters of spline regression for SNP-set analysis in genome-wide association study. Sci Technol Asia. 2021; 26(1): 39-52.

Sookkhee S, Kirdwichai P, Baksh F. The efficiency of SNP and SNP-set analysis in genome-wide association studies. Songklanakarin J Sci Technol. 2021; 43(1): 243-251.

Sukhumsirichart W. Polymorphisms. In: Yamin L, editor. Genetic Diversity and Disease Susceptibility [serial on the Internet]. InTech; 2018. Available from: http://doi.org/10.5772/intechopen.76728

Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1): 82-93.