An Efficient Association Test for High Dimensional Data, with Application in Genetic Studies

Pianpool Kirdwichai

***Abstract*—Developed in this paper is a statistically and computationally efficient method for association analysis of data with a sample size n that is small relative to the number p of explanatory variables. This so-called "curse of dimensionality" is a feature of many high-dimensional studies, such as human genomewide association studies (GWAS) concerned with elucidating the role of genes in biological pathways of complex diseases and other medical conditions. The proposed method is a novel approach to multiple testing that implements nonparametric spline regression models that inherently account for the correlation structure present in the data to identify association patterns between subsets of the explanatory variables and a response of interest. Additionally, a simple, computationally efficient algorithm for identifying significant regions of association is presented. Simulation results show that the spline regression approach is not only more powerful but also leads to substantial reduction in false positive findings compared with existing methods. The method is illustrated using data from the Wellcome Trust Case-Control Consortium (WTCCC) study of Crohn's disease.**

***Index Terms*—association study, genomewide, multiple testing.**

I. Introduction

UNARGUABLY, the most important statistical challenge in the analysis of high-dimensional data is to reliably identify true patterns of relationship in the data in a computationally efficient manner, yet simultaneously minimizing the number of spurious findings. This problem is compounded in studies with small sample sizes (n) relative to the number of observed explanatory variables; the so called "curse of dimensionality" [1]. Methodological publications in this area can broadly be placed into either multiple hypothesis testing (multiple comparison), dimension reduction or simulation-based methods, with some overlap between groups. The gold standard for controlling the error rate in testing multiple hypotheses is to compare each observed test statistic with an empirical distribution constructed from the observed data. In the exact test the distribution is based on each possible data permutation while in the Monte Carlo test, the distribution is based only on a randomly selected subset of permutations. The benefits of a permutation test are that no assumptions regarding the null distribution of the test statistic are required and the data's correlation structure is potentially inherently accounted for in the empirical distribution. However, while resampling is suitable for smaller studies, substantial computational effort is still required in high-dimensional studies, in spite of effort, see for example [6], to reduce this burden.

Bonferroni correction α/m, where m is the number of hypothesis tests conducted, is the most basic method for controlling the type I error α but it is well known to be conservative when the independence assumption between tests do not hold. Furthermore, due to the large number of tests in high-dimensional studies, there is no gain in efficiency from using traditional modifications such as Šidák correction [17] and Holm procedure [11]. The problem of identification of a less conservative threshold based on formal calculation of the effective number of independent tests M_{eff}, among all the tests conducted, have been considered [3], [13], [14], [8]. This number M_{eff} replaces m in the Bonferroni correction. Cheverud [3] proposed that the eigenvalues of the correlation matrix for the explanatory variables be used to estimate M_{eff}. However, Cheverud's method is still overly conservative when there is high correlation [13]. Nyholt [14] suggested a modification of the Cheverud's approach to improve the adjustment by excluding all variables in perfect correlation except one, but this was still found to be overly conservative. Li and Ji [13] proposed a method based on partitioning each eigenvalue into integral and fractional components. They argue that each integer represents identical tests and should be counted as one in M_{eff}. On the other hand, the fractional part represents a partially correlated test that should be counted as a number between 0 and 1. The authors present results on a small number of variables only and did not clarify how to perform the method in large datasets. Gao et. al. [8] proposed a principal components approach, which they call SimpleM, based on use of clusters or small subsets of the independent variables, in an attempt to filter out the correlation among tests. M_{eff} is derived from the number of principal components that explains a certain percentage of the variability in the data. In practice, this percentage is subjectively determined by the researcher.

Another technique for reducing the multiple testing burden aggregates the hypothesis tests within a joint test. Classic aggregation approaches include combining the p-values from single hypothesis tests [7] and assessing the smallest p-value within the set of m values [18]. In situations where the tests are not independent, Fisher's method tend to be anti-conservative while Tippett's method has well-controlled type I error rate but has been shown to have low power to identify small effect sizes, such as in genetic studies [4]. Modifications of Fisher's method proposed for genomewide studies, based on use of only a subset of the p-values in calculating the test statistic, include the threshold truncated product [23], the rank truncated product [5] and the adaptive

Manuscript received March 22, 2016; revised April 07, 2016. This research was funded by King Mongkut's University of Technology North Bangkok. Contract no. KMUTNB-GEN-59-37.

The author is Assistant Professor with the Faculty of Applied Science, KMUTNB, Thailand e-mail: pianpool.k@sci.kmutnb.ac.th.

rank truncated product [22] methods. The difference between the threshold and rank truncated methods is that the test statistic in the former is calculated using only those p-values smaller than some pre-specified threshold while in the latter it is calculated using a pre-specified quantity of the smallest p-values. For dependent single hypothesis tests, the overall significance level is obtained by a permutation method. The adaptive rank truncated method extends the rank truncated method by first calculating test statistics for a pre-specified range of possible truncation points in the ordered set of all p-values from the single hypothesis tests. These test statistics may then be compared with the relevant distributions and the p-value for the global hypothesis test is then the smallest among the set of p-values obtained.

A limitation of the above modifications of Fisher's method is the need to pre-specify truncation points or thresholds, the choice of which can be somewhat arbitrary and can affect the power to detect true associations in, for example, genomewide studies [2]. To avoid the threshold selection problem, Chen et. al. [2] proposed a sequential method for rejecting the global null hypothesis based on the cumulative product of the ordered p-values. More specifically, p-values from the single hypothesis tests are first ordered from smallest to largest and a sequence of test statistics are constructed. The first element of this sequence is the smallest p-value, the second element is the product of the first element and second smallest p-value, the third element of the sequence is the product of its second element and the third smallest p-value, and so on. The distributions of these test statistics are obtained using a permutation procedure and the thresholds for declaring significance are determined numerically. The authors present simulation results suggesting that the type I error rate will be close to the nominal value and that the power of the test procedure is comparable to the above modifications of Fisher's method. However, while the aggregation methods described herein attempt to capitalise on the correlation between explanatory variables, it is not always straight-forward to properly define the grouping structure. Additionally, these methods tend to be computationally burdensome with test statistics that have complex null distributions of which very little is currently known.

This paper proposes a novel spline regression method for interpreting results from single hypothesis tests that capitalises on the correlation between explanatory variables to reliably identify true associations between subsets of the explanatory variables and a response of interest and, at the same time, to reduce the number of false positive signals. The proposed methodology is premised on the fact that the majority of explanatory variables observed are not associated with the response and hence the collection of p-values from single hypothesis tests will mainly comprise of non-significant results, or noise, possibly interspersed with true signals of association. Then the challenge is to distinguish these rare signals from the noise. Framed in this context, the problem is similar to other application areas requiring signal cleaning such as microarray experiments where nonparametric regression is frequently used to remove systematic biases arising from the technology used to obtain the data, for example see [12] for details.

II. Method

The general nonparametric regression model with one predictor variable may be written as

$$u_i = b(x_i) + \varepsilon_i\,, \tag{1}$$

where u_i and x_i are the i^{th} response and predictor, respectively, $i = 1, \ldots, m$ and $b(\cdot)$ is an arbitrary function of x. In this paper u_i is the $-\log_{10}$ transformation of the p-value p_i from the association test of the i^{th} explanatory variable while x_i marks the position of this variable relative to the others in the observed set. As in linear regression, it is standard to assume that the ε_i's are independent identically distributed $N(0, \sigma^2)$. Unlike linear regression, the function $b(\cdot)$ is not specified in advance via a set of parameters and hence fitting the model involves estimating $b(\cdot)$ directly, rather than via parameter estimation. Most methods implicitly assume that $b(\cdot)$ is a smooth, continuous function and thus nonparametric regression may be viewed as nonlinear regression, but without explicitly stating the form of the function to be fitted.

In this paper, the function $b(x)$ is approximated by a spline and the resulting model is fitted using least squares. The cubic spline is a popular choice as it often has 2 continuous derivatives and therefore provide a reasonably smooth approximation to most non-linear functions [9]. The other reason why the cubic spline is popular is related to the number of used knots, or points at which the splines are joined. While increasing the number of knots tend to improve the fit of the data [9], fitting a spline with knots at every data point will result in over-parametrisation as the regression model will now consist of $p + 1 + n$ parameters and only n observations. A solution is however provided by standard penalised least squares. It turns out that a cubic spline is the solution to the problem of minimising the penalised sum of squares $\sum[y_i - f(x_i)]^2 + h\int_{x_{min}}^{x_{max}}[f''(u)]^2 du$ over all functions f that are twice continuously differentiable. This leads to the problem of selecting the number of knots being reduced to finding a value of the smoothing parameter h. Details of this approach to spline regression can be found in [9], for example.

The value of the smoothing parameter h is very important when fitting a nonparametric regression model as too large a value results in oversmoothing and potential loss in information while too small a value leads to insufficient accounting of the noise in the data. Unfortunately, the consensus is that the optimal smoothing parameter is difficult to find, is data specific, and existing methods [21], [10] are computationally expensive. Additionally, the guiding principle underlying commonly used cross validation methods is optimisation of the predictive power of the fitted regression curve. This is not the objective in this paper. Rather, of interest is identification of significant regions or contiguous sections of the fitted curve that are highly unlikely to confirm to any of the possible patterns under the global null hypothesis. This is achievable by finding the smoothing parameter that produces a good estimate of the noise. In this light, it is proposed to select a smoothing parameter satisfying the condition that the average squared residuals from the fitted model equals the difference estimate [19] of the variance σ^2 defined as

$$s_d^2 = \frac{1}{2m}\sum_{i=1}^{m-1}(u_{i+1} - u_i)^2\,. \tag{2}$$

The use of s_d^2 is justifiable on the basis that it is adjusted for the correlation structure in the explanatory variables. Indeed, s_d^2 is expected to be a slight underestimate of the true noise, which is appropriate for the stated objective of finding local structures in the data. Details of the theoretical development and evaluation of this method is provided in Kirdwichai [15].

A. Algorithm for identifying significant regions

An efficient algorithm that can perform the task of identifying significant regions is necessary due to the large dimension of the data and one was therefore developed using the R software [16]. The primary objective was to avoid use of conditional statements and loops as these are well-known to be computationally time consuming. Assuming that data points with values greater than some predefined threshold are coded to 1 and those below the threshold are coded to 0. This provides a sequence of ones and zeros with each significant region being represented by a sub-sequence of contiguous ones. An example of five regions is given below.

$$0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\,.$$

The objective of the algorithm is to identify the number of sub-sequences within this sequence. One approach would be to use conditional ”if” statements in R. Alternatively, it is noticed that all that is required is to be able to identify the borders of each sub-sequence of ones, which can be achieved by subtraction. The proposed algorithm first aligns two identical copies of the sequence as follows:

$$0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ \mathbf{0}$$

$$\mathbf{0}\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0$$

Notice that the first copy has a trailing zero added while the second copy has a leading zero added (shown in bold). This ensures that regions at the ends of the sequence are correctly identified.

Next, using the vector data structure in R, subtract the second copy from the first. In the example, this produces,

$$0\ 0\ 1\ -1\ 0\ 0\ 1\ 0\ -1\ 0\ 0\ 0\ 1\ -1\ 0\ 0\ 1\ 0\ -1\ 0\ 0\ 0\ 1\ 0\ 0\ -1\ 0\ 0$$

As can be seen in this sequence of differences, left borders of regions are now marked by a 1 while right borders are marked by a -1, and all other values are 0. Additionally, summing the absolute values of the elements of this sequence of differences and dividing by 2 will provide the number of regions.

III. Results

The proposed method is evaluated by simulating data under logistic regression models with correlated binary explanatory variables on which a metric is defined; an example is single nucleotide polymorphisms (SNPs) on a chromosome. Evaluations are conducted by simulating data assuming two of the explanatory variables, denoted X_1 and X_2, act independently on response so that each success is generated with probability

$$\pi_{x_1,x_2} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}\,, \tag{3}$$

where x_1 and x_2 are binary indicators and β_1, β_2 are the effect sizes as well as by simulating data under the interaction model

$$\pi_{x_1,x_2} = \frac{e^{\beta_0+\beta_{int} x_1 \times x_2}}{1+e^{\beta_0+\beta_{int} x_1 \times x_2}}\,. \tag{4}$$

Logistic regression models are next separately fitted for X_1 and X_2 as well as for each of approximately 14,000 other explanatory variables not associated with the response and the $-\log_{10}$ p-values obtained are then used as the responses in the nonparametric spline regression models fitted using the smoothing parameter selection procedure described in the Methods section. The estimated true positive (TP) rate is the proportion of studies for which at least one of the two regions containing X_1 and X_2 were detected while the false positive (FP) rate is the number of regions found to be significant but do not contain either X_1 or X_2. Shown in Figures 1-4 are ROC curves of the estimated TP against FP rates for varying significance threshold values. Simulation results for the proposed method are compared with Bonferroni correction for various false positive rates α,

$$\alpha'_{B,\alpha} = \frac{\alpha}{m}\,. \tag{5}$$

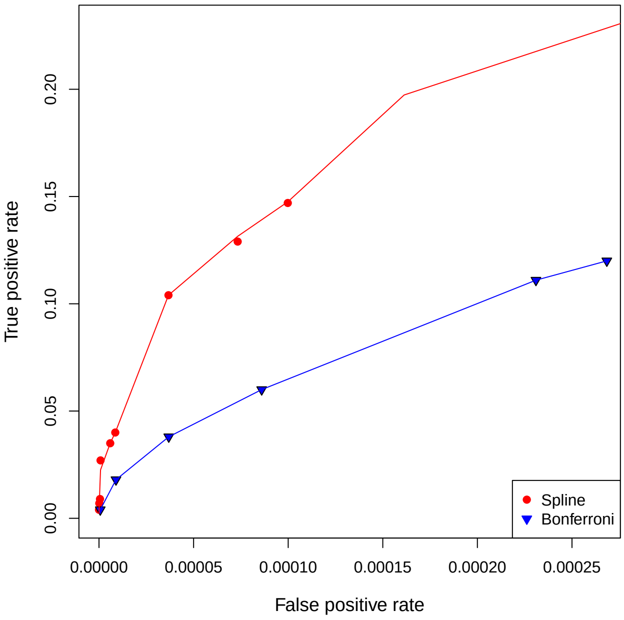

Fig. 1. ROC curves of false positive (FP) and true positive (TP) detection rates based on finding at least one of X_1, X_2 as significant by using spline regression and Bonferroni correction when correlation r^2 between X_1 and X_2 is **low**, in studies of size $n = 1000$ with true effect sizes $\boldsymbol{\beta_1 = \beta_2 = 0.2}$.

Shown in Figures 1 and 2 are rates when the effect size is 0.2 for X_1 and X_2 acting independently and with correlations $r^2 = 3.15 \times 10^{-9}$ and $r^2 = 0.4045$, respectively. Clearly the proposed spline regression approach outperforms the Bonferroni method with comparable FP rates but with TP rates that are much higher. Figures 3 and 4 presents the error rates when data is simulated under the interaction model with effect size $\beta_{int} = 0.2$. The TP rates are generally extremely low but, nevertheless, the proposed nonparametric regression method again clearly produces the better result.

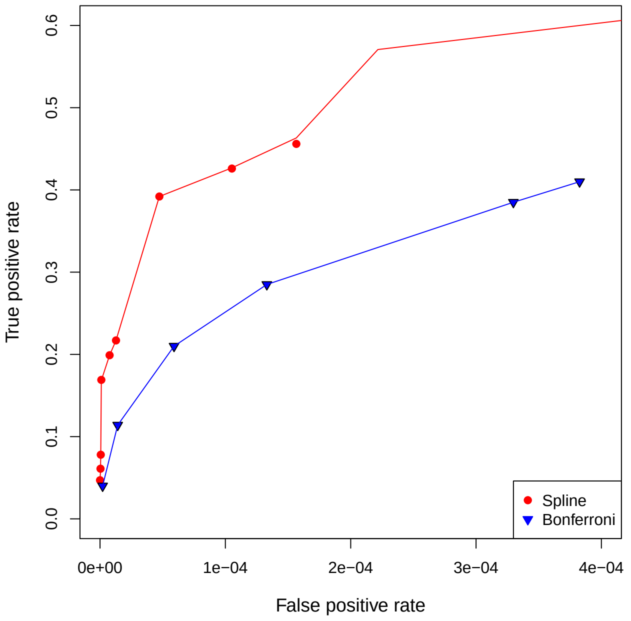

Fig. 2. ROC curves of false positive (FP) and true positive (TP) detection rates based on finding at least one of X_1, X_2 as significant by using spline regression and Bonferroni correction when correlation r^2 between X_1 and X_2 is **moderate**, in studies of size $n = 1000$ with true effect sizes $\boldsymbol{\beta_1 = \beta_2 = 0.2}$.

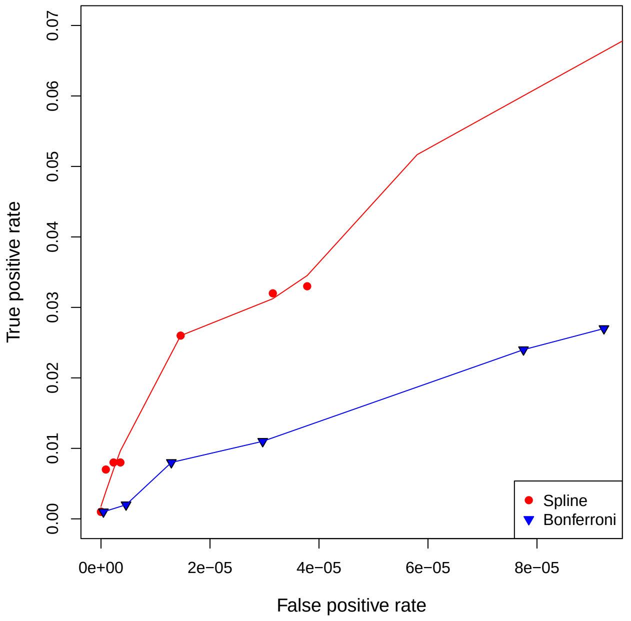

Fig. 4. ROC curves of false positive (FP) and true positive (TP) detection rates based on finding at least one of X_1, X_2 as significant by using spline regression and Bonferroni correction when correlation r^2 between X_1 and X_2 is **moderate**, in studies of size $n = 1000$ with true effect sizes effect size $\boldsymbol{\beta_{int} = 0.2}$.

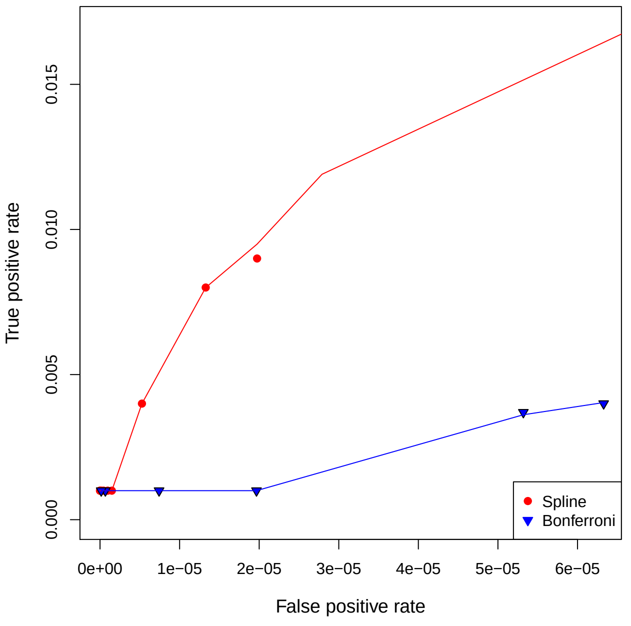

Fig. 3. ROC curves of false positive (FP) and true positive (TP) detection rates based on finding at least one of X_1, X_2 as significant by using spline regression and Bonferroni correction when correlation r^2 between X_1 and X_2 is **low**, in studies of size $n = 1000$ with true effect sizes effect size $\boldsymbol{\beta_{int} = 0.2}$.

The method was applied to data from the WTCCC study of Crohn's disease [20]. The dataset used has 14,292 SNPs on Chromosome 16 and comprise $2,005$ individuals with Crohn's disease and $3,004$ without the disease. The WTCCC study reports evidence for disease-gene association at SNP rs17221417 located on gene NOD2 and cites significant region of size $1,250,000$ basepairs to either side of rs17221417. The boundaries of this region, which was pointed out to experience high levels of recombination, were chosen to coincide with SNPs for which the $-\log_{10}$ p-values were deemed to have returned to expected levels under no genetic effect. Two significant regions were found when the proposed method was applied to this dataset. The larger of the two regions found is located at a distance of around 4.9×10^7 basepairs and contains SNP rs17221417 at $16,508$ basepairs from its left boundary. The second region detected consists of only three SNPs located within an intron at locus NR-002453.4. These findings concur with the results in [20].

IV. Conclusion

This paper tackled the current, unresolved problem of the statistical analysis of high dimensional data such as obtained in many genetic association studies of complex diseases. A novel method utilising concepts borrowed from nonparametric regression was developed and evaluated. The proposed method was shown to reduce the number of false positives found but still is more efficient than existing multiple comparison methods in detecting true positives. Unlike existing methods which attempt to increase power by aggregating the effects of tests, the approach in this paper was premised on treating true associations as rare signals to be identified among the noise generated by the large number of explanatory variables that are not associated with the response. The benefits of this approach over aggregation methods are that it is efficient and less computationally demanding, and the grouping is driven entirely by the data.

It is generally agreed that finding the optimal smoothing parameter for any given problem is difficult and, as often happens when using cross validation methods, is also computationally demanding and data specific. A novel selection method based on the use of an internal estimate of noise that inherently accounts for correlation in the data was proposed. The methodology is easily implemented via a simple search for the optimal bandwidth over a range of bandwidth values. The nonparametric regression approach is a promising alternative to existing methods in terms of improved efficiency and lower false positive rates. In contrast to most proposed methods in this area, the approach was seen to be more powerful than Bonferroni correction and with lower false positive findings. Perhaps the method's greatest appeal is as an exploratory analysis tool for association patterns that can subsequently be used to guide development of more in-depth, targeted studies. Finally, it is hoped that the method can be used to alleviate the problem of efficiently finding association in, for example, genomewide studies and can serve as a precursor to the further development of multiple hypothesis testing methods.

References

[1] R. E. Bellman, "Dynamic Programming," *Dover Books on Computer Science Series*, New York: Dover Publications, 2003.

[2] H-S. Chen and R. M. Pfeiffer and S. Zhang, "A powerful method for combining p-values in genomic studies," *Genetic Epidemiology*, vol. 37, no. 8, pp. 814-819, 2013.

[3] J. M. Cheverud, "A simple correction for multiple comparisons in interval mapping genome scans," *Heredity*, vol. 87, pp. 52-58, 2001.

[4] H. Dai and R. Charnigo and T. Srivastava and Z. Talebizadeh and S. Q. Ye, "Integrating p-values for genetic and genomic data analysis," *Biometrica and Biostatistics*, 3, e117. doi:10.4172/2155-6180.1000e117, 2012.

[5] F. Dudbridge and B. P. C. Koeleman, "Rank truncated product of p values, with application to genomewide association scans," *Genetic Epidemiology*, vol. 25, pp. 360-366, 2003.

[6] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genetic Epidemiology*, vol. 32, no. 3, pp. 227-234, 2008.

[7] R. A. Lee, *Statistical methods for research workers*, London: Oliver and Boyd, 1932.

[8] X. Gao and L. C. Becker and and D. M. Becker and J. D. Starmer and M. A. Province, "Avoiding the high bonferroni penalty in genome-wide association studies," *Genetic Epidemiology*, vol. 34, pp. 100-105, 2010.

[9] P. J. Green and B. W. Silverman, *Nonparametric regression and generalized linear models*, London: Chapman and Hall, 1994.

[10] W. Härdle, *Applied nonparametric regression*, Cambridge: Cambridge University Press, 1990.

[11] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65-70, 1979.

[12] M. Lee, *Analysis of microarray gene expression data*, Belgium: Springer-Verlag, 2004.

[13] J. Li and L. Ji, "Adjust multiple testing in multilocus analyses using the eigenvalues of a correlation matrix," *Heredity*, vol. 95, pp. 221-227, 2005.

[14] D. R. Nyholt, "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other," *American Journal of Human Genetics*, vol. 74, no. 4, pp. 765-769, 2004.

[15] P. Kirdwichai, "A nonparametric regression approach to the analysis of genomewide association studies," *PhD thesis, University of Reading*, 2014.

[16] R Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, Vienna, Austria, 2012.

[17] Z. Šidàk, "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, vol. 62, pp. 626-633, 1967.

[18] L. H. Tippett, *The methods of statistics*, London: Williams and Norgate, 1931.

[19] J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *Annals of Mathematical Statistics*, vol. 12, pp. 153-162, 1941.

[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 Publishing Group*, vol. 447, no. 7, pp. 661-678, 2007.

[21] M. P. Wand and M. C. Jones, *Kernel smoothing*, USA: Chapman & Hall/CRC, 1995.

[22] K. Yu and Q. Li and A. Bergen and R. M. Pfeiffer and P. Rosenberg and N. Caporaso and P. Kraft and N. Chatterjee, "Pathway analysis by adaptive combination of p-values," *Genetic Epidemiology*, vol. 33, pp. 700-709, 2009.

[23] D. V. Zaykin and L. A. Zhivotovsky and W. Czika and S. Shao and R. D. Wolfinger, "Combining p-values in large scale genomics experiments," *Pharmmceutical Statistics*, vol. 6, pp. 217-226, 2007.