The Journal of Applied Science **Vol. 18 No. 1: 20-30 [2019]**
วารสารวิทยาศาสตร์ประยุกต์ doi: 10.14416/j.appsci.2019.02.003
ISSN 1513-7805 Printed in Thailand ISSN 2586-9663 (Online)

Research Article

The analysis of genomewide SNP data using nonparametric and kernel machine regression

Pianpool Kirdwichai[1]*and Mohamed Fazil Baksh[2]

[1]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road, Bangsue, Bangkok 10800, Thailand
[2]Department of Mathematics and Statistics, School of Mathematical and Physical Sciences, University of Reading,
Mathematics Building, Whiteknights, Reading, RG6 6AX, UK
***E-mail**: pianpool.k@sci.kmutnb.ac.th
Received: 21/11/2018; Revised: 14/02/2019; Accepted: 26/02/2019

Abstract

This paper illustrates novel use of nonparametric regression in the challenging problem of reliably identifying true association patterns in high dimensional data without the cost, inherent in existing methods, of increasing the false positives. The proposed nonparametric association test (NPAT) treats p-values from multiple hypothesis tests as summaries of association that preserves the correlation in the data and capitalises on this correlation to increase power while minimising false discoveries, relative to existing methods. Distributional results are used to support estimation of the tuning parameter and significance thresholds for NPAT. The method is applied to the WTCCC study of Crohn's disease and results compared with a sequence kernel association test (SKAT) that conversely uses nonparametric regression techniques to group sets of explanatory variables, prior to association testing. Results show that NPAT is efficient, computationally tractable and produces findings comparable with Bonferroni correction while SKAT misses a strong association signal in the data.

Keywords: correlation structure, genomewide, multiple testing, nonparametric regression

Introduction

A key challenge in the statistical analysis of high dimensional datasets with a small number of individuals studied relative to the number of explanatory variables is the so called "curse of dimensionality" (Bellman, 2003). Although multiple hypothesis test procedures provide a solution and are computationally feasible with stable algorithms, the associated lack of power can mitigate against their usage in, for example, genomic studies where small effect sizes are the norm. This paper addresses this lack of power by novel use of nonparametric regression in multiple hypothesis tests of data from a genomewide association study (GWAS). GWAS is an approach which involves rapidly scanning a large number of SNP markers across the genome to find genetic variations associated with a particular disease or condition. A single nucleotide polymorphism, or SNP, is a variation at a single position in a DNA sequence among individuals.

Permutation tests control the error rate in multiple hypothesis test procedures by comparing each observed test statistic with an empirical distribution constructed from the observed data. In the "exact" test, this 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. 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 in still required in high-dimensional studies despite effort, (see for example Dudbridge & Gusnanto, 2008), 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 in multiple hypothesis tests 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 a GWAS, there is no gain in efficiency from using traditional modifications such as Šidák (1967) and Holm, Hochberg and Rom procedures (Olejnik et al., 1997). Risch & Merikangas (1996) recommended that Bonferroni correction should be modified to $\alpha = \alpha \times 10^{-6}$. This ad-hoc recommendation, which has been claimed (Johnson et al., 2010) to be the standard in GWAS, is based on an estimated 10^5 genes in the human genome (now accepted as an overestimation) and 10 independent tests for each gene.

Cheverud (2001) proposed that the eigenvalues of the correlation matrix for observed SNPs in the GWAS be used to estimate the effective number of independent tests M_{eff}. This number then replaces m in the Bonferroni correction. However, Cheverud's method is still overly conservative when there is high LD among SNPs (Li & Ji, 2005). Nyholt (2004) suggested a modification of the Cheverud's approach to improve the adjustment by excluding all SNPs in perfect LD except one, but this was still found to be overly conservative. Li & Ji (2005) 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 only present results on a small number of SNPs (< 15 for each gene) and did not clarify how to perform the method in large datasets. (Gao et al., 2010) proposed a principal components approach, which they call SimpleM, based on use of clusters or small subsets of SNPs, 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. Gao et al. (2010) evaluated their approach with a related approach (Moskvina & Schmidt, 2008) that estimates the number of independent SNPs using pairwise linkage disequilibrium (LD) directly. Using two real GWAS datasets, they found that the simpleM approach gave the best approximation (the smallest distance) to the thresholds obtained using permutation whereas the method by Moskvina gave more conservative estimates and Bonferroni gave the most conservative results. The authors further claim that the simpleM approach greatly reduced the computer time and resources required.

As in the above independent number of tests approach, multi-marker methods attempt to alleviate the multiple testing burden by pooling information within the observed dataset. A commonly employed technique is to aggregate the effects of markers within a known genomic region and then testing for joint association with the disease. Classic aggregation approaches include combining the p-values from single hypothesis tests (Fisher, 1932) and assessing the smallest p-value within the set of m values (Tippett, 1931). In situations where the statistical 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 multiple genes with small effects (Dai et al., 2012). 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 (Zaykin et al., 2007), the rank truncated product (Dudbridge & Koeleman, 2003) and the adaptive rank truncated product (Yu et al.,

Vol. 18 No. 1: 20-30 [2019]
doi: 10.14416/j.appsci.2019.02.003

2009) 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 SNP 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 SNP 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.

An obvious 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 for genomewide studies and can affect the power to detect true associations (Chen et al., 2013). To avoid the threshold selection problem, Chen et al. (2013) 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 SNP 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.

Alternatively, Wu et al. (2010) proposed grouping SNPs based on prior biological knowledge of their proximity to genes or haplotype blocks and evaluating the joint effects of each group. The authors argue that grouping in these so-called SNP sets permits the borrowing of information from correlated SNPs and increases the power to detect individual SNP effects. A logistic kernel-machine regression model is used to allow for modelling of epistatic and nonlinear effects. Proposed kernels include a measure of the number of alleles shared by a pair of individuals in the study and can be viewed as providing measures of genetic similarity between two individuals in the study. The authors propose use of the score statistic in testing association between SNP set and disease. The null distribution of this statistic is a complex mixture of χ^2 distributions which the authors approximate by a scaled χ^2 distribution. The scale parameter and degrees of freedom of this approximate distribution are estimated using the method of moments (see Wu et al., 2010, for details). Simulation results presented by the authors suggest that the type I error rate of their sequence kernel association test (SKAT) is protected and that SKAT has greater power than Bonferroni-corrected single SNP tests. This semiparametric method for modelling the joint effect of markers was proposed by Liu et al. (2007) while theory underpinning the approach utilised by the authors can be found in Hastie et al. (2009).

This paper illustrates novel use of nonparametric kernel regression for interpreting results from single SNP tests by capitalising on the location and correlation of SNPs to reliably identify true disease-gene associations and, at the same time, reduces the number of false positives. The proposed nonparametric association test (NPAT) avoids the aforementioned computational and other issues associated with existing methods and is premised on the fact that the majority of SNPs in a GWAS cannot be associated with the disease of interest and hence the collection of p-values from single SNP tests will mainly comprise of non-significant results, or noise, possibly interspersed with true signals of disease-gene 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, (see for example Lee, 2004, for details). Results of the method applied to The Wellcome Trust Case-Control (2007) study of Crohn's disease is compared with results from SKAT and multiple testing with Bonferroni correction.

Materials and methods

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

$$u_i = b(x_i) + \varepsilon_i, \qquad (1)$$

where u_i and x_i are the i^{th}response and predictor, respectively, $i = 1, \dots, m$ 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} SNP while x_i marks the position of this SNP. As in linear regression, it is standard to assume that ε_i ~IID(0, σ^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.

Nadaraya (1964) and Watson (1964) proposed to estimate $b(\cdot)$ as a locally weighted average of the responses u_i, $i = 1, \dots, m$ by using

$$\hat{b}(x) = \frac{\sum_{i=1}^{m} w(x,x_i) y_i}{\sum_{i=1}^{m} w(x,x_i)}, \qquad (2)$$

where the assigned weights $w(x, x_i) = K(\frac{x, x_i}{h})/h$ are determined by a kernel function K and the size of the weights are determined by the smoothing parameter h. Common kernels are usually functions of the distance between SNPs with the property that SNPs in close proximity to the i^{th} SNP contribute more to the fitted value $\hat{u}_i$ than SNPs further away. The normal kernel with weights

$$w(x, x_i) = \frac{e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}}{\sum_{i=1}^{m} e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}}, \qquad (3)$$

is used in this paper.

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. Existing methods (Wand & Johns, 1995; Härdle, 1990) for finding the optimal value for h are computationally expensive to implement in large datasets and additionally, the guiding principle underlying commonly used cross-validation methods is optimisation of the predictive power of the fitted regression curve, which is not the objective in this paper. Rather, of interest is identification of sections of the fitted curve that are highly unlikely to confirm to any of the possible patterns under the global null hypothesis of no association between disease and gene.

This is achievable by finding the smoothing parameter that produces a good estimate of the noise by allowing for the correlation in the data. 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 (von Neumann, 1941) of the variance σ^2 defined as

$$s_d^2 = \frac{1}{2m}\sum_{i=1}^{m-1}(u_{i+1}-u_i)^2. \tag{4}$$

This method is justifiable on the basis that

$$E(s_d^2) \approx \frac{1}{(\ln 10)^2}\left(1-\sum_{i=1}^{m-1} r_{i,i+1}^2\right), \tag{5}$$

and so s_d^2 is adjusted for the correlation of adjacent SNPs in the data. 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. Furthermore, there is precedence for use of a differencing approach (see for example Hart, 1991) in methodology for estimating the smoothing parameter in kernel regression problems.

In order to prove equation (5) above, first note that the distribution of the random variable U = $-\log_{10}(p)$ is exponential with rate parameter λ = ln(10) and furthermore that $Cor(U_i, U_{i'}), i \neq i'$ is approximately equal to the correlation $r_{i,i'}^2$ between the j^{th} and $(i')^{\text{th}}$ SNPs. (This result in proved in a forthcoming manuscript). Assuming $b(x_{i+1}) \approx b(x_i)$ gives s_d^2 as approximately $\frac{1}{2m}\sum_{i=1}^{m-1}(\varepsilon_{i+1}-\varepsilon_i)^2$ which, by expanding the square and summing, becomes $(MSR) - \frac{1}{m}\sum_{i=1}^{m-1}\varepsilon_{i+1}\varepsilon_i$, where $MSR = \frac{1}{m}\sum_{i=1}^{m-1}\varepsilon_i^2$. Taking expectations

$$E(s_d^2) \approx E(MSR) - \frac{1}{m}\sum_{i=1}^{m} E(\varepsilon_i\varepsilon_{i+1}). \tag{6}$$

Equation (5) follows by noting that from the distribution of U, $E(MSR) = \frac{1}{\ln(10)^2}$, and by definition of covariance, the assumption $E(\varepsilon_i) = 0$ and using the property that covariances are preserved under linear transformations, $E(\varepsilon_i\varepsilon_{i+1}) \approx \frac{r_{i,i+1}^2}{\ln(10)^2}$

Evidence for disease-gene association is likely to be provided by SNPs that are correlated with the disease causing gene(s) and pooling this evidence should increase the chance of finding true associations while spurious associations are less likely to likewise supported. In other words, utilising the correlation structure in the data is likely to improve both true positive (TP) and false positive (FP) rates of the test procedure. The kernel in equation (2) assumes that that physical proximity is a good proxy for correlation. This is a reasonable assumption for genomic data. In addition, NPAT identifies regions of the genome that show significance with the disease which is relevant as there is no guarantee that disease pre-disposing SNPs will be studied in the GWAS. Under the proposed method, significant regions are genomic regions for which the plot of the fitted nonparametric regression curve against basepair position of SNPs lie above some significance threshold.

The fitted value $\widehat{U}_i = \sum_{j=1}^{m} w(x_j, x_i)U_j$ of the nonparametric model is a linear combination of exponential random variables with different rate parameters $w(x_j, x_i)\ln(10)$; $j = 1, \ldots, m$ and so follows a hypoexponential or generalized Erlang distribution, ignoring the dependency structure. It follows that a significance threshold that is adjusted for multiple testing and assumes independence of tests is $q_{1-\frac{\alpha}{m}}$ given by

$$\int_{q_{1-\frac{\alpha}{m}}}^{\infty} f(u)du = 1 - \frac{\alpha}{m},$$

where $f(u)$ is the density of the hypoexponential distribution with rates $w(x_j, x_i)\ln(10)$; $j = 1, \dots, m$ and α is the global type I error of the multiple hypothesis test procedure. In situations where evaluation of this threshold is time consuming, the following normal approximation may be used. As the expected value of $\widehat{U}_i$ is clearly 1/ln10 while the variance is given by

$$Var(\widehat{U}_i) \approx \frac{1}{(\ln 10)^2}\left(\sum_{j=1}^{m} w(x_j, x_i)^2 + 2\sum_{j<j'} w(x_j, x_i) w(x'_j, x_i)\, r_{jj'}^2\right), \quad (7)$$

it follows that an approximate significance threshold is provided by

$$\frac{1}{\ln 10}\left(1 + q_{1-\frac{\alpha}{m}}\sqrt{\left(\sum_{j=1}^{m} w(x_j, x_i)^2 + 2\sum_{j<j'} w(x_j, x_i) w(x'_j, x_i)\, r_{jj'}^2\right)}\right), \quad (8)$$

where $q_{1-\frac{\alpha}{m}}$ is the $1 - \frac{\alpha}{m}$ quantile of the standard normal distribution.

Results

Simulation study

In a simulation evaluation of the proposed approach, each of the 1,504 genotypes in the UK Blood Services dataset is first used to construct two haplotypes, to give a total of 3,008 haplotypes. The constructed haplotypes are randomly paired to simulate genotypes of 13,479 SNPs each, and having LD pattern consistent with that of the UK Blood Services Cohort. Next a disease pre-disposing SNP, rs3789038 located at position 4486587bp on Chromosome 16, is selected and used to simulate a total of 3000 diseased individuals and 3000 unaffected individuals based on a log-odds ratio for disease of 0.2. A plot of the LD pattern in the region containing SNP rs3789038 is produced in Figure 1. This data, comprising 6000 randomly generated genotypes of 13,479 SNPs each, is then analysed using the proposed NPAT and SKAT with 5% significance level. Whether a region containing the disease SNP is correctly identified as associated with the disease as well as the number of false positive regions is recorded for both tests. Simulation results of the true positive (TP) and the false positive (FP) detection rates based on 1,500 replications of the above is shown in Table 1. As the results show, the FP rate of NPAT is approximately 10% that of SKAT whereas the TP rates are comparable.

Table 1. The TP and FP rates of NPAT and SKAT with disease SNP rs3789038 and effect size 0.2

Method	TP rate	FP rate
NPAT	0.83	0.00087
SKAT (default)	0.85	0.00939

WTCCC study

The proposed NPAT is illustrated using WTCCC (The Wellcome Trust Case-Control, 2007) genotype data at 13,479 SNPs on Chromosome 16 for 2005 individuals with Crohn's disease and 1,504 controls from UK Blood Services. All analyses were done in R (R Core Team, 2016). The WTCCC study reports evidence for disease-gene association at SNP rs17221417 located on gene NOD2 and cites the significant region 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.

A manhattan plot of the $-\log_{10}$ p-values for single SNP tests assuming an additive genetic model is provided in Figure 2. Application of NPAT next involves smoothing this plot using the method described in the above section with predictors the SNPs' basepair positions scaled to lie within the interval [0,1]. Using equation (4), the value of $s_d^2 = 0.2873933$ was calculated and the value of the smoothing parameter h = 1.26635 × 10^{-7} producing mean square residual $\frac{1}{m}\sum_{i=1}^{m}(u_i - \hat{u}_i)^2$ equal to s_d^2 was found using a simple search algorithm. The fitted curve (solid line) and threshold values (dotted line) provided by equation (8) with type I error α = 0.05 are superimposed on the manhattan plot. A group of SNPs will be declared significant if the fitted curve exceeds the threshold.

Grouping the WTCCC SNP data by genes and intragene regions produces a total of 914 groups. For purposes of comparison, these groups were also analysed by applying SKAT with its default linear weighted kernel and the default parameter for the correlation structure kernels. The SKAT p-values (squares) and its Bonferroni threshold (dot-dashed line) based on 914 tests are also plotted in Figure 2, as is a Bonferroni threshold (dashed line) based on the 13,479 single SNP tests with a global type I error α = 0.05.

As is evident in Figure 2, NPAT finds two significant regions, the larger of which is 110,000 basepairs wide on gene NOD2, located at around 4.93 × 10^7 basepairs and contains a cluster of 24 SNPs (rs16948715-rs11076540) that includes SNP rs17221417 at 50,000 basepairs from rs16948715. The second significant region found by NPAT is a cluster of two SNPS (rs4471699, rs11644392) located in the intron (non-coding) region of locus NR_002453.5 and has not been reported in the WTCCC (The Wellcome Trust Case-Control Consortium, 2007) study. SKAT also finds this second region and a cluster of three SNPs (rs16957197, rs13312720, rs868213) in intron regions around 6.71 × 10^7 basepairs but fails to detect the NOD2 gene (p-value = 0.096). All regions declared as significant by NPAT and SKAT were also found to be significant using Bonferroni multiple testing correction.

Discussion

NPAT isolates regions of the genome that present true signals of disease-gene association by enhancing these signals and simultaneously suppressing the noise generated by the large number of genetic markers that are not associated with the disease. Evidence for this is provided by a simulation evaluation and the results of the analysis of the WTCCC data. Note that there is biological support for a link between NOD2 mutations and Crohn's disease (Philpott et al., 2014; Nabatov, 2015). Unlike existing multi-marker methods which attempt to increase power by aggregating the effects of markers within predetermined genomic regions, NPAT is shown to use the information contained in the data to determine the pooling of the evidence and the form of the fitted curve. Furthermore unlike SKAT, NPAT is less computationally demanding, the association patterns do not depend on pre-selected groupings of SNPs and the actual effect of causal SNPs or SNPs linked to the causal SNPs on disease are readily available.

This latter point is invaluable in quantifying the disease-gene relationship and in the design of follow-up and replication studies.

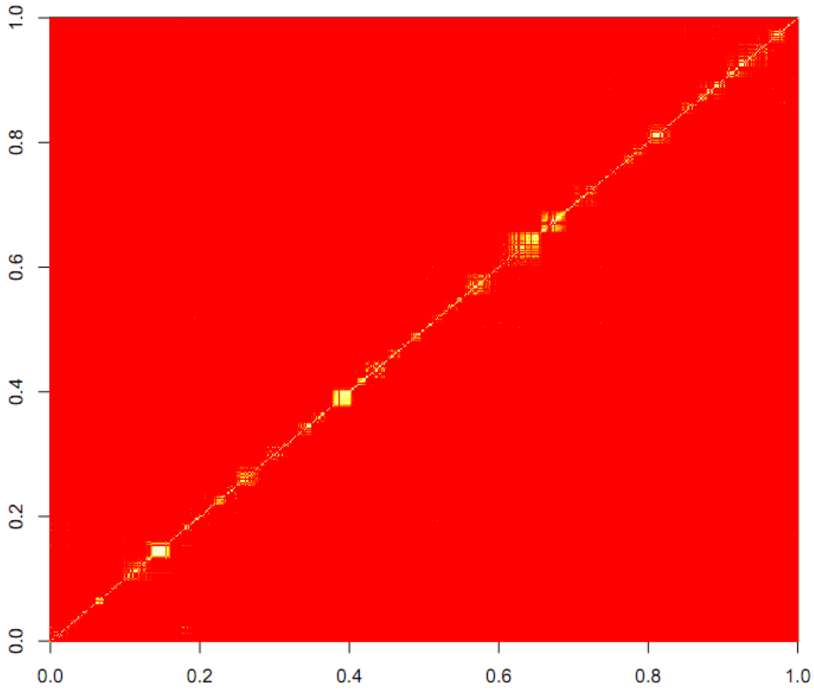

Figure 1. Heat spectrum of pairwise values r^2 for 1,000 SNPs within a region of Chromosome 16 containing SNP rs3789038. The observed pattern of small rectangular blocks are of contiguous SNPs in strong LD.

The non-parametric regression model in this paper was fitted using the Nadaraya-Watson estimator which has been criticised for its tendency to produce a nonlinear relationship when the true relationship is linear and for the property that the fitted value for data points at the boundaries of the dataset are based largely on observations on one side only. In contrast, local linear estimators preserves linearity and has much better properties at the boundary. However, the Nadaraya-Watson estimator is simpler to calculate and, for GWAS data, there is no reasonable expectation that the relationship is linear. Additionally, this estimator is preferable as it is not influenced by outliers as much as local linear estimators. A simulation comparison of methods based on Nadaraya-Watson and local linear estimators for disease SNPS at the edges of the chromosome would nevertheless be interesting. Also of interest is exploring the threshold error rates with the view of developing more computationally efficient improvements that do not require the normality and independence assumptions. Furthermore, it is generally agreed that finding the optimal tuning parameter for any given problem is difficult and, as often happens when using cross validation methods, is also computationally demanding and data specific. Additionally, cross validation methods have been shown to perform poorly when the data are correlated (see Altman, 1990, for example). In contrast, NPAT is easily implemented via a simple search for the tuning parameter over a range of values but further investigations of the effectiveness of the differencing approach is warranted.

The proposed nonparametric regression approach is a promising alternative to existing methods in terms of improved efficiency and lower false positive rates. Extensions of the method to non-genetic explanatory variables is straightforward as these can simply be included in the single SNP models. Finally, it is hoped that the method can serve as a precursor to the further development of multiple hypothesis testing methods in realm of high-dimensional data analysis.

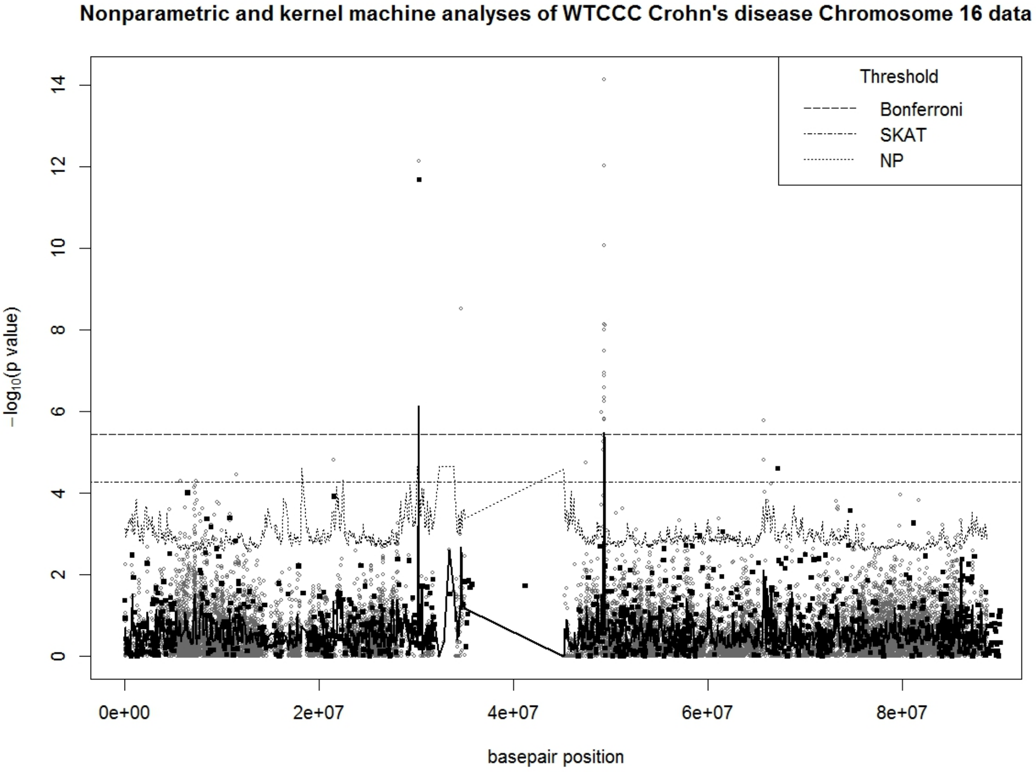

Figure 2. Manhattan plots of p-values from single SNP tests and SKAT p-values (squares) against their median basepair position, along with the fitted nonparametric regression curve (solid line) obtained in the analysis of WTCCC study on Crohn’s disease

Acknowledgements

Funding was provided by Faculty of Applied Science, King Mongkut's University of Technology North Bangkok. This support is gratefully acknowledged. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association, 85*(411), 749-759.

Bellman, R. E. (2003). *Dynamic Programming (Dover Books on Computer Science Series).* New York: Dover Publications.

Chen, H., Pfeiffer, R. M. & Zhang, S. (2013). A powerful method for combining p-values in genomic studies. *Genetic Epidemiology, 37*(8), 814-819. doi: 10.1002/gepi.21755

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity, 87*, 52-58. doi:10.1046/j.1365-2540.2001.00901

Dai, H., Charnigo, R., Srivastava, T., Talebizadeh, Z. & Ye, S. Q. (2012). Integrating p-values for genetic and genomic data analysis. *Biometrica and Biostatistics, 3*(7). doi: 10.4172/2155-6180.1000e117

Dudbridge, F. & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology, 32*(3), 227-234. doi: 10.1002/gepi.20297

Dudbridge, F. & Koeleman, B. P. C. (2003). Rank truncated product of p values, with application to genomewide association scans. *Genetic Epidemiology, 25*(4), 360-366. doi:10.1002/gepi.10264

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. (2010). Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology, 34*(1), 100-105. doi: 10.1002/gepi.20430

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, 53*(1), 173-187.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A. et al. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BioMed Central Genomics, 11*, 6 pp. doi: 10.1186/1471-2164-11-724

Lee, M. (2004). *Analysis of Microarray Gene Expression Data.* Belgium: Springer-Verlag.

Li, J. & Ji, L. (2005). Adjust multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity, 95*, 221-227. doi: 10.1038/sj.hdy.6800717

Liu, D., Lin, X. & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least squares kernel machines and linear mixed models. *Biometrics*, *63*(4), 1079-1088. doi: 10.1111/j.1541-0420.2007.00799

Moskvina, V. & Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology, 32*(6), 567-573. doi: 10.1002/gepi.20331

Nabatov, A. A. (2015). The vesicle-associated function of nod2 as a link between crohns disease and mycobacterial infection. *Gut Pathogens, 7*(1). doi: 10.1186/s13099-015-0049-1

Nadaraya, E. (1964). On estimation regression. *Theory of Probability and Its Applications*. *9*(1), 141-142. https://doi.org/10.1137/1109020

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics, 7*(4), 765-769. doi: 10.1086/383251

Olejnik, S., Li, J., Supattathum, S. & Huberty, C. J. (1997). Multiple testing and statistical power with modified bonferroni procedures. *Journal of Educational and Behavioral Statistics*, *22*(4), 389-406.

Philpott, D. J., Sorbara, M. T., Robertson, S. J., Croitoru, K. & Girardin, S. E. (2014). Nod proteins: regulators of inflammation in health and disease. *Nature Reviews Immunology*, *14*(1), 9-23. doi:10.1038/nri3565

R Core Team. (2016). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human disease. *Science, 273*(5281), 1516-1517.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariance normal distributions. *Journal of the American Statistical Association, 62*(318), 626-633. doi: 10.1080/01621459.1967.10482935

The Wellcome Trust Case-Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common disease and 3,000 shared controls. *Nature, 447*(7), 661-678. doi: 10.1038/nature05911

Tippett, L. H. (1931). *The Methods of Statistics.* London: Williams and Norgate.

von Neumann, J. (1941). Distribution of the ration of the mean square successive difference to the variance. *Annuals of Mathematical Statistics, 12*(4), 153-162. doi: 10.1214/aoms/1177731677

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman & Hall/CRC.

Watson, G. S. (1964), Smooth regression analysis. Sankhyā: *The Indian Journal of Statistics, Series A, 26*(4), 359-372.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. et al. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, *86*(6), 929-942. doi: 10.1016/j.ajhg.2010.05.002

Yu, K., Li, Q., Bergen, A., Pfeiffer, R. M., Rosenberg, P., Caporaso, N. et al. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology, 33*(8), 700-709. doi: 10.1002/gepi.20422

Zaykin, D. V., Zhivotovsky, L. A. Czika, W., Shao, S. & Wolfinger, R. D. (2007). Combining p-values in large scale genomics experiments. *Pharmaceutical Statistics, 6*(3), 217-226. doi: 10.1002/pst.304