Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 217-225

A Multivariate Normal Block Versus a Principal Components Approach: Competing Strategies for Multiple Testing in a Genome-wide Case-Control Association Framework

Arunabha Majumdar and Saurabh Ghosh
Indian Statistical Institute, Kolkata

Received 08 August 2013; Revised 05 April 2014; Accepted 19 April 2014

SUMMARY

The Genome-wide association studies have been partially successful in identifying novel variants involved in complex disorders. However, correcting for multiple testing in such studies becomes inevitable to maintain the appropriate overall false positive error rate. In this article, we consider a block wise strategy *MVNblock* of multiple testing correction based on an asymptotic multivariate normal framework for performing tests of association at correlated SNPs in a case-control study design. We investigate few of its important theoretical properties and using extensive simulations, compare its performance with a principal components analysis (PCA) based approach *simpleM*. We find that *MVNblock* behaves less conservatively than *simpleM* with respect to controlling for *FWER*. Moreover, *MVNblock* consistently produces a lower estimate of the effective number of independent SNPs compared to *simpleM*, and hence is expected to produce higher power compared to *simpleM*.

Keywords: Genome-wide association analyses, Family-wise error rate, Linkage disequilibrium.

1. INTRODUCTION

Genome-wide case-control association studies (GWAS) have provided an ideal platform for identifying novel variants involved in the pathogenesis of common genetic disorders. Such disorders are complex in nature and are controlled by multiple genes, each having moderate to small effect sizes. Since the genome-wide approach involves an unbiased screening of single nucleotide polymorphisms (SNPs) without any biological prior, one can obtain novel association findings in genes that would not have been considered in candidate gene studies. However, given the large number of SNPs that are screened in genome-wide studies, association analyses without correcting for multiple testing would result in inflated rates of false positives. Moreover, since the SNPs are in linkage disequibrium, development of an efficient multiple testing correction for correlated association tests remains a major statistical challenge.

Various methods have been proposed for controlling the probability of at least one false positive [known as the family wise error rate (*FWER*)]. Sidak's method provides an exact correction for controlling *FWER* when the tests for association are carried out at uncorrelated SNPs while Bonferroni's correction is the most standard approach when testing for correlated

Corresponding author: Saurabh Ghosh
E-mail address: saurabh@isical.ac.in

SNPs. However, both the approaches become almost equivalent to each other when the number of SNPs is very large as encountered in genome-wide association studies and tend to be over-conservative, that is, result in extremely low false positive rates and hence, lead to substantial loss in power. In the last decade, different methods controlling for *FWER* in genome-wide association studies have been explored. While permutation testing (Westfall and Young 1993) is a popular method for multiple testing correction due of its adaptability to arbitrary correlation structures among the SNPs, the computational load increases rapidly with the large number of SNPs at the genome-wide level. In order to circumvent this problem, two other prudent strategies have been investigated: one based on an asymptotic multivariate normal distribution framework for the test statistics at the correlated SNPs (Lin 2005, Seaman and Muller-Myhsok 2005, Conneely and Boehnke 2007, Han *et al*. 2009), and the other based on an estimation of the effective number of independent SNPs using a principal components analysis at the correlated SNPs (Cheverud 2001, Nyholt 2004, Li and Ji 2005, Gao *et al*. 2008).

If the sample size is sufficiently large, most association test statistics at correlated SNPs would asymptotically follow a multivariate normal distribution and hence, the adjusted p-values can be suitably computed to keep the desired control over *FWER*. In order to estimate the multivariate normal probabilities, Lin (2005) and Seaman and Muller-Myhsok (2005) used a simulation-based approach while Conneely and Boehnke (2007) employed numerical integration. Han *et al.* (2009) developed the method *SLIDE* based on a sliding window approach for locally inter correlated SNPs. On the other hand, Cheverud (2001), Nyholt (2004), Li and Ji (2005), and Gao *et al*. (2008) have investigated the strategy of using an estimate of the effective number of independent tests in the denominator of the classical Bonferroni correction instead of the total number of SNPs. Among these approaches, the method *simpleM* proposed by Gao *et al*. (2008), that estimates the effective number of independent SNPs based on the eigenvalues of the correlation matrix of composite linkage disequilibrium (CLD) structure among the correlated SNPs, is considered to be an efficient strategy at the genome-wide level. It has been shown that *simpleM* performs better than *SLIDE* for imputed SNPs at the genome wide level (Gao 2011).

In this article, we consider a block wise strategy *MVNblock* to adjust for multiple testing while testing association at correlated SNPs in a case-control set-up using an asymptotic multivariate normal framework. We investigate some of its important theoretical properties and also compare its performance with that of *simpleM* using extensive simulations. We find that *MVNblock* is not only less conservative with respect to controlling for *FWER,* but also consistently yields a lower estimate of the effective number of independent SNPs than *simpleM* subject to controlling *FWER* at the desired level. We also evaluate the relative performances of the two methods using real data on cardiovascular disease.

2. DATA DESCRIPTION

Consider a case-control study design comprising $2n$ unrelated individuals with equal number of cases and controls. Suppose that genotype data are available at L SNPs (not necessarily uncorrelated) for all $2n$ individuals. The alleles at the l^{th} locus are A_l and a_l, $l = 1, 2, ..., L$. We denote the genotype data of i^{th} case by $X_i = \{X_{i1}, X_{i2}, ..., X_{il}, ..., X_{iL}\}$, $i = 1, 2, ..., n$, where, X_{il} denotes the number of A_l alleles at the l^{th} locus, and hence, assumes value 0, 1 or 2. Similarly, the genotype data of i^{th} control individual is denoted by $Y_i = \{Y_{i1}, Y_{i2}, ..., Y_{il}, ..., Y_{iL}\}$, for $i = 1, 2, ..., n$. Suppose X_{il1} and X_{il2} are indicator random variables, each assuming values 1 or 0, according as the allele A_l is present or not at the l^{th} locus on the two chromosomes, respectively, of the i^{th} case individual, $i = 1, ..., n$, $l = 1, ..., L$. Similarly, Y_{il1} and Y_{il2}, $i = 1, ..., n$, $l = 1, ..., L$, are defined for the control individuals. Thus, $X_{il} = X_{il1} + X_{il2}$, $Y_{il} = Y_{il1} + Y_{il2}$, $i = 1, ..., n$, $l = 1, ..., L$.

3. MVNblock

Suppose the allele frequency at the l^{th} locus is p_{1l} among cases and p_{2l} among controls, $l = 1, 2, ..., L$. Each of the SNPs is assumed to be in Hardy-Weinberg Equilibrium. Thus, X_{il} and Y_{il} are distributed, respectively, as Binomial$(2, p_{1l})$ and Binomial$(2, p_{2l})$, $l = 1, 2, ..., L$, $i = 1, 2, ..., n$. We also assume that the linkage disequilibrium (LD) structure between SNPs within cases is identical to that within controls and denote the coefficient of linkage disequilibrium between the k^{th} SNP and the l^{th} SNP in the whole population by δ_{kl}, $k, l \in \{1, 2, ..., L\}$. Hence, $Cov(X_{ik}, X_{il}) = Cov(X_{ik1} + X_{ik2}, X_{il1} + X_{il2}) = 2\delta_{kl} = Cov(Y_{ik}, Y_{il})$, $k, l \in \{1, 2, ..., L\}$; $i = 1, 2, ..., n$. We shall discuss later a simple

modification of *MVNblock* to incorporate separate LD structures between the SNPs among cases and controls.

In order to test for possible association between each of the L SNPs and the disease phenotype, the null hypothesis of no difference in allele frequencies between the cases and controls is considered at each SNP. Thus, we test $H_{0l} : p_{1l} = p_{2l}$ versus $H_{1l} : p_{1l} \neq p_{2l}$, $l = 1, 2, ..., L$. Suppose that the common value of p_{1l} and p_{2l} under H_{0l} be denoted by p_l (the overall allele frequency at the l_{th} SNP in the whole population), $l = 1, ..., L$.

Suppose H_0 denotes the intersection of the L null hypotheses, *i.e.*, $H_0 = \bigcap_{l=1}^{L} H_{0l}$. The difference in the estimated allele frequency of A_l between cases and control at the l^{th} SNP is given by: $V_{ln} = \frac{1}{2n}\sum_{i=1}^{n} X_{il} - \frac{1}{2n}\sum_{i=1}^{n} Y_{il}$, $l = 1, ..., L$. The variance-ovariance matrix of $V_{1n}, ..., V_{Ln}$ can be easily derived and is given by:

$$Var(V_{ln}) = \frac{p_{1l}(1-p_{1l})}{2n} + \frac{p_{2l}(1-p_{2l})}{2n},$$

$$Cov(V_{kn}, V_{ln}) = \frac{\delta_{kl}}{n}; \; k,l \in \{1, 2, ..., L\}$$

Thus, under H_{0l}, $s.d.(V_{ln}) = \sqrt{\frac{p_l(1-p_l)}{n}}$ and S_{ln}, the sample standard deviation of V_{ln}, based on the combined set of cases and controls, is a consistent estimator of $s.d.(V_{ln})$. In order to test H_{0l} versus H_{1l}, the test statistic $T_{ln} = \frac{V_{ln}}{S_{ln}}$, is considered for $l = 1, 2, ..., L$. Using the multivariate central limit theorem, one can show that for large n, under H_0,

$$\sqrt{n} \times (T_{1n}, T_{2n}, ..., T_{Ln})' \sim N_L(\underline{0}, \delta^*) \qquad (1)$$

where, δ^* is a $L \times L$ matrix representing the correlation structure of the L SNPs with its $(k, l)^{th}$ element given by: $\delta_{kl^*} = \frac{\delta_{kl}}{\sqrt{p_k(1-p_k)\,p_l(1-p_l)}}$, for $k, l \in \{1, 2, ..., L\}$.

In order to test $H_{01}, ..., H_{0L}$ simultaneously, we consider a uniform critical region of the form $\{|T_{ln}| > C\}$, $l = 1, ..., L$, where, C is an unknown positive real number that needs to be determined subject to controlling the *FWER*, P_{H_0} (at least one false positive), at the desired level. We note that since the genome-wide association approach involves an unbiased scan of SNPs without any biological prior, it is a natural choice to consider the same threshold for the test statistic at each SNP. Thus, the critical region for testing $H_{01}, H_{02}, ..., H_{0L}$ simultaneously is $\{|T_{1n}| > C\}$, $\{|T_{2n}| > C\}$, ..., $\{|T_{Ln}| > C\}$, such that, C satisfies the condition $P_{H_0}\left(\bigcup_{l=1}^{L}\{|T_{ln}| > C\}\right) \leq \alpha$, where, the *FWER* is to be controlled at the level α. An useful expression of *FWER* corresponding to this critical region can be obtained as follows:

$$P_{H_0}\left(\bigcup_{l=1}^{L}\{|T_{ln}| > C\}\right) = 1 - P_{H_0}\left(\bigcap_{l=1}^{L}\{|T_{ln}| \leq C\}\right)$$

$$= 1 - P_{H_0}\left(\max_{1\leq l\leq L}|T_{ln}| \leq C\right)$$

Note that, $FWER \leq \alpha \Rightarrow P_{H_0}(\max_{1\leq l\leq L}|T_{ln}| \leq C) \geq 1 - \alpha$. In fact, since $\max_{1\leq l\leq L}|T_{ln}|$ is a continuous random variable, C can be chosen to attain the *FWER* exactly at the level α. Hence, the critical region for testing $H_{01}, ..., H_{0L}$ simultaneously, with the *FWER* controlled at the level α, is $\{(|T_{ln}| > C), l = 1, ..., L\}$, where, C is the $(1 - \alpha)$% quantile of the one dimensional random variable $\max_{1\leq l\leq L}|T_{ln}|$.

4. THEORETICAL PROPERTIES OF MVNblock

We shall state and prove two important properties of this method. The proofs of the properties are provided in the Appendix.

Property 1: The method controls the *FWER* strongly, that is, it controls the *FWER* under any configuration of the true and false null hypotheses.

Property 2: The power yielded by the method cannot be less than that produced by Sidak's correction.

5. IMPLEMENTATION OF MVNblock

Next, we discuss the implementation of the multiple testing method. We estimate the correlation structure (δ^*) between pairs of SNPs from the data. Since information on haplotype phase is likely to be unknown, we execute the *EM* algorithm (Dempster *et al.* 1977) to estimate two-locus haplotype frequencies corresponding to each pair of SNPs and hence, estimate the coefficient of linkage disequilibrium between them.

Since it is extremely difficult to determine C analytically, it is estimated based on Monte Carlo simulations as follows. We generate a large number (say, N) of random observations of T_{1n}, T_{2n}, ..., T_{Ln} from the L dimensional multivariate normal distribution with a zero mean vector and the estimate of correlation structure δ^* using its spectral decomposition and hence, obtain a random sample (of size N) of $\max_{1\le l\le L}|T_{ln}|$. Since, the sample quantile is a consistent estimator of the theoretical quantile for a continuous distribution, C is estimated to be the sample quantile based on the random sample of size N of $\max_{1\le l\le L}|T_{ln}|$. Our simulations based on a choice of $N = 100K$ have provided desirable results.

When this method is implemented at the genome-wide level (where, L in the order of thousands), the computation of the exact spectral decomposition of the entire ($L \times L$) correlation matrix becomes infeasible. Hence, we consider partitioning the total set of SNPs into smaller linkage disequilibrium blocks (LD blocks) of some fixed size. In other words, δ^* is constructed as consecutive nonzero diagonal blocks of a fixed size with all the other elements of the matrix set to zero. Gao *et al*. (2008) had also adopted a similar strategy and had set the fixed size of each LD block as 133 in their software *simpleM.* In our simulations, we set the size of each LD block as 150 for both *simpleM* and *MVNblock*. It is obvious that the size of the last LD block (the last diagonal block of δ^*) may be smaller than the fixed size of the other blocks. We note that, since, a correlation matrix with nonzero diagonal blocks represents a series of uncorrelated blocks of SNPs, the two important properties of *MVNblock* stated earlier also hold in the genome-wide set-up.

6. simpleM

Given that, λ_1, ..., λ_L are the eigenvalues of the CLD correlation matrix of L correlated SNPs arranged in descending order of magnitude, the effective number of independent SNPs is estimated as L_0 satisfying the conditions: $\frac{\sum_{l=1}^{L_0}\lambda_l}{\sum_{l=1}^{L}\lambda_l} \ge A$ and $\frac{\sum_{l=1}^{L_0-1}\lambda_l}{\sum_{l=1}^{L}\lambda_l} < A$, where, A is usually set at 0.995. In other words, the underlying idea in *simpleM* is to use the minimum number of eigenvalues explaining at least 99.5% of the total variance as an estimate of the effective number of independent SNPs.

7. MVNBLOCK UNDER SEPERATE LINKAGE DISEQUILIBRIUM STRUCTURES IN CASES AND CONTROLS

In the preceding sections, we have described *MVNblock* assuming that the LD structures are identical in cases and controls. However, it is possible that while the allele frequencies at the two loci are equal in cases and controls, the two-locus haplotype frequencies may be unequal resulting in different LD structures in cases and controls. It is interesting to note that *MVNblock* requires a simple modification in this scenario. Suppose $\delta_{kl}^{(1)}$ and $\delta_{kl}^{(2)}$ denote the coefficient of linkage disequilibrium between the k^{th} and l^{th} markers in cases and controls, respectively. It can be easily shown that, $Cov(V_{kn}, V_{ln}) = \frac{1}{n}\frac{\delta_{kl}^{(1)}+\delta_{kl}^{(2)}}{2}, k,l \in \{1, 2, \ldots, L\}$. However, under H_0, variance of (V_{ln}) remains as $\frac{p_l(1-p_l)}{n}$, for $l = 1, ..., L$. Thus, we only need to replace δ_{kl} in Equation (1) by $\frac{\delta_{kl}^{(1)}+\delta_{kl}^{(2)}}{2}$. The implementation of the method, of course, requires that the LD structures of cases and controls are estimated separately. We wish to highlight that both *Property* 1 and *Property* 2 hold under this set-up as well.

8. SIMULATIONS

In order to compare *MVNblock* with *simpleM* with respect to the estimation of the effective number of independent SNPs and hence, controlling for *FWER* in a case-control study, we carry out extensive simulations by generating genotype data on 1000 cases and 1000 controls at L SNPs not associated with the disease under a wide spectrum of LD structures among the SNPs. The minor allele frequencies at the L SNPs are generated from the Uniform(0.05, 0.5) distribution. In order to obtain the correlation structure of the SNPs, we simulate the standardized LD coefficient (D') between pairs of consecutive SNPs according to a Beta(a, b) distribution. This allows us to create different LD structures by varying the choice of the parameter values of a and b.

For each individual, the genotype data at the L SNPs are obtained sequentially: in the first step, we generate the genotype at SNP 1 according to Hardy-Weinberg Equilibrium; in the second step, we generate

the genotype data at SNP 2 conditioned on the generated genotype at SNP 1 using the D' value between the SNPs 1 and 2; in the third step, we generate the genotype data at SNP 3 conditioned on the genotype at SNP 2 using the D' value between SNPs 2 and 3; and so on. Thus, an overall correlation structure among the L SNPs is induced by generating the genotype data sequentially as described above. Since, none of the SNPs is associated with the disease, the genotype data for a case and a control at any SNP are generated using the same parameters. While generating the LD structure, it is important to ensure that the coefficient of linkage disequilibrium between a pair of SNPs is a decreasing function of the distance between them. We observe from the rows of the estimated correlation matrix of the SNPs that the strength of linkage disequilibrium between a pair of SNPs decays with increase in the number of SNPs present between them.

We note that, the expectation and the variance of a random variable distributed as *Beta*(a, b), are $\frac{a}{a+b}$ and $\frac{ab}{(a+b)^2(a+b+1)}$, respectively. Thus, choosing a larger value of a compared to b induces a larger value of the expectation and a smaller value of the variance. Using this feature of the Beta distribution, we vary the strength of the LD structure. We consider four different values of L : 100, 150, 1000, 10000. For each choice of L, we use five different choices of a : 2, 20, 50, 100, 200 but fix the value of b at 2. Thus, the strength of the LD structure increases as the value of a increases. For L = 10000, we have considered four different choices of (a, b) with a being chosen as 2, 50, 100, 200 and b being fixed at 2.

We estimate the *FWER* at the level 0.05 based on 1000 replications for both the methods. We also estimate the effective number of independent SNPs at each of the 1000 replications for both the methods. We estimate the effective number of independent SNPs for *MVNblock* as $\frac{0.05}{2\times(1-\Phi(C))}$, where, Φ is the c.d.f. of a standard normal distribution. However, since C is estimated based on Monte Carlo simulations, it is possible that the estimate of the effective number of independent SNPs may exceed L, in which case, it is estimated as L. We present a summary of the estimates of the effective number of independent SNPs in the 1000 replications using the sample mean, the sample median and the percentage of replications in which the estimate obtained for *MVNblock* is less than or equal to that obtained for *simpleM.*

9. RESULTS

The results of our simulations are presented in Tables 1 and 2 for two LD blocks comprising 100 and 150 SNPs, respectively. In Table 3, we present the results when a single LD block of size 1000 SNPs is considered; while in Table 4, the results pertain to a set of 1000 SNPs distributed in seven uncorrelated LD blocks with the first six blocks comprising 150 SNPs each and the last block comprising 100 SNPs. We note here that the coefficient of linkage disequilibrium between the terminating SNP of a LD block and the starting SNP of the next LD block in the simulations

Table 1. Comparison of *MVNblock* and *simpleM* with respect to the estimation of *FWER* and effective number of independent SNPs for a LD block of 100 SNPs

shape1	*FWER*				effective L				
	MVN	*siM*	*Bonf*	*Sidak*	mean		median		L_{MVN}
					MVN	*siM*	*MVN*	*siM*	$\le L_{siM}$
2	0.052	0.050	0.050	0.050	94.5	97.9	94.5	98	99.3%
20	0.051	0.046	0.042	0.043	85.0	95.0	85.0	95	100%
50	0.047	0.045	0.041	0.043	81.9	93.0	81.9	93	100%
100	0.055	0.045	0.044	0.044	75.9	89.9	75.9	90	100%
200	0.049	0.042	0.038	0.038	78.5	87.2	78.5	87	100%

MVN: MVNblock, siM: simpleM, L_{MVN}: effective number of independent SNPs for *MVNblock*, L_{siM} : effective number of independent SNPs for *simpleM.*

Table 2. Comparison of *MVNblock* and *simpleM* with respect to the estimation of *FWER* and effective number of independent SNPs for a LD block of 150 SNPs

shape1	*FWER*				effective L				
	MVN	*siM*	*Bonf*	*Sidak*	mean		median		L_{MVN}
					MVN	*siM*	*MVN*	*siM*	$\le L_{siM}$
2	0.045	0.044	0.042	0.044	143.6	143.6	146.7	147	92.8%
20	0.049	0.044	0.042	0.042	126.9	126.9	139.9	140	100%
50	0.053	0.045	0.041	0.042	124.7	138.8	124.8	139	100%
100	0.053	0.046	0.040	0.041	122.9	136.0	122.9	136	100%
200	0.054	0.043	0.039	0.040	119.2	134.8	119.2	135	100%

MVN: *MVNblock*, *siM: simpleM,* L_{MVN}: effective number of independent SNPs for *MVNblock*, L_{siM} : effective number of independent SNPs for *simpleM.*

Table 3. Comparison *of MVNblock* and *simpleM* with respect to the estimation of *FWER* and effective number of independent SNPs for a LD block of 1000 SNPs

shape1	*FWER*				effective *L*				
	MVN	*siM*	*Bonf*	*Sidak*	mean		median		L_{MVN}
					MVN	*siM*	*MVN*	*siM*	$\leq L_{siM}$
2	0.052	0.052	0.050	0.052	959.7	959.9	972.3	972	82.2%
20	0.048	0.045	0.043	0.043	882.0	928.1	881.9	928	100%
50	0.042	0.037	0.035	0.035	840.3	840.6	898.4	899	100%
100	0.053	0.049	0.042	0.045	838.3	897.4	838.7	897	100%
200	0.044	0.041	0.035	0.035	823.3	875.7	823.9	876	100%

MVN: MVNblock, *siM: simpleM*, L_{MVN}: effective number of independent SNPs for *MVNblock*, L_{siM} : effective number of independent SNPs for *simpleM.*

Table 4. Comparison of *MVNblock* and *simpleM* with respect to the estimation of *FWER* and effective number of independent SNPs for a set of 1000 SNPs composed of seven uncorrelated LD blocks

shape1	*FWER*				effective *L*				
	MVN	*siM*	*Bonf*	*Sidak*	mean		median		L_{MVN}
					MVN	*siM*	*MVN*	*siM*	$\leq L_{siM}$
2	0.055	0.053	0.053	0.053	961.1	973.6	960.6	974	81.4%
20	0.053	0.047	0.044	0.044	882.0	928.1	882.4	928	99.9%
50	0.055	0.053	0.044	0.047	843.4	902.0	842.8	902	100%
100	0.053	0.049	0.043	0.046	838.9	897.4	839.7	897	100%
200	0.047	0.043	0.037	0.037	822.8	875.7	823.3	876	100%

MVN: MVNblock, siM: simpleM, L_{MVN}: effective number of independent SNPs for *MVNblock*, L_{siM} : effective number of independent SNPs for *impleM.*

Table 5. Comparison of *MVNblock* and *simpleM* with respect to the estimation of *FWER* and effective number of independent SNPs for a *LD* block of 10000 SNPs

shape1	*FWER*				effective *L*				
	MVN	*siM*	*Bonf*	*Sidak*	mean		median		L_{MVN}
					MVN	*siM*	*MVN*	*siM*	$\leq L_{siM}$
2	0.048	0.048	0.047	0.047	9657.3	9729.8	9654.7	9730	71.1%
50	0.053	0.050	0.045	0.045	8773.7	9048.4	8776.1	9048	98.9%
100	0.049	0.048	0.043	0.045	8621.1	8904.4	8618.0	8904	98.1%
200	0.044	0.042	0.037	0.037	8539.4	8813.6	8539.9	813.5	98.8%

MVN: MVNblock, siM: simpleM, L_{MVN}: effective number of independent SNPs for *MVNblock*, L_{siM} : effective number of independent SNPs for *simpleM.*

pertaining to Table 4 is set as zero to generate uncorrelated LD blocks. We also simulate genoptype data on 10,000 SNPs to assess the relative performances of the two methods when the number of SNPs is very large. The results are provided in Table 5.

It is clear from all the tables that both *MVNblock* and *simpleM* control for *FWER* satisfactorily. However, we observe that *simpleM* controls *FWER* more conservatively compared to *MVNblock*. If we refer to the first row of each table, we observe that when Sidak's and Bonferroni's corrections are implemented, both *MVNblock* and *simpleM* estimate *FWER* almost at the same level. This suggests that the choice of *Beta*(2, 2) distribution for the LD structure induces SNPs that are practically uncorrelated. The tables also show that Sidak's and Bonferroni's corrections become more and more conservative with respect to controlling for *FWER* as the value of a increases, indicating that, the correlations between the SNPs increase with *a*.

We also observe from all the tables that, the sample mean and median of the estimated effective number of independent SNPs for *MVNblock* is consistently less than that for *simpleM* in all the simulations considered. Moreover, in most of the replications, the effective number of independent SNPs estimated by *MVNblock* is smaller than or equal to that estimated by *simpleM*. In fact, this phenomenon is observed in more than 98% of the replications for all choices of (*a*, *b*), except (2, 2) (that is the situation where a weak LD structure is induced). Hence, *MVNblock* is, in general, expected to yield higher power than *simpleM.*

As observed from the tables, the estimates of the effective number of independent SNPs for both the methods, on an average, decrease with increase in the value of *a*. Moreover, the difference between the estimated effective number of independent SNPs for *MVNblock* and *simpleM*, on an average, increases with *a*.

9. AN APPLICATION USING REAL DATA

We use data from an ongoing genetic study on coronary artery disease collected on a random sample of 1248 individuals from a North Indian population. We compare the performances of *simpleM* and *MVNblock* using genotype data at 4330 SNPs on Chromosome 1 on a set of 578 individuals selected after some initialltering of the original data.

Our results are based on 1500 replications. In each replication, we divide the set of individuals at random into two equal groups (each comprising 289

individuals) while the set of SNPs are decomposed into 28 non-overlapping blocks each comprising 150 SNPs and one block comprising 130 SNPs. We estimate the effective number of independent SNPs for both *simpleM* and *MVNblock* in each replication. We also estimate the *FWER* based on all the replications. We find that the median and the mean of the effective number of independent SNPs for *MVNblock* are 3356.9 and 3358.1, respectively, while both these measures for *simpleM* are found to be 2931 (since the covariance matrix and its spectral decomposition remains unchanged over replications, a fixed estimate of the effective number of independent SNPs is obtained in each replication). The *FWER* obtained by Bonferroni's correction, Sidak's correction, *MVNblock*, and *simpleM*, are 0.041, 0.043, 0.053, and 0.059, respectively. As expected, the corrections by Bonferroni and Sidak are extremely conservative. We note that, while *simpleM* provides marginally lower estimates of the effective number of independent tests compared to *MVNblock*, it is at the expense of a highly inated overall type I error rate. On the other hand, *MVNblock* maintained the overall false positive rate very close to the desired level.

10. DISCUSSION

We have compared a block-wise multiple testing strategy *MVNblock* and a prinicipal components analysis based method *simpleM*, both of which can be implemented at the genome-wide level. An immediate consequence of formulating *MVNblock* for obtaining a common cut-off, C, to be used while testing association at the individual SNP level is that C can be shown to be the $(1 - \alpha)$% quantile of the one-dimensional test statistic $\max_{1\le l\le L} |T_{ln}|$, and can be estimated easily as the $(1 - \alpha)$% sample quantile of the test statistic.

The global null hypothesis H_0 can also be tested in the *MVNblock* framework based on the minimum p-value, $p_{min} = P(\max_{1\le l\le L} |T_{ln}| > t)$; where t is the observed value of $\max_{1\le l\le L} |T_{ln}|$. We note that the quantity p_{min} can be estimated empirically based on a large random sample of $\max_{1\le l\le L} |T_{ln}|$. In fact, Seaman and Muller-Myhsok (2005) executed this strategy [direct simulation approach (DSA)] to compute p_{min}.

Among the multivariate normal approaches for multiple testing (Lin 2005, Seaman and Muller-Myhsok 2005, Conneely and Boehnke 2007, Han *et al.* 2009), *SLIDE* has been implemented at the genome-wide level and seems to be the most efficient method (Han *et al*. 2009). On the other hand, *simpleM* seems to be the most effcient method among the PCA based approaches (Gao *et al*. 2008). A comparison between *SLIDE* and *simpleM* at the genome-wide level (Gao 2011) showed that *simpleM* performed better than *SLIDE*. However, *SLIDE* is a sliding window based approach, but *simpleM* considers a fixed windows approach (that is, consecutive non-overlapping LD blocks of some fixed size). Thus, it was of interest to evaluate the performance of a multivariate normal based approach using the same fixed windows along the genome as the PCA based *simpleM.* Thus, we partitioned a series of markers into a set of consecutive LD blocks of some fixed size in our simulations and compared the relative performances of *simpleM* and *MVNblock* based on the same partition of LD blocks.

In addition to having a compact theoretical support, our simulation studies show that *MVNblock* controls for *FWER* less conservatively and produces lower estimates of the effective number of independent SNPs. Since this estimate is used as the denominator in the Bonferroni correction while testing association at the SNPs, *MVNblock* is expected to yield higher power compared to *simpleM* in case-control association studies. However, given the inherent differences in the theoretical frameworks of multivariate normal based approaches and PCA based approaches, it is intuitively difficult to explain the less conservative behavior of *MVNblock* compared to *simpleM.*

However, *simpleM* is computationally faster than *MVNblock*. For example, if genotype data are available at 10000 SNPs for 1000 cases and 1000 controls as in Table 5, *MVNblock* has a runtime of four and half minutes, while *simpleM* requires less than a minute to run on a desktop with processor frequency of 3.0 GHz and 3.9 GB of RAM. Hence, even though *simpleM* is computationally faster, the runtime of *MVNblock* is quite managable. While we have performed all our simulations with null markers, we would like to emphasize that *Property* 1 ensures that *MVNblock* controls for *FWER* even when the data comprises one or more associated SNPs. Thus, one can utilize all available SNPs in a genome-wide association study to correct for multiple testing.

The method implemented in *simpleM* models the composite linkage disequilibrium (CLD) structure as the correlation matrix for the correlated SNPs. The coefficient of CLD between two biallelic loci with alleles (A, a) and (B, b), respectively, first introduced by Weir (1979) and denoted by Δ_{AB}, is defined as $P_{AB}+P_{A/B} - 2P_AP_B$, where P_{AB} is the frequency of the gamete AB, $P_{A/B}$ is the joint frequency of alleles A and B at two different gametes, P_A and P_B are the frequencies

of alleles A and B at the two loci. Estimation of Δ_{AB} does not require haplotype phase information and hence, is computationally simpler. Suppose x_i and y_i denote the number of A and B alleles, respectively, in the i^{th} individual (both assuming the values 0, 1, 2). It can be shown that, $E(x_i) = 2P_A$, $E(y_i) = 2P_B$ and $Covariance(x_i, y_i) = 2\Delta_{AB}$ (Zaykin 2004).

We note that, the difference between $\frac{1}{2n}\sum_{i=1}^{n} x_i$ in cases and controls can be used as a test statistic to test for $H_0 : P_{A|case} = P_{A|control}$. We note that *MVNblock* can also be implemented using the CLD matrix as the correlation matrix. It can be easily verified that both the theoretical properties of *MVNblock* discussed earlier will also hold in this case. We have also found in our simulation study that considering the CLD matrix as the correlation matrix in *MVNblock* yields very similar results with respect to the estimated *FWER* and effective number of independent tests and requires almost the same runtime compared to an analysis considering the LD matrix as the correlation matrix, where we have executed the EM algorithm to estimate the LD coefficient between pair-wise SNPs.

We have considered 100K random samples to estimate C in *MVNblock* throughout our simulation study. However, given the huge number of SNPs involved in genome-wide association studies, a larger sample size is required for accurate estimation of C. It may be more optimal to use $500K$ random samples when data are available on $100K$ SNPs. In such situations, it is clear that the computational runtime will increase but will still be feasible to implement.

It has been discussed in Han *et al.* (2009) that, at the extreme tails, the true null distribution of the test statistic may deviate from the approximated multivariate normal distribution, especially when a large number of SNPs have rare variants (minor allele frequency < 0.01). In such situations, the approximated multivariate normal distribution needs to be scaled to fit the true null distribution of the test statistics (Han *et al*. 2009).

ACKNOWLEDGEMENTS

We are grateful to Professor Partha Pratim Majumder for initiating the statistical issues pertaining to the problem and to Dr. Indranil Mukhopadhyay for helpful methodological discussions. We sincerely thank Dr. Shantanu Sengupta of the Institute of Genomics and Integrative Biology (IGIB), New Delhi for providing access to data collected in the project on coronary artey disease. This work was partially supported by the Council of Scientific and Industrial Research (CSIR) fellowship 09/093(0112)/2008-EMR-I to Arunabha Majumdar.

REFERENCES

Cheverud, J.M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heridity*, **87,** 52-58.

Conneely, K.N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of P values for multiple correlated tests. *Am. J. Hum. Genet*., **81,** 1158-1168.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B,* **39***,* 1-38.

Gao X., Starmer J. and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol*., **32**, 361-369.

Gao, X. (2011). Multiple testing corrections for imputed SNPs. *Genet. Epidemiol*., **35,** 154-158.

Han, B., Kang, H.M. and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*., **5,** 1-13.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heridity*, **95,** 221-227.

Lin, D.Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, **21,** 781-787.

Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet*., **74,** 765-769.

Seaman, S.R. and Muller-Myhsok, B. (2005). Rapid simulation of Pvalues for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.,* **76,** 399-408.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.,* **62,** 626-633.

Weir, B.S. (1979). Inferences about linkage disequilibrium. *Biometrics*, **31,** 235-254.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based Multiple Testing*. Wiley, New York.

Zaykin, D.V. (2004). Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet. Epidemiol.,* **27**, 252-257.

Appendix

Proof of Property 1: Suppose that, m_0 SNPs out of the L SNPs under consideration are in reality not associated with the disease, that is, m_0 null hypotheses among $\{H_{01}, ..., H_{0L}\}$ are true. Suppose that the index set of these m_0 unassociated SNPs is denoted by $\{i_1, i_2, ..., i_{m_0}\}$, that is, $H_{i_1}, H_{i_2}, ..., H_{i_{m_0}}$ are true and the remaining null hypotheses are false. Under this configuration of true and false null hypotheses, *FWER* of the above simultaneous critical region is as follows:

$$
\begin{aligned}
FWER &= P_{H_{i_1}, ..., H_{i_{m_0}}} \text{ (at least one false positive)} \\
&= 1 - P_{H_{i_1}, ..., H_{i_{m_0}}} \text{ (no false positive)} \\
&= 1 - P_{H_{i_1}, ..., H_{i_{m_0}}} (\{|T_{ln}| \le C\}, \forall l = i_1, i_2, ..., i_{m_0}) \\
&= 1 - P_{H_{i_1}, ..., H_{i_{m_0}}} (\max_{l \in \{i_1, ..., i_{m_0}\}} |T_{ln}| \le C) \\
&\le 1 - P_{H_0} (\max_{1 \le l \le L} |T_{ln}| \le C) \le \alpha
\end{aligned}
$$

We have used the fact that, under H_0, $\max_{1 \le l \le L} |T_{ln}|$ is always greater than or equal to $\max_{l \in \{i_1, ..., i_{m_0}\}} |T_{ln}|$, under $\bigcap_{l=i_1}^{i_{m_0}} H_{0l}$.

Proof of Property 2: The first theorem of Sidak (1967) states that, if, $Z = (Z_1, Z_2, ..., Z_k)$ is a vector of random variables having a k-dimensional normal distribution with mean vector zero and an arbitrary variance-covariance matrix, then, for any positive real numbers $c_1, c_2, ..., c_k$,

$$
\begin{aligned}
P(|Z_1| \le c_1, |Z_2| \le c_2, ..., |Z_k| \le c_k) \ge P(|Z_1| \le c_1) \times \\
P(|Z_2| \le c_2, ..., |Z_k| \le c_k)
\end{aligned}
$$

It trivially follows from the above theorem that:

$$
\begin{aligned}
P(|Z_1| \le c_1, |Z_2| \le c_2, ..., |Z_k| \le c_k) \ge P(|Z_1| \le c_1) \times \\
P(|Z_2| \le c_2) \times ... \times P(|Z_k| \le c_k)
\end{aligned}
$$

Using the above inequality in our set-up, we can easily see that, for a positive real number c,

$$
\begin{aligned}
&P_{H_0} (\max_{1 \le l \le L} |T_{ln}| \le c) = P_{H_0} (|T_{1n}| \le c, |T_{2n}| \le c, ..., |T_{Ln}| \le c) \\
&\ge P_{H_{01}} (|T_{1n}| \le c) \times P_{H_{02}} (|T_{2n}| \le c) \times ... \times P_{H_{0L}} (|T_{Ln}| \le c)
\end{aligned}
$$

We note that the right hand side (R.H.S.) of the above inequality is identical to P_{H_0} (no false positive) in Sidak's correction that assumes all the SNPs to be independent. In order to control the *FWER* at level α by Sidak's correction, c in the above inequality needs to be chosen such that R.H.S. = $(1 - \alpha)$ (we denote this choice of c as C_{Sidak}). Thus, the above inequality implies that, $P_{H_0} (\max_{1 \le l \le L} |T_{ln}| \le C_{Sidak}) \ge (1 - \alpha)$. Since, C in the proposed multiple testing method is the $(1 - \alpha)$% quantile of $\max_{1 \le l \le L} |T_{ln}|$, it follows that $C \le C_{Sidak}$.