Journal of Statistical Software
February 2010, Volume 33, Issue 8. http://www.jstatsoft.org/

Simple Algorithms to Calculate Asymptotic Null Distributions of Robust Tests in Case-Control Genetic Association Studies in R

Yong Zang
University of Hong Kong

Wing Kam Fung
University of Hong Kong

Gang Zheng
National Heart, Lung and Blood Institute

Abstract

The case-control study is an important design for testing association between genetic markers and a disease. The Cochran-Armitage trend test (CATT) is one of the most commonly used statistics for the analysis of case-control genetic association studies. The asymptotically optimal CATT can be used when the underlying genetic model (mode of inheritance) is known. However, for most complex diseases, the underlying genetic models are unknown. Thus, tests robust to genetic model misspecification are preferable to the model-dependant CATT. Two robust tests, MAX3 and the genetic model selection (GMS), were recently proposed. Their asymptotic null distributions are often obtained by Monte-Carlo simulations, because they either have not been fully studied or involve multiple integrations. In this article, we study how components of each robust statistic are correlated, and find a linear dependence among the components. Using this new finding, we propose simple algorithms to calculate asymptotic null distributions for MAX3 and GMS, which greatly reduce the computing intensity. Furthermore, we have developed the R package **Rassoc** implementing the proposed algorithms to calculate the empirical and asymptotic p values for MAX3 and GMS as well as other commonly used tests in case-control association studies. For illustration, **Rassoc** is applied to the analysis of case-control data of 17 most significant SNPs reported in four genome-wide association studies.

Keywords: algorithm, asymptotic distributions, dependence of trend tests, genetic model selection, MAX3, robust tests.

1. Introduction

The case-control association study is a useful design for testing genetic association (Risch and Merikangas 1996). In such a design, case-control samples for a diallelic marker are summarized

in a 2×3 contingency table, where the rows correspond to case-control status and the columns to three genotypes. The null hypothesis of no association is equivalent to no association in the contingency table. Under the alternative hypothesis, if one of the two alleles confers a high risk of the disease, an individual's risk having the disease increases with the number of risk alleles in the genotype. In other words, the alternative is ordered or penetrances are ordered, where a penetrance is the probability of the disease given one of the three genotypes. Four genetic models named recessive (REC), multiplicative (MUL), additive (ADD) and dominant (DOM) models are commonly used (Sasieni 1997; Freidlin *et al.* 2002).

Two tests, Pearson's χ^2 test (Pearson's test) and the Cochran-Armitage trend test (CATT), are commonly used for the analysis of case-control genetic association studies (Sasieni 1997; Balding 2006). Pearson's test ignores the ordered alternative but the CATT takes the order into account. To apply the CATT, increasing scores are specified a priori for the three genotypes depending on the genetic model. The three CATTs optimal for REC, ADD/MUL and DOM models are obtained (Freidlin *et al.* 2002; Zheng *et al.* 2003). When the underlying genetic model can be approximately pre-specified, the CATT with the optimal score is always more powerful than Pearson's test. However, misspecification of the genetic model may result in a substantial loss of power for the CATT. In this case, Pearson's test is more robust (score independent). Therefore, there is a trade-off of robustness and efficiency between Pearson's test and the CATT (Yamada and Okada 2009; Zheng *et al.* 2009).

In real data analysis, the true genetic model is often unknown. Hence, robust tests are preferable to the CATT or Pearson's test (Freidlin *et al.* 2002). Robust tests for case-control genetic association studies include the maximin efficiency robust test (MERT, Gastwirth 1966, 1985) with the recently developed R package **lawstat** (Hui *et al.* 2008), MAX3 (Freidlin *et al.* 2002; Gonzalez *et al.* 2008), the constrained likelihood ratio test (CLRT, Wang and Sheffield 2005), the GMS (Zheng and Ng 2008), and the optimal dose-effect trend test (Yamada and Okada 2009). A detailed review of robust tests and their applications to genetic linkage and association studies can be found in Joo *et al.* (2009a). In the following we focus on commonly used robust tests for case-control studies.

Suppose a family of scientifically plausible models is defined. Corresponding to each model, an asymptotically optimal, normally distributed test is obtained. Hence, a family of normally distributed tests is formed. When the model is uncertain, a pre-specified test from this family is not fully efficient. The minimum efficiency (e.g., Pitman asymptotic relative efficiency; Noether (1955)) of each test can be obtained over the family of models. A test with higher minimum efficiency is more robust. The MERT achieves the maximum minimum efficiency (Gastwirth 1966). Under some conditions, the MERT can be written as the weighted average of two normally distributed tests with the minimum correlation (called the extreme pair) (Gastwirth 1985). In case-control association studies, the extreme pair corresponds to the CATTs under the REC and DOM models. On the other hand, MAX3 of Freidlin *et al.* (2002) takes the maximum of the absolute values of three CATTs respectively optimal for the REC, ADD and DOM models. The MERT is often less powerful than MAX3 for case-control studies (Freidlin *et al.* 2002). However, the MERT is easier to use because it asymptotically follows a normal distribution under the null hypothesis. The MAX3 of Gonzalez *et al.* (2008) is similar to that of Freidlin *et al.* (2002). Gonzalez *et al.* (2008) considered the maximum of the likelihood ratio tests (LRTs) for the three genetic models. A test with performance similar to MAX3 is the CLRT (Wang and Sheffield 2005). It is a LRT but restricts the alternative space to REC, ADD and DOM models. Yamada and Okada (2009) found Pearson's test is a special

trend test with a data-driven score, which was also noticed by Zheng *et al.* (2009). Based on this finding, Yamada and Okada (2009) proposed an optimal dose-effect mode trend test where the genetic effect of the heterozygous genotype is restricted between two homozygous genotypes. The performance of the test of Yamada and Okada (2009) is similar to that of the CLRT. Another recent proposed robust procedure is the GMS (Zheng and Ng 2008), which is a two-phase adaptive test. In the first phase, the underlying genetic model is selected using the Hardy-Weinberg disequilibrium (HWD) trend test between cases and controls (Song and Elston 2006). In the second phase, the CATT with the selected genetic model is applied to test for association. Zheng and Ng (2008) studied an asymptotic null distribution of the GMS.

Among all the tests we considered above, MAX3 and GMS have greater efficiency robustness than Pearson's test and single CATT (Freidlin *et al.* 2002; Zheng and Ng 2008). On the other hand, MAX3, the CLRT, and the optimal dose-effect mode trend test have comparable power. The GMS seems to be slightly more powerful than MAX3 (Zheng and Ng 2008). However, there is no direct comparison between the GMS and the optimal dose-effect trend test. The asymptotic distributions of the CLRT, the MAX3 of Gonzalez *et al.* (2008), and the optimal dose-effect mode trend test have been studied. If we directly derive the asymptotic distributions for MAX3 of Freidlin *et al.* (2002) and GMS, our work would be similar to those of Wang and Sheffield (2005); Gonzalez *et al.* (2008); Zheng and Ng (2008) and Yamada and Okada (2009), who derived different asymptotic distributions for their robust tests. In addition, there would have no computation benefits. However, we identify a linear dependence structure of the CATTs and use this finding to derive the asymptotic distribution of MAX3 and provide algorithms to obtain the p values of MAX3 and GMS. By doing this way, we greatly simplify the computation and make computation of MAX3 and GMS more efficient. Our finding cannot be applied to the robust tests of Wang and Sheffield (2005); Gonzalez *et al.* (2008) and Yamada and Okada (2009). Thus, we only focus on MAX3 and GMS.

One common feature of MAX3 and GMS is that their asymptotic null distributions do not follow a standard normal distribution. In practice, a parametric bootstrap procedure may be used to obtain their empirical null distributions or approximate p values. In the parametric bootstrap approach, case-control data are resampled based on the observed case-control genotype counts under the null hypothesis. For each of the m replicates, the two robust test statistics are calculated. The m calculated statistics for each robust test based on the bootstrapped data form its empirical null distribution. In this article, we study correlations among the three trend tests and identify a new asymptotic linear dependence among them. Using this finding, we consider a simple Monte-Carlo approach to approximate null distributions for MAX3 and GMS. This approach does not need to generate case-control data, calculate trend tests and form a robust test. Instead, it only needs to generate bivariate normal distribution with a given correlation and form the robust test. Furthermore, using the same finding, we propose to use the asymptotic null distributions to obtain the p values for MAX3. We find that the proposed method gives a very good approximation to the results of the Monte-Carlo approaches, but the proposed algorithm is much more computationally efficient.

Finally, we have developed a package called **Rassoc** in the R system (R Development Core Team 2009) for the Monte-Carlo algorithms and the asymptotic algorithms of MAX3 and GMS as well as other commonly used tests in case-control association studies. For illustration, we apply the developed R package to genetic markers reported in four genome-wide association studies (Klein *et al.* 2005; Hunter *et al.* 2007; Yeager *et al.* 2007; The Wellcome Trust Case Control Consortium 2007). These data sets are also incorporated as a data frame in **Rassoc**.

The rest of this article is organized as follows. In Section 2, the two robust procedures, MAX3 and GMS, with new algorithms are presented. The R package is presented in Section 3 with illustrations. Conclusions are given in Section 4.

2. Two robust procedures: MAX3 and GMS

Consider a diallelic marker, e.g., a single nucleotide polymorphism (SNP) with alleles D and d and assume D is the allele conferring high risk of the disease. Denote the three genotypes by $G_0 = dd$, $G_1 = Dd$ and $G_2 = DD$ with genotype frequencies $g_i = \mathsf{P}(G_i)$ for $i = 0, 1, 2$. Denote the allele frequencies by $\mathsf{P}(D) = p$ and $\mathsf{P}(d) = 1 - p = q$. When Hardy-Weinberg equilibrium (HWE) proportions hold in the population, g_i can be written as $g_0 = q^2$, $g_1 = 2pq$ and $g_2 = p^2$. Denote the penetrance by $f_i = \mathsf{P}(case|G_i)$, the disease prevalence by $k = \mathsf{P}(case) = \sum_{i=0}^{2} f_i g_i$ and the genotype frequencies in cases and controls by $p_i = \mathsf{P}(G_i|case) = f_i g_i/k$ and $q_i = \mathsf{P}(G_i|control) = (1 - f_i)g_i/(1 - k)$ respectively for $i = 0, 1, 2$. Define genotype relative risks (GRRs) as $\lambda_i = f_i/f_0$ for $i = 1, 2$ ($f_0 > 0$). A genetic model is REC, MUL, ADD and DOM if $\lambda_1 = 1$, $\lambda_1 = \sqrt{\lambda_2}$, $\lambda_1 = (1 + \lambda_2)/2$ and $\lambda_1 = \lambda_2$, respectively. Under the null hypothesis of no association H_0: $\lambda_1 = \lambda_2 = 1$, i.e., $p_i = q_i$ for $i = 0, 1, 2$. Under the alternative hypothesis H_1 with D conferring the high risk, the alternative hypothesis can be expressed as H_1: $\lambda_2 \geq \lambda_1 \geq 1$ and $\lambda_2 > 1$. Thus, the alternative hypothesis is ordered.

Denote the genotype counts of (G_0, G_1, G_2) in r cases and s controls by (r_0, r_1, r_2) and (s_0, s_1, s_2), respectively. Thus, $r = \sum_{i=0}^{2} r_i$ and $s = \sum_{i=0}^{2} s_i$. Denote $n_i = r_i + s_i$ and $n = r + s$. The case-control samples for a SNP are summarized in Table 1.

The CATT has been employed to test association for the data in Table 1 (Sasieni 1997), which can be written as

$$Z_x = \frac{n^{1/2} \sum_{i=0}^{2} x_i (s r_i - r s_i)}{\left\{ rsn[n \sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2] \right\}^{1/2}}$$

where $(x_0, x_1, x_2) = (0, x, 1)$ are scores for (G_0, G_1, G_2) and x is a real number between 0 and 1. Under H_0, Z_x asymptotically follows a standard normal distribution $N(0, 1)$. Zheng *et al.* (2003) showed that Z_0, $Z_{1/2}$ and Z_1 are asymptotically optimal for REC, MUL/ADD and DOM models, respectively.

2.1. MAX3

When the genetic model is unknown, MAX3, denoted by $Z_{\text{MAX3}} = \text{MAX}\left\{|Z_0|, |Z_{1/2}|, |Z_1|\right\}$, has been proposed and used in practice (Freidlin *et al.* 2002; Sladek *et al.* 2007; Li *et al.*

	dd	Dd	DD	total
Case	r_0	r_1	r_2	r
Control	s_0	s_1	s_2	s
total	n_0	n_1	n_2	n

Table 1: Genotype data of a single SNP.

2008b,a). Because MAX3 covers a wider range of genetic models, it is more robust than any CATT.

Parametric bootstrap is commonly used to approximate the null distribution of MAX3. Given the observed data, a bootstrapped data set is simulated m times. For each bootstrapped data set, MAX3 is calculated, denoted by $Z_{\text{MAX3},j}$ for $j = 1 \dots m$. Finally, $Z_{\text{MAX3},i}, i = 1, \dots, m$, form an empirical null distribution for MAX3, which can be used to approximate its critical value or the p value. This bootstrap approach is denoted as `boot`.

The `boot` method is computation intensive, in particular for a large m, because the bootstrap case-control samples are sampled each time. The choice of m is at least 10,000 for genome-wide association studies (GWAS) (Sladek *et al.* 2007; Li *et al.* 2008a,b). An asymptotic method which reduces the computation is given as follows.

Asymptotic null distribution and p value

Before we present the asymptotic null distribution of MAX3, we first report the following result whose proof is given in Appendix A.

Lemma 1. Under H_0, Z_0, $Z_{1/2}$ and Z_1 are asymptotically linearly dependent. In addition, $Z_{1/2} = \omega_0 Z_0 + \omega_1 Z_1$ where $\omega_0 = (\rho_{0,1/2} - \rho_{0,1}\rho_{1/2,1})/(1-\rho_{0,1}^2)$ and $\omega_1 = (\rho_{1/2,1} - \rho_{0,1}\rho_{0,1/2})/(1-\rho_{0,1}^2)$ and $\rho_{i,j}$ is an asymptotic null correlation between Z_i and Z_j for $i, j = 0, 1/2, 1$, which are given in Appendix A.

Lemma 1 can be used to derive the asymptotic null distribution of MAX3. Denote the joint density function of (Z_0, Z_1) by $f(z_0, z_1, \Sigma)$ where Σ is the variance-covariance matrix of (Z_0, Z_1). Then

$$\begin{aligned}
\mathsf{P}(Z_{\text{MAX3}} < t) &= \mathsf{P}(|Z_0| < t, |Z_{1/2}| < t, |Z_1| < t) \\
&= \mathsf{P}(|Z_0| < t, |\omega_0 Z_0 + \omega_1 Z_1| < t, |Z_1| < t) \\
&= 2\int_0^{t(1-\omega_1)/\omega_0} \int_{-t}^{t} f(z_0, z_1, \Sigma) dz_1 dz_0 \\
&\quad +2\int_{t(1-\omega_1)/\omega_0}^{t} \int_{-t}^{(t-\omega_0 z_0)/\omega_1} f(z_0, z_1, \Sigma) dz_1 dz_0.
\end{aligned}$$

Note that

$$f(z_0, z_1; \Sigma) = f(z_0) f(z_1|z_0; \rho_{0,1}),$$

where $f(z_0) = \phi(z_0)$ is the density of $N(0,1)$ and $f(z_1|z_0;\rho_{0,1})$ is the density of $N(\rho_{0,1}z_0, 1-\rho_{0,1}^2)$. That is,

$$f(z_1|z_0;\rho_{0,1}) = \frac{1}{\sqrt{1-\rho_{0,1}^2}} \phi\left(\frac{z_1 - \rho_{0,1}z_0}{\sqrt{1-\rho_{0,1}^2}}\right).$$

Applying the above densities, we have

$$\begin{aligned}
&\int_0^{t(1-\omega_1)/\omega_0} \int_{-t}^{t} f(z_0, z_1, \Sigma) dz_1 dz_0 \\
= &\int_0^{t(1-\omega_1)/\omega_0} \left\{ \Phi\left(\frac{t - \rho_{0,1}z_0}{\sqrt{1-\rho_{0,1}^2}}\right) - \Phi\left(\frac{-t - \rho_{0,1}z_0}{\sqrt{1-\rho_{0,1}^2}}\right) \right\} \phi(z_0) dz_0
\end{aligned}$$

and

$$\begin{aligned}
&\int_{t(1-\omega_1)/\omega_0}^{t} \int_{-t}^{(t-\omega_0 z_0)/\omega_1} f(z_0, z_1, \Sigma) dz_1 dz_0 \\
= &\int_{t(1-\omega_1)/\omega_0}^{t} \left\{ \Phi\left(\frac{(t-\omega_0 z_0)/\omega_1 - \rho_{0,1} z_0}{\sqrt{1-\rho_{0,1}^2}} \right) - \Phi\left(\frac{-t-\rho_{0,1} z_0}{\sqrt{1-\rho_{0,1}^2}} \right) \right\} \phi(z_0) dz_0.
\end{aligned}$$

Finally

$$\begin{aligned}
\mathsf{P}(Z_{\mathrm{MAX3}} < t) &= 2\int_0^{t(1-\omega_1)/\omega_0} \Phi\left(\frac{t-\rho_{0,1} z_0}{\sqrt{1-\rho_{0,1}^2}} \right) \phi(z_0) dz_0 \\
&+ 2\int_{t(1-\omega_1)/\omega_0}^{t} \Phi\left(\frac{(t-\omega_0 z_0)/\omega_1 - \rho_{0,1} z_0}{\sqrt{1-\rho_{0,1}^2}} \right) \phi(z_0) dz_0 \\
&- 2\int_0^{t} \Phi\left(\frac{-t-\rho_{0,1} z_0}{\sqrt{1-\rho_{0,1}^2}} \right) \phi(z_0) dz_0,
\end{aligned}$$

where $\rho_{0,1}$ can be estimated under H_0. We use the function `integrate` in R to approximate the integration numerically. According to the help documentation in R, the numerical solution is calculated based on a global adaptive interval subdivision in connection with extrapolation by the Epsilon algorithm (Piessens *et al.* 1983). Hence, if the observed statistical value t^* is obtained, the p value is given by $\mathsf{P}(Z_{\mathrm{MAX3}} > t^*) = 1 - \mathsf{P}(Z_{\mathrm{MAX3}} < t^*)$. This approach is denoted by `asy`.

The `asy` method is more computation efficient because it does not require generating case-control samples. Note that Gonzalez *et al.* (2008) studied the asymptotic distribution of MAX3 of three LRTs under the null hypothesis and proposed formula to calculate the associated p value. They studied the correlations among three LRTs using the Delta method and employed three-fold integrations to calculate the asymptotic distribution. We identified and used the linear dependence of the CATTs and only require double integrations.

Using Lemma 1, a simplified algorithm to approximate the null distribution of MAX3 can also be obtained as follows. First, generate (Z_0, Z_1) from the bivariate normal distribution and calculate $Z_{1/2}$ using lemma 1. Second, calculate Z_{MAX} based on the three CATTs obtained. Last, repeat the above procedure m times to simulate an empirical null distribution of MAX3. Note that all correlations are estimated under H_0. This method is denoted as `bvn`. Notice that both `boot` and `bvn` methods require simulations of m replications to approximate the p values. However, unlike the `boot` method, the `bvn` method does not require to generate case-control data.

Numerical results

A simulation is conducted to compare three methods: the existing `boot` method, and the `bvn` and `asy` methods. The empirical cumulative distribution functions based on the three methods are reported based on 1,000,000 replicates. In each replicate, $r = 500$ cases and $s = 500$ controls are used under HWE. The results are summarized in Figure 1 with different minor allele frequencies (MAFs). The plots show that the asymptotic null distribution (`asy`) approximates the two simulation-based empirical null distributions (`boot` and `bvn`) very well.

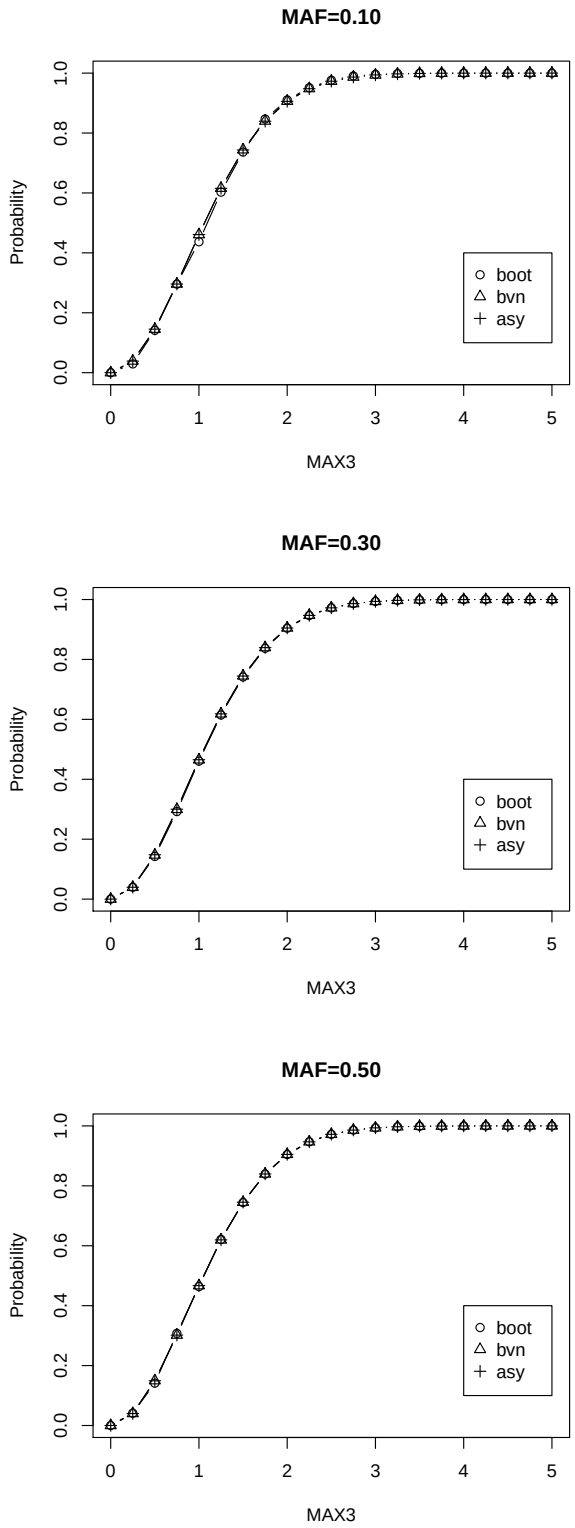

Figure 1: Cumulative distribution functions for MAX3 with different minor allele frequencies (MAFs) using the parametric bootstrap (`boot`), bivariate normal distribution (`bvn`), and asymptotic distribution (`asy`) methods.

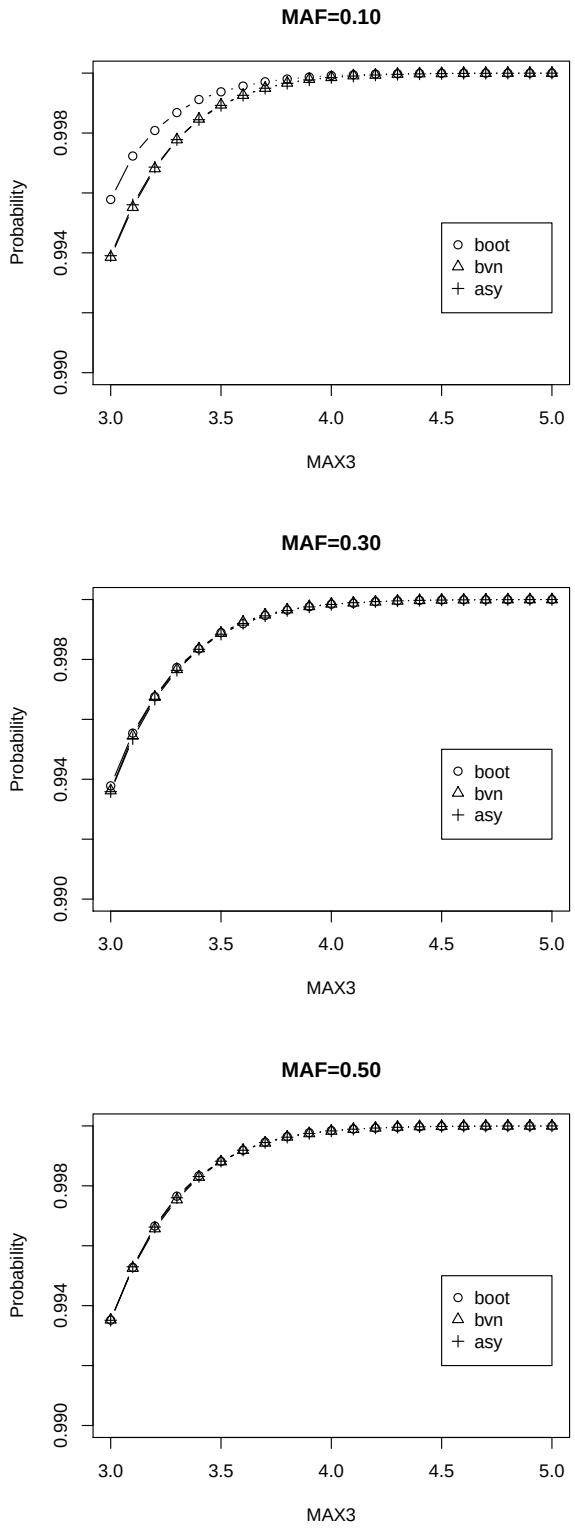

Figure 2: Cumulative distribution functions for MAX3 with different minor allele frequencies (MAFs) using the parametric bootstrap (`boot`), bivariate normal distribution (`bvn`), and asymptotic distribution (`asy`) methods. Critical values are greater than 3.0.

		Critical values					Critical values		
MAF	α	`boot`	`bvn`	`asy`	MAF	α	`boot`	`bvn`	`asy`
0.1	0.05	2.237	2.266	2.266	0.3	0.05	2.277	2.273	2.274
	0.01	2.731	2.842	2.842		0.01	2.850	2.852	2.857
	1e-3	3.375	3.520	3.520		1e-3	3.536	3.531	3.539
	1e-4	3.911	4.094	4.095		1e-4	4.114	4.114	4.116
	1e-5	4.434	4.580	4.604		1e-5	4.622	4.677	4.625
0.2	0.05	2.266	2.272	2.271	0.4	0.05	2.278	2.275	2.275
	0.01	2.842	2.855	2.852		0.01	2.864	2.855	2.859
	1e-3	3.511	3.536	3.532		1e-3	3.539	3.537	3.543
	1e-4	4.025	4.112	4.108		1e-4	4.124	4.107	4.120
	1e-5	4.626	4.571	4.617		1e-5	4.673	4.711	4.629
0.25	0.05	2.279	2.272	2.273	0.5	0.05	2.273	2.276	2.276
	0.01	2.854	2.851	2.855		0.01	2.855	2.855	2.860
	1e-3	3.535	3.533	3.536		1e-3	3.541	3.543	3.544
	1e-4	4.101	4.104	4.113		1e-4	4.146	4.105	4.122
	1e-5	4.647	4.691	4.622		1e-5	4.725	4.713	4.631

Table 2: Critical values for MAX3 with different MAFs using the parametric bootstrap (`boot`), simplified bivariate normal simulation (`bvn`), and asymptotic null distribution (`asy`) methods under HWE proportions. There are 500 cases and 500 controls in each of 1,000,000 replicates.

Furthermore, we are interested in the most significant p values in GWAS. Hence, we re-plot in Figure 2 the region of Figure 1 with critical values greater than 3.0 (the region corresponding to small p values). The results show that when MAF = 0.1, the `asy` and `bvn` method is slightly different from the `boot` methods. All three procedures match very well for other MAFs.

The critical values of MAX3 using the above three methods can also be obtained. The results are presented in Table 2. Overall, the critical values calculated using the above three methods match very well except when MAF = 0.1. Besides, the critical values are not sensitive to the change of MAFs.

We also investigate the accuracy of the asymptotic distribution of MAX3 for small smaple sizes ($n = 100$ and 200) by simulations with 1 million replicates. Table 3 reports the empirical type I error rates of MAX3 using simulations. The results show that the test MAX3 based on the asymptotic null distribution is conservative when the MAF and sample sizes are small, i.e., MAF = 0.1 and $n = 100$ or 200. In fact, with MAF = 0.1, the expected total counts n_0 in the 2×3 tables (Table 1) are 1 and 2 for $n = 100$ and 200 respectively, and the tables are highly unbalanced. Under this situation, the asymptotic theory of MAX3 may not work well when the sample size is small.

2.2. GMS

The GMS is another robust method which has two phases (Zheng and Ng 2008). In phase 1, the Hardy-Weinberg disequilibrium trend test (HWDTT) proposed by Song and Elston (2006) is used to detect the underlying genetic model. In phase 2, an optimal CATT corresponding

		Empirical type I errors	
MAF	α	$n = 100$	$n = 200$
0.1	0.05	0.020	0.024
	0.01	0.003	0.004
	1e-3	1e-4	3e-4
	1e-4	8e-6	2e-5
	1e-5	0	1e-6
0.3	0.05	0.043	0.049
	0.01	0.007	0.009
	1e-3	6e-4	8e-4
	1e-4	4e-5	5e-5
	1e-5	6e-6	9e-6
0.5	0.05	0.053	0.051
	0.01	0.010	0.010
	1e-3	8e-4	1e-3
	1e-4	7e-5	8e-5
	1e-5	4e-6	6e-6

Table 3: Empirical type I error rates for MAX3 with small sample sizes n and equal numbers of cases and controls. The nominal level α is based on the asymptotic distribution.

to the selected genetic model is used for testing association. Since the GMS needs the HWE proportions to hold in the population, we assume HWE proportions to hold throughout this section.

The Hardy-Weinberg disequilibrium (HWD) coefficients in cases and controls are denoted by $\Delta_P = P_2 - (P_2 + P_1/2)^2$ and $\Delta_Q = Q_2 - (Q_2 + Q_1/2)^2$ where $P_i = \mathsf{P}(G_i|case)$ and $Q_i = \mathsf{P}(G_i|control)$, $i = 1, 2$. Zheng and Ng (2008) showed that, under HWE and when D confers a high risk of the disease, $\Delta_P - \Delta_Q > 0$ under the REC model and $\Delta_P - \Delta_Q < 0$ under the DOM model. The HWDTT of Song and Elston (2006) is the standardized test of $\widehat{\Delta}_P - \widehat{\Delta}_Q = \{\widehat{p}_2 - (\widehat{p}_2 + \widehat{p}_1/2)^2\} - \{\widehat{q}_2 - (\widehat{q}_2 + \widehat{q}_1/2)^2\}$, where $\widehat{p}_i = r_i/r$ and $\widehat{Q}_i = s_i/s$ for $i = 0, 1, 2$. The HWDTT can be written as (Song and Elston 2006)

$$Z_{\mathrm{HWDTT}} = \frac{(rs/n)^{1/2}(\widehat{\Delta}_P - \widehat{\Delta}_Q)}{\{1 - n_2/n - n_1/(2n)\}\{n_2/n + n_1/(2n)\}}.$$

Under the null hypothesis of no association, $Z_{\mathrm{HWDTT}} \sim N(0, 1)$. With a pre-specified threshold $c > 0$ (commonly set at $\Phi^{-1}(0.95) = 1.645$), Zheng and Ng (2008) classified the underlying genetic model as REC if $Z_{\mathrm{HWDTT}} > c$, DOM if $Z_{\mathrm{HWDTT}} < -c$ and MUL/ADD otherwise. Then, when the underlying genetic model is selected, Z_x optimal for the selected genetic model is used for testing association in phase 2.

Note that we assume D conferring a high risk of the disease in the above discussion. Although the sign of Z_{HWDTT} is independent of which allele confers a high risk, the optimal CATT for the REC or DOM models does depend on such an assumption. If D confers a high risk as we assume, Z_0 and Z_1 are optimal for the REC and DOM models respectively. On the

other hand, if d confers a high risk, then Z_0 and Z_1 are optimal for DOM and REC models, respectively. In this case, the expected values of Z_0 and Z_1 are negative. Which allele confers a high risk can be detected using $Z_{1/2}$ (Joo *et al.* 2009b). That is, if $Z_{1/2} > 0$, Z_0, $Z_{1/2}$ and Z_1 are optimal for REC, MUL/ADD and DOM models. On the other hand, if $Z_{1/2} < 0$, $-Z_1$, $-Z_{1/2}$ and $-Z_0$ are optimal for REC, MUL/ADD and DOM models. Finally, the test statistic for GMS can be written as (Joo *et al.* 2009b)

$$\begin{aligned} Z_{\text{GMS}} &= Z_0 I(Z_{1/2} > 0) I(Z_{\text{HWDTT}} > c) + Z_{1/2} I(Z_{1/2} > 0) I(|Z_{\text{HWDTT}}| \leq c) \\ &+ Z_1 I(Z_{1/2} > 0) I(Z_{\text{HWDTT}} < -c) - Z_1 I(Z_{1/2} \leq 0) I(Z_{\text{HWDTT}} > c) \\ &- Z_{1/2} I(Z_{1/2} \leq 0) I(|Z_{\text{HWDTT}}| \leq c) - Z_0 I(Z_{1/2} \leq 0) I(Z_{\text{HWDTT}} < -c). \end{aligned} \quad (1)$$

Z_{GMS} does not follow $N(0,1)$ under H_0.

Like MAX3, the previous bootstrap (`boot`) method can be applied similarly. The case-control data are resampled and the GMS is applied to each replicate. In addition, a similar asymptotic method can be proposed as follows.

Asymptotic null distribution and p value

Denote the joint density function (Z_0, $Z_{1/2}$, Z_{HWDTT}) and (Z_1, $Z_{1/2}$, -Z_{HWDTT}) by $f(z, \Sigma_1)$ and $f(z, \Sigma_2)$ respectively, where f is the density function of a trivariate normal distribution and Σ_1 and Σ_2 are the variance-covariance matrixes. From (1), follow Joo *et al.* (2009b), for any $t \geq 0$,

$$\begin{aligned} Pr(Z_{\text{GMS}} \leq t) &= 1 - Pr(Z_{\text{GMS}} > t) \\ &= 1 - 2\{Pr(Z_0 > t, Z_{1/2} > 0, Z_{\text{HWDTT}} > c) \\ &\quad + Pr(Z_1 > t, Z_{1/2} > 0, -Z_{\text{HWDTT}} > c) + 0.9 Pr(Z_{1/2} > t)\} \\ &= 1.8\Phi(t) - 2\int_t^{+\infty}\int_0^{+\infty}\int_c^{+\infty} \{f(z, \Sigma_1) + f(z, \Sigma_2)\} dz - 0.8 \end{aligned}$$

We use the function `pmvnorm` from function **mvtnorm** (Genz *et al.* 2009) in R to numerically calculate the normal probabilities. According to the R documentation, we adopt the GenzBretz algorithm to approximate the numerical solution. The methodology is described in Genz (1992, 1993). Substituting the correlations by the estimates under H_0, we can use this formula to calculate the critical values and p values of the GMS.

Besides, we discuss a similar simplified bivariate normal distribution simulation. First, we present the following lemma whose proof is given in Appendix B.

Lemma 2. Under H_0, Z_0, Z_{HWDTT} and Z_1 are asymptotically linearly dependent. In addition, $Z_{\text{HWDTT}} = \mu_0 Z_0 + \mu_1 Z_1$ where $\mu_0 = (\rho_0 - \rho_{0,1}\rho_1)/(1 - \rho_{0,1}^2)$ and $\mu_1 = (\rho_1 - \rho_{0,1}\rho_0)/(1 - \rho_{0,1}^2)$, where $\rho_{0,1}$ is an asymptotic null correlation between Z_0 and Z_1 and ρ_i is an asymptotic null correlation between Z_i and Z_{HWDTT} for $i = 0, 1$, which are given in Appendix B.

Using Lemma 2, the algorithm `bvn` for MAX3 can be obtained similarly to calculate Z_{HWDTT} and obtain empirical null distribution of the GMS.

Numerical results

Simulation is used to compare the above three methods for the GMS. Analogy to Figures 1 and 2, the empirical cumulative distribution functions for the GMS are reported in Figures 3

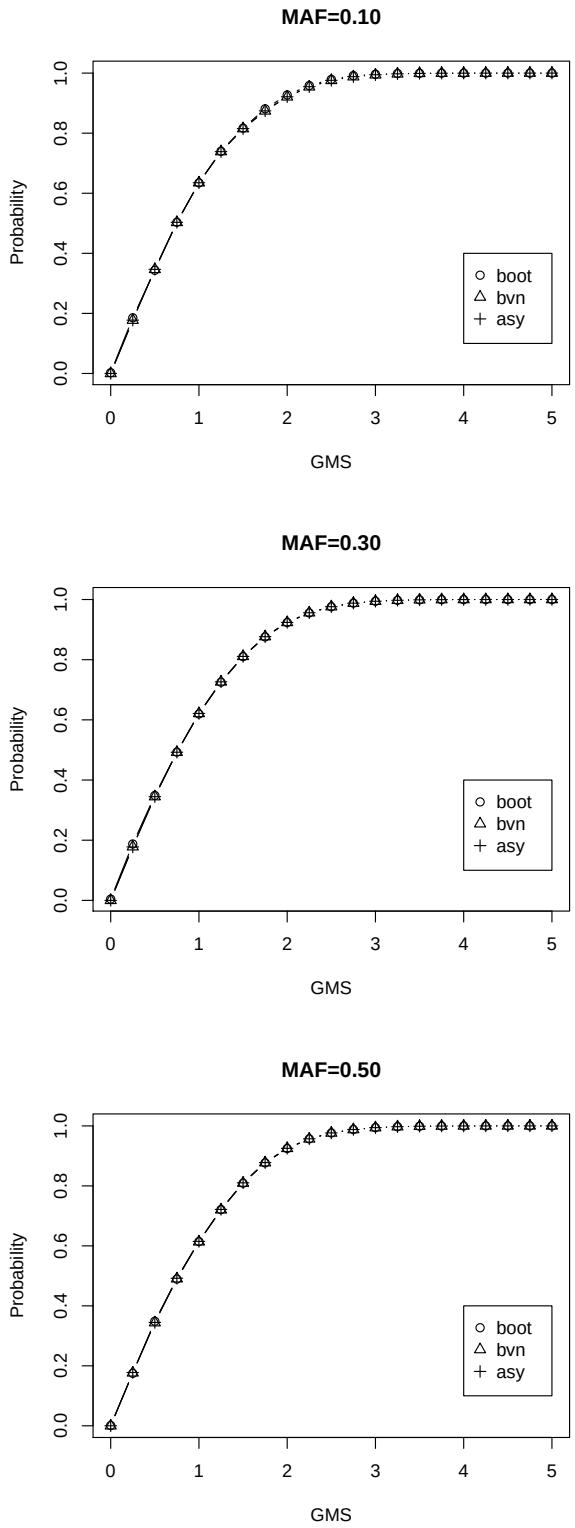

Figure 3: Cumulative distribution functions for the GMS with different minor allele frequencies (MAFs) using the parametric bootstrap (`boot`), bivariate normal distribution (`bvn`), and asymptotic distribution (`asy`) methods.

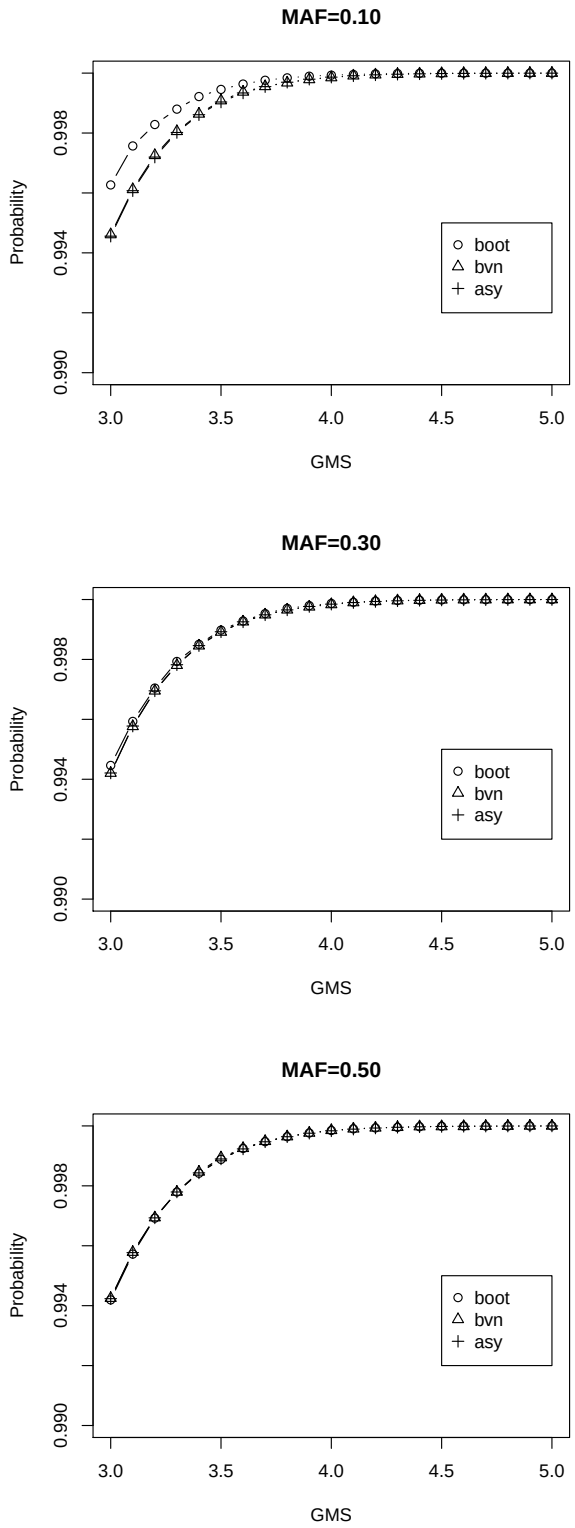

Figure 4: Cumulative distribution functions for the GMS with different minor allele frequencies (MAFs) using the parametric bootstrap (`boot`), bivariate normal distribution (`bvn`), and asymptotic distribution (`asy`) methods. Critical values are greater than 3.0.

		Critical values					Critical values		
MAF	α	`boot`	`bvn`	`asy`	MAF	α	`boot`	`bvn`	`asy`
0.1	0.05	2.153	2.207	2.207	0.3	0.05	2.191	2.198	2.194
	0.01	2.717	2.803	2.805		0.01	2.817	2.824	2.818
	1e-3	3.341	3.492	3.489		1e-3	3.505	3.514	3.520
	1e-4	3.884	4.078	4.070		1e-4	4.079	4.159	4.103
	1e-5	4.387	4.570	4.582		1e-5	4.680	4.716	4.616
0.2	0.05	2.202	2.203	2.204	0.4	0.05	2.186	2.187	2.186
	0.01	2.804	2.824	2.818		0.01	2.812	2.815	2.815
	1e-3	3.484	3.512	3.509		1e-3	3.528	3.528	3.525
	1e-4	4.007	4.091	4.089		1e-4	4.115	4.061	4.113
	1e-5	4.501	4.571	4.601		1e-5	4.508	4.576	4.626
0.25	0.05	2.197	2.201	2.199	0.5	0.05	2.185	2.184	2.184
	0.01	2.807	2.822	2.819		0.01	2.818	2.820	2.813
	1e-3	3.482	3.513	3.515		1e-3	3.531	3.511	3.527
	1e-4	4.059	4.080	4.097		1e-4	4.070	4.098	4.116
	1e-5	4.490	4.547	4.609		1e-5	4.498	4.565	4.630

Table 4: Critical values for GMS with different minor allele frequencies (MAFs) using the parametric bootstrap (`boot`), bivariate normal distribution (`bvn`), and asymptotic distribution (`asy`) methods under HWE proportions. There are 500 cases and 500 controls in each of 1,000,000 replicates.

		Empirical type I errors	
MAF	α	$n = 100$	$n = 200$
0.1	0.05	0.016	0.024
	0.01	0.002	0.004
	1e-3	9e-5	3e-4
	1e-4	3e-6	3e-5
	1e-5	0	0
0.3	0.05	0.050	0.048
	0.01	0.007	0.009
	1e-3	5e-4	8e-4
	1e-4	4e-5	7e-5
	1e-5	3e-6	8e-6
0.5	0.05	0.050	0.050
	0.01	0.009	0.010
	1e-3	8e-4	1e-3
	1e-4	7e-5	9e-5
	1e-5	4e-6	9e-6

Table 5: Empirical type I error rates for GMS with small sample sizes n and equal numbers of cases and controls. The nominal level α is based on the asymptotic distribution.

and 4 for different regions of probabilities. The parameter settings are the same as those in Figure 1 for MAX3. The plots in Figures 3 and 4 show that these three methods match very well except for small MAF (MAF = 0.1). For example, in Figure 4, when MAF = 0.1, the cumulative distribution functions for the GMS at 3.0 are around 0.994 using the `asy` and `bvn` methods while around 0.996 using the `boot` method.

Next we report the critical values of the GMS using the above three methods based on 1,000,000 replicates (Table 4). The results show that the critical values calculated from different methods match well except for MAF = 0.1. For example, in Table 4, when MAF = 0.1 and the significance level was 0.05, the critical value calculated from the `boot` method was 2.153 while those calculated from the `asy` and `bvn` methods were both 2.207. Finally, analogy to MAX3, we also report the empirical type I error rates of the GMS when the asymptotic null distribution is used. The results for small sample sizes $n = 100$ and 200 are summarized in Table 5. It is observed that, with a small sample size, the test GMS based on the asymptotic null distribution is generally conservative especially when MAF is small (MAF = 0.1). For example, when $n = 100$ and MAF = 0.1 with nominal level equal to 0.05 and 0.01, the empirical type I error rates were 0.016 and 0.002, respectively.

3. R package and examples

3.1. R package: Rassoc

In this section we describe the R package **Rassoc** developed for calculating the p values of MAX3 and GMS based on the three methods introduced above. This package also contains some other commonly used tests in case-control association studies. The package is available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=Rassoc. After it has been installed, the package may be loaded with:

```
R> library("Rassoc")
```

Rassoc contains the following functions:

`ABT(data)`: Calculate the statistic and associated p value of the allelic based test (ABT) in case-control association studies (Sasieni 1997). The ABT detects association by comparing the allele frequencies between cases and controls. Under the null hypothesis of no association, the ABT follows the standard normal distribution $N(0, 1)$.

`CATT(data, x)`: Calculate the statistic and associated p value of the Cochran-Armitage trend test (CATT) in case-control association studies for a given genetic model ($x = 0$, 1/2, or 1). Under the null hypothesis of no association, the CATT follows the standard normal distribution $N(0, 1)$.

`MERT(data)`: Calculate the statistic and associated p value of the maximin efficiency robust test (MERT) in case-control association studies (Freidlin *et al.* 2002). Under the null hypothesis of no association, the MERT follows the standard normal distribution $N(0, 1)$.

`MAX3(data, method, m)`: Calculate the statistic and associated p value of the MAX3 test in case-control association studies. The p value is calculated based on two empirical Monte-Carlo methods or the proposed asymptotic method.

`GMS(data, method, m)`: Calculate the statistic and associated p value of the GMS test in case-control association studies. The p value is calculated based on two empirical Monte-Carlo methods or the proposed asymptotic method.

The arguments in the functions are defined below:

`data`: The 2×3 contingency table for analysis. The rows represent disease status and the columns represent genotypes. See the data formats in Table 1

`x`: The score for the CATT. It can be any real number between 0 and 1.

`method`: The methods for calculating the p values of MAX3 and the GMS. `"boot"` represents the parametric bootstrap method; `"bvn"` represents the simulation from the bivariate normal distribution; `"asy"` represents the asymptotic null distribution method.

`m`: The number of replicates for `"boot"` and `"bvn"`. It can be any positive integer for `"asy"`.

Since the `ABT`, `CATT` and `MERT` follow the standard normal distribution $N(0, 1)$ under the null hypothesis, we focus on `MAX3` and `GMS` in this package. For each variable that has been imputed, `MAX3` or `GMS` creates a list containing the name of the method selected, the values of test statistics and their corresponding p values. Therefore, using these two functions, we can calculate the p values of `MAX3` and `GMS` for a dataset. For example, a simulated case-control dataset called `a` under the null hypothesis is given below:

```
R> p <- c(0.25, 0.5, 0.25)
R> ca <- rmultinom(1, 500, p)
R> co <- rmultinom(1, 500, p)
R> a <- matrix(rbind(ca, co), nrow = 2, byrow = TRUE)
R> a

139 249 112
136 244 120
```

We can apply our programs to this dataset to calculate the p values of `MAX3` and `GMS` respectively as follows:

```
R> MAX3(a, "boot", 100000)

The MAX3 test using the boot method
data: a
statistic=0.5993  p-value=0.7907

R> MAX3(a, "bvn", 100000)
```

```
The MAX3 test using the bvn method
data: a
statistic=0.5993  p-value=0.7935

R> MAX3(a, "asy", 1)

The MAX3 test using the asy method
data: a
statistic=0.5993  p-value=0.7933

R> GMS(a, "boot", 100000)

The GMS test using the boot method
data: a
statistic=0.4894  p-value=0.6608

R> GMS(a, "bvn", 100000)

The GMS test using the bvn method
data: a
statistic=0.4894  p-value=0.6609

R> GMS(a, "asy", 1)

The GMS test using the asy method
data: a
statistic=0.4894  p-value=0.6621
```

From the above output, the test statistics for `MAX3` and `GMS` are 0.5993 and 0.4894 respectively. The corresponding p values of `MAX3` for the simulated dataset `a` using `"boot"`, `"bvn"` and `"asy"` methods are 0.7907, 0.7935 and 0.7933 respectively. On the other hand, for the `GMS`, they are 0.6608, 0.6609 and 0.6621 respectively. Note that the number of replicates 1 in the functions `MAX3` and `GMS` can be any positive integer by using `"asy"` method. The empirical p values are calculated based on 100,000 replicates.

3.2. Real data analysis

For the purpose of illustration, we apply **Rassoc** to SNPs reported from four GWAS with 100,000 to 500,000 SNPs for age-related macular degeneration (AMD) (Klein *et al.* 2005), two cancer studies (Hunter *et al.* 2007; Yeager *et al.* 2007), and a hypertension study (The Wellcome Trust Case Control Consortium 2007). The datasets are summarized in Table 6 which were also given in Li *et al.* (2008a).

In fact, we also incorporate those datasets in **Rassoc** as a date frame named `caco` and it can be loaded easily in R with:

```
R> data("caco")
```

SNP ID	Case			Control		
	dd	Dd	DD	dd	Dd	DD
AMD						
rs380390	50	35	11	6	25	19
rs1329428	2	24	68	5	29	14
Prostate cancer						
rs1447295	25	283	864	10	218	929
rs6983267	223	598	351	301	579	277
rs7837688	27	283	861	11	206	939
Breast cancer						
rs10510126	10	180	955	14	272	854
rs12505080	50	477	608	99	408	628
rs17157903	18	316	777	26	220	862
rs1219648	250	543	352	170	538	433
rs7696175	187	605	353	249	496	396
rs2420946	242	546	357	165	537	440
Hypertension						
rs2820037	40	587	1,325	72	684	2,180
rs6997709	118	716	1,116	237	1,201	1,500
rs7961152	416	963	570	492	1,448	992
rs11110912	67	647	1,237	83	804	2,049
rs1937506	113	742	1,097	244	1,205	1,484
rs2398162	111	624	1,205	194	1,121	1,608

Table 6: Genotype distributions of the 17 SNPs selected from the four GWAS.

We code the genotype counts for 17 SNPs in `caco` by `d1` to `d17` in R as follows:

```
R> d1 <- matrix(caco[1,], nrow = 2, byrow = TRUE)
...
R> d17 <- matrix(caco[17,], nrow = 2, byrow = TRUE)
```

Then, `MAX3` and `GMS` can be applied to the above SNPs by the following programs:

```
R> c(MAX3(d1, "boot", 1000000)$p.value, MAX3(d1, "bvn", 1000000)$p.value,
+    MAX3(d1, "asy", 1)$p.value)
...
R> c(MAX3(d17, "boot", 1000000)$p.value, MAX3(d17, "bvn", 1000000)$p.value,
+    MAX3(d17, "asy", 1)$p.value)
R> c(GMS(d1, "boot", 1000000)$p.value, GMS(d1, "bvn", 1000000)$p.value,
+    GMS(d1, "asy", 1)$p.value)
...
R> c(GMS(d17, "boot", 1000000)$p.value, GMS(d17, "bvn", 1000000)$p.value,
+    GMS(d17, "asy", 1)$p.value)
```

SNP ID	MAX3			GMS		
	boot	bvn	asy	boot	bvn	asy
rs380390	0.10e-5	0.30e-5	0.09e-5	0.20e-5	0.10e-5	0.09e-5
rs1329428	0.10e-5	0.30e-5	0.22e-5	0.10e-5	0.30e-5	0.21e-5
rs1447295	8.30e-5	9.30e-5	10.90e-5	8.10e-5	10.20e-5	9.79e-5
rs6983267	3.00e-5	2.30e-5	2.16e-5	2.90e-5	1.60e-5	2.13e-5
rs7837688	0.50e-5	0.40e-5	0.67e-5	0.60e-5	1.00e-5	0.60e-5
rs10510126	0.10e-5	0.10e-5	0.14e-5	0.30e-5	0.30e-5	0.31e-5
rs12505080	7.40e-5	7.90e-5	8.46e-5	8.30e-5	6.50e-5	7.93e-5
rs17157903	5.30e-5	6.40e-5	6.17e-5	4.80e-5	6.00e-5	5.58e-5
rs1219648	0.40e-5	0.30e-5	0.50e-5	0.70e-5	0.50e-5	0.50e-5
rs7696175	210.00e-5	210.00e-5	207.00e-5	192.00e-5	194.00e-5	192.00e-5
rs2420946	0.70e-5	0.80e-5	0.53e-5	0.40e-5	0.30e-5	0.53e-5
rs2820037	0.10e-5	0.20e-5	0.32e-5	0.20e-5	0.10e-5	0.30e-5
rs6997709	2.80e-5	2.10e-5	2.07e-5	1.40e-5	2.60e-5	1.96e-5
rs7961152	1.40e-5	2.00e-5	2.01e-5	1.60e-5	1.60e-5	1.98e-5
rs11110912	0.50e-5	0.80e-5	0.82e-5	1.90e-5	2.00e-5	2.13e-5
rs1937506	3.40e-5	3.20e-5	2.43e-5	2.60e-5	2.20e-5	2.29e-5
rs2398162	0.10e-5	0.40e-5	0.24e-5	0.20e-5	0.20e-5	0.23e-5

Table 7: The p values of `MAX3` and `GMS` for the SNPs reported in Table 6 using the three approaches: the parametric bootstrap (`boot`), the bivariate normal distribution simulation (`bvn`), and the asymptotic distribution (`asy`). The number of replicates is 1 million for the simulation-based p values.

Note that all the empirical p values are calculated based on 1,000,000 replicates due to the computational burden. The results are summarized in Table 7.

Results show that when the analytical p values are greater than 10^{-5}, the analytical p values and simulated p values match well. However, when the analytical p values are of order of magnitude 10^{-6}, they do not match too well. One reason may be due to the fact that one million replicates may not be enough for p values of that magnitude. In conclusion, the parametric bootstrap and bivariate normal distribution methods based on simulation can be easily used in candidate-gene association studies in which several markers are tested, while the asymptotic distribution method is preferred in large-scale association studies such as GWAS. The computing times of the three approaches are also different. The asymptotic distribution method does not need any replication, so it takes least time to compute. The parametric bootstrap method takes more time to compute. The bivariate normal distribution method replicates directly on statistics while the parametric bootstrap method needs to generate case-control data in each replicate. For MAX3, using our Pentium(R) D CPU 3.00 GHZ, 1.00 GB of RAM computer, it took less than 1 second to calculate the asymptotic p value of the first row of the data in Table 6, about 1 minute using the bivariate normal distribution method, and nearly 18 minutes using the parametric bootstrap method. For the GMS, the three methods took about 1 second, 4 minutes, and 22 minutes, respectively. All empirical methods were based on 1,000,000 replicates.

4. Conclusion

Unlike the model-based CATT which assumes the underlying genetic model is known, MAX3 and GMS do not rely on any prior information of the underlying genetic model while perform robustly across a family of scientifically plausible genetic models. For many complex diseases, the underlying genetic models are unknown. Thus, MAX3 and GMS are preferable. However, their null distributions (asymptotic and empirical) have not been fully investigated. In this article, we studied the dependence structures among the CATTs for the three most common genetic models and identified an asymptotic linear relationship among them. Using this finding, we proposed to obtain asymptotic p values of the test of statistics MAX3 and GMS based on numerical integrations. The proposed method was compared to the bivariate normal simulation and the parametric bootstrap methods. Our simulation results demonstrated that the proposed asymptotic distribution method performs well. The proposed method is computationally efficient because no simulation is needed.

Moreover, we developed the R package **Rassoc** which incorporates all the methods discussed above as well as other commonly used tests in case-control association studies. We illustrated the use of **Rassoc** based on simulated datasets and some associated SNPs from four real GWAS datasets. In practice, empirical p values of MAX3 and GMS for a candidate SNP analysis or the corresponding analytical p values for GWAS can be obtained by simply setting the arguments in our R programs incorporated in the package. We also plan to update **Rassoc** by adding other recently developed tests in case-control genetic association studies in the future, including those of Gonzalez *et al.* (2008) and Yamada and Okada (2009). In this article, we only consider the null distributions of some robust tests. The asymptotic distributions under the alternative hypothesis are not considered. Li *et al.* (2009) studied the asymptotic power for MAX3 based on multivariate normal among the CATTs. Applying the linear dependence structure, their computation can be simplified. This may be also worth investigating in future.

Acknowledgments

The research of Y. Zang was partially supported by the China Natural Science Foundation grant 10701067 and the research of W. K. Fung was partially supported by the Croucher Foundation. The authors would like to thank two anonymous reviewers and an associate editor for their helpful suggestions to the manuscript and the R package.

References

Balding D (2006). "A Tutorial on Statistical Methods for Population Association Studies." *Nature Review Genetics*, **7**(10), 781–791.

Freidlin B, Zheng G, Gastwirth JL (2002). "Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness." *Human Heredity*, **53**(3), 146–152.

Gastwirth JL (1966). "On Robust Procedures." *Journal of American Statistical Association*, **61**(316), 929–948.

Gastwirth JL (1985). "The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survial Analysis." *Journal of American Statistical Association*, **80**(390), 380–384.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**(2), 141–150.

Genz A (1993). "Comparison of Methods for the Computation of Multivariate Normal Probabilities." *Computing Science and Statistics*, **25**, 400–405.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2009). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 0.9-8, URL http://CRAN.R-project.org/package=mvtnorm.

Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V (2008). "Maximizing Association Statistics over Genetic Models." *Genetic Epidemiology*, **32**(3), 246–254.

Hui W, Gel YR, Gastwirth JL (2008). "**lawstat**: An R Package for Law, Public Policy and Biostatistics." *Journal of Statistical Software*, **28**(3), 1–26. URL http://www.jstatsoft.org/v28/i03.

Hunter DL, Kraft P, Jacobs KB, Cox DG, Yeager N, Hankinson SE, Wacholder S, Wang Z, Welch R, et al AH (2007). "A Genome-Wide Association Study Identifies Alleles in FGFR2 Associated with Risk of Sporadic Postmenpausal Breast Cancer." *Nature Genetics*, **39**(7), 870–874.

Joo J, Kwak M, Chen Z, Zheng G (2009a). "Tutorial in Biostatistics: Efficiency Robust Statistics in Genetic Linkage and Association Studies under Genetic Model Uncertainty." *Statistics in Medicine*. To appear.

Joo J, Kwak M, Zheng G (2009b). "Improving Power for Testing Genetic Association in Case-Control Studies by Reducing Alternative Space." *Biometrics*. doi:10.1111/j.1541-0420.2009.01241.x. In press.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni J, Mane SM, Mayne ST, Bracken MB, et al FLF (2005). "Complement Factor H Polymorphism in Aged-Related Macular Degeneration." *Science*, **308**(5720), 385–389.

Li QZ, Yu K, Li Z, Zheng G (2008a). "MAX-rank: A Simple and Robust Genome-Wide Scan for Case-Control Association Studies." *Human Genetics*, **123**(6), 617–623.

Li QZ, Zheng G, Li Z, Yu K (2008b). "Efficient Approximation of P-value of the Maximum of Correlated Tests, with Applications to Genome-Wide Association Studies." *Annals of Human Genetics*, **72**(3), 397–406.

Li QZ, Zheng G, Liang X, Yu K (2009). "Power of the Robust Single-Marker Tests for case-Control Association Studies." *Annals of Human Genetics*, **73**(2), 245–252.

Noether GE (1955). "On a Theorem of Pitman." *Annals of Mathematical Statistics*, **26**(1), 64–68.

Piessens R, deDoncker Kapenga E, Uberhuber C, Kahaner D (1983). ***quadpack**: A Subroutine Package for Automatic Integration*. Springer-Verlag.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Risch N, Merikangas K (1996). "The Future of Genetic Studies of Complex Human Diseases." *Science*, **273**(5281), 1516–1517.

Sasieni PD (1997). "From Genotypes to Genes: Doubling the Sample Size." *Biometrics*, **53**(4), 1253–1261.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007). "A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes." *Nature*, **445**(7130), 881–885.

Song K, Elston RC (2006). "A Powerful Method of Combining Measures of Association and Hardy-Weinberg Disequilibrium for Fine-Mapping in Case-Control Studies." *Statistics in Medicine*, **25**(1), 105–126.

The Wellcome Trust Case Control Consortium (2007). "Genome-Wide Asssociation Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature*, **447**(7145), 661–683.

Wang K, Sheffield VC (2005). "A Constrained-Likelihood Approach to Marker-Trait Association Studies." *American Journal of Human Genetics*, **77**(5), 768–780.

Yamada R, Okada Y (2009). "An Optimal Dose-Effect Mode Trend Test For SNP Genotype Tables." *Genetic Epidemiology*, **33**(2), 114–127.

Yeager M, Orr N, Hayes R, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, et al NC (2007). "Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24." *Nature Genetics*, **39**(5), 645–649.

Zheng G, Freidlin B, Li Z, Gastwirth JL (2003). "Choice of Scores in Trend Tests for Case-Control Studies of Candidate-Gene Associations." *Biometrical Journal*, **45**(3), 335–348.

Zheng G, Joo J, Yang YN (2009). "Pearson's Test, Trend Test, and MAX Are All Trend Tests with Different Types of Scores." *Annals of Human Genetics*, **73**(2), 133–140.

Zheng G, Ng HKT (2008). "Genetic Model Selection in Two-Phase Analysis for Case-Control Association Studies." *Biostatistics*, **9**(3), 391–399.

A. Proof of Lemma 1

Let (p_0, p_1, p_2) be the probabilities of $(G_0 = dd, G_1 = Dd, G_2 = DD)$ in the population under H_0. Freidlin *et al.* (2002) derived asymptotic null correlations among the three trend tests. Their correlations for $(Z_0, Z_{1/2})$ and $(Z_1, Z_{1/2})$ seemed to be switched in their appendix. The correct expressions are given by

$$\begin{aligned}
\rho_{0,1/2} &= \frac{p_2(p_1 + 2p_0)}{\sqrt{p_2(1-p_2)}\sqrt{(p_1+2p_2)p_0 + (p_1+2p_0)p_2}}, \\
\rho_{1/2,1} &= \frac{p_0(p_1 + 2p_2)}{\sqrt{p_0(1-p_0)}\sqrt{(p_1+2p_2)p_0 + (p_1+2p_0)p_2}}, \\
\rho_{0,1} &= \sqrt{\frac{p_0 p_2}{(1-p_0)(1-p_2)}}.
\end{aligned}$$

Define $\tilde{\Sigma} = \begin{pmatrix} 1 & \rho_{0,1/2} & \rho_{0,1} \\ \rho_{0,1/2} & 1 & \rho_{1/2,1} \\ \rho_{0,1} & \rho_{1/2,1} & 1 \end{pmatrix}$ as the variance-covariance matrix of $(Z_0, Z_{1/2}, Z_1)$ under H_0. With simple algebra and substituting the above correlations, we obtain the determinant:

$$|\Sigma| = 1 + 2\rho_{0,1/2}\rho_{1/2,1}\rho_{0,1} - \rho_{0,1/2}^2 - \rho_{1/2,1}^2 - \rho_{0,1}^2 = 0.$$

This shows that Z_1, $Z_{1/2}$ and Z_1 are asymptotically linearly dependent. That is, there exists a non-zero vector $\mathbf{a}$ such that $\mathbf{a}^t\mathbf{Z} = 0$ where $\mathbf{Z} = (Z_0, Z_{1/2}, Z_1)$. Let $\mathbf{a} = (a_1, a_2, a_3) \neq 0$. Then $a_1 Z_0 + a_2 Z_{1/2} + a_3 Z_1 = 0$. If $a_2 = 0$, then $a_1 \neq 0$ and $a_3 \neq 0$. So there exists a real number $c^* = -\frac{a_3}{a_1} \neq 0$ satisfy $Z_0 = c^* Z_1$. Since under the null hypothesis, $\mathrm{Var}(Z_0) = (c^*)^2 = 1$. Hence $c^* = \pm 1$, which contradicts the correlation between Z_0 and Z_1. Thus, $a_2 \neq 0$. Without loss of generality, we write

$$Z_{1/2} = \omega_0 Z_0 + \omega_1 Z_1.$$

Then

$$\begin{aligned}
\rho_{0,1/2} &= \omega_0 + \omega_1 \rho_{0,1}, \\
\rho_{1/2,1} &= \omega_0 \rho_{0,1} + \omega_1.
\end{aligned}$$

Solving the above equations, we obtain

$$\begin{aligned}
\omega_0 &= \frac{\rho_{0,1/2} - \rho_{0,1}\rho_{1/2,1}}{1 - \rho_{0,1}^2}, \\
\omega_1 &= \frac{\rho_{1/2,1} - \rho_{0,1}\rho_{0,1/2}}{1 - \rho_{0,1}^2}.
\end{aligned}$$

B. Proof of Lemma 2

Use the same notation as in Appendix A. Let $p = Pr(D)$. When HWE proportions hold in the population, according to Zheng and Ng (2008),

$$\begin{aligned}
\rho_0 &= \sqrt{\frac{1-p}{1+p}}, \\
\rho_1 &= -\sqrt{\frac{p}{2-p}}.
\end{aligned}$$

Define $\Sigma^* = \begin{pmatrix} 1 & \rho_0 & \rho_{0,1} \\ \rho_0 & 1 & \rho_1 \\ \rho_{0,1} & \rho_1 & 1 \end{pmatrix}$ as the variance-covariance matrix of $(Z_0, Z_{\text{HWDTT}}, Z_1)$ under the null hypothesis. With simple algebra and substituting the above correlations, we obtain

$$|\Sigma^*| = 1 + 2\rho_0\rho_1\rho_{0,1} - \rho_0^2 - \rho_1^2 - \rho_{0,1}^2 = 0,$$

which shows that Z_0, Z_{HWDTT} and Z_1 are linearly dependent. Then, using the same argument as in Appendix A, we have

$$Z_{\text{HWDTT}} = \mu_0 Z_0 + \mu_1 Z_1$$

and

$$\begin{aligned} \mu_0 &= \frac{\rho_0 - \rho_{0,1}\rho_1}{1 - \rho_{0,1}^2}, \\ \mu_1 &= \frac{\rho_1 - \rho_{0,1}\rho_0}{1 - \rho_{0,1}^2}. \end{aligned}$$

Affiliation:

Yong Zang, Wing Kam Fung
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong, China
E-mail: zangyong@hku.hk, wingfung@hku.hk
URL: http://www.hku.hk/statistics/staff/wingfung/

Gang Zheng
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD, United States of America
E-mail: zhengg@nhlbi.nih.gov
URL: http://www.statisticalsource.com/GZ.htm

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 33, Issue 8 *Submitted:* 2009-05-05
February 2010 *Accepted:* 2009-11-10