Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 4

Testing for association between a disease and a multi-allelic marker: a powerful score test

R. el Galta, T. Stijnen and J.J. Houwing-Duistermaat

Abstract

To study association between a candidate gene and a complex genetic disease, Pearson's χ^2 statistic can be applied to a m-by-2 contingency table, where the m categories correspond to m haplotypes or marker alleles. For $m > 2$, two alternative approaches for Pearson's χ^2 can be followed that are more powerful if one haplotype or marker allele is associated. For the first approach, various 2-by-2 tables are formed by combining various categories and the maximum of the corresponding chi-square statistics is considered as the final statistic. The second approach takes the average over the possible associated categories by writing down an overall likelihood. For the latter approach we propose a new score statistic, which gives more weight to haplotypes or marker alleles that are common. Since the disease allele is often not observed, the power of the various statistics depends both on the linkage disequilibrium pattern as well as the frequencies of the associated haplotype or marker allele in the cases and the controls. We heuristically compare various statistics within the two approaches and present the results of a simulation that compares the performance of all considered statistics. Finally we apply the statistics to a case control study on the association between COL2A1 gene and radiographic osteoarthritis. Our conclusion is that overall the new proposed score statistic has good power.

Submitted for publication

4.1 Introduction

As more and more single nucleotide polymorphisms (SNPs) are discovered, candidates genes will be saturated with SNPs and the focus on haplotype based analysis will increase. When the homologous chromosomes are independently transmitted to the next generation, i.e. when Hardy-Weinberg equilibrium holds, the haplotype counts can be summarized in a m-by-2 contingency table, whose columns refer to m haplotypes and whose rows refer to the disease status. When phase is unknown, the haplotype counts have to be estimated. When phase is known (see for example Uitte de Willige et al. (2005)) or when multi allelic markers are used (see for example Kizawa et al. (2005)), summarizing the data in a m-by-2 table is straightforward. A classical test statistic for the m-by-2 table is Pearson's χ^2 statistic. For a large m this statistic has low power and when the assumption can be made that one haplotype is associated with the binary trait (Terwilliger, 1995), a more specific statistic may be preferred. In this paper we consider various statistics for these m-by-2 tables.

Genetic association is a powerful approach for common associated variants (Wang et al., 2005). Usually the disease allele is not observed and the power of the study will depend on the linkage disequilibrium between the disease locus and the marker loci and on the frequencies of the associated marker allele or haplotype in the cases and the controls (Zondervan and Cardon, 2004). Note that if the disease allele is rare, it will be detectable if the unobserved disease allele has a rather large effect on the trait and is sometimes present on a common haplotype. For sake of simplicity, we describe the methods and simulations in terms of haplotypes, but they can be applied to any m-by-2 table.

To deal with the fact that the associated haplotype is unknown, two approaches may be followed. (1) For each haplotype a statistic is computed by combining the other haplotyes and the maximum of these statistics is taken as the final statistic (maximizing approach). (2) For each haplotype a conditional likelihood given that this haplotype is associated is computed. The overall likelihood is the weighted sum over all haplotypes of all these conditional likelihoods with weights equal to the prior probabilities that a haplotypes is associated. These approaches can also be followed if one allows for a few haplotypes to be associated. Then the maximum is taken over all possible 2-by-2 tables and the likelihood is computed over all possible sets of associated haplotypes.

The maximizing approach was considered by several authors. Ewens et al. (1992) proposed to use the maximum of the χ^2 statistics of 2-by-2 tables each of which compares one variant against the rest ($\hat{Z}_{max}$), when at most one variant is associated. Sham and Curtis (1995) proposed to use the maximum of χ^2 statistics corresponding to all possible 2-by-2 tables, comparing any combination of variants against the rest ($\hat{Z}_{clump}$). From a rather small simulation study, they concluded that Pearson's χ^2 and $\hat{Z}_{clump}$ should be preferred above $\hat{Z}_{max}$ for highly polymorphic markers. Intuitively, when one haplotype is associated with a disease, $\hat{Z}_{max}$ should be more powerful than Pearson's χ^2 and $\hat{Z}_{clump}$, while if more than one associated haplotype exists $\hat{Z}_{clump}$ should have more power than $\hat{Z}_{max}$. More simulations are needed to study the performance of these test statistics.

An alternative to taking the maximum is to take the sum over all possibilities. When one variant is associated Terwilliger (1995) proposed to model the excess of the associated variant in cases by the parameter λ which is the population attributable risk (Clayton, 2000). Since it is unknown which variant is associated with the disease, the likelihood corresponding to this model is a weighted sum over all variants i of conditional likelihoods given that variant i is over-represented in the set of cases. These weights represent the prior probability that a haplotype is associated to the disease. In line with the common disease common variant hypothesis (Reich and Lander, 2001) and in line with the method of Terwilliger (1995), the haplotype frequencies in controls can be used as weights. To test for association the likelihood ratio test can be used. Maximizing the log likelihood function over the haplotype frequencies and λ appears not straightforward, because the weights are equal to the haplotype frequencies, and these same haplotype frequencies are also unknown parameters in the conditional likelihood functions. In this paper we propose the corresponding score statistic and we use Monte-Carlo permutation to derive p-values (Sham and Curtis, 1995).

We first compare heuristically the power of the Pearson's χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. We then derive the new score test and describe the results of a simulation study which we performed to compare the performance of the new score test, χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. In the simulations we assumed that phase is known. In the discussion we describe how to derive p-values for the case of phase ambiguity. As an illustration we apply these test statistics to a published case-control study on association between COL2A1 gene and radiographic osteoathritis (Meulenbelt et al., 1999).

4.2 The maximising approach

Assume that Hardy-Weinberg equilibrium holds and that we have a sample of n_1 case chromosomes and of n_2 control chromosomes. Let $p = (p_1, \cdots, p_m)$ be the vector of frequencies of the m haplotypes in controls. Let $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ be the vector of haplotype counts in the cases and the controls, respectively and let n be equal to $n_1 + n_2$. Let $\hat{Z}_i$ be equal to the observed minus expected ith haplotype count $x_i - n_1\hat{p}_i$ with $\hat{p}_i = \frac{x_i+y_i}{n}$, the estimate of haplotype frequency in combined sample. Let $\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_m)'$. Throughout the text the hat symbol^refers to two samples statistics emphasizing the fact that haplotype frequencies are estimated under the null hypothesis.

Testing the null hypothesis of no disease-marker association is classically performed by means of Pearson's χ^2 statistic

$$\chi^2 = \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{n_1\hat{p}_j} + \sum_{j=1}^{m} \frac{(y_j - n_2\hat{p}_j)^2}{n_2\hat{p}_j} = \frac{n}{n_1 n_2} \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{\hat{p}_j}.$$

An alternative test statistic is $\hat{Z}_{max}$, defined as

$$\hat{Z}_{max} = \max_{i=1\cdots m} \frac{\hat{Z}_i^2}{Var(\hat{Z}_i)}.$$

Sham and Curtis (1995) proposed the largest value of all possible χ^2 statistics of 2-by-2 tables each obtained by testing a combination of haplotype against the rest. We denote this statistics by $\hat{Z}_{clump}$ according to the program they use for the computation. In addition, they proposed to use Monte-Carlo methods to derive the empirical p-values of $\hat{Z}_{clump}$, χ^2 and $\hat{Z}_{max}$.

In order to compare these three test statistics heuristically, we rewrite them as maxima of the same expression where the maximum is taken over different sets. Pearson's χ^2 can be rewritten as:

$$\chi^2 = \max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where R is the set of vectors with m coordinates (see appendix 4.7). Since $\sum_{i=1}^{m} \hat{Z}_i = 0$, the Pearson's χ^2 test is the Hotelling's test statistic applied to any $m-1$ coordinates of the vector $\hat{Z}$.

Now $\hat{Z}_{max}$ is the maximum value of all Pearson's χ^2 tests on 2-by-2 tables obtained by comparing any haplotype against the rest. $\hat{Z}_{max}$ can be reexpressed by

$$\hat{Z}_{max} = \max_{u \in A} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where A is the set of the m different permutations of the vector $(1, 0, .., 0)'$.

The $\hat{Z}_{clump}$ statistic can be given by

$$\hat{Z}_{clump} = \max_{s \subseteq (1,\ldots,m)} \frac{(\sum_{i \in s} x_i - n_1 \sum_{i \in s} \hat{p}_i)^2}{n_1(1 - \sum_{i \in s} \hat{p}_i) \sum_{i \in s} \hat{p}_i} = \max_{u \in S} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u},$$

where S is the set of vectors whose k coordinates set to 1 and $m - k$ coordinates set to 0, with $k = 1, ..., m - 1$ and s any subset of $(1, 2, ..., m)$. This implies that under the alternative hypothesis of the presence of an association, all associated haplotypes are assumed to have the same effect sizes in terms of relative risk.

Note that A is a subset of S, which in turn, is a subset of R. Hence, if only one haplotype is associated with the disease, the alternative hypothesis is properly specified by A, and then $\hat{Z}_{max}$ is likely to have more power than $\hat{Z}_{clump}$ and χ^2. However if two or more marker haplotypes are associated with the disease and have the same effect size in terms of relative risk the $\hat{Z}_{clump}$ is expected to provide more power than $\hat{Z}_{max}$ and χ^2 as the alternative hypothesis is better specified by the set S. In case of associated haplotypes with unequal effect sizes χ^2 is expected to perform the best unless the number of haplotypes is large.

4.3 The averaging approach

Assume that one of the haplotypes is over-represented in the cases. Denote this haplotype with index i. The haplotype frequencies in the cases can be modelled as $q_i = p_i + \lambda(1 - p_i)$ with $0 \leq \lambda \leq 1$ for the associated haplotype i and as $q_j = p_j - \lambda p_j$ for the remaining haplotypes with $j = 1, \cdots, m$ and $j \neq i$. Here λ is the population attributable risk. Then the conditional likelihood of data given that haplotype i is over-represented in cases is

$$L_i(x, y|\lambda, p) = (p_i + \lambda(1 - p_i))^{x_i}(1 - \lambda)^{n_1 - x_i} \prod_{j \neq i}^{m} p_j^{x_j} \prod_{j=1}^{m} p_j^{y_j}. \tag{4.1}$$

Terwilliger (1995) proposed the following likelihood, assuming that the prior probability of a marker haplotype i being associated with the disease is equal to the haplotype frequency p_i

$$L(x, y|\lambda, p) = \sum_{j=1}^{m} p_j L_j, \tag{4.2}$$

with L_j given in formula (4.1). The corresponding score statistic is an alternative to the likelihood ratio test proposed by Terwilliger (1995) for $H_0 : \lambda = 0$ versus $H_a : \lambda > 0$. Since the first derivative of the likelihood L with respect to λ at $\lambda = 0$ is equal to zero, we propose to use the second derivative of the log-likelihood with respect to λ to derive the score statistic (Dudoit and Speed, 2000; Tritchler et al., 2003). The second derivative of the log-likelihood with respect to λ evaluated for $\lambda = 0$ is

$$\frac{\partial^2}{\partial \lambda^2} \log(L(x, y|\lambda = 0, p)) \quad = \quad \sum_{j=1}^{m} \frac{(x_j - n_1 p_j)^2}{p_j} - \sum_{j=1}^{m} \frac{x_j - n_1 p_j}{p_j} - n_1(m-1).$$

Derivation of the second derivative is given in the appendix (4.8). By dividing the second derivative by n_1 and then taking the stochastic part of it, the score statistic can be given by

$$S_p = X^2 - \frac{U}{n_1}, \tag{4.3}$$

with $X^2 = \sum_{j=1}^{m} \frac{(x_j - n_1 p_j)^2}{n_1 p_j}$, the one sample Pearson's χ^2 statistic (haplotype frequencies in controls are known), and $U = \sum_{j=1}^{m} \frac{x_j - n_1 p_j}{p_j}$, the score statistic obtained by replacing the weights in the likelihood (4.2) by equal weights. For equally frequent haplotypes the statistic $U = 0$, hence $S_p = X^2$. Under the null hypothesis, S_p has mean $\mathrm{E}[S_p] = m - 1$ and variance $\mathrm{Var}(S_p) = 2n_1^{-1}(n_1 - 1)(m-1)$ (see Appendix 4.8). Hence, asymptotically the statistics S_p and X^2 have the same expectation and the same variance.

When the haplotype frequencies p_i are unknown, the score statistic can be estimated by replacing the haplotype frequencies p_i by their maximum likelihood estimators under the null hypothesis $\hat{p}_i = \frac{x_i + y_i}{n}$. Now after some algebra the score statistic can be given by

$$\hat{S}_p = \chi^2 - \frac{n}{n_1 n_2} \hat{U}, \tag{4.4}$$

with χ^2, the two samples Pearson's χ^2 statistic on the m-by-2 table and $\hat{U} = \sum_{j=1}^{m} \frac{x_j - n_1 \hat{p}_j}{\hat{p}_j}$. It can be shown by means of the δ-method that the score statistic $\hat{S}_p$ and Pearson's χ^2 have asymptotically the same expectation and the same variance under the null hypothesis (see appendix 4.8). For m large, the score test $\hat{S}_p$ (S_p) follows approximately a normal distribution under the null hypothesis. To ensure the validity of the asymptotic distribution, the number of cases and control chromosomes should be much larger than the number of marker haplotypes m. Nevertheless, the empirical distribution under the null hypothesis of the statistic $\hat{S}_p$ (S_p) can easily be derived by using Monte-Carlo methods (Sham and Curtis, 1995).

Under the alternative hypothesis of the presence of one positively associated haplotype i, it can be shown that the expectations of U and $\hat{U}$ are

$$\begin{aligned} \mathrm{E}[U] &= n_1\lambda(\frac{1}{p_i} - m) \\ \mathrm{E}[\hat{U}] &\approx \frac{n_1 n_2 \lambda}{n - n_1\lambda}(\frac{n}{np_i + n_1\lambda(1-p_i)} - m) \leq \frac{n_1 n_2\lambda}{n - n_1\lambda}(\frac{1}{p_i} - m). \end{aligned}$$

This implies that the expectations of U and $\hat{U}$ are negative if the frequency of the associated haplotype p_i is larger than the inverse of the number of marker haplotypes $\frac{1}{m}$. Consequently, the score statistic $\hat{S}_p$ (S_p) becomes larger in expectation than $\chi^2(X^2)$ if the frequency of the associated haplotype is larger than $\frac{1}{m}$. Hence, for common associated haplotypes $(p_i > \frac{1}{m})$ the score statistic is expected to have higher power than Pearson's χ^2.

Terwilliger (1995) discussed the presence of more than one associated haplotype. For two positively associated haplotypes i and k he proposed the following model with two free parameters λ_1 and λ_2 with $\lambda_1 + \lambda_2 \leq 1$

$$\begin{aligned} q_i &= p_i(1 - \lambda_1 - \lambda_2) + \lambda_1, \\ q_k &= p_k(1 - \lambda_1 - \lambda_2) + \lambda_2 \text{ for } i \neq k, \\ q_j &= p_j(1 - \lambda_1 - \lambda_2) \text{ for } j \neq i \text{ and } j \neq k. \end{aligned}$$

The likelihood for two associated haplotypes given by Terwilliger (1995) was incorrect since the weights are prior probabilities and did not sum to 1. Therefore, we propose the following likelihood for two associated variants

$$L(x, y|\lambda_1, \lambda_2, p) = \sum_{i=1}^{m}\sum_{k=1}^{m} p_i p_k \prod_{j}^{m} q_j^{x_j} p_j^{y_j},$$

assuming $q_i = p_i(1 - \lambda_1 - \lambda_2) + \lambda_1 + \lambda_2$ for $i = k$. Since i and k are interchangeable with respect to λ_l for $l = 1, 2$ the following derivatives are

$$\begin{aligned}
\frac{\partial}{\partial\lambda_l} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= 0, \\
\frac{\partial^2}{\partial\lambda_1\partial\lambda_2} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= 0, \text{and} \\
\frac{\partial^2}{\partial\lambda_l^2} L(x, y|\lambda_1 = 0, \lambda_2 = 0, p) &= n_1 X^2 - mU - n_1(m - 1).
\end{aligned}$$

In contrast to the Terwilliger's likelihood ratio, $\hat{S}_p$ (equation 4.4) is the score statistic of testing no disease-marker association regardless of the potential number of associated variants.

4.4 Simulation study

The aim of the simulation study is to evaluate empirically the power of the score test $\hat{S}_p$ in comparison with the Pearson's χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$ tests. We generated at least 1000 replicates from the multinomial distributions according to the models described previously. Without loss of generality we assumed that the first or first two haplotypes are associated with the disease. The remaining haplotypes were equally frequent. We varied the number of variants m from 3 to 20. The p-values of the test statistics were calculated empirically by means of 1000 Monte-Carlo permutations using a program based on the program Clump (Sham and Curtis, 1995). We used a nominal p-value of 0.05.

Type I error rate

To verify whether Monte-Carlo yields the right type I error rate of these test statistics, data sets were generated under the null model ($\lambda = 0$) each time for markers with 5, 7, 9, 11, 16 and 20 alleles. The frequency of the first allele was set to 0.5, whereas the remaining alleles were equally frequent. The results are shown in Table 4.1. The type I error rate is approximately equal to the nominal rate for the score $\hat{S}_p$, Pearson's χ^2, and $\hat{Z}_{clump}$ tests, regardless of the number of alleles m at the marker locus, whereas the $\hat{Z}_{max}$ becomes somewhat conservative as the number of marker alleles m increases (Sham and Curtis, 1995).

TABLE 4.1: The type I error rates based on 10000 simulated m-by-2 tables for $\lambda = 0$ and $p_1 = 0.5$.

α	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$
0.05	5	0.053	0.053	0.047	0.054	11	0.047	0.046	0.039	0.046
0.01		0.011	0.011	0.009	0.011		0.010	0.010	0.008	0.010
0.001		0.001	0.001	0.001	0.001		0.001	0.001	0.000	0.001
0.05	7	0.051	0.048	0.045	0.052	16	0.049	0.049	0.040	0.047
0.01		0.011	0.010	0.009	0.010		0.011	0.011	0.007	0.010
0.001		0.001	0.001	0.000	0.001		0.001	0.001	0.000	0.001
0.05	9	0.051	0.052	0.044	0.052	20	0.052	0.052	0.035	0.053
0.01		0.001	0.011	0.008	0.010		0.011	0.011	0.008	0.010
0.001		0.002	0.002	0.001	0.001		0.001	0.002	0.001	0.002

Single associated variant

To study the power of the statistics we first considered the model used by Terwilliger (1995) for one positively associated common haplotype. The frequency p_1 of this haplotype was 0.5 in controls. The parameter λ was fixed to 0.5, which corresponds to a haplotype frequency of 0.75 in the cases and a relative risk γ of 3. We considered 100 case chromosomes (n_1) and 100 control chromosomes (n_2). The results are shown in Table 4.2. For $m \leq 5$ all test statistics performed well; however $\hat{S}_p$ had slightly higher power than other test statistics. For $m > 5$ the score test $\hat{S}_p$ and $\hat{Z}_{max}$ tests appeared to perform better than the Pearson's χ^2 and $\hat{Z}_{clump}$ tests regardless of the number of haplotypes at the marker locus. Especially for the significant level of 0.05, $\hat{S}_p$ and $\hat{Z}_{max}$ had similar power, while for lower significant levels $\hat{S}_p$ had somewhat lower power than $\hat{Z}_{max}$. The power of Pearson's χ^2 decreased as the number of haplotypes increased.

Second, we studied the power of the test statistics for various values of the frequency of the associated haplotype (0.06 to 0.5). We chose λ so that the relative risk γ of the associated variant with respect to its absence was about 2. Because of low λ, the number of chromosomes n_1 and n_2 were now set to 200. The results are depicted in figure 4.1. Almost overall $\hat{Z}_{max}$ outperformed the other test statistics. $\hat{S}_p$ had the second best power. It performed better than Pearson's χ^2 especially when $m \geq 10$. Further it had higher power than Z_{clump} for $p_1 = 0.06$ and 0.1 while for $p_1 = 0.15$, 0.2, 0.3 and 0.4, the performances of

TABLE 4.2: The power based on 10000 simulated m-by-2 tables for $\lambda = 0.5$ and $p_1 = 0.5$.

α	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$	m	χ^2	$\hat{Z}_{clump}$	$\hat{Z}_{max}$	$\hat{S}_p$
0.05	3	0.92	0.93	0.93	0.94	9	0.76	0.85	0.91	0.88
0.01		0.79	0.80	0.80	0.83		0.55	0.68	0.79	0.73
0.001		0.54	0.55	0.55	0.58		0.30	0.42	0.57	0.48
0.05	4	0.89	0.90	0.91	0.92	11	0.72	0.84	0.93	0.88
0.01		0.73	0.76	0.77	0.79		0.51	0.67	0.82	0.72
0.001		0.47	0.5	0.53	0.54		0.27	0.40	0.60	0.48
0.05	5	0.86	0.89	0.90	0.91	16	0.64	0.82	0.95	0.88
0.01		0.69	0.74	0.77	0.77		0.42	0.63	0.85	0.72
0.001		0.43	0.49	0.54	0.53		0.20	0.36	0.61	0.46
0.05	7	0.80	0.87	0.89	0.89	20	0.60	0.82	0.95	0.88
0.01		0.61	0.70	0.76	0.74		0.38	0.62	0.85	0.72
0.001		0.35	0.45	0.55	0.50		0.20	0.35	0.62	0.46

$\hat{S}_p$ and Z_{clump} were comparable. Pearson's χ^2 performed well when $m = 5$ or when $p_1 = 0.06$.

Two associated variants

To study the performance of the test statistics when there are two haplotypes positively associated with the disease, we generated data according to the model given by (4.5). We simulated data sets for $(p_1, p_2) = (0.06, 0.06)$, (0.06,0.1), (0.1, 0.15), (0.15,0.2), (0.2,0.3), (0.3,0.4) and their corresponding excess frequencies $(\lambda_1, \lambda_2) = (0.05, 0.05)$, (0.05,0.08), (0.08,0.1), (0.1,0.15), (0.15,0.2), (0.18,0.25), respectively. The total relative risk of the two associated variants with respect to their absence was again about two (between 1.9 and 2.1). The number of chromosomes n_1 and n_2 were set to 100. The power curves are shown in figure 4.2. In contrast to the case of one associated haplotype, $\hat{S}_p$ and $\hat{Z}_{clump}$ performed now better than Z_{max}. For $m \leq 8$ and $p_1 = 0.06$, $\hat{S}_p$ appeared to have somewhat less power than Pearson's χ^2 and $\hat{Z}_{clump}$. Whereas for $m \geq 10$ $\hat{S}_p$ had somewhat more power than Pearson's χ^2 and $\hat{Z}_{clump}$. For the remaining situations ($p_1 \geq 0.1, p_2 \geq 0.15$), $\hat{S}_p$ had the best power and $\hat{Z}_{clump}$ had the second best power. The power of Pearson's χ^2 was comparable to that of $\hat{Z}_{clump}$ for $m = 5$ and it decreased with the increase of m. For

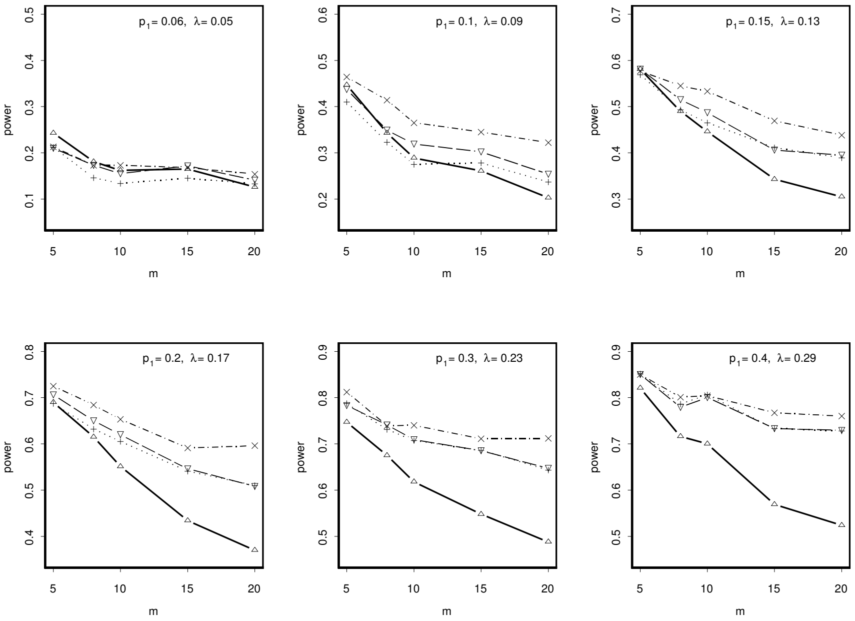

FIGURE 4.1: *Power curves of χ^2 (—△—), $\hat{Z}_{clump}$ (····+····), $\hat{Z}_{max}$ (–·–·×·–·–) and $\hat{S}_p$ (——▽——) for 200 case and 200 control chromosomes and a haplotypic relative risk of about 2.*

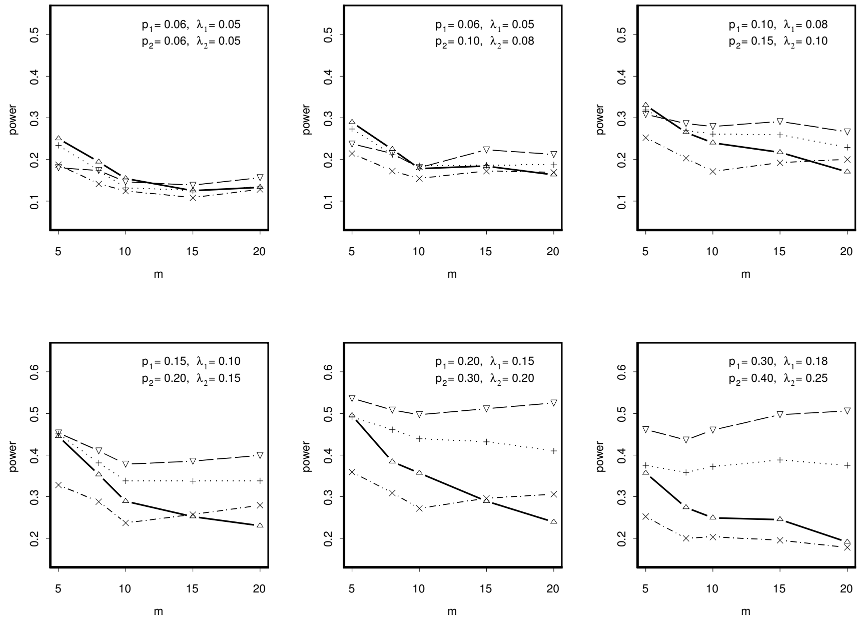

FIGURE 4.2: *Power curves at the nominal level $\alpha = 0.05$ of χ^2 (—△—), $\hat{Z}_{clump}$ (····+····), $\hat{Z}_{max}$ (–·–·×·–·–) and $\hat{S}_p$ (——▽——) for 100 case and 100 control chromosomes and a relative risk of about 2 for the combined associated haplotypes.*

most situations, it had higher power than $\hat{Z}_{max}$, which had low power.

4.5 Application to real data

To illustrate the score test with real data we used data obtained from a published study on association between the collagen type II gene(COL2A1) and radiographic osteoarthritis (ROA) (Meulenbelt et al., 1999). Osteoarthritis is a degenerative disease of the joints. The VNTR marker next to COL2A1 was typed in 820 subjects aged 50-70 years from a population-based cohort study, the Rotterdam study. Radiographs of knees, hips, hands, and spine were scored for the presence of ROA. 123 cases had ROA in at least 3 joints groups. 697 remaining subjects were used as controls. Five variants had frequencies $\geq$ 0.05. The other variants were combined into one group. Preliminary in-

TABLE 4.3: *Results of analysis of association between ROA and COL2A1 gene.*

COL2A1 haplotypes	2# cases $n_1 = 246$	2# controls $n_2 = 1394$	χ^2	$\hat{Z}_{max}$	$\hat{Z}_{clump}$	$\hat{S}_p$
13R1	114 (0.46)	581 (0.42)				
14R1	54 (0.22)	385 (0.28)				
11R1	22 (0.09)	149 (0.11)				
$14R2^+$	27 (0.11)	78 (0.06)				
13R2	12 (0.05)	85 (0.05)				
Others pooled	17 (0.08)	116 (0.08)				
P-value			0.013	0.009	0.044	0.012

spection of data suggested the existence of one positively associated variant. The frequency of the potential associated variant, 14R2, was 0.06 ($< 1/m$ with $m = 6$) in controls. Its frequency was increased in the cases with a relative risk of about 1.94 ($\lambda \approx 0.05$). Hence one might expect that Pearson's χ^2 will perform slightly better than $\hat{S}_p$ (see Figure 4.2). (Meulenbelt et al., 1999) used Terwilliger's *LR*. They found evidence of association between ROA and COL2A1 (P=0.03). We also applied the score $\hat{S}_p$, Pearson's χ^2, $\hat{Z}_{max}$, and $\hat{Z}_{clump}$ tests to these data. Table 4.3 shows the distribution of the variants among cases and controls and summarizes the results of applied test statistics. All test statistics indicated the presence of a significant association at the 0.05 level. The score test $\hat{S}_p$, Pearson's χ^2, and $\hat{Z}_{max}$ gave quite similar p-values of about 0.01. The $\hat{Z}_{clump}$ test yielded a quite higher p-value of about 0.04.

4.6 Discussion

In this paper, we derived a new score statistic $\hat{S}_p$, which corresponds to the likelihood ratio statistic of Terwilliger (1995). The score test is easy to compute and is asymptotically locally most powerful (Cox and Hinkley, 1974). For a single common positively associated haplotype (frequency $> 1/m$), we showed heuristically that the score test would provide more power than Pearson's χ^2 on the m-by-2 table. Further in contrast to the likelihood ratio statistic of Terwilliger, the same score test is obtained regardless of the number of associated haplotypes. For large m, the score test follows approximately a normal distribution under the null hypothesis. For small sample sizes or small m, the empirical p-values can be derived by means of Monte-Carlo methods.

By means of simulations we compared the performance of this new statistic to the existing statistics χ^2, $\hat{Z}_{max}$ and $\hat{Z}_{clump}$. The $\hat{S}_p$ gives more weight to common haplotypes, but for one associated haplotype it had similar or slightly less power than $\hat{Z}_{max}$. The power of $\hat{Z}_{max}$ was dramatically low for many considered models of two associated haplotypes. The power of Pearson's χ^2 decreased with the number of observed haplotypes, due to the increasing number of degrees of freedom.

In the simulation study we assumed that phase is known. When phase is ambiguous, the haplotype counts have to be estimated and the uncertainty in phase has to be taken into account when computing the p-value of the statistics. This can easily be incorporated when Monte Carlo methods are used by estimating the haplotype frequencies in each permutation step (see for example Becker et al. (2005)). This adjustment is less efficient than maximizing the likelihood over the haplotype frequencies and the unknown parameter λ simultaneously, but in many situations the loss of information due to phase-uncertainty is small. We recommend to consider smaller blocks or single marker methods when the loss of information due to phase uncertainty is rather large (Uh et al., 2005).

In this paper, we used the chromosome as unit of analysis and not the individual. By doing so we have to assume Hardy Weinberg equilibrium and the methods correspond to a multiplicative model for diplotypes (Sasieni, 1997). Therefore the power will decrease when the true model deviates from this multiplicative model as is the case for a recessive model. For this model we recommend to use other methods (see also Cordell and Clayton (2005)).

Like the approach of Terwilliger (1995), $\hat{S}_p$ gives more weight to common haplotypes. A positive association may be due to a common causal variant, due to a rare mutation with a rather large impact which is sometimes present on a common haplotype or due to multiple mutations on the associated haplotype. When mutations are present on more than one haplotype, the score statistic $\hat{S}_p$ can still detect this association, but identification of the causal variants will be difficult. Genetic association studies have no power to identify genes with multiple rare mutations on rare haplotypes (Zondervan and Cardon, 2004).

The advantages of the parameter λ are that it is directly related to recombination fraction and is less sensitive to haplotype frequencies than other measures (Devlin and Risch, 1995). However, when allelic association is modelled by means of λ it is not straightforward to adjust for other covariates. Houwing-Duistermaat and Elston (2001) discussed various ways to quantify allelic association and estimate the location of the gene responsible for the disease using logistic regression models. As an alternative to λ, the log relative risk as measured by the regression coefficient in the logistic model may be used to allow for adjustment of other covariates. More research is needed to build this kind of flexible models.

We conclude that overall the score statistic $\hat{S}_p$ has good power regardless of the number of observed haplotypes. When one haplotype is associated, $\hat{Z}_{max}$ performs better, but when two haplotypes are associated, $\hat{Z}_{max}$ performs dramatically bad and $\hat{S}_p$ performs well.

Software

The method described in this paper is implemented in software written in C programming language and it is based on the source of Clump program (Sham and Curtis, 1995). The C program will be available from our Web site (http://clinicalresearch.nl/personalpage/)

4.7 Appendix 1

Pearson's chi-square

Let R be the set of vectors with m coordinates and R_{-m} be the set of vectors with $m-1$ coordinates. Let $\hat{Z}_{-m}$ the vector of the first $m-1$ centered allele counts. Since $\sum_{i=1}^{m} \hat{Z}_i = 0$, it follows

$$\max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u} = \max_{v \in R_{-m}} \frac{(v'\hat{Z}_{-m})^2}{v'Var(\hat{Z}_{-m})v}.$$

The covariance matrix $Var(\hat{Z}_{-m})$ is positive definite, hence applying the extend Cauchy-Schwarz inequality (Johnson and Wichern, 1998) and noting that the maximum is attained when $v \propto Var(\hat{Z}_{-m})^{-1}\hat{Z}_{-m}$ give

$$\max_{v \in R_{-m}} \frac{(v'\hat{Z}_{-m})^2}{v'Var(\hat{Z}_{-m})v} = \hat{Z}'_{-m} Var(\hat{Z}_{-m})^{-1}\hat{Z}_{-m}.$$

After some algebra it follows

$$\max_{u \in R} \frac{(u'\hat{Z})^2}{u'Var(\hat{Z})u} = \sum_{j=1}^{m} \frac{(x_j - n_1\hat{p}_j)^2}{n_1\hat{p}_j} + \sum_{j=1}^{m} \frac{(y_j - n_2\hat{p}_j)^2}{n_2\hat{p}_j}.$$

4.8 Appendix 2

Derivation of the score statistic

The first derivative of the likelihood is

$$\begin{aligned}\frac{\partial}{\partial\lambda}L(x,y|\lambda,p) &= \sum_{i=1}^{m} p_i \frac{\partial}{\partial\lambda}L_i(x,y|\lambda,p)\\ &= \sum_{i=1}^{m}\{x_i(1-p_i)(p_i+\lambda(1-p_i))^{-1}-(n_1-x_i)(1-\lambda)^{-1}\}p_iL_i(x,y|\lambda,p).\end{aligned}$$

Hence the score statistic is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log(L(x,y|\lambda=0,p)) &= \frac{\frac{\partial}{\partial\lambda}(L(x,y|\lambda=0,p))}{L(x,y|\lambda=0,p)}\\ &= \sum_{i=1}^{m}\frac{p_i(x_i-n_1p_i)}{p_i}=0\end{aligned}$$

Therefore, the score statistic can be now obtained by evaluating the second derivative of the log-likelihood with respect to λ at $\lambda = 0$ and using the fact that the first derivative is zero

$$\frac{\partial^2}{\partial\lambda^2}\log(L(x,y|\lambda=0,p)) = \frac{\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda=0,p)}{L(x,y|\lambda=0,p)} \tag{4.5}$$

The second derivative of the likelihood is

$$\begin{aligned}\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda,p) &= \sum_{i=1}^{m}\{-x_i(1-p_i)^2(p_i+\lambda(1-p_i))^{-2}-(n_1-x_i)(1-\lambda)^{-2}\}p_iL_i(x,y|\lambda,p)\\ &+ \sum_{i=1}^{m}\{x_i(1-p_i)(p_i+\lambda(1-p_i))^{-1}-(n_1-x_i)(1-\lambda)^{-1}\}^2p_iL_i(x,y|\lambda,p).\end{aligned}$$

Therefore the second derivative at $\lambda = 0$ is

$$\frac{\partial^2}{\partial\lambda^2}L(x,y|\lambda=0,p)=\{\sum_{i=1}^{m}(x_i^2-x_i)p_i^{-1}+n_1-n_1^2\}L(x,y|\lambda=0,p). \tag{4.6}$$

Combining (4.5) and (4.6)

$$\frac{\partial^2}{\partial\lambda^2}\log(L(x,y|\lambda=0,p)) = \sum_{i=1}^{m}\frac{(x_i-n_1p_i)^2}{p_i}-\sum_{i=1}^{m}\frac{x_i-n_1p_i}{p_i}-n_1(m-1)$$

Derivation of expectation and variance of S_p

The expectation of S_p is

$$E[S_p] = E[X^2] = k - 1.$$

and the variance of $n_1 S_p$ is

$$\begin{aligned}
Var[n_1 S_p] &= \sum_{i,j=1}^{m} \frac{Cov[(x_i - n_1 p_i)^2, (x_j - n_1 p_j)^2]}{p_i p_j} + \sum_{i,j=1}^{m} \frac{E[(x_i - n_1 p_i)(x_j - n_1 p_j)]}{p_i p_j} \\
&- 2\sum_{i,j=1}^{m} \frac{E[(x_i - n_1 p_i)^2 (x_j - n_1 p_j)]}{p_i p_j} \\
&= \sum_{i,j=1}^{m} \frac{-n_1 p_i p_j (1 - 2p_i - 2p_j + 6p_i p_j)}{p_i p_j} + 2\sum_{i,j=1}^{m} \frac{(n_1 p_i p_j)^2}{p_i p_j} \\
&+ \sum_{i=1}^{m} \frac{n_1 p_i (1 - 6p_i + 8p_i^2)}{p_i^2} + 2\sum_{i=1}^{m} \frac{n_1^2 p_i^2 (1 - 2p_i)}{p_i^2} + \sum_{i,j=1}^{m} \frac{-n_1 p_i p_j}{p_i p_j} \\
&+ \sum_{i=1}^{m} \frac{n_1 p_i}{p_i^2} - 2\sum_{i,j=1}^{m} \frac{-n_1 p_i p_j (1 - 2p_i)}{p_i p_j} - 2\sum_{i=1}^{m} \frac{n_1 p_i (1 - 2p_i)}{p_i^2} \\
&= 2n_1 (n_1 - 1)(m - 1)
\end{aligned}$$

Note that

$$\begin{aligned}
Cov[(x_i - n_1 p_i)^2, (x_j - n_1 p_j)^2] &= -n_1 p_i p_j \{(1 - 2p_i - 2p_j + 6p_i p_j) - 2n_1 p_i p_j\} \\
&+ I_{\{j=i\}} n_1 p_i \{(1 - 6p_i + 8p_i^2) + 2n_1 p_i (1 - 2p_i)\} \\
E[(x_i - n_1 p_i)^2 (x_j - n_1 p_j)] &= -n_1 p_i p_j (1 - 2p_i) + I_{\{j=i\}} n_1 p_i (1 - 2p_i) \\
E[(x_i - n_1 p_i)(x_j - n_1 p_j)] &= -n_1 p_i p_j + I_{\{j=i\}} n_1 p_i,
\end{aligned}$$

with $I_{\{j=i\}} = 1$ if $i = j$ otherwise 0. Hence the variance of S_p is

$$Var[S_p] = 2(m - 1)\frac{n_1 - 1}{n_1}$$

Derivation of the asymptotic expectation and variance of $\hat{S}_p$

The expectation and the variance of $\hat{S}_p$ can be given as follows

$$\begin{aligned}
E[\hat{S}_p] &= E[\chi^2] - \frac{n}{n_1 n_2} E[\hat{U}] \\
Var[\hat{S}_p] &= Var[\chi^2] - 2\frac{n}{n_1 n_2} Cov[\chi^2, \hat{U}] + (\frac{n}{n_1 n_2})^2 Var[\hat{U}]
\end{aligned}$$

By means of δ-method it can be shown that

$$\begin{aligned} E[\hat{U}] &\approx 0, \\ Var[\hat{U}] &\approx n^{-1} n_1 n_2 (\sum_{j=1}^{m} \hat{p}_j^{-1} - m^2), \\ Cov[\chi^2, \hat{U}] &= E[\chi^2 \hat{U}] \approx 0. \end{aligned}$$

Hence for $\frac{n}{n_2} \ll \infty$ and $n_1 \to \infty$

$$\begin{aligned} E[\hat{S}_p] &= E[\chi^2] = m - 1 \\ Var[\hat{S}_p] &= Var[\chi^2] = 2(m - 1) \end{aligned}$$