Behavior Genetics, Vol. 34, No. 2, March 2004 (© 2004)

Copulas in QTL Mapping

Bojan Basrak,[1,6] Chris A. J. Klaassen,[2] Marian Beekman,[3] Nick G. Martin,[4] and Dorret I. Boomsma[5]

The standard variance components method for mapping quantitative trait loci is derived on the assumption of normality. Unsurprisingly, statistical tests based on this method do not perform so well if this assumption is not satisfied. We use the statistical concept of copulas to relax the assumption of normality and derive a test that can perform well under any distribution of the continuous trait. In particular, we discuss bivariate normal copulas in the context of sib-pair studies. Our approach is illustrated by a linkage analysis of lipoprotein(a) levels, whose distribution is highly skewed. We demonstrate that the asymptotic critical levels of the test can still be calculated using the interval mapping approach. The new method can be extended to more general pedigrees and multivariate phenotypes in a similar way as the original variance components method.

KEY WORDS: Quantitative trait loci; variance components; normal distribution; copulas; genome scan.

INTRODUCTION

In human genetics the linkage analysis of quantitative trait loci (QTL) tries to detect a connection between genetic similarity at a given marker (commonly measured by identity by descent [IBD] status) and similarity of phenotypes (measured in many different ways). Performing a statistical analysis in such a study, we typically cannot influence the way genetic similarity is measured, but we can choose the way to measure similarity of phenotypes. Most popular procedures use the notion of linear correlation to do so. The correlation is the canonical measure of dependence in the world of (multivariate) normal distributions, but it can be less suitable when the normality assumption is not met. The most general way of expressing stochastic dependence between variables is via copulas. We show how this well-established statistical tool can be applied in QTL linkage analysis with a little extra effort and potentially many benefits. One particular copula, the bivariate normal copula, is discussed in some detail below. In particular, we demonstrate how a statistical analysis based on the normal copula model deals with problems of nonnormality that appear in many practical studies.

Suppose we are given data from a study based on n sib-pairs. We denote the trait values of sib-pairs (phenotypes) by $(Y_{i,1}, Y_{i,2})$ with $i = 1, \ldots, n$. Their IBD status at a marker t is a random variable with values in $\{0, 1, 2\}$ denoted by $X_i(t)$ with $i = 1, \ldots, n$ again. Observe that in the genetics literature $X_i(t)$ are frequently denoted as $2\hat{\pi}_i(t)$. In the sequel we concentrate on one fixed marker (hence we ignore the variable t and just write X_i). Moreover, we ignore uncertainties concerning the measurements of the X_is. The classical method of QTL linkage analysis is due to Haseman and Elston (1972). It suggests regressing the squared difference $(Y_{i,1} - Y_{i,2})^2$ on X_i and declaring linkage whenever one finds evidence for a negative slope of the regression line. One can easily see (as in Sham [1998] for instance) that this boils down to a test whether the correlation corr $(Y_{i,1}, Y_{i,2} \mid X_i)$ can be linearly regressed on X_i with a positive coefficient.

In the last decade, likelihood models have been introduced to obtain more powerful tests for the presence of QTLs when data satisfy additional assumptions. An

[1] University of Zagreb, Eurandom.

[2] University of Amsterdam, Eurandom.

[3] Leiden University Medical Center.

[4] Queensland Institute of Medical Research.

[5] Free University Amsterdam.

[6] To whom correspondence should be addressed at Department of Mathematics, University of Zagreb, Bijenicka 30, 10000 Zagreb, Croatia. E-mail: bbasrak@math.hr

0001-8244/04/0300-0161/0 © 2004 Plenum Publishing Corporation

example of the univariate likelihood model is given in Kruglyak and Lander (1995). Somewhat later, Fulker and Cherny (1996) showed an example of a bivariate model; this approach is commonly known as the variance components method. Both of these likelihood methods test essentially for the very same regression as the Haseman-Elston method, but assuming more about the data, namely univariate or multivariate normality of the trait values. Naturally, these methods have optimal power when their assumptions are met. However, when the trait distribution deviates from normality, neither their power nor their significance level can be guaranteed unless some adjustments are made. This has been an important topic of research in the last couple of years (see for instance Blangero *et al.* [2000] and Sham *et al.* [2000]). For an interesting viewpoint that relates Haseman-Elston and similar methods with variance components see Putter *et al.* (2002).

Remark 1.1 Observe that all of the methods above consider it safe to assume that the marginal distribution of the phenotypes does not change with IBD status, and that it is *only dependence* between them that does. And it is this change in dependence between traits that we want to detect. Moreover, in sib-pair studies it is reasonable to assume that the sibs are randomly ordered, so that the marginal distributions of the traits are equal; that is, $Y_{i,1}$ and $Y_{i,2}$ have the same distribution function. If they are ordered by sex, age, or some other factor, we assume that the factor does not influence the phenotype.

DISCUSSION

Copulas

We have explained how the classical methods of linkage analysis measure dependence between the traits using correlation coefficients. If the multivariate normality assumption does not hold, this is not such a natural idea anymore. It is (almost always) reasonable to assume that we do not have to worry about a change in marginal distribution; thus we can apply an extremely useful tool that statistical theory uses to separate the marginal distributions from the dependence structure—copulas.

We restrict attention to sib-pair studies and hence to the case of bivariate distributions and bivariate copulas (for the more general theory see Joe [1997] or Nelsen [1999]). Let us denote by F the joint distribution function of the random variables Y_1 and Y_2

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2), \qquad y_1, y_2 \in \mathbb{R}.$$

This joint distribution function completely describes the dependence structure as well as the marginal distributions of the pair (Y_1, Y_2).

Assume now that the random variables Y_1 and Y_2 have marginal distribution functions F_1 and F_2, respectively. The copula of the pair (Y_1, Y_2) is defined as the joint distribution function C of the pair $[F_1(Y_1), F_2(Y_2)]$. By the definition of distribution function it follows that if F_1 and F_2 are continuous (which we will assume throughout), then the transformed random variables $F_1(Y_1)$ and $F_2(Y_2)$ both have a uniform distribution on the interval [0, 1]. Consequently, any distribution function of a random vector with values in the unit square $[0, 1] \times [0, 1]$ and with uniform marginal distributions can be viewed as a copula. Note that

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2)), \qquad y_1, y_2 \in \mathbb{R}. \quad (1)$$

From this formula we can see how a joint distribution function "splits into" three parts: the copula C and the marginal distribution functions F_1 and F_2.

Remark 2.1 It is straightforward to show that the copula does not change if we transform each component by a strictly increasing function. In other words, the copula of the random vector $[h_1(Y_1), h_2(Y_2)]$ is the same as the copula of (Y_1, Y_2) for strictly increasing functions h_1 and h_2. The marginal distributions change, however, from (F_1, F_2) to $(F_1 \circ h_1^{-1}, F_2 \circ h_2^{-1})$. For any function h, by h^{-1} we denote its inverse.

One of the most important copulas is the independence copula

$$C_0(u_1, u_2) = u_1 u_2, \qquad u_1, u_2 \in [0, 1],$$

which is obtained whenever the two random variables Y_1 and Y_2 are independent. On the opposite end of the spectrum we have the copula of positive dependence

$$C_+(u_1, u_2) = \min\{u_1, u_2\}, \qquad u_1, u_2 \in [0, 1],$$

which, for instance, can be obtained when $Y_1 = g(Y_2)$ for some strictly increasing function g. Similarly we can define the copula of negative dependence C_-. Observe that copula C_0 has constant (uniform) density on the unit square. On the other hand, copulas C_+ and C_- do not have densities. Their distributions concentrate on the diagonals $u_2 = u_1$ and $u_2 = 1 - u_1$, respectively.

As stated earlier, one can frequently assume that the phenotypic traits of a pair of sibs have the same marginal distribution, which means that we can set $F_1 = F_2$. This restricts the class of copulas we have to consider in our applications even further to the case of the so called exchangeable copulas. Their distributions are symmetric around the diagonal $u_2 = u_1$.

Roughly speaking, in sib-pair studies we expect (in the vicinity of QTLs) that the copula of a pair of phenotypes (Y_1, Y_2) conditioned on their IBD status $X = x$ gets closer and closer to C_+ (and more distant from C_0) as x increases from 0 to 2. But it is still not obvious how to measure this distance in general. This is one of the reasons why we restrict our attention to parametric families of copulas.

The most prominent place in our applications is dedicated to the family of bivariate normal copulas. They arise, in the way explained above, from a random vector (Y_1, Y_2) that has a multivariate normal distribution. These copulas do not depend on the mean and variance of the Y_is but only on their mutual correlation coefficient, ρ. They are equal to C_- and C_+ when $\rho = -1$ or 1, respectively. For $-1 < \rho < 1$, we denote them by $C_N^{\rho}(u_1, u_2)$ and observe that by (1)

$$C_N^{0}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \times \exp\left(\frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\right) \mathrm{d}s\, \mathrm{d}t, \quad (2)$$

where by Φ we denote the standard normal distribution function. This copula has a density as well. Two examples of this density are shown in Figure 1, namely for $\rho = 1/4$ and $\rho = 4/5$.

Recall that the variance components method assumes that the phenotypes $(Y_{i,1}, Y_{i,2})$ conditioned on the IBD values have a bivariate normal distribution. For simplicity we assume further that the random variables $Y_{i,j}, i = 1, \ldots, n, j = 1, 2$, are standardized so that they all have a mean of 0 and variance of 1. It can be shown (see Tang and Siegmund [2002]) that if we estimate expectation and variance of the traits in real-life studies, this does not influence the asymptotic theory of the test statistic (see also the Appendix). To make the assumptions behind the variance components approach more precise, we denote by $F(\cdot, \cdot\,|\,x)$ the conditional distribution of the phenotypes (Y_1, Y_2), given that their IBD status X equals x, that is, $F(y_1, y_2\,|\,x) = P(Y_1 \leq y_1, Y_2 \leq y_2\,|\,X = x)$, and assume

Condition (A): *The conditional distribution function* $F(\cdot, \cdot\,|\,x)$ *is a bivariate normal distribution function with a mean of* 0, *and variance of* σ^2 *(assumed to be equal to* 1 *unless stated otherwise) and a correlation coefficient that depends on* x *as* $\rho(x) = \rho + \gamma(x-1), x = 0, 1, 2$.

Consequently, there is a straightforward likelihood ratio test for the null hypothesis $\gamma = 0$ against the alternative $\gamma > 0$. To make $\rho(x)$ a proper correlation coefficient we need $|\rho| + |\gamma| \leq 1$.

In real-life studies, however, the normality assumption frequently fails to hold even for the univariate variables $Y_{i,1}, Y_{i,2}$. Trying to correct for this, researchers frequently apply some (usually continuous but nonlinear, for instance, logarithmic) transformation to the data to bring them more in line with this assumption. By doing so, they implicitly assume that the bivariate distribution of the traits comes from the normal copula model. In other words, they assume that there is a (strictly monotone) transformation g such that $(Y_{i,1}, Y_{i,2}) = [g^{-1}(W_{i,1}), g^{-1}(W_{i,2})]$ where the pairs $(W_{i,1}, W_{i,2})$ satisfy condition (A). This leads to the following generalization of the previous condition.

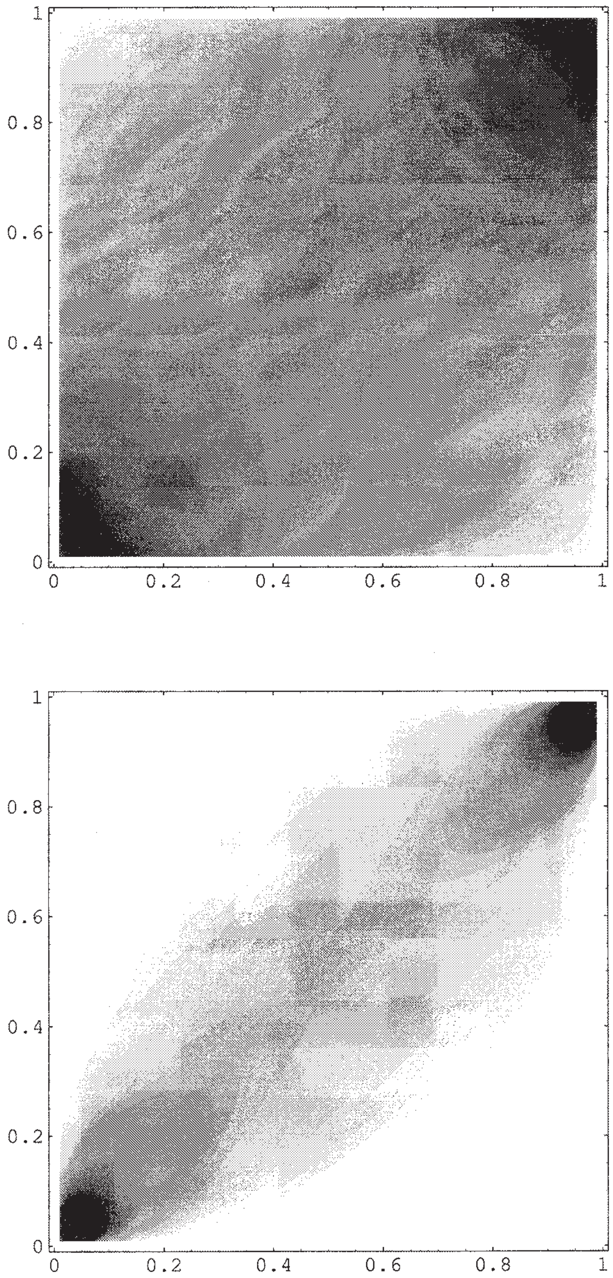

Fig. 1. Grey level intensity plots of densities for copulas $C_N^{0.25}$ and $C_N^{0.8}$.

Condition (B): *There exists a strictly monotone function g such that the distribution function of the random vectors*

$$(W_{i,1}, W_{i,2}) = [g(Y_{i,1}), g(Y_{i,2})]. \qquad (3)$$

conditional on $X_i = x$ satisfies Condition (A).

It follows that the copula $C_{\mathbf{Y}|x}$ of the pair $(Y_{i,1}, Y_{i,2})$ when conditioned on $X_i = x$ is the same as the one for $(W_{i,1}, W_{i,2})$, that is, using the notation of (2) we can write

$$C_{\mathbf{Y}|x} = C_N^{\rho+\gamma(x-1)}. \qquad (4)$$

The marginal distribution of both $Y_{i,1}$ and $Y_{i,2}$ has the form

$$F_1(y) = P(Y_1 \le y) = \Phi[g(y)], \qquad y \in \mathbb{R}. \qquad (5)$$

By (1), the last two formulas completely specify the joint distribution of $(Y_{i,1}, Y_{i,2})$ conditioned on X_i.

Hence the bivariate normal copula model is widely used already. We make it our main assumption in the rest of the article. Note that this model includes the standard variance components model when $g(x) = x$. But it also allows any continuous marginal distribution of the phenotypes. The only assumption it makes concerns the dependence structure between them. Still, there are situations in which such an assumption may not be appropriate. In such circumstances the dependence between traits should be better modeled by some other family of copulas. (Many examples can be found in Nelsen [1999]).

Observe further that by choosing this one-parameter family of copulas, we can measure similarity between phenotypes $Y_{i,1}$ and $Y_{i,2}$ given $X_i = x_i$ by one number again, namely $\rho_i = \rho + \gamma(x_i - 1)$. However, ρ_i represents the correlation between W values and not between Y values. For the latter ones it has an interpretation as the maximum correlation coefficient (see the last paragraph of the Appendix).

In real-life studies the function g in (3) is unknown. One may try to guess g, as one frequently does in practice, but there is another option. If we would know the marginal distribution F_1 of the trait, we could use relation (5) to obtain

$$g(y) = \Phi^{-1}[F_1(y)], \qquad y \in \mathbb{R}. \qquad (6)$$

Hence knowing F_1 means knowing g too. In some cases, assuming that we know F_1 is not unrealistic because the marginal distribution of the traits can be estimated from the larger population that contains the sibs and not only from the data in the study. Frequently F_1 is not known and has to be estimated from the data. An obvious estimator of F_1 is the empirical distribution of all of the $2n$ values $Y_{i,1}, Y_{i,2}, i = 1, \ldots, n$, of the phenotypic trait. Details of this procedure will be explained in the next section.

One can give an alternative explanation for the procedure we advocate, using the concept of van der Waerden normal scores rank correlation coefficient. Readers familiar with this notion will realize that we essentially use this coefficient now to measure similarity between phenotypic traits given their IBD status and not the ordinary linear correlation. Apart from that we leave the variance component approach basically unaltered.

There are other families of copulas that one could, and in some cases should, use in practice. However, the bivariate normal copulas have some obvious advantages: most researchers are familiar with them, even more, they implicitly use them in many studies. Moreover, the commonly used procedures, software, and significance levels can be applied directly.

Copulas in Linkage

Recall that the variance components method assumes that the data satisfy condition (A) and that it tests the hypothesis $\gamma > 0$ using the log-likelihood ratio test statistic

$$2\left(\max_{\rho,\gamma} l(\rho, \gamma) - \max_{\rho} l(\rho, 0)\right) \qquad (7)$$

where l denotes the logarithm of the likelihood of the phenotypes given the values of their IBD status. Sib-pairs are assumed to be independent; thus l is the sum of the contributions of each pair

$$l(\rho, \gamma) = \sum_{i=1}^{n} l(\mathbf{Y}_i \,|\, X_i; \rho, \gamma).$$

Let us denote by $l_\gamma(\cdot\,|\,\cdot; \rho, \gamma)$ the score function of the log-likelihood (i.e., its partial derivative with respect to γ). It is known (see van der Vaart [1998] for instance) that the likelihood ratio test in (7) is locally asymptotically equivalent to the test based on the score statistic

$$Z_n^0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\gamma(\mathbf{Y}_i \,|\, X_i; \hat{\rho}_n, 0)/\sqrt{I_\gamma},$$

where $\hat{\rho}_n$ is the maximum likelihood estimator of ρ, and I_γ denotes the diagonal entry of the Fisher information matrix corresponding to the parameter γ (see Putter *et al.* [2002] or Tang and Siegmund [2002]). In practice, I_γ above is also replaced by an appropriate estimate. It gives a suitable normalization when the assumptions of the model hold. However, in practice it may be advisable to use a “robustified” version of the statistic Z_n^0, that is

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\gamma(\mathbf{Y}_i \,|\, X_i;\, \hat{\rho}_n, 0) \Bigg/ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l_\gamma^2(\mathbf{Y}_i \,|\, X_i;\, \hat{\rho}_n, 0)}. \qquad (8)$$

For a detailed derivation of this statistic see for instance Tang (2000) or Putter *et al.* (2002). Observe that the statistic Z_n has a standard normal distribution asymptotically, even if condition (A) does not hold, as long as l_γ has finite variance and the same mean for each value x_i of the IBD status. Linkage is now concluded whenever Z_n is sufficiently large.

Under the bivariate normal copula model, that is, condition (B), this same procedure can be applied to appropriately transformed phenotypes, that is, to the values [cf. (6)]

$$\mathbf{Y}_i^* = (Y_{i,1}^*, Y_{i,2}^*) = \{\Phi^{-1}[F_1(Y_{i,1})], \Phi^{-1}[F_1(Y_{i,2})]\}, \qquad i = 1, \ldots, n. \qquad (9)$$

Observe that the values $(Y_{i,1}^*, Y_{i,2}^*)$ and X_i satisfy assumption (A) directly, because by Remark 2.1 they have the same copula and the same marginal distribution as the values $(W_{i,1}, W_{i,2})$ given in (3).

As mentioned earlier, if the marginal distribution F_1 of the Ys must be estimated, it is natural to take $\hat{F}_{2n}$ the empirical distribution function of all $2n$ trait values (multiplied by $2n/(2n+1)$ to avoid that it takes the value 1, which would result in $\Phi^{-1}(1) = \infty$) as the estimator. It has the form

$$\hat{F}_{2n}(y) = \frac{1}{2n+1} \#\{Y_{i,k} \le y : i = 1, \ldots, n, k = 1, 2\}.$$

Under our conditions we have with probability one

$$\hat{F}_{2n}(y) \to F_1(y) \qquad \text{for all } y \in \mathbb{R}, \text{ as } n \to \infty,$$

which follows by the strong law of large numbers. The accuracy of $\hat{F}_{2n}$ in estimating F_1 is maximal if all $Y_{i,k}$s are independent. The variance of $\hat{F}_{2n}(y)$ equals $2n(2n+1)^{-2}F_1(y)(1 - F_1(y))$ then. In the other extreme case $Y_{i,1} = Y_{i,2}, i = 1, \ldots, n$, holds and the variance of $\hat{F}_{2n}(y)$ is two times larger. In any case, this justifies the application of the variance components method on the transformed phenotypes

$$\mathbf{Y}_i' = (Y_{i,1}', Y_{i,2}') = \{\Phi^{-1}[\hat{F}_{2n}(Y_{i,1}]), \Phi^{-1}(\hat{F}_{2n}[Y_{i,2})]\}, \qquad i = 1, \ldots, n. \qquad (10)$$

The formula above is not difficult to implement in any software package for data analysis. In particular, an Excel macro performing this transformation is available from the corresponding author on request. It is important to stress that if any of the statistics introduced in (7) or (8) is calculated with these new values, asymptotic significance levels (as those in Dupuis and Siegmund [1999]) stay the same as in the original variance components model (see Proposition 6.3 in the Appendix). They will also give us efficient tests asymptotically. We demonstrate applicability and usefulness of this approach by a small simulation study in the next section.

Real Data and Simulations

We apply the method introduced in the previous section to one particular data set. The phenotypic trait measured is lipoprotein level Lp(a) and the sibs involved are dizygotic twins. This data set is a part of a larger data set produced in an international study involving twins from Australia. The Netherlands, and Sweden. Details of the study can be found in Beekman *et al.* (2002). To illustrate the normal copula method we restrict ourselves to the Australian sample and chromosomes 1 and 6. We ignore the sex of the sibs, because Lp(a) levels and variances do not systematically vary with sex. The first histogram in Figure 2 shows that the Lp(a) levels have a distribution that is extremely skewed. Therefore the levels have been transformed by a classical device—the natural logarithm. The resulting histogram (see Figure 2[b]) seems to indicate that skewness is not a serious problem anymore, but the distribution of the transformed values is still far from normal. This can be checked by a rigorous test but it is also clear just from looking at the *QQ*-plot in Figure 2(c). If we perform the transformation by the empirical distribution function given in (10) the marginal distribution of the data is very close to normal; see the histogram in Figure 2(d). In fact, the ordered components of the transformed data are the deterministic numbers $\Phi^{-1}[1/(2n+1)], \ldots, \Phi^{-1}[2n/(2n+1)]$. The remaining randomness in (10) is in the pairing of these numbers.

We have performed three tests over a given set of markers. The first one is the classical Haseman-Elston test performed on the logarithms of the original data, the second one is the log-likelihood ratio test performed on the same values, and the third test is the same as the second one, but it uses the normal copula approach to transform the data. For this illustration, we have used the estimated expectation of the IBD status (usually called $\hat{\pi}$ values) of the twins and not the estimated IBD probabilities.

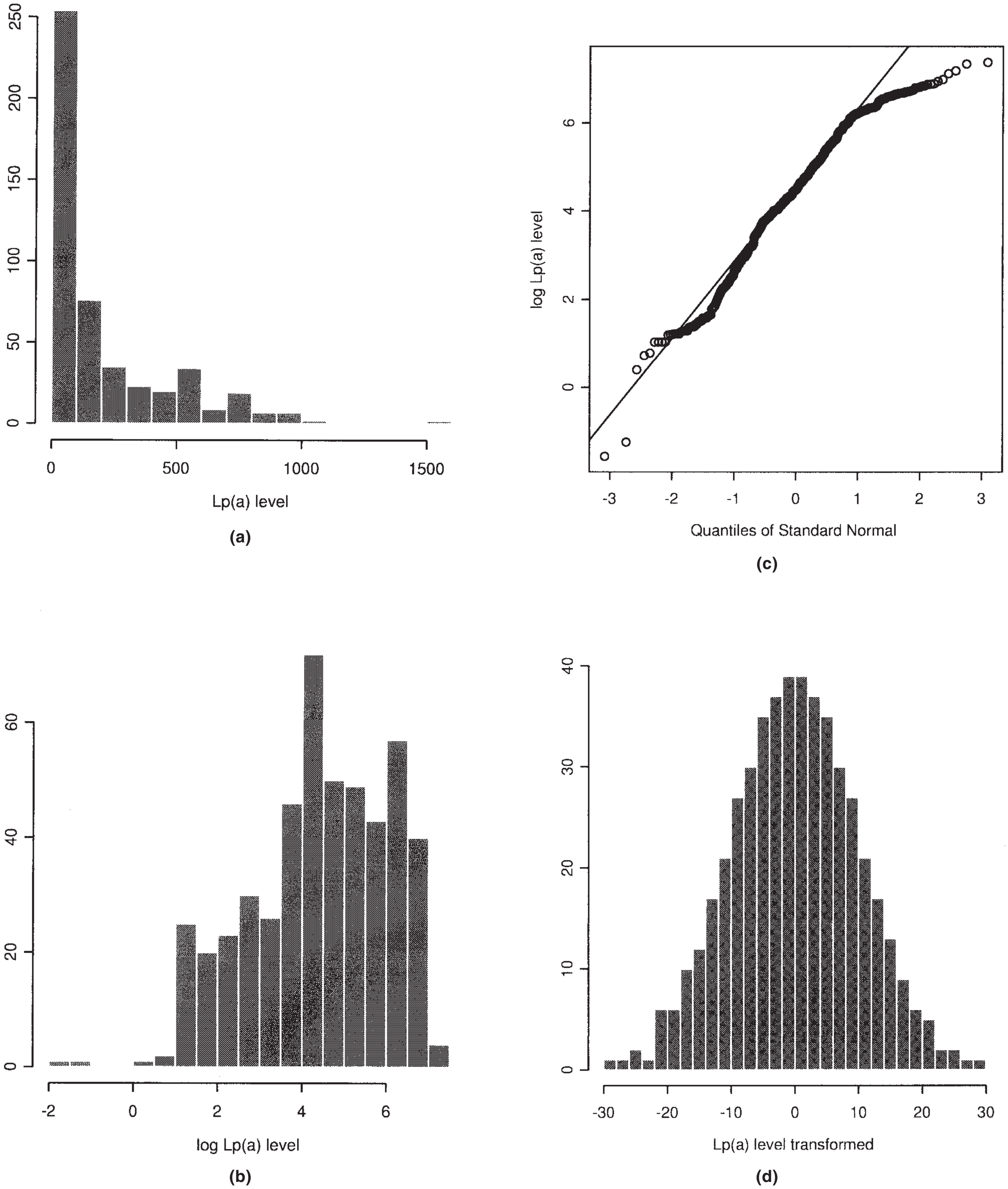

Fig. 2. (a) Histograms of lipoprotein levels, (b) histogram of their logarithms, (c) *QQ* plot of the logarithms against the normal distribution, and (d) histogram of the values transformed nonparametrically using formula (10).

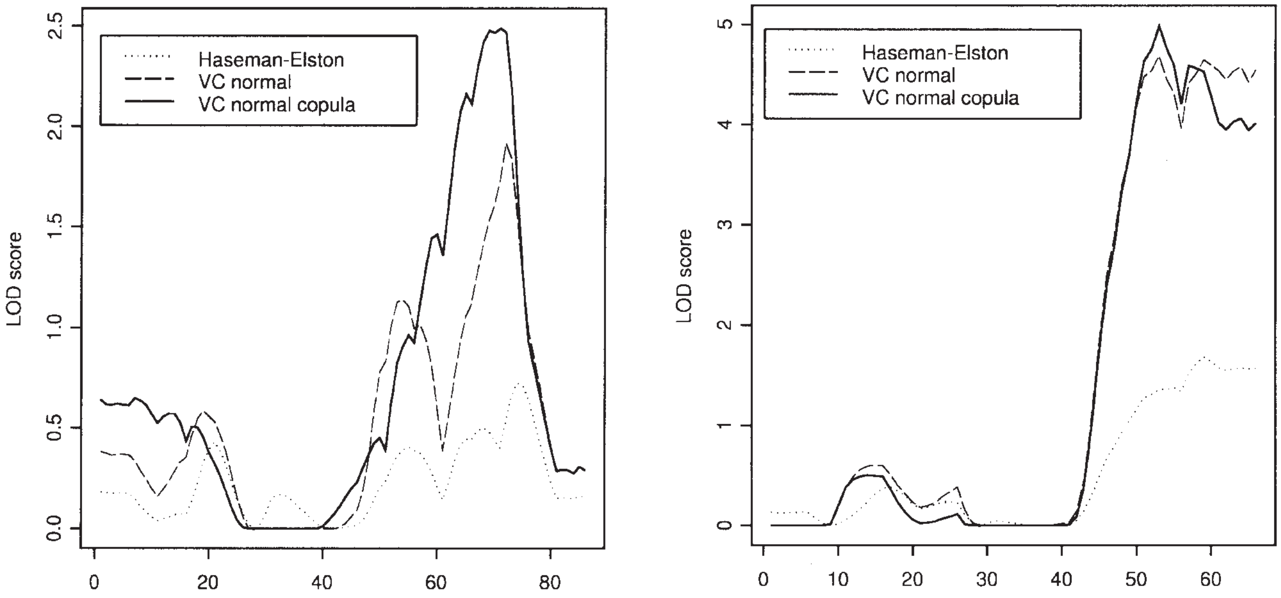

Fig. 3. Three test statistics plotted on the LOD scale over chromosome 1 (left) and chromosome 6 (right).

For both chromosomes all three tests achieve their maximum at approximately the same location, as can be seen in Figure 3. In both cases the copula–based test has the highest LOD score at the location of suspected QTL (i.e., the location of the maximum). Note that it also gives less significance (i.e., the smaller LOD score) to the second largest local maximum of the LOD score based on the usual variance component test. Loosely speaking, this might mean that the copula–based test distinguishes better between "true" and "false" QTLs. We would like to stress that these results change if we calculate LOD scores conditionally on the QTL at the other chromosome. In that case, only the known Lp(a) locus at chromosome 6 appears to be significant (Figure 4).

We have also performed a small simulation study to compare the powers of the different test procedures. It is based on 1000 simulations of 200 pairs of phenotypes

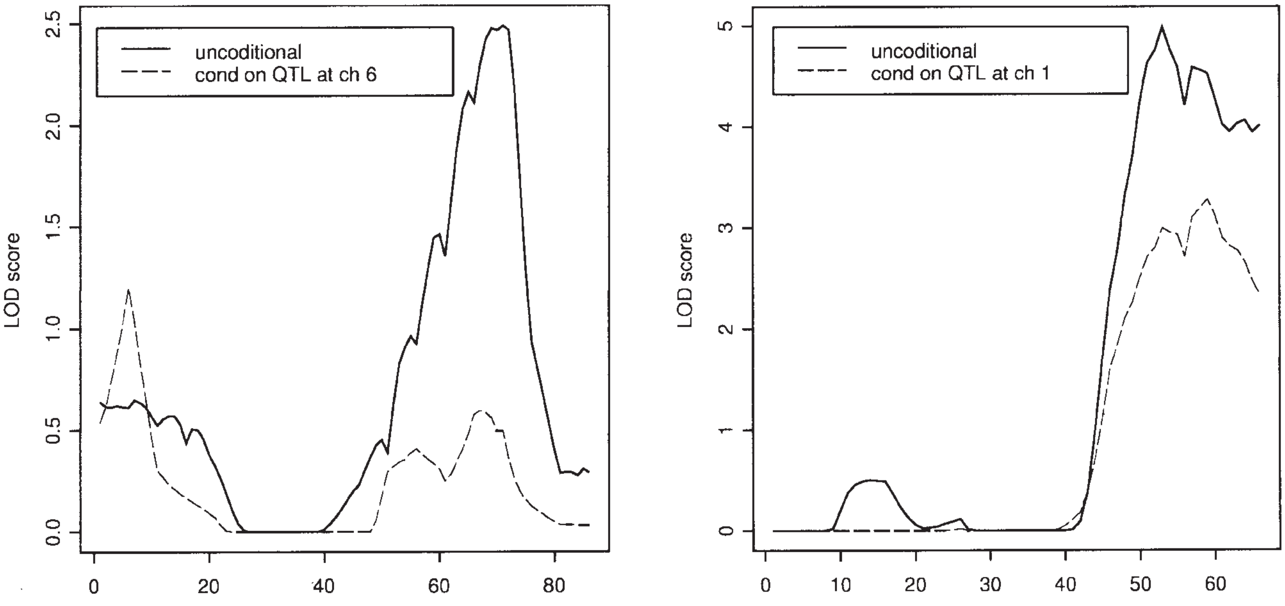

Fig. 4. Conditional test statistics plotted on the LOD scale over chromosome 1 (left) and chromosome 6 (right).

Table I. Power Estimates from 1000 Independent Simulations

	S_a	S_b	S_c	mLOD	S_a	S_b	S_c	mLOD	S_a	S_b	S_c	mLOD
	$\rho = 0.2, \gamma = 0.0$				$\rho = 0.3, \gamma = 0.1$				$\rho = 0.4, \gamma = 0.2$			
LLR	4.5	0.0	0	0.097	31.0	7.4	0.1	0.492	82.7	53.7	6.8	1.732
C-LLR	5.9	0.1	0	0.114	32.3	7.8	0.2	0.507	82.1	52.5	7.9	1.723
H-E	5.8	0.1	0	0.116	30.9	5.8	0.0	0.474	80.9	40.6	1.0	1.382
Z score	4.5	0.4	0	0.100	27.2	5.8	0.1	0.450	75.0	38.5	2.6	1.348
H-E after g_1	6.1	0.3	0	0.114	23.0	3.8	0.0	0.387	64.4	24.1	0.5	0.997
Z after g_1	4.5	0.4	0	0.104	26.5	4.8	0.0	0.416	66.7	32.4	2.1	1.184
H-E after g_2	4.7	0.2	0	0.110	10.9	0.8	0.0	0.235	15.7	0.9	0.0	0.272
Z after g_2	4.7	0.1	0	0.112	14.8	0.4	0.0	0.281	29.1	1.4	0.0	0.422

Note: All test statistics are calculated on the LOD scale. Columns S_a, S_b, and S_c contain the percentages of LOD scores that exceed levels $a = 0.59$, $b = 1.5$, and $c = 3.62$, respectively. The column mLOD contains the mean LOD score in each case.
LLR, log-likelihood ratio statistic (7); *C-LLR*, copula–based log-likelihood ratio statistic; *Z*, score test statistic (8); H-E, Haseman-Elston test statistic calculated on the original data. The last two are recalculated after two nonlinear transformations (g_1 and g_2) of the same data.

and their IBD values at a fixed QTL. They are generated from the standard variance components model for three different sets of parameters. More precisely, the distribution of the pairs satisfies Condition (A) with different values of ρ and γ. After that, we performed the usual tests: the log-likelihood ratio test, see (7), the Haseman-Elston test, and the score test, see (8). We present the results based on the 1000 simulation runs in Table I. It gives the percentages of the LOD scores that exceed levels $a = 0.59$, $b = 1.5$, and $c = 3.62$, respectively. Note that a and c are asymptotic critical thresholds at the significance level $\alpha = 0.05$ for the single marker test and the genome-wide scan. The first set of parameters ($\rho = 0.2$ and $\gamma = 0$) is chosen to explore behavior of different test statistics under the null hypothesis of no linkage.

Finally, we transformed the simulated data using two nonlinear functions. We did this by taking the cube root and the cube of the generated phenotypes and restandardizing them to have mean 0 and variance 1. Observe that the transformed phenotypes come from the bivariate copula model, that is, they satisfy Condition (B). On the transformed data we applied the Haseman-Elston method and the "robustified" score test (8). They both exhibit a decrease in power to detect this QTL now. However, for the copula–based approach this is not a problem because its results stay the same when the data are transformed by an increasing function. One can see this by comparing the rows of the table corresponding to the log-likelihood ratio (LLR) test statistic based on normality and the same statistic applied on the nonparametrically transformed phenotypes (C-LLR) [see (10)]. Observe that under the null hypothesis $\gamma = 0$ all tests have similar empirical type 1 error rates. This suggests that by estimation of the marginal distribution function, we do not inflate the type 1 error, at least when the sample size is about 200 or more. Moreover, Table I shows that after a transformation like g_2 the performance of the Haseman-Elston method and the score test Z can be rather poor.

Observe that all of our samples satisfy Condition (B). Admittedly, it is also important to investigate the behavior of the new method when this assumption fails. However, the class of distributions for which Condition (B) does not hold is extremely large and disordered. Moreover, simulation from a general copula is not a completely trivial issue. On the other hand, choosing only copulas from which one can easily simulate may not be very illustrative. This is certainly a topic that deserves more attention.

CONCLUSION

The bivariate normal copula model suggested in Condition (B) is well studied in the statistics literature (e.g., Klaassen and Wellner [1999]). We are convinced that it can be successfully applied in practical QTL analysis, in particular when the traits have marginal distributions that are very far from normal. Researchers who perform ad hoc transformations of the traits to make them comply with the model behind the variance components method in fact implicitly accept the validity of the normal copula model. The normal copula model includes the variance components model, but it also allows any (continuous) marginal distribution of the phenotypes. Its only restrictions concern the dependence structure between traits. Note, however, that the assumptions of Condition (B) are not always justified. In such a case, one might explore other families of copulas. Finally, the marginal distribution function could be more precisely estimated using not only genotyped sib-pairs but all available phenotype data from the population, thus improving on $\hat{F}_{2n}$ from (10). When such an estimator is available, it should be applied as in (10), and the resulting copula–based analysis is even more powerful then. In particular, this method might be very useful in the case of selected samples.

We have illustrated application of the copula based method in the case of independent sib-pair studies, but the method is readily extendable to different pedigrees in the same way as the variance components method. The assumption of additivity of the trait can be relaxed by including a dominance effect as well. The method performs a simple ranks-based transformation of the data and then applies the usual test procedures; therefore it can be easily applied using any statistical software that supports the variance components approach.

In linkage analysis of QTLs we typically need to adjust the critical values because of multiple testing issues. Recall that we usually test by checking if $\max_t Z_n(t) > b$ where $Z_n(t)$ are test statistics, where the values t belong to a given set of markers, and where b is a suitably chosen critical value. For a dense set of markers, the asymptotic theory of Lander and Botstein (1989) (see also Dupuis and Siegmund [1999]) relates probabilities of exceedance of score statistics over large thresholds with the distribution of maxima of a certain stochastic process (Ornstein-Uhlenbeck process) under usual assumptions. Because one can show that the convergence in Proposition 6.3 in the Appendix holds not only for each fixed marker, but also at the level of processes, it follows that these asymptotic thresholds and p values apply unaltered to the same statistic applied to the transformed data. In particular, asymptotic critical values for the score statistic (Z_n' in the Appendix) in genome-wide human studies with significance level $\alpha = 0.05$ stay at $b = 4.08$ or 3.62 on the LOD scale. Similarly, when the markers are equally spaced, the theory of Feingold *et al.* (1993) applies directly. Of course, one can apply Monte Carlo simulations to obtain more precise p values empirically.

APPENDIX

The main result of this section is contained in Proposition 6.3. Roughly speaking, it states that asymptotically the critical values that are used for the test in the variance components method remain the same if we apply the more general bivariate normal copula discussed in the text. Observe first that the score function $l_\gamma(\mathbf{Y}\,|\,x; \rho, 0)$ used in the statistic Z_n defined in (8) has the following form

$$l_\gamma(\mathbf{Y}_i\,|\,X_i; \rho, 0) = (X_i - 1)h(\mathbf{Y}_i, \rho),$$

where h is defined by

$$\begin{aligned} h(\mathbf{Y}_i, \rho) &= h[(Y_{i,1}, Y_{i,2}), \rho] \\ &= \frac{\rho}{1-\rho^2} + \frac{S_i^2}{2(1+\rho)^2} - \frac{D_i^2}{2(1-\rho)^2} \end{aligned}$$

with $S_i = (Y_{i,1} + Y_{i,2})/\sqrt{2}$ and $D_i = (Y_{i,1} - Y_{i,2})/\sqrt{2}$. Write

$$Z_n(\mathbf{Y}, \mathbf{X}, \rho) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\gamma(\mathbf{Y}_i\,|\,X_i; \rho, 0) \Bigg/ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l_\gamma^2(\mathbf{Y}_i\,|\,X_i; \rho, 0)}.$$

and recall that $Z_n = Z_n(\mathbf{Y}, \mathbf{X}, \hat{\rho}_n)$. Hence, in the statistic Z_n we approximate ρ by its sample version $\hat{\rho}_n$. Our first lemma claims that this does not influence the asymptotic behavior of the statistic Z_n. By $\xrightarrow{P}$ we denote convergence in probability.

Lemma 6.1: *Let the conditional distribution of the traits* $\mathbf{Y}_i$ *satisfy Condition (A) with* $\gamma = 0$ *and* $|\rho| < 1$. *If* $\hat{\rho}_n$ *converges to* ρ *in probability, we have*

$$Z_n(\mathbf{Y}, \mathbf{X}, \hat{\rho}_n) - Z_n(\mathbf{Y}, \mathbf{X}, \rho) \xrightarrow{P} 0.$$

Proof: We observe that the statistic

$$\frac{1}{n} \sum_{i=1}^{n} l_\gamma^2(\mathbf{Y}_i\,|\,X_i; \rho, 0)$$

converges to the same constant if we substitute ρ by $\hat{\rho}_n$ as long as $\hat{\rho}_n \xrightarrow{P} \rho$, because l_γ is a differentiable function of ρ with a sufficiently well-behaved derivative for $|\rho| < 1$. So, it suffices to consider the numerator of $Z_n(\mathbf{Y}, \mathbf{X}, \rho)$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1) \left(\frac{\rho}{1-\rho^2} + \frac{S_i^2}{2(1+\rho)^2} - \frac{D_i^2}{2(1-\rho)^2} \right).$$

Observe now that we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1) S_i^2 \left(\frac{1}{2(1+\rho)^2} - \frac{1}{2(1+\hat{\rho}_n)^2} \right) \xrightarrow{P} 0,$$

as may be seen by considering the second moment of this sum and taking into account the independence between X_is and $Y_{i,j}$s. Because the other terms in the difference $Z_n(\mathbf{Y}, \mathbf{X}, \hat{\rho}_n) - Z_n(\mathbf{Y}, \mathbf{X}, \rho)$ can be treated similarly the statement of the lemma follows.

Subsequently, we have to show that by using the values $\mathbf{Y}_i'$ from (10) instead of $\mathbf{Y}_i^*$ from (9) we do not change the asymptotic behavior of the test statistic.

Lemma 6.2: *Under Condition (B) and the null hypothesis* $\gamma = 0$

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}(X_i - 1)h(\mathbf{Y}'_i, \rho) - \sum_{i=1}^{n}(X_i - 1)h(\mathbf{Y}^*_i, \rho)\right) \xrightarrow{P} 0.$$

Proof: The statement of the lemma follows immediately if we can show that the second moment of the expression on the left-hand side above converges to 0. This second moment equals

$$\begin{aligned} &E(X_1 - 1)^2 E[h(\mathbf{Y}'_1, \rho) - h(\mathbf{Y}^*_1, \rho)]^2 \\ &= \tfrac{1}{2}E[h(\Phi^{-1}(\hat{F}_{2n}(Y_{1,1})), \Phi^{-1}(\hat{F}_{2n}(Y_{1,2}))) \\ &\quad - h(\Phi^{-1}(F_1(Y_{1,1})), \Phi^{-1}(F_1(Y_{1,2})))]^2, \end{aligned}$$

where we have used that under the null hypothesis, the following holds

$$\begin{aligned} &E[(X_i - 1)(X_j - 1)(h(\mathbf{Y}'_j, \rho) \\ &\quad - h(\mathbf{Y}^*_j, \rho))(h(\mathbf{Y}'_j, \rho) - h(\mathbf{Y}^*_j, \rho))] = 0, \quad \text{for } i \neq j. \end{aligned}$$

To show that the expectation above converges to 0, note again that $\hat{F}_{2n} \to F_1$ pointwise with probability 1. Therefore we just need to show the uniform square integrability of

$$h(\Phi^{-1}[\hat{F}_{2n}(Y_{1,1})], \Phi^{-1}[\hat{F}_{2n}(Y_{1,2})])$$

under the null hypothesis. Because of the form of the function h, it is sufficient to show that the random variables

$$\Phi^{-1}[\hat{F}_{2n}(Y_{1,1})]\Phi^{-1}[\hat{F}_{2n}(Y_{1,2})] \quad \text{and} \quad \Phi^{-1}[\hat{F}_{2n}(Y_{1,1})]$$

are uniformly square integrable. Let us consider only the first of these because the second one is easier to analyze. Uniform integrability follows if we can show

$$\sup_n E\,|\,\Phi^{-1}[\hat{F}_{2n}(Y_{1,1})]\Phi^{-1}[\hat{F}_{2n}(Y_{1,2})]\,|^{2+\varepsilon} < \infty,$$

for some $\varepsilon > 0$. By the Cauchy-Schwarz inequality, it is sufficient to show

$$\sup_n E(\Phi^{-1}[\hat{F}_{2n}(Y_{1,1})])^{2(2+\varepsilon)} < \infty.$$

Observe that $\hat{F}_{2n}(Y_{1,1})$ is a random variable with a uniform distribution on the values $[k/(2n + 1) : k = 1, \dots, 2n]$. The claim now follows from the fact that

$$\frac{1}{2n}\sum_{k=1}^{2n}[\Phi^{-1}(k/2n + 1)]^{2(2+\varepsilon)} \to EN^{2(2+\varepsilon)} < \infty$$

for a standard normal random variable N and any $\varepsilon > 0$.

If we calculate the statistic Z_n using the values $\mathbf{Y}^*_i$ and $\mathbf{Y}'_i$, respectively, it follows from the two lemmas above that these two statistics have the same limiting behavior. To see this, denote the sample correlations based on the sequences $(\mathbf{Y}^*_i)$ and $(\mathbf{Y}'_i)$ by

$$\hat{\rho}'_n = \frac{n^{-1}\sum_{i=1}^{n} Y'_{i,1}Y'_{i,2}}{\sqrt{n^{-1}\sum_{i=1}^{n}(Y'_{i,1})^2 \cdot n^{-1}\sum_{i=1}^{n}(Y'_{i,2})^2}}$$

and

$$\hat{\rho}^*_n = \frac{n^{-1}\sum_{i=1}^{n} Y^*_{i,1}Y^*_{i,2}}{\sqrt{n^{-1}\sum_{i=1}^{n}(Y^*_{i,1})^2 \cdot n^{-1}\sum_{i=1}^{n}(Y^*_{i,2})^2}}.$$

Observe that by the strong law of large numbers $\hat{\rho}^*_n \to \rho$ with probability 1. To show that the same holds for $\hat{\rho}'$, we can use a similar argument as in the proof of Lemma 6.2. Note, for instance, that the sample covariances

$$\hat{c}'_n = n^{-1}\sum_{i=1}^{n} Y'_{i,1}Y'_{i,2} \quad \text{and} \quad \hat{c}^*_n = n^{-1}\sum_{i=1}^{n} Y^*_{i,1}Y^*_{i,2}$$

satisfy $\hat{c}'_n - \hat{c}^*_n \xrightarrow{P} 0$, simply because

$$E|Y'_{i,1}Y'_{i,2} - Y^*_{i,1}Y^*_{i,2}|^2 \to 0$$

holds by the proof of Lemma 6.2. A similar result holds for the sample variances. So we may conclude $\hat{\rho}'_n \xrightarrow{P} \rho$.

Proposition 6.3: *Under condition (B) and the null hypothesis* $\gamma = 0$

$$Z_n(\mathbf{Y}^*, \mathbf{X}, \hat{\rho}^*_n) - Z_n(\mathbf{Y}', \mathbf{X}, \hat{\rho}'_n) \xrightarrow{P} 0. \tag{11}$$

Proof: As we have shown above $\hat{\rho}^*_n, \hat{\rho}'_n \xrightarrow{P} \rho$. So by Lemma 6.1 we can use ρ instead of its estimators in the definition of Z^*_n and Z'_n. In Lemma 6.2 we have shown that the difference of the numerators in the two statistics converges to 0 in probability. It is enough to show that the denominators satisfy

$$\frac{1}{n}\left(\sum_{i=1}^{n} l^2_\gamma(\mathbf{Y}'_i\,|\,X_i; \rho, 0) - \sum_{i=1}^{n} l^2_\gamma(\mathbf{Y}^*_i\,|\,X_i; \rho, 0)\right) \xrightarrow{P} 0.$$

But this follows by exactly the same method as used in the proof of Lemma 6.2.

It is possible to give yet another interpretation of the correlation ρ that is estimated by $\hat{\rho}'_n$ above. For random variables Y_1 and Y_2 we denote the correlation between them by $\rho(Y_1, Y_2)$. However, we may also consider the correlation of $a(Y_1)$ and $b(Y_2)$ for any real transformations a and b such that $0 < \text{var}[a(Y_1)]$, $\text{var}[b(Y_2)] < \infty$. If we take a supremum over all these transformations we get the maximum correlation coefficient of the pair Y_1 and Y_2, namely

$$\rho_M(Y_1, Y_2) = \sup_{a,b} \rho[a(Y_1), b(Y_2)].$$

It is known that for the bivariate normal copula model given in (3) we have $\rho_M = |\rho| = |\rho(W_1, W_2)|$. In other words, the van der Waerden normal scores rank correlation coefficient ρ'_n is also an estimator of the maximum correlation coefficient between phenotypic traits. The properties of this estimator are studied in Klaassen and Wellner (1997). They also show that ρ'_n is an asymptotically efficient estimator of ρ.

ACKNOWLEDGMENTS

We are grateful to L. Beem and M.C.M. de Gunst for careful reading of the manuscript and many interesting and useful suggestions.

REFERENCES

Beekman, M., Heijmans, B. T., Martin, N. G., Pedersen, N. L., Whitfield, A. B., DeFaire, U., van Baal, G. C., Snieder, H., Vogler, G. P. *et al.* (2002). Heritabilities of apolipoprotein and lipid levels in three countries. *Twin Res.* **5**:87–97.

Blangero, J., Williams, J. T., and Almasy, L. (2000). Robust LOD scores for variance component-based linkage analysis. *Genet. Epidemiol.* **19**(Suppl. 1):S8–S14.

Dupuis, J., and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**:373–386.

Feingold, E., Brown, P. O., and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* **53**: 234–251.

Fulker, D. W., and Cherny, S. S. (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behav. Genet.* **26**: 527–532.

Haseman, J. K., and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**:3–19.

Joe, H. (1997). *Multivariate models and dependence concepts.* Monographs on Statistics and Applied Probability, 73, London: Chapmann & Hall.

Klaassen, C. A. J., and Wellner, J. A. (1997). Efficient estimation in the bivariate copula model: Normal margins are least favorable. *Bernoulli* **3**:55–77.

Kruglyak, L., and Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**:439–454.

Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**:185–199.

Nelsen, R. B. (1999). *An introduction to copulas.* Lecture notes in statistics, Vol. 139. New York: Springer-Verlag.

Putter, H., Sandkuijl, L. A., and van Houwelingen, J. C. (2002). Score test for detecting linkage to quantititative traits. *Genet. Epidemiol.* **22**:345–355.

Sham, P. (1998). *Statistics in human genetics.* London: Arnold.

Sham, P. C., Zhao, J. H., Cherny, S. S., and Hewitt, J. K. (2000). Variance components QTL linkage analysis of selected and non-normal samples: Conditioning on trait values. *Genet. Epidemiol.* **19**:S22–S28.

Tang, H. K. (2000). Using variance components to map quantitative trait loci in human. Ph.D. thesis. Stanford, CA: Stanford University.

Tang, H. K., and Siegmund, D. (2002). Mapping multiple genes for quantitative or complex traits. *Genet. Epidemiol.* **22**:313–327.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge: Cambridge University Press.