Thorsten Dickhaus*, Jens Stange and Haydar Demirhan

On an extended interpretation of linkage disequilibrium in genetic case-control association studies

DOI 10.1515/sagmb-2015-0024

Abstract: We are concerned with statistical inference for $2\times C\times K$ contingency tables in the context of genetic case-control association studies. Multivariate methods based on asymptotic Gaussianity of vectors of test statistics require information about the asymptotic correlation structure among these test statistics under the global null hypothesis. In the case of $C=2$, we show that for a wide variety of test statistics this asymptotic correlation structure is given by the standardized linkage disequilibrium matrix of the K loci under investigation. Three popular choices of test statistics are discussed for illustration. In the case of $C=3$, the standardized composite linkage disequilibrium matrix is the limiting correlation matrix of the K locus-specific Cochran-Armitage trend test statistics.

Keywords: asymptotic Gaussianity; chi-squared statistic; Cochran-Armitage trend test; contingency table; correlation structure; Delta method; Fisher's exact test; odds ratio.

1 Introduction

Multivariate statistical methods based on asymptotic Gaussianity of test statistics are receiving more and more attention in the context of multiple test problems in genetics; see, e.g. Conneely and Boehnke (2007), Moskvina and Schmidt (2008), Dickhaus and Stange (2013), and Part II of Dickhaus (2014). The reason is that incorporating the (asymptotic) correlation structure of test statistics in the statistical analysis leads to an improvement of statistical power in comparison with a locus-by-locus analysis in combination with a (for instance, Bonferroni) correction for multiplicity. In genetics, correlations between the (expected) allele frequencies at genomic positions (loci) in the same chromosome are technically described by linkage disequilibrium (LD); see, for example, Chapter 10 of Ziegler and König (2006).

Several measures of LD have been introduced in the literature; cf., e.g. Devlin and Risch (1995). In this work, we focus on the standardized coefficient of LD of two loci i and j, which is also commonly referred to as the signed square root of the r^2 LD measure, and denote it by LD(i, j) throughout the remainder. LD matrices for several target populations are publicly available from databases like those of The International HapMap Consortium (2005) or The 1000 Genomes Consortium (2010). Hence, they may be regarded as external structural information in the context of frequentist inference or as prior information in the context of Bayesian inference. We show that LD(i, j) has a much broader interpretation when testing for association in $2\times 2\times K$ contingency tables occurring in allelic case-control studies, where the total number of loci under consideration is equal to K. Namely, under the null hypothesis of no associations between allelic status and binary phenotype of interest, it coincides with the asymptotic Pearson correlation coefficient of T_i and T_j [denoted by $\rho(T_i, T_j)$] for a wide variety of different test statistics T_i and T_j, which are commonly used for testing association in the

***Corresponding author: Thorsten Dickhaus,** Institute for Statistics, University of Bremen, P. O. Box 330 440, D-28344 Bremen, Germany, e-mail: dickhaus@uni-bremen.de
Jens Stange: Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
Haydar Demirhan: Hacettepe University, Department of Statistics, Ankara, Turkey

marginal contingency tables i and j, respectively. The argumentation is based on an application of the Delta method. Furthermore, we apply this method also to the case of $2\times3\times K$ contingency tables, where associations of bi-allelic genotypes (instead of alleles) with the phenotype of interest are analyzed. Here, the composite (standardized) LD matrix of the K loci under consideration describes the limiting correlation structure.

The rest of the paper is structured as follows. In Section 2, we introduce basic notation and general assumptions. Section 3 contains our results regarding allelic association test problems as well as three examples. Section 4 covers the case of genotypic association tests, with particular emphasis on Cochran-Armitage trend test statistics. We conclude with a discussion in Section 5. Some auxiliary results needed for Example 2 are deferred to the Appendix.

2 Notation and preliminaries

Throughout the work, we let i and j denote two genomic positions in the same chromosome. We assume that the data collected for testing association between the allelic status of the respective locus and a given binary phenotype can be summarized in two contingency tables which are as in Table 1 (allelic association test problem) or in Table 2 (genotypic association test problem), respectively, where A (B) denotes the major allele at locus i (j) and a (b) the corresponding minor allele. The numbers $n_{1.}$ (phenotype 1, corresponding to cases) and $n_{2.}$ (phenotype 0, corresponding to controls) do not depend on the genomic position and are fixed by experimental design. Furthermore, we assume that all observational units (alleles or genotypes, respectively) have been sampled independently of each other from the same target population. For alleles, such sampling requires the Hardy-Weinberg equilibrium assumption, so that sampling an individual is equivalent to randomly sampling two haplotypes.

In all asymptotic considerations, we assume for convenience that

$$\lim_{N\to\infty} n_{1.}/N=\tau\in(0,\,1).$$

3 Allelic association test problems

Notice that, conditional to all four marginal counts $n_{1.}$, $n_{2.}$, $n_{.1}^{(\gamma)}$, and $n_{.2}^{(\gamma)}$, the (2×2) contingency table for locus γ can be reconstructed from $x_{11}^{(\gamma)}$ alone, $\gamma\in\{i,\,j\}$; cf. Table 1. Hence, conditionally to these marginal counts $X_{11}^{(\gamma)}$ is a (marginally) sufficient statistic for contingency table γ, where the capitalized notation indicates that the cell entry is regarded as a random variable. This is why essentially all (marginal) allelic asso-

Table 1: Schematic representation of data for an allelic association test problem at two genetic loci i and j.

Allele	A	a	Total	B	b	Total
Phenotype 1	$x_{11}^{(i)}$	$x_{12}^{(i)}$	$n_{1.}$	$x_{11}^{(j)}$	$x_{12}^{(j)}$	$n_{1.}$
Phenotype 0	$x_{21}^{(i)}$	$x_{22}^{(i)}$	$n_{2.}$	$x_{21}^{(j)}$	$x_{22}^{(j)}$	$n_{2.}$
Absolute count	$n_{.1}^{(i)}$	$n_{.2}^{(i)}$	N	$n_{.1}^{(j)}$	$n_{.2}^{(j)}$	N

Table 2: Schematic representation of data for a genotypic association test problem at two genetic loci i and j.

Genotype	AA	Aa	aa	Total	BB	Bb	bb	Total
Phenotype 1	$x_{11}^{(i)}$	$x_{12}^{(i)}$	$x_{13}^{(i)}$	$n_{1.}$	$x_{11}^{(j)}$	$x_{12}^{(j)}$	$x_{13}^{(j)}$	$n_{1.}$
Phenotype 0	$x_{21}^{(i)}$	$x_{22}^{(i)}$	$x_{23}^{(i)}$	$n_{2.}$	$x_{21}^{(j)}$	$x_{22}^{(j)}$	$x_{23}^{(j)}$	$n_{2.}$
Absolute count	$n_{.1}^{(i)}$	$n_{.2}^{(i)}$	$n_{.3}^{(i)}$	N	$n_{.1}^{(j)}$	$n_{.2}^{(j)}$	$n_{.3}^{(j)}$	N

ciation tests which are commonly used in practice employ some transformation of $X_{11}^{(\gamma)}$ as a test statistic in contingency table γ. Letting f_γ denote such a (smooth) transformation, we will show in this section that asymptotically ($N\to\infty$) the correlation coefficient of the test statistics $T_i=f_i(X_{11}^{(i)})$ and $T_j=f_j(X_{11}^{(j)})$ equals LD(i, j), provided that the null hypothesis of no association is true for both loci. This result has important consequences, because it enables one to carry out the multiple association test for all K loci simultaneously as a multivariate procedure which takes the asymptotic correlation structure among the locus-specific test statistics into account in the calibration with respect to type I error control.

Lemma 1. *Assume that the null hypothesis of no association between phenotype and allelic status is fulfilled at both loci i and j. Let $\gamma\in\{i, j\}$ and denote by p_γ the (expected) major allele frequency in the target population at locus γ. Finally, denote by p_{ij} the expected frequency of the occurrence of the major alleles at both loci i and j. Then, the following assertions hold true.*

(*a*) $X_{11}^{(\gamma)}\sim Bin(n_{1.}, p_\gamma)$.

(*b*) $Cov(X_{11}^{(i)}, X_{11}^{(j)})=n_{1.}\mathscr{D}_{AB}$, *where* $\mathscr{D}_{AB}=p_{ij}-p_ip_j$.

(*c*) $\rho(X_{11}^{(i)}, X_{11}^{(j)})=LD(i, j)=\dfrac{\mathscr{D}_{AB}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$.

(*d*) *Let*

$$\tilde{X}_{11}^{(\gamma)}=\sqrt{n_{1.}}\left(\frac{X_{11}^{(\gamma)}}{n_{1.}}-p_\gamma\right).$$

Then, the bivariate random vector $(\tilde{X}_{11}^{(i)}, \tilde{X}_{11}^{(j)})^\top$ is asymptotically jointly normally distributed with zero expectation and covariance matrix

$$\Sigma^*=\begin{pmatrix} p_i(1-p_i) & \mathscr{D}_{AB} \\ \mathscr{D}_{AB} & p_j(1-p_j)\end{pmatrix}. \tag{1}$$

Proof. Assertion (a) is obvious. For assertion (b), we employ the representation

$$X_{11}^{(\gamma)}=\sum_{k=1}^{n_{1.}}\mathbf{1}\{\text{Case } k \text{ exhibits the major allele at locus } \gamma\}.$$

This entails that $\mathbb{E}[X_{11}^{(i)}X_{11}^{(j)}]=n_{1.}p_{ij}+(n_{1.}^2-n_{1.})p_ip_j$. Combining this with assertion (a) implies (b). Assertion (c) follows immediately from (a) and (b). Assertion (d) is an application of the binomial central limit theorem of de Moivre and Laplace in combination with the Cramér-Wold device [see, e.g. Shorack and Wellner (1986), p 862]. □

Remark 1. *The statistic $X_{11}^{(\gamma)}$ is the test statistic employed by Fisher's exact test in contingency table γ.*

Theorem 1. *Let $f=(f_i, f_j):\mathbb{R}^2\to\mathbb{R}^2$ be a smooth transformation such that its Jacobian (matrix of partial derivatives) ∇f, evaluated at the point $(\mathbb{E}[n_{1.}^{-1}X_{11}^{(i)}]=p_i, \mathbb{E}[n_{1.}^{-1}X_{11}^{(j)}]=p_j)$, is a positive definite diagonal matrix. Then, under the assumptions of Lemma 1, the random vector $\sqrt{n_{1.}}f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$ is asymptotically ($N\to\infty$) bivariate normally distributed, and the correlation coefficient of its asymptotic normal distribution is equal to LD(i, j).*

Proof. We apply the bivariate Delta method in analogy to Section 3 of Wei and Higgins (2013). In combination with part (d) of Lemma 1, it entails asymptotic bivariate normality of $\sqrt{n_{1.}}f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$. To calculate the correlation coefficient of its asymptotic normal distribution, we let $\nabla f(u, v)$ denote the entry at position (u, v) of ∇f, evaluated at (p_i, p_j), where $1\le u, v\le 2$, and let Σ stand for the asymptotic covariance matrix of the two components of $\sqrt{n_{1.}}f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$. Making use of the assumption regarding ∇f and of Σ^* from (1), the bivariate Delta method yields that

$$\Sigma=\begin{pmatrix}\nabla f(1,\,1) & 0\\ 0 & \nabla f(2,\,2)\end{pmatrix}\ \Sigma^{*}\begin{pmatrix}\nabla f(1,\,1) & 0\\ 0 & \nabla f(2,\,2)\end{pmatrix}=\begin{pmatrix}\nabla f(1,\,1)\Sigma^{*}_{11}\nabla f(1,\,1) & \nabla f(1,\,1)\Sigma^{*}_{12}\nabla f(2,\,2)\\ \nabla f(1,\,1)\Sigma^{*}_{21}\nabla f(2,\,2) & \nabla f(2,\,2)\Sigma^{*}_{22}\nabla f(2,\,2)\end{pmatrix}.$$

Hence, the correlation coefficient among the two components of $f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$ is asymptotically ($N\to\infty$) equal to

$$\frac{\nabla f(1,\,1)\mathscr{D}_{AB}\nabla f(2,\,2)}{\nabla f(1,\,1)\sqrt{\Sigma^{*}_{11}}\nabla f(2,\,2)\sqrt{\Sigma^{*}_{22}}}=\frac{\mathscr{D}_{AB}}{\sqrt{p_i(1-p_i)\,p_j(1-p_j)}}=\mathrm{LD}(i,j).$$ □

Plainly phrased, the assertion of Theorem 1 means that the asymptotic correlation structure among the K marginal association test statistics is exactly given by the (standardized) LD matrix among the K loci under consideration, provided that each test statistic T_γ is a smooth transformation of $X_{11}^{(\gamma)}$ only, without utilizing data from other loci.

Example 1 (Logarithmic odds ratios). *One widely applied marginal test statistic is the Wald statistic based on the (empirical) logarithmic odds ratio, say* $\hat{\lambda}_\gamma$ *for marginal contingency table* γ. *To avoid pathologies, assume that* $x_{rc}^{(\gamma)}>0$ *for* $1\le r,\ c\le 2$ *and* $\gamma\in\{i,\,j\}$. *In terms of* $x_{11}^{(\gamma)}$, *one can then write* $\hat{\lambda}_\gamma=\log(x_{11}^{(\gamma)})+\log(n_{.2}^{(\gamma)}-n_{1.}+x_{11}^{(\gamma)})-\log(n_{1.}-x_{11}^{(\gamma)})-\log(n_{.1}^{(\gamma)}-x_{11}^{(\gamma)})=f_\gamma(x_{11}^{(\gamma)})$. *Considering the so-defined function* $f=(f_i,f_j)$, *where we artificially include* $x_{11}^{(j)}(x_{11}^{(i)})$ *as a second argument to* f_i (f_j), *it is straightforward to check that the assumptions of Theorem 1 are fulfilled. Hence, we have that* $\rho(\hat{\lambda}_i,\hat{\lambda}_j)$ *is asymptotically equal to LD*(*i*, *j*). *The application of the univariate Delta method to prove asymptotic Gaussianity of* $\hat{\lambda}_\gamma$ *is mentioned in Section 3.1.7 of Agresti* (*2002*).

We may remark here that the exact finite-sample correlation coefficient of $\hat{\lambda}_i$ *and* $\hat{\lambda}_j$ *has been derived in Bagos* (*2012*). *As a sanity check for our asymptotic result, we derived Figure 1. The data for this figure have been taken from a genetic association study regarding a* (*dichotomized*) *behavioral measure of impulsiveness* (*yet unpublished data*), *consisting of* $n_{1.}=299$ *cases* (*highly impulsive individuals*) *and* $n_{2.}=2430$ *controls. The data points in Figure 1 correspond to* $K=10$ *correlated genomic loci, leading to* $\binom{10}{2}=45$ *pairwise* (*standardized*) *LD coefficients. Although it is not guaranteed that the global null hypothesis of no genetic association with impulsiveness holds for the* $K=10$ *loci displayed in Figure 1, the* 10 *estimated odds ratios were very close to* 1, *such that this assumption seemed justified. Qualitatively, the obvious agreement of abscissas and ordinates in Figure 1 has been confirmed by many other analogous graphs which we omit here.*

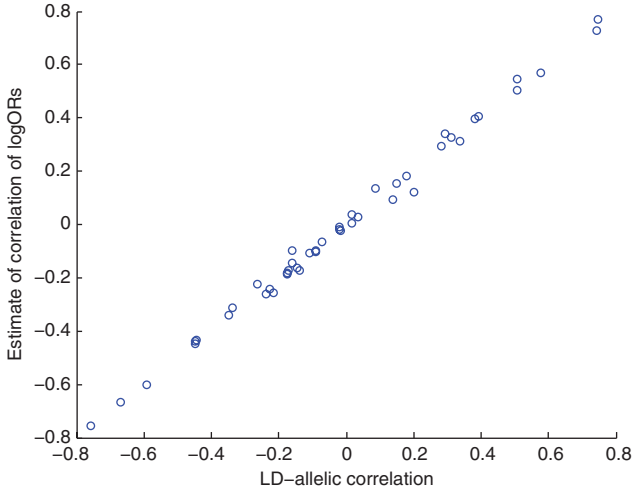

Figure 1: Empirical comparison of LD(*i*, *j*) and $\rho(\hat{\lambda}_i,\hat{\lambda}_j)$ on the basis of $K=10$ correlated genomic loci, leading to 45 pairwise data points. The abcissas are given by pairwise standardized LD coefficients, while the ordinates have been calculated by the exact finite-sample formulas given in Bagos (2012). The sample consisted of $n_{1.}=299$ cases and $n_{2.}=2430$ controls.

Example 2 (Chi-squared statistics). *Let* $e_{rc}^{(\gamma)}=n_{r.}n_{.c}^{(\gamma)}/N$, $1\leq r, c\leq 2$, *and denote by*

$$Q^{(\gamma)}=\sum_{r=1}^{2}\sum_{c=1}^{2}\frac{(X_{rc}^{(\gamma)}-E_{rc}^{(\gamma)})^2}{E_{rc}^{(\gamma)}}$$

the chi-squared statistic for testing association in contingency table γ. *It is well-known that* $Q^{(\gamma)}$ *is asymptotically chi-square distributed under the null with one degree of freedom*.

Theorem 1 is not directly applicable in this case, because the representation (5) *given in Lemma 3 below shows that the assumption of a positive definite Jacobian is violated here* (*diagonal elements of* ∇f *are equal to zero*). *However, Lemma 1 in combination with equation* (*13*) *of Bohrnstedt and Goldberger* (*1969*) *yields that the correlation coefficient between* $Q^{(i)}$ *and* $Q^{(j)}$ *is asymptotically given by* $LD^2(i, j)$. *We verify this result in Figure 2, in analogy to Figure 1*.

4 Genotypic association test problems

The Delta method can also be employed to work out the asymptotic correlation structure of test statistics in $2\times C\times K$ tables for $C>2$. The special case of $C=3$ is relevant in association studies if the locus-specific diploid allele pairs are considered instead of the mere alleles. This setup was referred to in Dickhaus and Stange (2013) as a multiple genotypic association test problem and the authors derived the asymptotic correlation structure of the K marginal chi-squared statistics in that case; cf. the discussion around their Definition 4.2 and Lemma 4.2. In the case of a (2×3)-contingency table at each locus γ the bivariate vector $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})^\top$ is sufficient conditional to all marginals (cf. Table 2), and a multinomial central limit theorem holds for $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})^\top$. Hence, considering two such loci i and j, the four-variate Delta method can be applied. To this end, we consider the expected (joint) genotype frequencies given in Table 3.

The following lemma can be proved in analogy to Lemma 1.

Lemma 2. *Assume that the null hypothesis of no association between phenotype and genotypic status is fulfilled at both loci* $\gamma\in\{i, j\}$. *Then, letting* $\mathcal{M}_3(n_{1.}, \mathbf{p})$ *denote the multinomial distribution with three categories, total sample* $n_{1.}$ *and vector of probabilities* $\mathbf{p}$, *the following assertions hold true*.

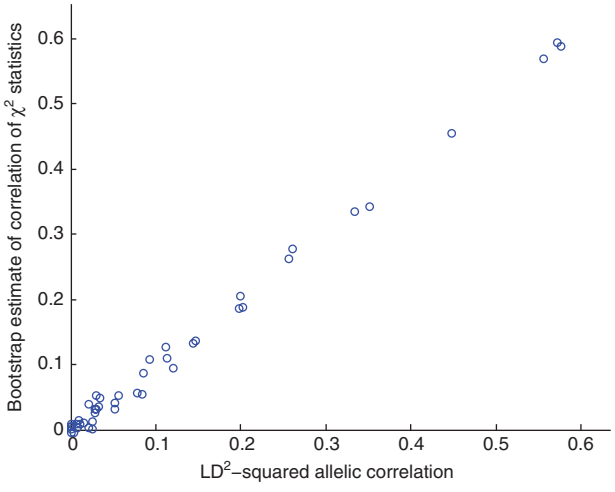

Figure 2: Empirical comparison of $LD^2(i, j)$ and $\rho(Q^{(i)}, Q^{(j)})$ on the basis of $K=10$ correlated genomic loci, leading to 45 pairwise data points. The abcissas are given by pairwise squared standardized LD coefficients, while the ordinates have been calculated by nonparametric bootstrap. The sample consisted of $n_{1.}=299$ cases and $n_{2.}=2430$ controls.

Table 3: Expected (joint) genotype frequencies for two bi-allelic markers *i* and *j*.

	AA	*Aa*	*aa*	Σ
BB	π_{00}	π_{01}	π_{02}	p_j
Bb	π_{10}	π_{11}	π_{12}	q_j
bb	π_{20}	π_{21}	π_{22}	$1-p_j-q_j$
Σ	p_i	q_i	$1-p_i-q_i$	1

(*a*) $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)}, X_{13}^{(\gamma)}) \sim \mathcal{M}_3(n_{1.}, (p_\gamma, q_\gamma, 1-p_\gamma-q_\gamma)^\top)$, *hence* $Cov(X_{11}^{(\gamma)}, X_{12}^{(\gamma)}) = -n_{1.} p_\gamma q_\gamma$.

(*b*) $Cov(X_{11}^{(i)}, X_{11}^{(j)}) = n_{1.}(\pi_{00} - p_i p_j)$, $Cov(X_{11}^{(i)}, X_{12}^{(j)}) = n_{1.}(\pi_{01} - p_i q_j)$,
$Cov(X_{12}^{(i)}, X_{11}^{(j)}) = n_{1.}(\pi_{10} - q_i p_j)$, $Cov(X_{12}^{(i)}, X_{12}^{(j)}) = n_{1.}(\pi_{11} - q_i q_j)$.

(*c*) *Let*

$$\tilde{X}_{11}^{(\gamma)} = \sqrt{n_{1.}}\left(\frac{X_{11}^{(\gamma)}}{n_{1.}} - p_\gamma\right), \quad \tilde{X}_{12}^{(\gamma)} = \sqrt{n_{1.}}\left(\frac{X_{12}^{(\gamma)}}{n_{1.}} - q_\gamma\right).$$

Then the vector $(\tilde{X}_{11}^{(i)}, \tilde{X}_{12}^{(i)}, \tilde{X}_{11}^{(j)}, \tilde{X}_{12}^{(j)})^\top$ *is asymptotically jointly normal with zero expectation and covariance matrix*

$$\Sigma^* = \begin{pmatrix} p_i(1-p_i) & -p_i q_i & \pi_{00} - p_i p_j & \pi_{01} - p_i q_j \\ -p_i q_i & q_i(1-q_i) & \pi_{10} - q_i p_j & \pi_{11} - q_i q_j \\ \pi_{00} - p_i p_j & \pi_{10} - q_i p_j & p_j(1-p_j) & -p_j q_j \\ \pi_{01} - p_i q_j & \pi_{11} - q_i q_j & -p_j q_j & q_j(1-q_j) \end{pmatrix}.$$

Lemma 2, in combination with the four-variate Delta method, leads to the following result about the asymptotic correlation structure among locus-specific test statistics T_γ, provided that T_γ is a smooth transformation of $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})^\top$ only, without utilizing data from loci not equal to γ. In Theorem 2 below, f_γ operates on the pair $(x_{11}^{(\gamma)}/n_{1.}, x_{12}^{(\gamma)}/n_{1.})^\top$, $\gamma \in \{i, j\}$.

Theorem 2. *Let* $f = (f_i, f_j): \mathbb{R}^4 \to \mathbb{R}^2$ *be a smooth transformation such that its Jacobian* ∇f*, evaluated at the point* $(p_i, q_i, p_j, q_j)^\top$*, has the following structure*:

$$\nabla f(p_i, q_i, p_j, q_j) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(p_i, q_i, p_j, q_j) & 0 \\ \frac{\partial f}{\partial x_2}(p_i, q_i, p_j, q_j) & 0 \\ 0 & \frac{\partial f}{\partial x_3}(p_i, q_i, p_j, q_j) \\ 0 & \frac{\partial f}{\partial x_4}(p_i, q_i, p_j, q_j) \end{pmatrix}$$

Then it holds

$$\sqrt{n_{1.}}\{(T_i, T_j)^\top - f(p_i, q_i, p_j, q_j)\} \xrightarrow{d} \mathcal{N}_2(0, \Sigma),$$

where $T_\gamma = f_\gamma(X_{11}^{(\gamma)}/n_{1.}, X_{12}^{(\gamma)}/n_{1.})$ *and*

$$\Sigma = \nabla f(p_i, q_i, p_j, q_j)^\top \Sigma^* \nabla f(p_i, q_i, p_j, q_j) \tag{2}$$

only depends on the quantities listed in Table 3. *The asymptotic correlation of* $(T_i, T_j)^\top$ *is consequently given by* $\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$.

One very popular class of test statistics T_γ fulfilling the requirements of Theorem 2 is the family of Cochran-Armitage trend test (CATT) statistics; cf., among others, Langaas and Bakke (2013) and references therein.

Example 3 (CATT statistics). *The CATT statistic with weights $w=(w_1, w_2, w_3)$ for locus γ is given by*

$$T_\gamma=\frac{\sum_{c=1}^{3}(w_c-\bar{w}^{(\gamma)})X_{1c}^{(\gamma)}}{\sqrt{\hat{\tau}(1-\hat{\tau})\sum_{c=1}^{3}n_{.c}^{(\gamma)}(w_c-\bar{w}^{(\gamma)})^2}},$$

where $\bar{w}^{(\gamma)}=N^{-1}\sum_{c=1}^{3}n_{.c}^{(\gamma)}w_c$ and $\hat{\tau}=n_{1.}/N$.

Equivalently, one can write

$$T_\gamma=f_\gamma(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})=\frac{(w_1-w_3)X_{11}^{(\gamma)}+(w_2-w_3)X_{12}^{(\gamma)}+n_{1.}(w_3-\bar{w}^{(\gamma)})}{\sqrt{\hat{\tau}(1-\hat{\tau})\sum_{c=1}^{3}n_{.c}^{(\gamma)}(w_c-\bar{w}^{(\gamma)})^2}}$$

with Jacobian

$$\nabla f(x_1, x_2, x_3, x_4)\equiv\begin{pmatrix}(w_1-w_3)/D_i & 0\\ (w_2-w_3)/D_i & 0\\ 0 & (w_1-w_3)/D_j\\ 0 & (w_2-w_3)/D_j\end{pmatrix}$$

of $f=(f_i, f_j)$, where $D_\gamma^2=\hat{\tau}(1-\hat{\tau})\sum_{c=1}^{3}n_{.c}^{(\gamma)}(w_c-\bar{w}^{(\gamma)})^2$, $\gamma\in\{i, j\}$.

Based on Theorem 2, the asymptotic covariance matrix of $(T_i, T_j)^\top$ has entries

$$\begin{aligned}\Sigma_{12}&=(D_iD_j)^{-1}[(w_1-w_3)^2(\pi_{00}-p_ip_j)+(w_1-w_3)(w_2-w_3)(\pi_{01}-p_iq_j)+(w_2-w_3)(w_1-w_3)(\pi_{10}-q_ip_j)\\&\quad+(w_2-w_3)^2(\pi_{11}-q_iq_j)],\\ \Sigma_{11}&=D_i^{-2}[p_i(1-p_i)(w_1-w_3)^2-2p_iq_i(w_1-w_3)(w_2-w_3)+q_i(1-q_i)(w_2-w_3)^2],\\ \Sigma_{22}&=D_j^{-2}[p_j(1-p_j)(w_1-w_3)^2-2p_jq_j(w_1-w_3)(w_2-w_3)+q_j(1-q_j)(w_2-w_3)^2].\end{aligned}$$

If the weights w_1, w_2, w_3 are chosen such that

$$(w_1-w_2)/(w_2-w_3)=1 \text{ or, equivalently, } w_2=(w_1+w_3)/2, \tag{3}$$

then $\rho(T_i, T_j)=\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ is equal to

$$CLD(i, j)=\frac{2(\pi_{00}-p_ip_j)+(\pi_{01}-p_iq_j)+(\pi_{10}-q_ip_j)+(\pi_{11}-q_iq_j)/2}{2\sqrt{(p_i+q_i/4-(p_i+q_i/2)^2)(p_j+q_j/4-(p_j+q_j/2)^2)}}. \tag{4}$$

Condition (3) *is for instance fulfilled for the popular choice* $(w_1, w_2, w_3)=(0, 1, 2)$, *corresponding to the assumption of an additive risk allele contribution.*

Remark 2. *An alternative representation of CLD(i, j) from* (4) *is given by*

$$CLD(i, j)=\frac{2\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}/2-2p_Ap_B}{\sqrt{(p_A(1-p_A)+p_i-p_A^2)(p_B(1-p_B)+p_j-p_B^2)}},$$

where $p_A=p_i+q_i/2$ and $p_B=p_j+q_j/2$ denote the (*expected*) *allele frequencies of the major allele at locus i or j, respectively. This coefficient is referred to as the standardized composite LD coefficient of loci i and j, see Weir* (*1979*) *and Weir and Cockerham* (*1979*).

Finally, we verify the result of Example 3 by means of real data as in Section 3, leading to Figure 3.

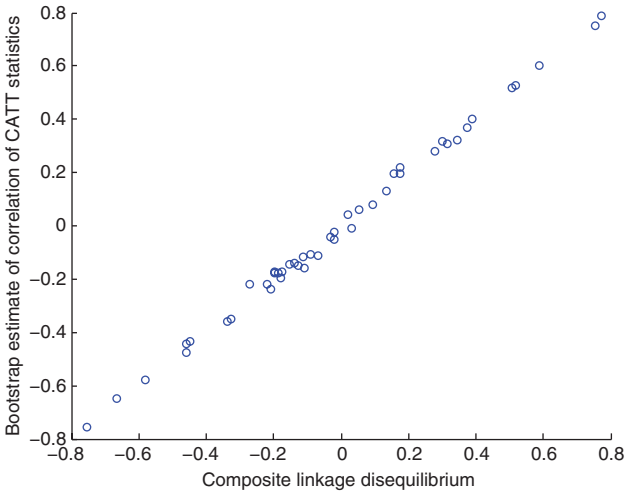

Figure 3: Empirical comparison of CLD(i, j) and $\rho(T_i, T_j)$ on the basis of K=10 correlated genomic loci, leading to 45 pairwise data points. The abscissas are given by pairwise standardized CLD coefficients, while the ordinates have been calculated by nonparametric bootstrap. The sample consisted if $n_{1.}$=299 cases and $n_{2.}$=2430 controls.

5 Discussion

We have drawn a connection between the correlation structure among (marginal) test statistics for association in $2\times C\times K$ contingency tables, $C\in\{2, 3\}$, and the standardized (composite) LD matrix pertaining to the K loci under investigation in genetic case-control studies. Our respective results can be exploited for multivariate statistical inference, because external (composite) LD information can be used to approximate the correlation matrix of the locus-specific test statistics, provided that the sample size N is large.

One concrete application of our Example 1, which we want to work out in a detailed manner in future research, consists in Bayesian inference for $2\times2\times K$ contingency tables based on the K-dimensional vector of logarithmic odds ratios as considered in Demirhan and Hamurkaroglu (2008). Here, the (standardized) LD matrix can be used as an informative prior for the correlation structure among the logarithmic odds ratios (which are regarded as random objects in the Bayesian setup).

Finally, one limitation of our current approach is that all results are derived under the assumption that the null hypotheses of no association are true at all loci under consideration. Hence, the proposed multivariate methods can only be employed for type I error calibration, but not for power or sample size calculations. It may be interesting to consider the correlation structure of association test statistics also under well-defined alternatives, but this is beyond the scope of the present work and devoted to future research.

Acknowledgments: We thank Mette Langaas and Øyvind Bakke for some useful hints regarding Lemma 3. This research was partly supported by the Deutsche Forschungsgemeinschaft via grant No. DI 1723/3-1 (Jens Stange). We are grateful to two anonymous referees for their careful reading of the manuscript and constructive comments which have improved the presentation.

Appendix I: Auxiliary results

Lemma 3. *Let* $\mathbf{x}=\begin{pmatrix} x_{11}\dots x_{1C} \\ x_{21}\dots x_{2C} \end{pmatrix}$ *denote a* $(2\times C)$*-contingency table with row marginals* $n_{1.}$, $n_{2.}$, *column marginals* $n_{.1}$, …, $n_{.C}$, *and total sample size* $N=n_{1.}+n_{2.}$. *Define* $e_{rc}=n_{r.}n_{.c}/N$ *for* $1\le r\le 2$ *and* $1\le c\le C$, *and let*

$$Q(\mathbf{x})=\sum_{r=1}^{2}\sum_{c=1}^{C}\frac{(x_{rc}-e_{rc})^2}{e_{rc}}$$

denote the value of the chi-squared statistic for testing association based on $\mathbf{x}$. *Then the following two assertions hold true*.
(*a*)

$$Q(\mathbf{x})=\frac{N}{n_{2.}}\sum_{c=1}^{C}\frac{(x_{1c}-e_{1c})^2}{e_{1c}}.$$

(*b*) *In the special case of* $C=2$, *it holds that*

$$Q(\mathbf{x})=\left[\sqrt{\frac{N}{n_{2.}}}\frac{(x_{11}-e_{11})}{\sqrt{n_{1.}\hat{p}(1-\hat{p})}}\right]^2, \tag{5}$$

where $\hat{p}=n_{.1}/N$ *denotes the empirical major allele frequency*.

Proof. For proving part (a), it suffices to show that for any given column $1\le c\le C$, we have

$$\frac{(x_{1c}-e_{1c})^2}{e_{1c}}+\frac{(x_{2c}-e_{2c})^2}{e_{2c}}=\frac{N}{n_{2.}}\frac{(x_{1c}-e_{1c})^2}{e_{1c}},$$

which can be verified by elementary calculations. Part (b) follows from part (a) and by noticing that $(x_{12}-e_{12})^2=(x_{11}-e_{11})^2$. □

References

Agresti, A. (2002): *Categorical data analysis. 2nd ed.*, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Chichester: Wiley.

Bagos, P. G. (2012): "On the covariance of two correlated log-odds ratios," Stat. Med., 31, 1418–1431.

Bohrnstedt, G. and A. Goldberger (1969): "On the exact covariance of products of random variables," J. Am. Stat. Assoc., 64, 1439–1442.

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

Demirhan, H. and C. Hamurkaroglu (2008): "Bayesian estimation of log odds ratios from $R\times C$ and $2\times 2\times K$ contingency tables," Stat. Neerl., 62, 405–424, URL: http://dx.doi.org/10.1111/j.1467-9574.2008.00387.x.

Devlin, B. and N. Risch (1995): "A comparison of linkage disequilibrium measures for fine-scale mapping," Genomics, 29, 311–322.

Dickhaus, T. (2014): Simultaneous statistical inference with applications in the life sciences, Berlin, Heidelberg: Springer-Verlag.

Dickhaus, T. and J. Stange (2013): "Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate," Calcutta Stat. Assoc. Bull., 65, 123–144.

Langaas, M. and Ø. Bakke (2013): "Increasing power with the unconditional maximization enumeration test in small samples – a detailed study of the MAX3 test statistic," Statistics Preprint No. 1/2013, Trondheim: Norwegian University of Science and Technology.

Moskvina, V. and K. M. Schmidt (2008): "On multiple-testing correction in genome-wide association studies," Genet. Epidemiol., 32, 567–573.

Shorack, G. R. and J. A. Wellner (1986): *Empirical processes with applications to statistics*, Wiley Series in Probability and Mathematical Statistics. New York, NY: Wiley.

The 1000 Genomes Consortium (2010): "A map of human genome variation from population-scale sequencing," Nature, 467, 1061–1073.

The International HapMap Consortium (2005): "A haplotype map of the human genome," Nature, 437, 1299–1320.

Wei, Y. and J. P. Higgins (2013): "Estimating within-study covariances in multivariate meta-analysis with multiple outcomes," Stat. Med., 32, 1191–1205.

Weir, B. S. (1979): "Inferences about linkage disequilibrium," Biometrics, 35, 235–254.

Weir, B. S. and C. C. Cockerham (1979): "Estimation of linkage disequilibrium in randomly mating populations," Heredity, 42, 105–111.

Ziegler, A. and I. R. König (2006): A statistical approach to genetic epidemiology, Weinheim: Wiley.