Bernoulli **11**(6), 2005, 1093–1113

Spectral decomposition of score functions in linkage analysis

OLA HÖSSJER

Department of Mathematics, Stockholm University, S-106 91 Stockholm, Sweden
E-mail: ola@math.su.se

We consider stochastic processes occurring in nonparametric linkage analysis for mapping disease susceptibility genes in the human genome. Under the null hypothesis that no disease gene is located in the chromosomal region of interest, we prove that the linkage process converges weakly to a mixture of Ornstein–Uhlenbeck processes as the number of families N tends to infinity. Under a sequence of contiguous alternatives, we prove weak convergence towards the same Gaussian process with a deterministic non-zero mean function added to it. The results are applied to power calculations for chromosome- and genome-wide scans, and are valid for arbitrary family structures. Our main tool is the inheritance vector process v, which is a stationary and continuous-time Markov process with state space the set of binary vectors w of given length. Certain score functions are expanded as a linear combination of an orthonormal system of basis functions which are eigenvectors of the intensity matrix of v.

Keywords: continuous-time Markov process; inheritance vectors; invariance principle; linkage analysis; Ornstein–Uhlenbeck process; spectral decomposition

1. Introduction

Linkage analysis is a technique for localizing gene(s) that influence a certain trait, typically an inheritable disease. Measurements related to the disease (phenotypes) are collected for individuals in a number of families together with their DNA. Small segments of DNA, so-called genetic markers, are measured at a number of known positions (loci) along the genome. Markers close to the (unknown) disease locus cosegregate with the disease gene, meaning that the pattern of grandpaternal/grandmaternal DNA transmission in each family is correlated around the disease locus. Since phenotypes are indirect measurements of disease genes, there will also be correlation between inheritance of phenotypes and markers close to the disease locus. This correlation decays with distance between the disease locus and the marker. The reason for this is the occurrence of so-called crossovers, which are points of switching between grandpaternal and grandmaternal allele transmission. The usual procedure is to define a linkage score function $Z(t)$ as a function of the (genetic) map position t. A large value of $Z(t)$ indicates high correlation between inheritance of phenotypes and DNA at locus t. Hence, regions where the stochastic process Z is large are candidates for harbouring the disease gene. Formally, linkage analysis can be formulated as

1350–7265

a hypothesis-testing problem, with a null hypothesis of no disease gene located along the genomic region of interest. See Sham (1998) and Ott (1999) for more details.

When DNA is collected from all (or sufficiently many) pedigree members at a dense set of genetic markers, the marker data are perfect. This means that Z is piecewise constant, with discontinuities at the points of crossover. The most common model, due to Haldane (1919), assumes that crossovers occur randomly according to a Poisson process. This assumption implies that Z is a stationary process under the null hypothesis H_0 of no linkage. In this paper we consider nonparametric linkage score functions Z of the type considered in human genetics (Kruglyak *et al.* 1996). We prove that asymptotically, as the number of families N tends to infinity, Z converges weakly to a stationary Gaussian process which is a mixture of Ornstein–Uhlenbeck processes. Under a sequence of contiguous alternatives H_1, we establish weak convergence towards the same Gaussian process plus a deterministic drift function μ, which is a mixture of double exponential functions centred at the disease locus τ. Our results are valid for arbitrary scoring functions and family structures, and generalize previous work by Feingold *et al.* (1993) and Feingold and Siegmund (1997), where covariance and drift functions for sib pairs, half sibs, aunt–niece, first cousins and some other families are obtained. Analogous results in animal genetics for quantitative trait locus mapping were obtained by Lander and Bolstein (1989) and Dupuis and Siegmund (1999). In these cases Z is asymptotically a χ^2 process under H_0.

Our tool in this paper is the inheritance vector process v for each family (Donnelly 1983). This is a stationary continuous-time Markov process on the space $\mathbb{Z}_2^m$ of binary vectors of length m. The eigenvectors of its intensity matrix have eigenvalues that are integer multiples of -2 (Dudoit and Speed 1999). It turns out that these eigenvectors form an orthonormal basis on the space of functions $\mathbb{Z}_2^m \to \mathbb{R}$. Moreover, the coefficients when certain functions $\mathbb{Z}_2^m \to \mathbb{R}$ are expanded in this orthonormal-basis determine the covariance and drift functions of Z.

This paper is organized as follows. In Sections 2 and 3 we introduce the mathematical framework and establish spectral decomposition of the intensity matrix. In Section 4 we derive, for one pedigree, the covariance function of Z under H_0 and the mean function of Z under H_1. The invariance principle is proved in Section 5, and its consequences for significance levels and power are discussed. Examples of binary phenotypes and allele-sharing score functions are treated in Section 6. All proofs are given in a separate appendix.

2. Inheritance vectors and score functions

A pedigree $\mathcal{P}$ of n individuals can be represented as a graph as shown in Figure 1. Persons without parents in the pedigree are called founders and the remaining individuals non-founders. We assume that the pedigree is balanced in the sense that each non-founder has both of its parents in the pedigree. Consider a certain position (locus) t on one of the chromosomes. The DNA of a (short) segment surrounding this locus can have different forms, so-called alleles. According to Mendelian laws of segregation, each individual has two alleles at the locus of interest, one inherited from the father and one from the mother

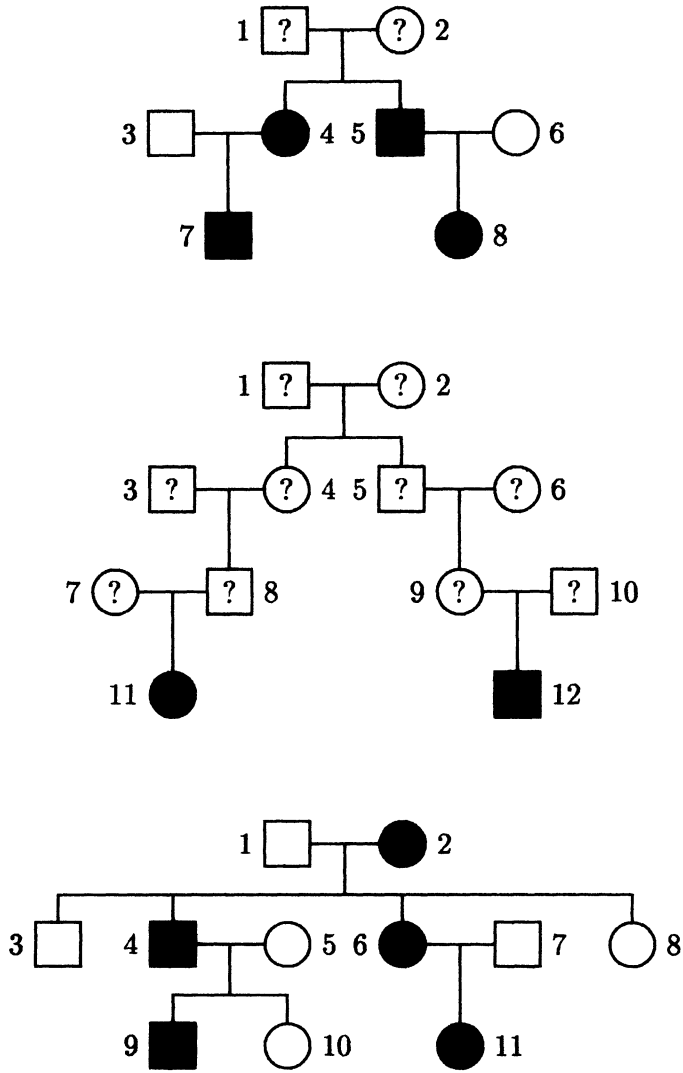

Figure 1. Pedigrees used in the simulations. Pedigree k ($k = 1, 2, 3, 4$), consists of two parents and $k + 1$ offspring, with the parents and offspring numbered 1, 2 and $3, \ldots, k + 3$, respectively. Pedigrees 5 (top), 6 (middle) and 7 (bottom) are shown with individual numbers. For each pedigree an example of phenotype vector Y is given. Males and females correspond to squares and circles. Affected individuals have black and unaffected ones have white symbols. Individuals with unknown phenotypes have question marks.

during formation of sperm and ovum cells. This pair of alleles forms a genotype at locus t, and represents the individual's DNA at this locus.

During the process of forming germ cells, called meiosis, each parent transmits one of its two alleles at locus t to the child. For each non-founder there are two meioses of interest during which alleles are transmitted from the father and mother, respectively. If f is the number of founders in the pedigree, the total number of meioses is $m = 2(n - f)$. Segregation of alleles in the pedigree at locus t can be represented by means of a binary inheritance vector $v = v(t)$ of length m; cf. Donnelly (1983). If meioses are numbered $1, \ldots, m$, we write $v = (v_1, \ldots, v_m)$, where $v_j = 0$ or 1 depending on whether a grandpaternal or grandmaternal allele was transmitted during the jth meiosis. We regard v as an element of $\mathbb{Z}_2^m$, the group of binary vectors of length m under component-wise modulo 2 addition.

A score function is a mapping $S : \mathbb{Z}_2^m \to \mathbb{R}$ which assigns to each inheritance vector v a score $S(v)$. In nonparametric linkage (cf. Kruglyak *et al.* 1996), $S(v) = S(v; \mathcal{P}, Y)$ is a function of v, the pedigree $\mathcal{P}$ and the vector $Y = (Y_1, \ldots, Y_n)$ of phenotypes in the pedigree. The latter are disease-related quantities (affection status, body mass index, insulin concentration, ...) observed for some or all individuals. For an inheritable disease, Y_k carries information about the kth individual's DNA at one or several disease genes. This implies that DNA cosegregates with Y at disease genes. A large value of $S(v; \mathcal{P}, Y)$ indicates high compatibility between Y and the inheritance vector $v = v(t)$. This in turn suggests that a disease gene is located in vicinity of t. We regard $\mathcal{P}$ and Y as fixed and hence often drop them in notation.

Let $\mathcal{A} = \mathcal{A}_m = \{S\}$ be the space of all mappings $\mathbb{Z}_2^m \to \mathbb{R}$. We endow $\mathcal{A}$ with the scalar product $(\cdot, \cdot) : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, defined as

$$(S, R) = 2^{-m} \sum_{w \in \mathbb{Z}_2^m} S(w)R(w).$$

Since $\mathbb{Z}_2^m$ consists of 2^m elements, $\mathcal{A}$ is isomorphic to the Euclidean space $\mathbb{R}^{2^m}$. For each fixed $w \in \mathbb{Z}_2^m$ we introduce $S_w \in \mathcal{A}$ as

$$S_w(u) = (-1)^{w \cdot u}, \tag{1}$$

where $w \cdot u$ is the vector dot product of w and u. Then the following property is easily established:

Proposition 1. *The collection $\{S_w\}_{w \in \mathbb{Z}_2^m}$ forms a complete orthonormal system of basis functions in $\mathcal{A}$, that is,*

$$(S_w, S_{w'}) = \begin{cases} 1, & \text{if } w = w', \\ 0, & \text{if } w \neq w', \end{cases}$$

and each $S \in \mathcal{A}$ can be written as a unique linear combination of elements S_w.

The coefficients of S in terms of the orthonormal-basis $\{S_w\}$ are written as

$$R_S(w) = (S, S_w), \qquad \forall w \in \mathbb{Z}_2^m.$$

Notice that $R_S \in \mathbb{Z}_2^m$. In fact, $2^m R_S$ equals the Fourier transform of S; cf. Diaconis (1988) and Kruglyak and Lander (1998). The latter authors apply the Fourier transform to multipoint linkage analysis for a different purpose than ours – to speed up computation of certain matrix products.

3. Crossovers and spectral decomposition

During each meiosis, there is switching between grandpaternal and grandmaternal DNA allele transmission along each chromosome. The switching points, which are called crossovers, occur randomly. If the average number of crossovers between two loci on the same chromosome is h, they are at genetic distance h Morgans from each other. The most widely used model for crossovers was introduced by Haldane (1919), assuming that crossovers occur randomly according to a Poisson process with intensity 1 when genetic distance is measured in Morgans.

Consider a chromosome of genetic length L Morgans. With Haldane's map function, the result of each single meioses j is described as a 'time'-homogeneous Markov process $\{v_j(t);\ 0 \leqslant t \leqslant L\}$ with state space $\mathbb{Z}_2 = \{0, 1\}$ and intensity matrix

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We make $v_j(\cdot)$ stationary by requiring that for some (and hence all) loci t $P(v_j(t) = 0) = P(v_j(t) = 1) = 0.5$. This is a consequence of Mendel's law of segregation that grandpaternal and grandmaternal allele transmissions occur with the same probability.

For a pedigree with m meioses we describe $\{v(t);\ 0 \leqslant t \leqslant L\}$ as a time-homogeneous Markov process on $\mathbb{Z}_2^m$ with intensity matrix $A = \{A(w, w');\ w, w' \in \mathbb{Z}_2^m\}$, where

$$A(w, w') = \begin{cases} -m, & w = w', \\ 1, & |w - w'| = 1, \\ 0, & |w - w'| \geqslant 2, \end{cases} \tag{2}$$

and $|w - w'| = \sum_{j=1}^m |w_j - w'_j|$ is the Hamming distance between w and w'; cf. Dudoit and Speed (1999) and Hössjer (2003a). The marginal distribution for m meioses is

$$\pi_0(w) := P(v(t) = w) = 2^{-m}, \qquad \forall w \in \mathbb{Z}_2^m, \tag{3}$$

at all loci t. We tacitly assumed in (2) and (3) that allele transmissions for different meioses are independent. Standard theory for continuous-time Markov processes implies that $Q_h = \exp(hA)$ is a transition matrix between two loci at distance h Morgans. Viewing $A = \{A(w, u);\ w, u \in Z_2^m\}$ and $Q_h = \{Q_h(w, u);\ w, u \in Z_2^m\}$ as self-adjoint operators on $\mathcal{A}$, the following result can be established:

Theorem 1 (Spectral theorem for A and Q_h). *The score functions S_w in* (1) *are eigenvectors of A and Q_h for all $w \in \mathbb{Z}_2^m$ with eigenvalues $-2|w|$ and* $\exp(-2|w|h)$, *respectively.*

Theorem 1 can be deduced from Propositions 1–3 of Dudoit and Speed (1999). Their proof is based on first establishing eigenvectors and eigenvalues for A through a certain adjacency matrix. Here, we give an alternative proof based on first establishing eigenvectors and eigenvalues for Q_h by induction with respect to m and then, by letting $h \to 0$, obtaining eigenvalues and eigenvectors for A. Yet another method of proof is used by Kruglyak and Lander (1998). They utilize the fact that Q_h is a convolution operator ($SQ_h = S * T_h$)) and then compute the Fourier transform of $T_h \in \mathcal{A}$ by direct combinatorial arguments.

4. Linkage process: One pedigree

The purpose of linkage analysis is to test the presence of a disease locus τ on the chromosome. This can be formulated as a hypothesis-testing problem

$$H_0 : \tau = \infty, \qquad H_1 : \tau \in [0, L].$$

Here $\tau = \infty$ means that a disease locus does not exist or is located on another chromosome. By extracting DNA at so-called genetic markers from pedigree members, we obtain information about the different individuals' DNA alleles on $[0, L]$. This in turn implies information about the inheritance vector process $v(\cdot)$. Assuming that DNA marker data is perfect, we can observe the stochastic process

$$Z(t) = S(v(t)), \qquad 0 \leqslant t \leqslant L, \tag{4}$$

where $S \in \mathcal{A}$ is a score function, introduced in Section 2. A large value of $Z(t)$ indicates that τ is close to t, and hence that H_0 should be rejected. In practice, we need many pedigrees in order for a formal test between H_0 and H_1 to have high power; see Section 5.

Under H_0, the distribution of the Markov process $v(\cdot)$ can be summarized by (2) and (3). Since v is stationary under H_0, so is Z. Let $1 = S_0 \in \mathcal{A}$ be a score function of ones. It is customary in nonparametric linkage analysis to standardize S in (4) so that $(1, S) = 0$ and $\|S\|^2 = (S, S) = 1$; see, for example, Kruglyak *et al.* (1996). This implies

$$\mathrm{E}_{H_0}(Z(t)) = 0, \qquad \mathrm{var}_{H_0}(Z(t)) = 1. \tag{5}$$

The following result gives an explicit expression for the covariance function $r_Z(h) = \mathrm{cov}_{H_0}(Z(t), Z(t+h))$.

Theorem 2. *The covariance function of Z in* (4) *under H_0 is given by*

$$r_Z(h) = \sum_{l=1}^{m} \kappa_l \exp(-2l|h|), \tag{6}$$

where $\kappa_l = \sum_{w;|w|=l} R_S^2(w)$ and $\sum_{l=1}^{m} \kappa_l = 1$.

Under H_1 the distribution of v on $[0, L]$ is different. The phenotype vector Y carries information about $v(\tau)$. Define

$$\pi(w) := P(v(\tau) = w|Y), \qquad \forall w \in \mathbb{Z}_2^m, \tag{7}$$

as the posterior distribution of $v(\tau)$. The stronger the genetic influence of the disease gene is, the more the posterior π differs from the prior π_0. We will write

$$\pi(w) = 2^{-m}(1 + \xi\tilde{S}(w)), \tag{8}$$

where $\tilde{S} \in \mathcal{A}$ is normalized so that $(\tilde{S}, 1) = 0$ and $\|\tilde{S}\| = 1$. In other words, $\tilde{S}$ is the direction, in $\mathcal{A}$, of a linear path leading from π_0 to π. The scalar $\xi = 2^m\|\pi - \pi_0\|$ measures how informative the pedigree $\mathcal{P}$, the phenotype vector Y and the genetic model are for detecting linkage. In fact, $\log_2(\xi^2 + 1)$ can be interpreted as an effective number of meioses, with $\log_2(\xi^2 + 1) = m$ in the ideal case where Y gives complete information about $v(\tau)$ (Hössjer 2004).

Of particular interest is the noncentrality parameter (Feingold *et al.* 1993)

$$\eta = \mu_Z(\tau) = \xi(\tilde{S}, S).$$

As will be seen in the next section, η is closely related to the power to detect τ. The factor $(\tilde{S}, S)$ is a number between 0 and 1 measuring how efficient the chosen score function S is. In Hössjer (2003a), $(\tilde{S}, S)^2$ is interpreted as the efficiency of S compared to the optimal score function $\tilde{S}$.

We assume that Y and $\{v(t); t \neq \tau\}$ are conditionally independent given $v(\tau)$. This implies that under H_1, $v(t)$ has marginal distribution π at $t = \tau$. Then, because of the Markov property, $\{v(t); 0 \leq t \leq l\}$ propagates as two independent Markov processes with intensity matrices A in either direction from τ. Using this, the following theorem can be established for $\mu_Z(t) = \mathrm{E}_{H_1}(Z(t))$:

Theorem 3. *The mean function of the linkage score* (4) *is given by*

$$\mu_Z(t) = \eta \sum_{l=1}^{m} \delta_l \exp(-2l|t - \tau|), \qquad \forall t \in [0, L],$$

under H_1, *where*

$$\delta_l = \sum_{w;|w|=l} R_{\tilde{S}}(w)R_S(w)/(\tilde{S}, S)$$

and hence $\sum_{l=1}^{m}\delta_l = 1$.

5. An invariance principle

Consider a collection of N pedigrees. We assume that these can be of K different types. The type ϕ of a pedigree includes both the pedigree structure and phenotype vector. Let $\mathcal{P}_\phi$, m_ϕ and $Y_\phi = (Y_{\phi 1}, \ldots, Y_{\phi n_\phi})$ denote a pedigree, number of meioses and phenotype vector of type ϕ, and $\phi_i \in \{1, \ldots, K\}$ the type of pedigree i. The score function $w \to S(w; \mathcal{P}_\phi, Y_\phi) \in \mathcal{A}_{m_\phi}$ we write more compactly as $S_\phi(w)$. To ensure that (5) holds for

each family score, we assume $(S_\phi, 1) = 0$ and $\|S_\phi\|^2 = 1$. The total linkage process is defined as

$$Z(t) = \frac{\sum_{i=1}^N \gamma_{\phi_i} S_{\phi_i}(v_i(t))}{\sqrt{\sum_{i=1}^N \gamma_{\phi_i}^2}}, \qquad 0 \leqslant t \leqslant L, \tag{9}$$

a weighted sum of familywise scores (4) which is normalized to ensure (5). Here $v_i(t) = (v_{i1}(t), \ldots, v_{im_{\phi_i}}(t))$ is the inheritance vector at locus t for the ith pedigree and γ_ϕ the weight assigned to a pedigree of type ϕ. By giving larger weights γ_ϕ to more informative pedigree types, it is possible to increase power; see, for example, Sham *et al.* (1997) and Hössjer (2003a).

As our test statistic for testing H_0 versus H_1 we use

$$Z_{\max} = \sup_{0 \leqslant t \leqslant L} Z(t) \geqslant T \Rightarrow \text{Reject } H_0, \tag{10}$$

where T is a predefined threshold. The significance level and power are

$$\begin{aligned} \alpha(T) &= P_{H_0}(Z_{\max} \geqslant T), \\ \beta(T) &= P_{H_1}(Z_{\max} \geqslant T). \end{aligned} \tag{11}$$

We will consider the asymptotic behaviour of $\alpha(T)$ and $\beta(T)$ as $N \to \infty$ and K is kept fixed. In order to avoid a trivial power limit 1, we define a sequence of contiguous alternatives (7) for each pedigree type. This means, if π_ϕ is the posterior (7) for a pedigree of type ϕ, that

$$\pi_\phi(w) = 2^{-m_\phi}(1 + \xi_\phi \tilde{S}_\phi(w)/\sqrt{N}) + o(1/\sqrt{N}), \qquad \forall w \in \mathbb{Z}_2^{m_\phi}, \tag{12}$$

where ξ_ϕ measures the strength of the pedigree type ϕ and $\tilde{S}_\phi \in \mathcal{A}_{m_\phi}$ has been standardized so that $(\tilde{S}_\phi, 1) = 0$ and $\|\tilde{S}_\phi\| = 1$. Notice that $\xi_\phi = 0$ corresponds to H_0 and then $\pi_\phi = \pi_0 \equiv 2^{-m_\phi}$.

We view Z in (9) as a random element of $D[0, L]$, the space of right-continuous functions on $[0, L]$ with left-hand limits. In order to define weak convergence ($\xrightarrow{\mathcal{L}}$) we endow $D[0, L]$ with the Skorohod topology (Billingsley, 1968). Then the following result holds:

Theorem 4. *Let N_ϕ be the number of pedigrees of type ϕ. Assume that $N_\phi/N \to \nu_\phi$ as $N \to \infty$, $\phi = 1, \ldots, K$. Then, assuming H_1 and a sequence* (12) *of contiguous alternatives,*

$$Z \xrightarrow{\mathcal{L}} \mu + W \qquad \text{as } N \to \infty. \tag{13}$$

Here μ is the mean function, defined as

$$\mu(t) = \sum_{l=1}^m \exp(-2l|t - \tau|) \sum_{\phi=1}^K \eta_\phi \sqrt{\nu_\phi \bar{\nu}_\phi} \delta_{\phi l},$$

where $\eta_\phi = \xi_\phi(\tilde{S}_\phi, S_\phi)$, $\delta_{\phi l} = \sum_{w;|w|=l} R_{\tilde{S}_\phi}(w) R_{S_\phi}(w)/(\tilde{S}_\phi, S_\phi)$, $m = \max_{1 \leqslant \phi \leqslant K} m_\phi$ *and* $\bar{\nu}_\phi = \gamma_\phi^2 \nu_\phi / \sum_{\phi'=1}^K \gamma_{\phi'}^2 \nu_{\phi'}$. *Further, W is a stationary and zero-mean Gaussian process on $[0, L]$ which is a finite mixture of Ornstein–Uhlenbeck processes with covariance function*

$$r_W(h) = \sum_{l=1}^{m} \exp(-2lh) \sum_{\phi=1}^{K} \bar{v}_\phi \kappa_{\phi l},$$

and $\kappa_{\phi l} = \sum_{w;|w|=l} R^2_{S_\phi}(w)$. *Under* H_0, (13) *holds with* $\mu = 0$ *instead.*

Equipped with the invariance principle (13), the continuous mapping theorem immediately implies the following:

Corollary 1. *Under the same assumptions as in Theorem 4, the significance level and power* (11) *satisfy*

$$\alpha(T) \to \alpha_\infty(T) := P(\sup_{0 \leqslant t \leqslant L} W(t) \geqslant T),$$

$$\beta(T) \to \beta_\infty(T) := P(\sup_{0 \leqslant t \leqslant L} (\mu(t) + W(t)) \geqslant T),$$

as $N \to \infty$.

Exact formulae for $\alpha_\infty(T)$ and $\beta_\infty(T)$ are complicated, although approximations can be obtained using extreme value theory for non-differentiable Gaussian processes. Define $\eta = \mu(\tau)$, $\rho = -r'_W(0)/2$ and $d = -\mu'(\tau)/(2\mu(\tau)\rho)$, where the last two derivatives are taken from the right. The approximations

$$\alpha_\infty(T) \approx 1 - \exp(-(1 - \Phi(T))(1 + 2\rho L T^2)) \tag{14}$$

and

$$\beta_\infty(T) \approx 1 - \Phi(T - \eta) + \varphi(T - \eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d - 1) + T}\right) \tag{15}$$

are defined in Lander and Kruglyak (1995) and Feingold *et al.* (1993), respectively. Here Φ and φ are the cumulative distribution and density of a standard normal random variable. Formula (15) requires $(1 - d)\eta < T$. The significance level $\alpha_\infty(T)$ depends on the crossover rate ρ. It measures the amount of fluctuations of W and hence the amount of multiple testing in (10). The non-centrality parameter η, and to some extent the normalized slope-to-noise ratio d, determine the power $\beta_\infty(T)$.

6. Examples

Consider a genetic model based on binary phenotypes. Then Y_k, the phenotype of the kth pedigree member, equals 0, 1 or ? depending on whether k is unaffected, affected or has unknown affections status. Let $G = (G_1, \ldots, G_n)$ be the set of genotypes at the disease locus, where $G_k = (a_{2k-1}a_{2k})$ is the genotype of the kth individual. It consists of two alleles, one inherited from the father (a_{2k-1}) and one from the mother (a_{2k}). In a biallelic model, we assume there are two alleles, a normal one (0) and one causing disease (1). Notice that $G = G(a, w)$, where w is the inheritance vector and $a = (a_1, \ldots, a_{2f})$ the set of

founder alleles (assuming founders are labelled $1, \ldots, 2f$). By Bayes' rule and the law of total probability the posterior distribution (7) of $v(\tau)$ can be calculated from

$$\pi(w) = 2^{-m} P(Y|v(\tau) = w)/P(Y) \propto P(Y|v(\tau) = w),$$

$$P(Y|v(\tau) = w) = \sum_a P(Y|a, w)P(a) = \sum_a P(Y|G)P(a),$$

where in the second equation we sum over all 2^{2f} possible founder allele combinations. Computational reductions are possible, especially for pedigrees without loops; cf. Kruglyak *et al.* (1996) and references therein. For a monogenic model without environmental effects one has $P(Y|G) = \prod_{k=1}^{n} P(Y_k|G_k)$, where $P(?|G_k) = 1$, $P(0|G_k) = 1 - P(1|G_k)$, $P(1|(00)) = \psi_0$, $P(1|(01)) = \psi_1$ and $P(1|(11)) = \psi_2$. The three penetrance parameters (ψ_0, ψ_1, ψ_2) are affection probabilities for an individual with 0, 1 or 2 disease-causing alleles. Let p be the probability of the disease-causing allele. Assuming random mating, the founder alleles are independent and hence $P(a) = \prod_{j=1}^{2f} P(a_j)$, where $P(0) = 1 - p$ and $P(1) = p$. Hence the four genetic model parameters $(p, \psi_0, \psi_1, \psi_2)$ determine the posterior distribution π.

We further need to define a score function $S = S(w; \mathcal{P}, Y)$. In nonparametric linkage analysis S measures the extent to which the affected individuals share the same founder alleles. Therefore, S is a function of Y only through the set of n_a affected individuals. Three commonly used score functions (Whittemore and Halpern 1994; McPeek 1999) are

$$\begin{aligned} S_{\text{pairs}}(w) &= \sum_{k<l} \text{IBD}_{kl}, \\ S_{\text{all}}(w) &= \sum_u \text{nrperm}(a(u)), \\ S_{\text{robdom}}(w) &= \sum_{j=1}^{2f} 7^{n_j}. \end{aligned} \tag{16}$$

The first sum ranges over all pairs k, l of affected individuals and $\text{IBD}_{kl} = \text{IBD}_{kl}(w)$ is the number of alleles that k and l share identical by descent, that is, from the same founder allele. In the middle equation u picks one allele from each affected individual, and the sum ranges over all 2^{n_a} ways to do this. Further, $a(u)$ is a vector of length n_a containing the founder alleles picked by u, and $\text{nrperm}(a(u))$ is the number of permutations of $a(u)$ that leaves it unchanged. In the definition of S_{robdom}, $n_j = n_j(w)$ is the number of affected pedigree members that share at least one copy of the jth founder allele. It is assumed that all score functions in (16) are standardized so that $(1, S) = 0$ and $\|S\| = 1$.

When all N pedigrees are of the same type it follows that the limit process in Theorem 4 has

$$\eta = \mu(\tau) = \xi(\tilde{S}, S),$$

$$\rho = \sum_{l=1}^{m} l\kappa_l,$$

$$d = \frac{\sum_{l=1}^{m} l\delta_l}{\sum_{l=1}^{m} l\kappa_l},$$

if family type index ϕ is omitted. Hence ρ is the 'average frequency size' of S and d depends on the frequencies of $\tilde{S}$ in relation to those of S.

When computing ρ and d, we utilize founder phase symmetry reduction of inheritance vectors and fast Fourier transforms of functions $S \in \mathcal{A}_m$ (Kruglyak and Lander 1998). Then $R_{\tilde{S}}$ and R_S are computed in $O(2^{m-f} \log 2^{m-f})$ steps for a single pedigree. In Tables 1–3,

Table 1. Values of κ_l and ρ for various score functions, pedigrees and phenotypes when either $N = 1$ or $N > 1$ but all pedigrees are of the same type. For pedigree numbers and labelling of individuals, see Figure 1. The possible phenotypes are 1 (affected) and $*$ (either unaffected, 0, or unknown, ?). S_{all} and S_{robdom} are not included for those combinations of $(\mathcal{P}, Y)$ where they give the same results as S_{pairs}

S	$\mathcal{P}$	Y	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	ρ
S_{pairs}	1	(*, *, 1, 1)	0	1	0	0	0	0	2
S_{pairs}	2	(*, *, 1, 1, 1)	0	1	0	0	0	0	2
S_{pairs}	3	(*, *, 1, 1, 1, 1)	0	1	0	0	0	0	2
S_{all}			0	0.9855	0	0.0423	0	0	2.0291
S_{robdom}			0	0.9499	0	0.0501	0	0	2.1002
S_{pairs}	4	(*, *, 1, 1, 1, 1, 1)	0	1	0	0	0	0	2
S_{all}			0	0.9577	0	0.0423	0	0	2.0847
S_{robdom}			0	0.8634	0	0.1366	0	0	2.2732
S_{pairs}	4	(*, *, *, 1, 1, 1, 1)	0	1	0	0	0	0	2
S_{all}			0	0.9855	0	0.0145	0	0	2.0291
S_{robdom}			0	0.9499	0	0.0501	0	0	2.1002
S_{pairs}	4	(*, *, *, *, 1, 1, 1)	0	1	0	0	0	0	2
S_{pairs}	4	(*, *, *, *, *, 1, 1)	0	1	0	0	0	0	2
S_{pairs}	5	(*, *, *, *, *, *, 1, 1)	0	0.5	0.3333	0.1667	0	0	2.6667
S_{pairs}	5	(*, *, *, 1, 1, *, 1, 1)	0	0.8137	0.1765	0.0098	0	0	2.1961
S_{all}			0	0.6556	0.2981	0.0462	0	0	2.3906
S_{robdom}			0	0.5236	0.3658	0.1106	0	0	2.5870
S_{pairs}	5	(*, 1, *, 1, 1, *, 1, 1)	0.1356	0.7034	0.1525	0.0085	0	0	2.0339
S_{all}			0.2166	0.4887	0.2513	0.0435	0	0	2.1216
S_{robdom}			0.2482	0.3626	0.2988	0.0904	0	0	2.2314
S_{pairs}	6	(*, ..., *, 1, 1)	0.1333	0.1667	0.2667	0.2667	0.1333	0.0333	3.2
S_{pairs}	7	see Figure 1	0.1356	0.7034	0.1525	0.0085	0	0	2.0339
S_{all}			0.2166	0.4887	0.2513	0.0435	0	0	2.1216
S_{robdom}			0.2482	0.3623	0.2988	0.0904	0	0	2.2314

Table 2. Values of δ_l, d, ξ, efficiency ($(S, \tilde{S})^2$) and η when $N = 1$, for score function S_{all}, various pedigrees and phenotype vectors. The genetic model is dominant ($\psi_0 = 0$, $\psi_1 = \psi_2 = 1$) with disease allele frequency 0.1. $\delta_l = 0$ when $l \geqslant 5$ except for pedigree 6, which has $\delta_5 = 0.1333$ and $\delta_6 = 0.0333$

$\mathcal{P}$	Y	δ_1	δ_2	δ_3	δ_4	d	ξ	$(S, \tilde{S})^2$	η
1	(?, ?, 1, 1)	0	1	0	0	1	0.4904	0.9986	0.4901
2	(?, ?, 1, 1, 1)	0	1	0	0	1	0.6916	0.9964	0.6907
3	(?, ?, 1, 1, 1, 1)	0	0.9586	0	0.0414	1.0265	0.7853	0.9370	0.7602
	(?, ?, 0, 1, 1, 1)	0	1	0	0	1	1.6972	0.4007	1.0743
	(?, ?, 0, 0, 1, 1)	0	1	0	0	1	1.8451	0.1337	0.6748
4	(?, ?, 1, 1, 1, 1, 1)	0	0.8896	0	0.1104	1.0653	0.7778	0.8694	0.7252
	(?, ?, 0, 1, 1, 1, 1)	0	0.9567	0	0.0433	1.0284	2.3266	0.4104	1.4904
	(?, ?, 0, 0, 1, 1, 1)	0	1	0	0	1	2.6302	0.1889	1.1431
5	(?, ?, ?, ?, ?, ?, 1, 1)	0	0.5	0.3333	0.1667	1	0.5811	1	0.5811
	(?, ?, ?, 1, 1, ?, 1, 1)	0	0.5698	0.3493	0.0809	1.0504	0.9532	0.8841	0.8962
	(?, ?, 0, 1, 1, 0, 1, 1)	0	0.5446	0.3564	0.0990	1.0685	1.1991	0.7822	1.0605
	(0, 1, 0, 1, 1, 0, 1, 1)	0.3141	0.3518	0.2570	0.0771	0.9884	2.3828	0.7733	2.0954
6	(?, ..., ?, 1, 1)	0.1333	0.1667	0.2667	0.2667	1	0.4340	1	0.4340
7	see Figure 1	0.2793	0.3575	0.2793	0.0838	1.0217	7.9373	0.0881	2.3563

Table 3. As Table 2, but for a recessive model ($\psi_0 = \psi_1 = 0$, $\psi_1 = 1$) with disease allele frequency 0.1. $\delta_l = 0$ when $l \geqslant 5$ except for pedigree 6, which has $\delta_5 = 0.1333$ and $\delta_6 = 0.0333$

$\mathcal{P}$	Y	δ_1	δ_2	δ_3	δ_4	d	ξ	$(S, \tilde{S})^2$	η
1	(?,?,1,1)	0	1	0	0	1	1.3368	0.7492	1.1571
2	(?,?,1,1,1)	0	1	0	0	1	2.2233	0.5818	1.6958
	(?,?,0,1,1)	0	1	0	0	1	1.8014	0.4835	1.2526
3	(?,?,1,1,1,1)	0	0.9432	0	0.0568	1.0417	2.7880	0.4793	1.9302
	(?,?,0,1,1,1)	0	1	0	0	1	3.2444	0.3776	1.9938
	(?,?,0,0,1,1)	0	1	0	0	1	2.2496	0.3327	1.2926
4	(?,?,1,1,1,1,1)	0	0.8557	0	0.1443	1.0978	2.7690	0.4421	1.8411
	(?,?,0,1,1,1,1)	0	0.9411	0	0.0589	1.0437	4.9665	0.2875	2.6631
	(?,?,0,0,1,1,1)	0	1	0	0	1	4.1454	0.2573	2.1027
5	(?,?,?,?,?,?,1,1)	0	0.5	0.3333	0.1667	1	1.1991	1	1.1991
	(?,?,0,0,0,0,1,1)	0	0.5203	0.2952	0.1845	0.9991	1.4460	0.9854	1.4354
	(0,0,0,0,0,0,1,1)	0	0.4821	0.3214	0.1964	1.0179	1.6248	0.9899	1.6166
6	(?, ..., ?,1,1)	0.1333	0.1667	0.2667	0.2667	1	1.3943	1	1.3943
7	see Figure 1	0	1	0	0	0.9427	3.8730	0.0056	0.2896

we have evaluated η, d, ρ, κ_l, δ_l, ξ and $(\tilde{S}, S)^2$ when $N = 1$ for the three score functions (16), two genetic models, the pedigrees in Figure 1 and various phenotype vectors. It is seen that ρ and d in most cases are quite close to 2 and 1, whereas η varies a lot. From Table 1 we find that the average frequency size, ρ, is highest for S_{robdom}, followed by S_{all}

and S_{pairs}. The efficiencies of all three score functions (16) compared to the optimal $\tilde{S}$ are generally smaller for larger pedigrees, especially when these have many unaffected individuals. Hence, there is often a considerable loss of performance with score functions based on affected individuals only. In Tables 2 and 3, we only included S_{all}. In general S_{robdom} is slightly more and S_{pairs} slightly less efficient than S_{all} for the dominant model, whereas the opposite is true for the recessive model. See also McPeek (1999), Feingold *et al.* (2000) and Sengul *et al.* (2001) for further comparisons between score functions.

For non-contiguous alternatives (8), the non-centrality parameter is

$$\eta = \mu(\tau) = \sqrt{N}\xi(\tilde{S}, S) \tag{17}$$

when all pedigrees are of the same type. More generally, η is essentially a weighted average of (17), with $\xi = \xi_\phi$ and $\tilde{S} = \tilde{S}_\phi$, for different pedigree types ϕ. Hence, η depends a lot on sample size N (in addition to dependence on genetic model, pedigree and score function, as reported in Tables 2–3). Therefore, it will vary a lot between data sets. The quantities d and ρ, on the other hand, are essentially weighted averages of the corresponding quantities d_ϕ and ρ_ϕ included in the sample. They depend very little on N and not on ξ. For this reason, d and ρ will usually not vary much between data sets and are quite stable around values 1 and 2–3, respectively. In Figures 2 and 3, we have plotted the (approximate) power (15) as a function of the (approximate) significance level (14) for chromosome- and genome-wide scans. In the plots, we have chosen values of η large enough (corresponding to large enough samples) to yield reasonable power. It is seen from these curves that η has a large effect on power but ρ and d do not.

7. Discussion

In this paper we have shown that spectral decomposition of score functions is a valuable tool when covariance and mean functions for linkage score functions Z are computed. Under the assumption of perfect marker information, the results in this paper hold for general pedigree structures. We derived an invariance principle for Z asymptotically as the number of pedigrees tends to infinity. These results were applied to compute power β_∞ and significance levels α_∞ when the presence of a disease locus is tested. By plotting β_∞ as function of α_∞ (rather than the threshold T), we demonstrated that the non-centrality parameter essentially determines the strength of the test.

In principle, our results are valid for score functions S based on more or less arbitrary (that is, not necessarily binary) phenotypes and genetic models. See Commenges (1994), Tang and Siegmund (2001) and Hössjer (2003c; 2005) for examples of score functions for quantitative and other phenotypes. Technically, they are more realistic with a continuum of possible pedigree types $\phi = (\mathcal{P}, Y)$ for quantitative phenotypes. In fact, at the expense of more technical arguments, it is possible to generalize Theorem 4 to $K = \infty$, using a similar approach to that in Hössjer (2003a).

The perfect marker assumption used in this paper requires a dense set of genetic markers. In view of the current availability of several million single nucleotide polymorphism (SNP)

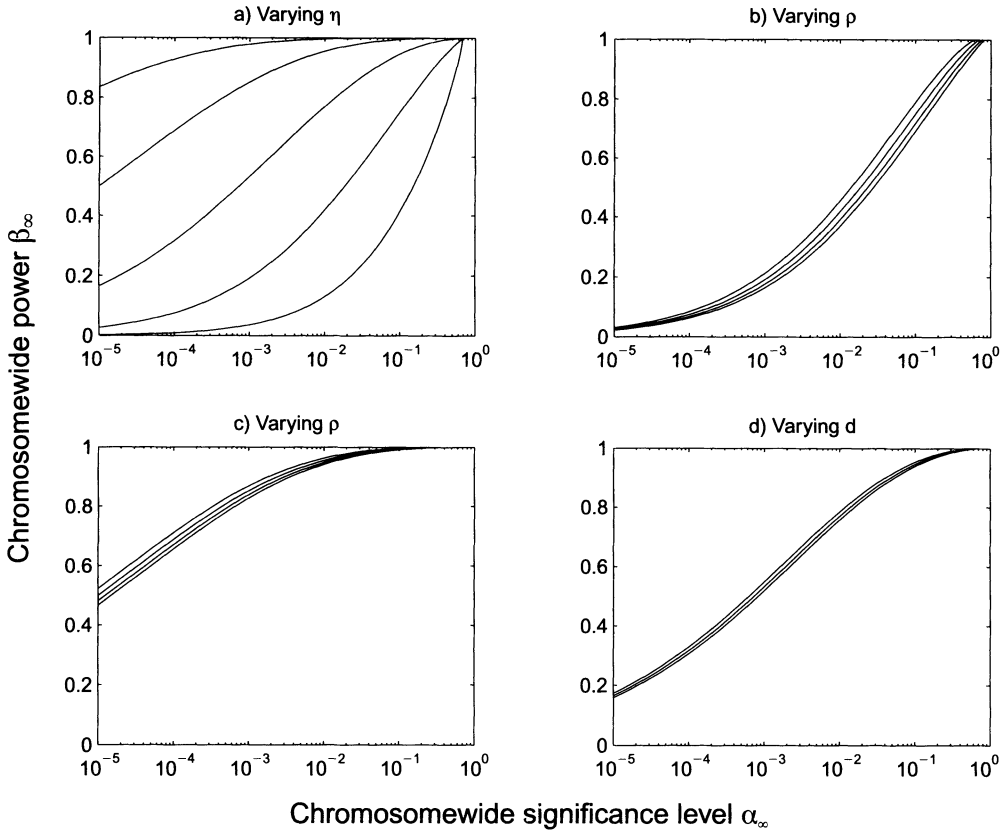

Figure 2. Plot of β_∞ as function of α_∞ for one chromosome when T varies, using approximations (14)–(15). Chromosome length is 1.5 Morgans and parameter values (a) $\eta = 2, 3, 4, 5, 6$, $\rho = 2$, $d = 1$, (b) $\rho = 1.5, 2, 2.5, 3$, $\eta = 3$, $d = 1$, (c) $\rho = 1.5, 2, 2.5, 3$, $\eta = 5$, $d = 1$ and (d) $d = 0.9, 1, 1.1$, $\eta = 4$, $\rho = 2$. To distinguish curves, note that β_∞ is an increasing function of η and a decreasing function of ρ and d when α_∞ is kept fixed.

markers in the human genome, this is not unrealistic. For large multigenerational pedigrees, the assumption that all (or sufficiently many) pedigree members are genotyped for DNA is not realistic though.

As a final remark, we notice that quite a different kind of asymptotics occurs when the focus is on *estimating* the position of τ rather than on *testing* for its presence. Then, assuming H_1, we know that $\tau \in [0, L]$ and rescale Z locally around τ. For a fixed sequence of alternatives (8) (rather than contiguous alternatives (12)), the rescaled process

$$\tilde{Z}(s) = \sqrt{N}(Z(\tau + N^{-1}s) - Z(\tau)) \tag{18}$$

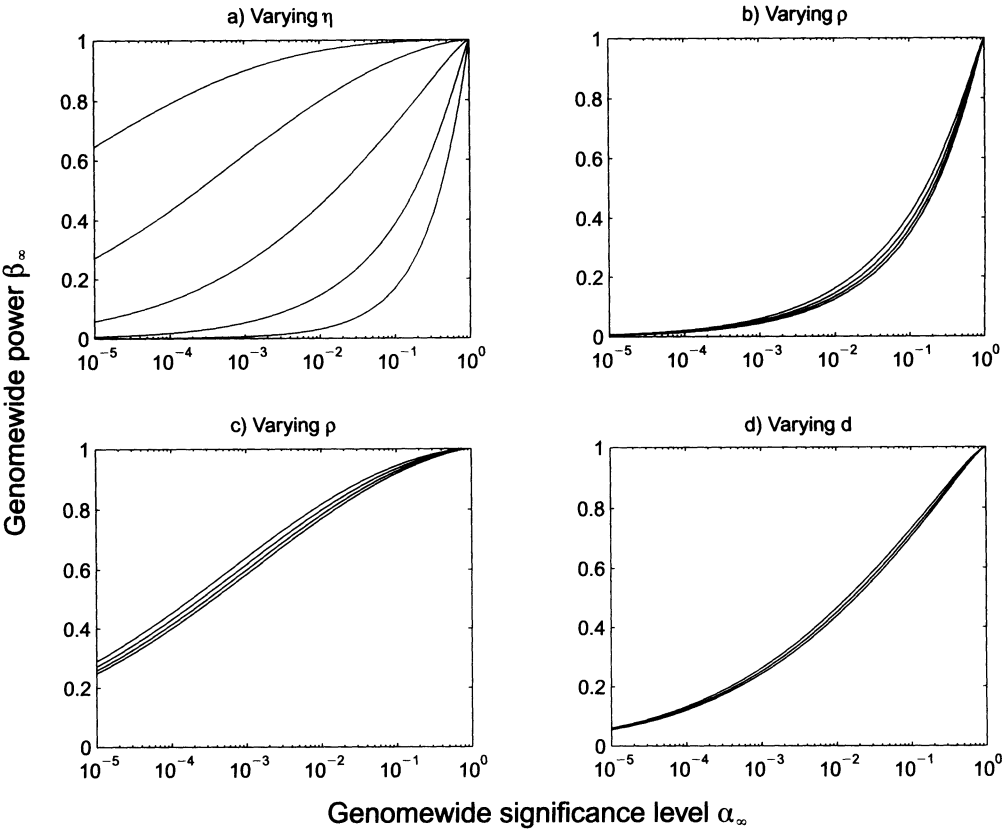

Figure 3. Plot of β_∞ as function of α_∞ for 22 autosomes when T varies, using $\alpha_\infty(T) = 1 - \prod_{s=1}^{22}(1 - \alpha_{\infty,s}(T))$ and $\beta_\infty(T) = 1 - (1 - \beta_{\infty,t}(T))\prod_{s\neq t}(1 - \alpha_{\infty,s}(T))$. Here t is the chromosome where τ is located and $\beta_{\infty,s}(T)$ and $\alpha_{\infty,s}(T)$ approximations for chromosome s using (14)–(15). The 22 chromosome lengths are taken from Table 1.2 of Ott (1999) and $t = 1$ is assumed. The values of η, ρ and d are as in Figure 2.

is defined over local neighbourhoods of τ of size $O(N^{-1})$. The N^{-1} rescaling implies that the length of confidence regions of τ tends to zero at a fast rate N^{-1}. As $N \to \infty$, $\tilde{Z}$ converges weakly to a compound Poisson process. The reason for a non-Gaussian limit is that the number of crossovers in a window of size $O(N^{-1})$ does not grow with N. For this reason, individual crossovers are visible in the limit process as jump points of the Poisson process. See Kong and Wright (1994), Kruglyak and Lander (1995), Dupuis and Siegmund (1999) and Hössjer (2003a; 2003b) for more details.

Appendix: Proofs

Proof of Theorem 1. We start by proving the spectral decomposition for Q_h and proceed by induction with respect to m. For $m = 1$ we have

$$Q_h = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix},$$

where $\theta = \theta(h) = (1 - \exp(-2h))/2$ is the recombination fraction between two loci at distance h Morgans. That is, θ is the probability of an odd number of crossovers between the two loci. Now $S_0(0) = S_0(1) = 1$ and $S_1(0) = 1$, $S_1(1) = -1$. We write $S_0 = (1, 1)$ and $S_1 = (1, -1)$ as row vectors. Inspection shows that S_0 has eigenvalue 1 and S_1 eigenvalue $1 - 2\theta = \exp(-2h)$, and this completes the proof for $m = 1$.

For the induction step, let superscripts denote the number of meioses and notice that

$$Q_h^{m_1+m_2} = Q_h^{m_1} \otimes Q_h^{m_2},$$

$$S_w = S_{w_1} \otimes S_{w_2},$$

where $w_1 \in \mathbb{Z}_2^{m_1}$, $w_2 \in \mathbb{Z}_2^{m_2}$, $w = (w_1, w_2) \in \mathbb{Z}_2^{m_1+m_2}$ and $\otimes$ is the tensor product. Hence, if S_{w_1} and S_{w_2} are eigenvectors of $Q_h^{m_1}$ and $Q_h^{m_2}$ with eigenvalues λ_{w_1} and λ_{w_2}, then S_w is an eigenvector of $Q_h^{m_1+m_2}$ with eigenvalue $\lambda_{w_1}\lambda_{w_2}$. By induction, we assume that $\lambda_{w_1} = \exp(-2|w_1|h)$ and $\lambda_{w_2} = \exp(-2|w_2|h)$. Then $\lambda_{w_1}\lambda_{w_2} = \exp(-2(|w_1| + |w_2|)h) = \exp(-2|w|h)$. This completes the proof for Q_h. Notice then that

$$A = \lim_{h \to 0} (Q_h - I)/h, \tag{A.1}$$

where the limit is taken from above and I is the identity operator on $\mathcal{A} = \mathcal{A}_m$. For each fixed $h > 0$, S_w is an eigenvector of the right-hand side of (A.1) with eigenvalue $(\exp(-2h|w|) - 1)/h$. By continuity, we can take the limit $h \to 0$ and conclude that S_w is an eigenvector of A with eigenvalue $-2|w|$. □

Proof of Theorem 2. Let $\lambda_w = \exp(-2h|w|)$ be the eigenvalue of S_w for the operator $Q_h = \{Q_h(w, w'); w, w' \in \mathbb{Z}_2^m\}$. Viewing $S = \{S(w); w \in \mathbb{Z}_2^m\}$ as a row vector in $\mathbb{Z}_2^m$ and letting S^T be the transpose of S, we obtain

$$
\begin{aligned}
r_Z(h) &= \mathrm{E}_{H_0}(Z(t)Z(t+h)) \\
&= \textstyle\sum_{w,w'} P_{H_0}(v(t) = w)P(v(t+h) = w' | v(t) = w)S(w)S(w') \\
&= \textstyle\sum_{w,w'} \pi_0(w)Q_h(w, w')S(w)S(w') \\
&= 2^{-m} S Q_h S^{\mathrm{T}} \\
&= 2^{-m} (\textstyle\sum_w R_S(w)S_w) Q_h (\sum_{w'} R_S(w')S_{w'})^{\mathrm{T}} \\
&= 2^{-m} \textstyle\sum_{w,w'} R_S(w)R_S(w')S_w Q_h S_{w'}^{\mathrm{T}} \\
&= \textstyle\sum_w R_S^2(w)\lambda_w,
\end{aligned}
$$

where $S_w Q_h S_{w'}^{\mathrm{T}} = 2^m \lambda_w(S_w, S_{w'})$ was used in the last step. By collecting all w into groups with identical $|w|$, (6) follows from the last line. Finally, by Parseval's formula,

$$
\sum_{l=1}^{m} \kappa_l = \sum_{w \neq 0} R_S^2(w) = \sum_w R_S^2(w) = \|S\|^2 = 1,
$$

since $R_S(0) = (S, 1) = 0$. □

Proof of Theorem 3. Viewing π in (7) as a row vector in $\mathbb{Z}_2^m$, the marginal distribution of $v(t)$ under H_1 is $\pi Q_{|t-\tau|}$. Hence, using a similar expansion to that in the proof of Theorem 2, we obtain

$$
\begin{aligned}
\mu_Z(t) &= \textstyle\sum_w P_{H_1}(v(t) = w)S(w) \\
&= \pi Q_{|t-\tau|} S^{\mathrm{T}} \\
&= (\textstyle\sum_w R_\pi(w)S_w) Q_{|t-\tau|} (\sum_{w'} R_S(w')S_{w'})^{\mathrm{T}} \\
&= 2^m \textstyle\sum_w R_\pi(w)R_S(w)\lambda_w \\
&= \xi \textstyle\sum_w R_{\tilde{S}}(w)R_S(w)\lambda_w,
\end{aligned}
$$

where $\lambda_w = \exp(-2|w|\,|t-\tau|)$ is the eigenvalue of S_w for the operator $Q_{|t-\tau|}$. The proof is finished by grouping all terms with the same $|w|$. □

Proof of Theorem 4. We start proving convergence of the first two moments of $Z = Z_N$ in (9) towards those of the limit process $\mu + W$. Let $\pi_{\phi t} = \pi_\phi Q_{|t-\tau|}$ be the distribution of $v_i(t)$ under H_1 for a pedigree of type ϕ (that is, $\phi_i = \phi$). Then, since Q_h is a self-adjoint operator on $\mathcal{A}_{m_\phi}$ and $(1, Q_h S_\phi) = (1Q_h, S_\phi) = (1, S_\phi) = 0$, we obtain

$$\begin{aligned}\mu_\phi(t) := \sqrt{N}\mathrm{E}(S_\phi(v_i(t))) &= \sqrt{N}(2^{m_\phi}\pi_{\phi t}, S_\phi)\\ &= \sqrt{N}(2^{m_\phi}\pi_\phi, Q_{|t-\tau|}S_\phi)\\ &= \sqrt{N}(2^{m_\phi}\pi_\phi - 1, Q_{|t-\tau|}S_\phi)\\ &\to \xi_\phi(\tilde{S}_\phi, Q_{|t-\tau|}S_\phi)\\ &= \xi_\phi \textstyle\sum_w R_{\tilde{S}_\phi}(w)R_{S_\phi}(w)\exp(-2|w|\,|t-\tau|)\end{aligned}$$

as $N \to \infty$. Let $\nu_{\phi N} = N_\phi/N$ and $\bar{\nu}_{\phi N} = \gamma_\phi^2\nu_{\phi N}/\left(\sum_{\phi'=1}^K \gamma_{\phi'}^2\nu_{\phi' N}\right)$. Then write the first moment of Z_N as

$$\mathrm{E}(Z_N(t)) = \sum_{\phi=1}^K \sqrt{\nu_{\phi N}\bar{\nu}_{\phi N}}\mu_\phi(t).$$

By combining the last two displayed equations, collecting terms with the same $|w|$ and noticing $\nu_{\phi N} \to \nu_\phi$ and $\bar{\nu}_{\phi N} \to \bar{\nu}_\phi$, it follows that $\mathrm{E}(Z_N(t)) \to \mu(t)$ as $N \to \infty$.

Let $\pi_{\phi st} = P_{H_1}((v_i(s), v_i(t)) = (\cdot, \cdot))$ be the bivariate marginal distribution of $v_i(\cdot)$ for a pedigree of type ϕ (that is, $\phi_i = \phi$). Then, viewing $\pi_{\phi st}$ as an element of $\mathcal{A}_{2m_\phi}$, we have

$$\begin{aligned}r_\phi(t, t+h) &:= \mathrm{cov}(S_\phi(v_i(t)), S_\phi(v_i(t+h)))\\ &= (2^{2m_\phi}\pi_{\phi t,t+h}, S_\phi \otimes S_\phi) - (2^{m_\phi}\pi_{\phi t}, S_\phi)(2^{m_\phi}\pi_{\phi,t+h}, S_\phi).\end{aligned}$$

Since $2^{m_\phi}\pi_\phi \to 1$ as $N \to \infty$, it follows that $2^{m_\phi}\pi_{\phi t} \to 1$ and $2^{m_\phi}\pi_{\phi t,t+h} \to Q_h$. In the last limit we interpreted Q_h as an element of $\mathcal{A}_{2m_\phi}$. Hence,

$$\begin{aligned}r_\phi(t, t+h) &\to (2^{m_\phi}Q_h, S_\phi \otimes S_\phi) - (1, S_\phi)^2\\ &= (2^{m_\phi}Q_h, S_\phi \otimes S_\phi)\\ &= 2^{-m_\phi}SQ_hS^{\mathrm{T}}\\ &= \textstyle\sum_w R_{S_\phi}^2(w)\exp(-2|w|h).\end{aligned}$$

The covariance function of Z_N can be written as

$$\mathrm{cov}(Z_N(t), Z_N(t+h)) = \sum_{\phi=1}^K \bar{\nu}_{\phi N} r_\phi(t, t+h),$$

and the last two displayed equations imply convergence of covariances of Z_N towards those of W.

Convergence of finite-dimensional distributions of Z_N is proved in the standard way using the Cramér-Wold device and the Lindeberg–Feller central limit theorem for triangular arrays. We omit the details, but notice that linear combinations (with fixed weights and time indices t) of $S_{\phi_i}(v_i(t))$ are uniformly bounded random variables in i, and this enforces the central limit theorem.

It remains to prove tightness. According to Theorem 15.6 in Billingsley (1968), it suffices to find a constant $C > 0$ such that

$$I := \mathrm{E}((Z_N(t) - Z_N(t_1))^2(Z_N(t_2) - Z_N(t))^2) \leqslant C(t_2 - t_1)^2 \tag{A.2}$$

uniformly for all large enough N and $0 \leqslant t_1 < t < t_2 \leqslant L$. Write $\bar{\gamma}_i = \gamma_{\phi_i}/\sqrt{\sum_{i=1}^N \gamma_{\phi_i}^2}$, $U_i = S_{\phi_i}(v_i(t)) - S_{\phi_i}(v_i(t_1))$ and $V_i = S_{\phi_i}(v_i(t_2)) - S_{\phi_i}(v_i(t))$. Then

$$\begin{aligned} I &= \mathrm{E}\left(\left(\sum_{i=1}^N \bar{\gamma}_i U_i\right)^2 \left(\sum_{i=1}^N \bar{\gamma}_i V_i\right)^2\right) \\ &= \sum_{i=1}^N \bar{\gamma}_i^4 \mathrm{E}(U_i^2 V_i^2) \\ &\quad + \sum_{i \neq j} \bar{\gamma}_i^2 \bar{\gamma}_j^2 \mathrm{E}(U_i^2 V_j^2) \\ &\quad + 2 \sum_{i \neq j} \bar{\gamma}_i^2 \bar{\gamma}_j^2 \mathrm{E}(U_i U_j V_i V_j) \\ &=: i + ii + iii. \end{aligned}$$

Using $2|U_i U_j| \leqslant (U_i^2 + U_j^2)$ and $2|V_i V_j| \leqslant (V_i^2 + V_j^2)$, it follows that $|iii| \leqslant i + ii$. Hence $I \leqslant 2(i + ii)$. Because of the Markov property of all $v_i(\cdot)$ (also under H_1) and the uniform boundedness of the functions S_ϕ, it is possible to find constants C_1, $C_2 > 0$ such that

$$\mathrm{E}(S_{\phi_i}(v_i(t_2)) - S_{\phi_i}(v_i(t_1)))^2 \leqslant C_1|t_2 - t_1|,$$

$$\mathrm{E}((S_{\phi_i}(v_i(t)) - S_{\phi_i}(v_i(t_1)))^2(S_{\phi_i}(v_i(t_2)) - S_{\phi_i}(v_i(t)))^2) \leqslant C_2(t_2 - t_1)^2,$$

uniformly for all i and $0 \leqslant t_1 < t < t_2 \leqslant L$. Since $\sum_{i=1}^N \bar{\gamma}_i^2 = 1$, if follows that $ii \leqslant C_1^2(t_2 - t_1)^2$. Further, since $\max_{1 \leqslant i \leqslant N} \bar{\gamma}_i \leqslant 1$ and hence $\sum_{i=1}^N \bar{\gamma}_i^4 \leqslant 1$, it follows that $i \leqslant C_2(t_2 - t_1)^2$. But then (A.2) holds with $C = 2(C_1^2 + C_2)$, and weak convergence under H_1 is proved.

Weak convergence under H_0 is proved in the same way. We simply put $\xi_\phi = 0$ for all ϕ in the proof above. □

Acknowledgement

This research was supported by the Swedish Research Council, under contract no. 626–2002–6286. The author wishes to thank an anonymous referee for helpful comments.

References

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Commenges, D. (1994) Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic. *Genetic Epidemiology*, **11**, 189–200.

Diaconis, P. (1988) *Group Representations in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Donnelly, P. (1983) The probability that some related individuals share some section of the genome identical by descent. *Theoret. Popul. Biol.*, **23**, 34–64.

Dudoit, S. and Speed, T.P. (1999) A score test for linkage using identity by descent data from sibships. *Ann. Statist.*, **27**, 943–986.

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. and Siegmund, D. (1997) Strategies for mapping heterogeneous recessive traits by allele-sharing methods. *Amer. J. Hum. Genetics*, **60**, 965–978.

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

Feingold, E., Song, K.K. and Weeks, D.E. (2000) Comparison of allele-sharing statistics for general pedigrees. *Genetic Epidemiology*, **19** (Suppl. 1), S92–S98.

Haldane, J.B.S. (1919) The combination of linkage values and the calculation of distances between loci of unlinked factors. *J. Genetics*, **8**, 299–309.

Hössjer, O. (2003a) Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Ann. Statist*, **31**, 1075–1109.

Hössjer, O. (2003b) Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology*, **25**, 59–72.

Hössjer, O. (2003c) Determining inheritance distributions via stochastic penetrances. *J. Amer Statist Assoc.*, **98**, 1035–1051.

Hössjer, O. (2004) Information and effective number of meioses in linkage analysis. *J. Math. Biol.*, **50**(2), 208–232.

Hössjer, O. (2005) Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics*, **6**, 313–332. Supplementary material at http://biostatistics.oupjournals.org/.

Kong, A. and Wright, F. (1994) Asymptotic theory for gene mapping. *Proc. Natl. Acad. Sci. USA*, **91**, 9705–9709.

Kruglyak, L. and Lander, E.S. (1995) High-resolution gene mapping of complex traits. *Amer. J. Hum. Genetics*, **56**, 1212–1223.

Kruglyak, L. and Lander, E. (1998) Faster multipoint linkage analysis using Fourier transforms, *J. Comput. Biol.*, **5**(1), 1–7.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996) Parametric and nonparametric linkage analysis: A unified multipoint approach. *Amer. J. Hum. Genetics*, **58**, 1347–1363.

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

Lander, E.L. and Kruglyak, L. (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

McPeek, M.S. (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology*, **16**, 225–249.

Ott, J. (1999) *Analysis of Human Genetic Linkage*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.

Sengul, H., Weeks, D.E. and Feingold, E. (2001) A survey of affected-sibship statistics for nonparametric linkage analysis. *Amer. J. Hum. Genetics*, **69**, 179–190.

Sham, P. (1998) *Statistics in Human Genetics*. London: Arnold.

Sham, P., Zhao, J.H. and Curtis, D. (1997) Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann. Hum. Genetics*, **61**, 61–69.

Tang, H.-K. and Siegmund, D. (2001) Mapping quantitative trait loci in oligogenic models. *Biostatistics*, **2**, 147–162.

Whittemore A.S. and Halpern, J. (1994) A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

Received January 2004 and revised May 2005