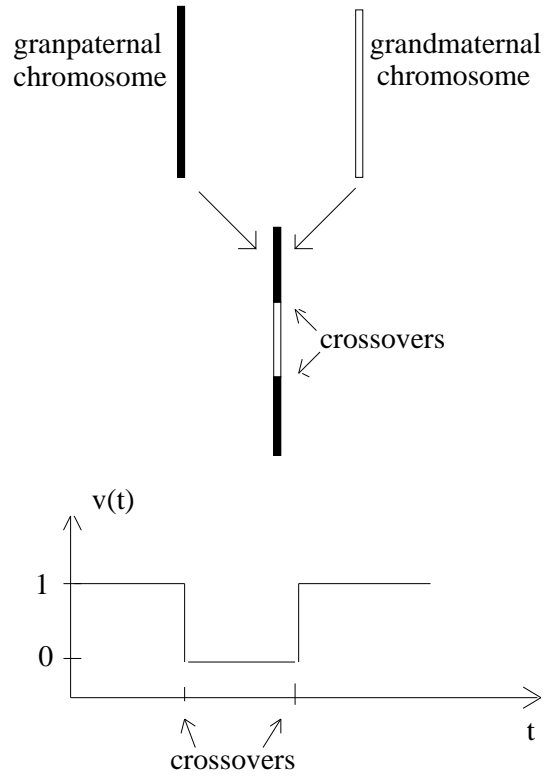


Invariance principles and spectral decomposition in genetics

Ola Hössjer
Dept. of Mathematics
Stockholm University

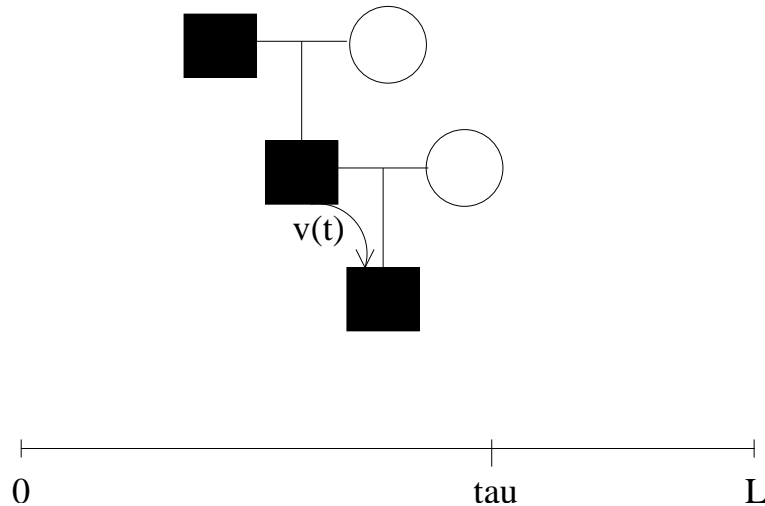
Crossovers



Haldane (1919): v stationary Markov process on $\{0, 1\}$ with intensity matrix

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

Grandparent-grandchild



Chromosome of length L

Disease gene at (unknown) position τ .

Boxes=male, circles=female, black=affected.

H_0 : τ located on another chromosome

H_1 : $\tau \in [0, L]$.

Under H_1 , information about $v(\cdot) = \{v(t); 0 \leq t \leq L\}$ changes:

1. Let $\pi(w) = P(v(\tau) = w)$. Assume $\pi(1) > 0.5$.
2. Given $v(\tau)$, $v(\cdot)$ proceeds as two independent Markov process to the left and right with intensity matrix A .

Monogenic disease

Two alleles at τ :

Normal (a) and disease causing (A).

Parameters:

$$\begin{aligned} p &= P(A) \\ \psi_0 &= P(\text{affected}|aa) \\ \psi_1 &= P(\text{affected}|Aa) \\ \psi_2 &= P(\text{affected}|AA) \end{aligned}$$

Then

$$\pi(1) = f(p, \psi_0, \psi_1, \psi_2).$$

Examples:

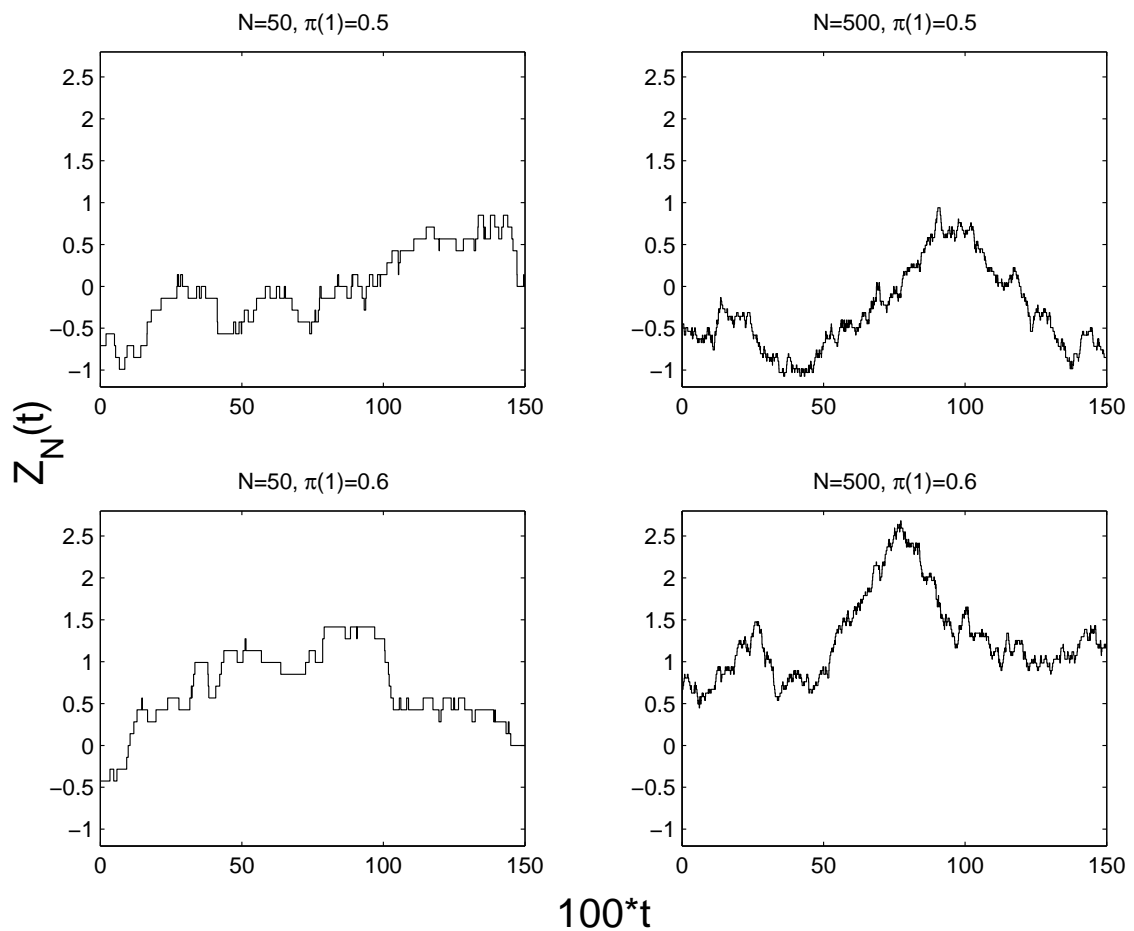
1. Dominant: $f(p, 0, 1, 1) = 1$
2. No genetic effect: $f(p, \psi, \psi, \psi) = 0.5$.

Scores for N families

Let $S(0) = -1$, $S(1) = 1$ and

$$Z_N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^N S(v_i(t)),$$

where $v_i(t)$ is inheritance indicator of family i at position t .



Disease gene at $\tau = 0.75$.

Asymptotics

Under H_0 ,

$$Z_N \xrightarrow{\mathcal{L}} Z \quad \text{on } D([0, L]),$$

as $N \rightarrow \infty$, where Z is a stationary Ornstein-Uhlenbeck process with covariance function

$$r_Z(h) = \text{Cov}(Z(t), Z(t+h)) = \exp(-2|h|).$$

Under sequence of alternatives

$$H_{1N} : \pi(1) = 0.5(1 + \xi/\sqrt{N}),$$

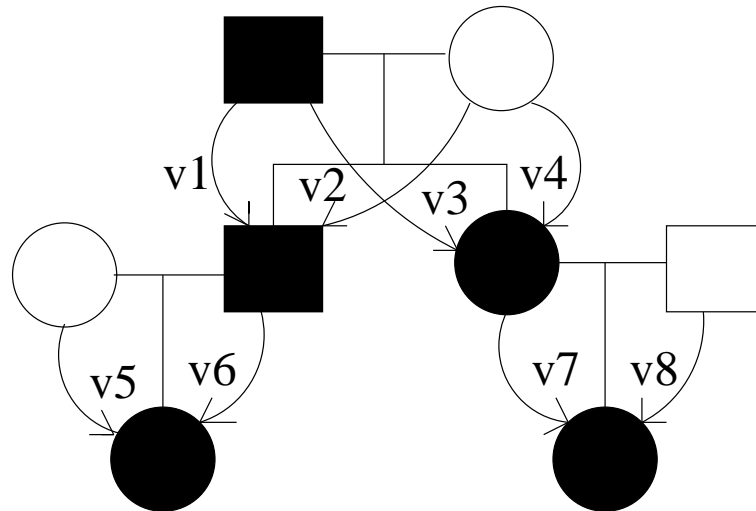
we have

$$Z_N \xrightarrow{\mathcal{L}} Z + \mu,$$

as $N \rightarrow \infty$, where

$$\mu(t) = \xi \exp(-2|t - \tau|).$$

General family structures



$n =$ nr. of individuals $= 8$

$f =$ nr. of founders $= 4$

$m =$ nr. of meioses $= 2(n - f) = 8$.

Inheritance vector at pos. t (Donnelly, 1983):

$$\mathbf{v}(t) = (v_1(t), \dots, v_m(t)).$$

Distribution of $\mathbf{v}(\cdot)$:

H_0 : $\mathbf{v}(\cdot)$ stationary Markov process on $\{0, 1\}^m$ with marginal distribution

$$P(\mathbf{v}(t) = \mathbf{w}) = \pi_0(\mathbf{w}) = 2^{-m},$$

$\forall \mathbf{w} \in \{0, 1\}^m$, and intensity matrix

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

where $|\mathbf{w} - \mathbf{w}'| = \sum_1^m |w_i - w'_i|$ is the Hamming distance between \mathbf{w} and \mathbf{w}' .

H_1 : At disease gene ($t = \tau$),

$$P(\mathbf{v}(\tau) = \mathbf{w}) = \pi(\mathbf{w})$$

Given $\mathbf{v}(\tau)$, $\mathbf{v}(\cdot)$ proceeds as two independent Markov processes with intensity matrix A to the left and right of τ .

Scores for general family structures

Define

$$S : \{0, 1\}^m \longrightarrow \mathbb{R}$$

where $S(\mathbf{w})$ is large if affected pedigree members share the same founder alleles. (Or $S(\mathbf{w})$ large when $\pi(\mathbf{w})$ is.)

Example:

$$S_{\text{pairs}}(\mathbf{w}) = \sum_{j, k \text{ affected}}^n I_{jk},$$

where

$$I_{jk} = I_{jk}(\mathbf{w}) \in \{0, 1, 2\}.$$

is number of founder alleles shared by affected individuals j and k .

Asymptotics

1) Standardize S : $E_{H_0}(S) = 0$, $\text{Var}_{H_0}(S) = 1$.

2) For N families, put

$$Z_N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^N S(\mathbf{v}_i(t)),$$

where $\mathbf{v}_i(t)$ is inheritance vector at pos. t for family i .

3) As $N \rightarrow \infty$

$$Z_N \xrightarrow{\mathcal{L}} \begin{cases} Z, & \text{under } H_0, \\ Z + \mu, & \text{under } H_{1N}, \end{cases}$$

where

$$H_{1N} : \pi(\mathbf{w}) = 2^{-m} (1 + \xi \tilde{S}(\mathbf{w}) / \sqrt{N}),$$

and \tilde{S} is standardized; $E_{H_0}(\tilde{S}) = 0$, $\text{Var}_{H_0}(\tilde{S}) = 1$.

4) Z is a mixture of OU-processes, and μ mixture of double exponentials

$$\begin{aligned} r_Z(h) &= \sum_{l=1}^m \kappa_l \exp(-2l|h|), \quad \sum_{l=1}^m \kappa_l = 1, \\ \mu(t) &= \sum_{l=1}^m \eta_l \exp(-2l|t - \tau|). \end{aligned}$$

Space of mappings

Let

$$\mathcal{A} = \{S; S : \{0, l\}^m \rightarrow \mathbb{R}\} = \mathbb{R}^{2^m}$$

with inner product

$$(S, \tilde{S}) = E_{H_0}(S\tilde{S}) = 2^{-m} \sum_{\mathbf{w}} S(\mathbf{w})\tilde{S}(\mathbf{w})$$

and ON-system of 2^m basis functions $S_{\mathbf{u}}$, $\forall \mathbf{u} \in \{0, 1\}^m$, where

$$S_{\mathbf{u}}(\mathbf{w}) = (-1)^{\mathbf{w} \cdot \mathbf{u}},$$

and $\mathbf{w} \cdot \mathbf{u} = \sum_{j=1}^m w_j u_j$ is vector dot product.

Expand $S \in \mathcal{A}$ as

$$S = \sum_{\mathbf{u}} F_S(\mathbf{u}) S_{\mathbf{u}},$$

where $F_S(\mathbf{u}) = (S, S_{\mathbf{u}})$. $F_S \in \mathcal{A}$ is (essentially) Fourier transform of S on the group $\{0, 1\}^m$ (see Diaconis, 1988, Kruglyak and Lander, 1998).

Basis functions

$$\underline{m = 1}$$

w	0	1
$S_0(w)$	1	1
$S_1(w)$	1	-1

$$\underline{m = 2}$$

w	(00)	(01)	(10)	(11)
$S_{00}(w)$	1	1	1	1
$S_{01}(w)$	1	-1	1	-1
$S_{10}(w)$	1	1	-1	-1
$S_{11}(w)$	1	-1	-1	1

Covariance function

- 1) Intensity matrix A and lag h transition matrix $P_h = \exp(|h|A)$ self-adjoint operators on \mathcal{A}
- 2) $S_{\mathbf{u}}$ eigenvector of A and P_h with eigenvalues $-2|\mathbf{u}|$ and $\exp(-2|h||\mathbf{u}|)$.
- 3) Covariance function under H_0 follows as

$$\begin{aligned} r_Z(h) &= E_{H_0}(S(v(t))S(v(t+h))) \\ &= (S, P_h S) \\ &= (\sum_{\mathbf{u}} F_S(\mathbf{u})S_{\mathbf{u}}, P_h \sum_{\mathbf{u}} F_S(\mathbf{u})S_{\mathbf{u}}) \\ &= \sum_{\mathbf{u}} \exp(-2|h||\mathbf{u}|) F_S^2(\mathbf{u}) \\ &= \sum_{l=1}^m \kappa_l \exp(-2l|h|), \end{aligned}$$

where

$$\kappa_l = \sum_{\mathbf{u}; |\mathbf{u}|=l} F_S^2(\mathbf{u}).$$

- 4) Covariance function under H_{1N} obtained similarly.

Mean Function

Recall that

$$\pi = P_{H_{1N}}(v(\tau) = \cdot) = 2^{-m}(1 + \xi \tilde{S}).$$

with $\xi \geq 0$ and $\pi, \tilde{S} \in \mathcal{A}$, $(1, \tilde{S}) = 0$ and $(\tilde{S}, \tilde{S}) = 1$.
Hence

$$\begin{aligned}\mu(t) &= \sqrt{N} E_{H_{1N}}(S(v(t))) \\ &= \sqrt{N} \sum_{\mathbf{w}} S(\mathbf{w}) P_{H_{1N}}(\mathbf{v}(t) = \mathbf{w}) \\ &= \sqrt{N} (S, P_{|t-\tau|} \pi) \\ &= \xi (S, P_{|t-\tau|} \tilde{S}) \\ &= \xi \sum_{\mathbf{u}} \exp(-2\|\mathbf{u}\||t-\tau|) F_S(\mathbf{u}) F_{\tilde{S}}(\mathbf{u}) \\ &= \sum_{l=1}^m \eta_l \exp(-2l|t-\tau|)\end{aligned}$$

where

$$\eta_l = \xi \sum_{\mathbf{u}; |\mathbf{u}|=l} F_S(\mathbf{u}) F_{\tilde{S}}(\mathbf{u}).$$

Significance level and power

Define

$$\begin{aligned}\eta &= \mu(\tau) \\ &= \sum_{l=1}^m \eta_l = \xi(S, \tilde{S}) \\ \rho &= -r'_Z(0)/2 \\ &= \sum_{l=1}^m l\kappa_l \\ d &= \mu'(\tau)/(\mu(\tau)r'_Z(0))\end{aligned}$$

Using extreme value theory of Gaussian processes

$$\begin{aligned}\alpha &= P_{H_0}(\sup_{0 \leq t \leq L} Z_N(t) \geq T) \\ &\approx 1 - (1 - \Phi(T))^{1+2\rho LT^2}\end{aligned}$$

and

$$\begin{aligned}\beta &= P_{H_{1N}}(\sup_{0 \leq t \leq L} Z_N(t) \geq T) \\ &\approx 1 - \Phi(T - \eta) + \varphi(T - \eta) \left(\frac{2}{\eta d} - \frac{1}{\eta(2d-1)+T} \right),\end{aligned}$$

cf. Leadbetter et al. (1983), Siegmund (1986), Aldous (1989), Feingold et al. (1993), Lander and Kruglyak (1995) and Ängquist and Hössjer (2005).

Example families

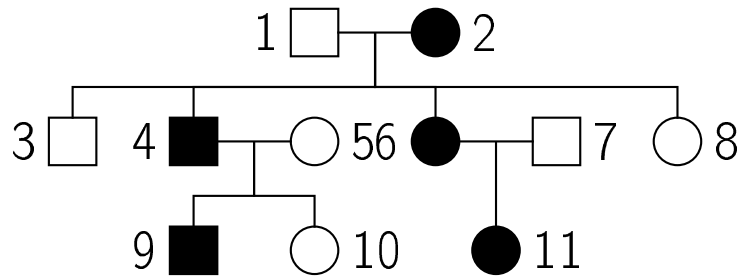
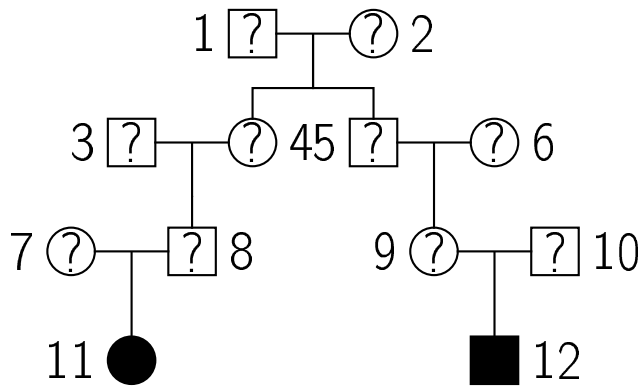
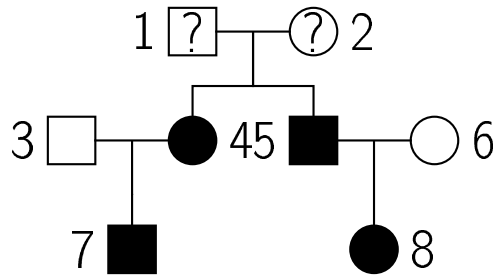


Figure 1:

Examples of covariance functions

Family	Affecteds	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	ρ
1	Children	0	1	0	0	0	0	2
2	All children	0	1	0	0	0	0	2
3	All children	0	1	0	0	0	0	2
4	All children	0	1	0	0	0	0	2
4	4 children	0	1	0	0	0	0	2
4	3 children	0	1	0	0	0	0	2
4	2 children	0	1	0	0	0	0	2
5	Cousins	0	0.5	0.333	0.167	0	0	2.667
5	Sibs, cousins	0	0.814	0.176	0.010	0	0	2.196
5	Grandmother, sibs, cousins	0.136	0.703	0.152	0.008	0	0	2.034
6	Second cousins	0.133	0.167	0.267	0.267	0.133	0.033	3.2
7	see figure	0.136	0.703	0.152	0.008	0	0	2.034

$$S = S_{\text{pairs}}$$

Family $i(=1,2,3,4)$: Two parents, $i - 2$ children.

Families 5-7: See figure.

Examples of mean functions 1

F	η_1/η	η_2/η	η_3/η	η_4/η	d	ξ	$(S, \tilde{S})^2$	η
1	0	1	0	0	1	0.490	0.999	0.490
2	0	1	0	0	1	0.692	0.996	0.691
3	0	0.959	0	0.041	1.026	0.785	0.937	0.760
	0	1	0	0	1	1.697	0.401	1.074
	0	1	0	0	1	1.845	0.134	0.675
4	0	0.890	0	0.110	1.065	0.778	0.869	0.725
	0	0.957	0	0.043	1.028	2.327	0.410	1.490
	0	1	0	0	1	2.630	0.189	1.143
5	0	0.5	0.333	0.167	1	0.581	1	0.581
	0	0.570	0.349	0.081	1.050	0.953	0.884	0.8962
	0	0.545	0.356	0.099	1.068	1.199	0.782	1.060
	0.314	0.352	0.257	0.077	0.988	2.383	0.773	2.095
7	0.279	0.357	0.279	0.084	1.022	7.937	0.088	2.356

$$S = S_{\text{all}}$$

Dominant model: $p = 0.1$, $\psi_0 = 0$, $\psi_1 = 1$, $\psi_2 = 1$

F = family number (as before)

Phenotypes (Y_1, \dots, Y_n) (1=affected, 0=unaffected, ?=unknown):

$F = 1$: (?, ?, 1, 1)

$F = 2$: (?, ?, 1, 1, 1)

$F = 3$: a) (?, ?, 1, 1, 1, 1), b) (?, ?, 0, 1, 1, 1), c) (?, ?, 0, 0, 1, 1)

$F = 4$: a) (?, ?, 1, 1, 1, 1, 1), b) (?, ?, 0, 1, 1, 1, 1), c) (?, ?, 0, 0, 1, 1, 1)

$F = 5$: a) (?, ?, ?, ?, ?, ?, 1, 1), b) (?, ?, ?, 1, 1, ?, 1, 1), c) (?, ?, 0, 1, 1, 0, 1, 1)

$F = 7$: see figure

Examples of mean functions 2

F	η_1/η	η_2/η	η_3/η	η_4/η	d	ξ	$(S, \tilde{S})^2$	η
1	0	1	0	0	1	1.337	0.749	1.157
2	0	1	0	0	1	2.223	0.582	1.696
	0	1	0	0	1	1.801	0.484	1.253
3	0	0.943	0	0.057	1.042	2.788	0.479	1.930
	0	1	0	0	1	3.244	0.378	1.994
	0	1	0	0	1	2.250	0.333	1.293
4	0	0.856	0	0.14463	1.098	2.769	0.442	1.841
	0	0.941	0	0.059	1.044	4.966	0.288	2.663
	0	1	0	0	1	4.145	0.257	2.103
5	0	0.5	0.333	0.167	1	1.199	1	1.199
	0	0.520	0.295	0.184	0.999	1.446	0.985	1.435
	0	0.482	0.321	0.196	1.018	1.625	0.990	1.617

$$S = S_{\text{all}}$$

Recessive model: $p = 0.1$, $\psi_0 = 0$, $\psi_1 = 0$, $\psi_2 = 1$

F = family number (as before)

Phenotypes (Y_1, \dots, Y_n) (1=affected, 0=unaffected, ?=unknown):

$F = 1$: (?, ?, 1, 1)

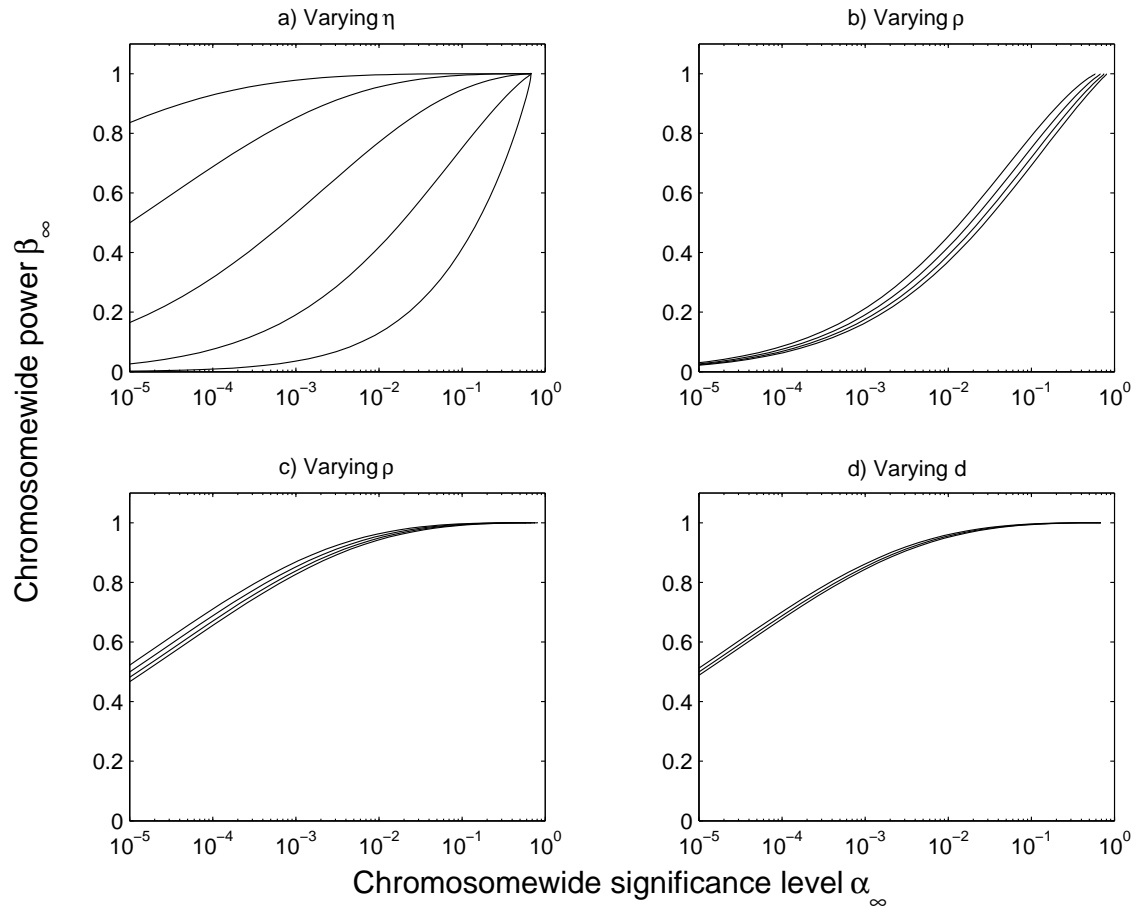
$F = 2$: a) (?, ?, 1, 1, 1), b) (?, ?, 0, 1, 1)

$F = 3$: a) (?, ?, 1, 1, 1, 1), b) (?, ?, 0, 1, 1, 1), c) (?, ?, 0, 0, 1, 1)

$F = 4$: a) (?, ?, 1, 1, 1, 1, 1), b) (?, ?, 0, 1, 1, 1, 1), c) (?, ?, 0, 0, 1, 1, 1)

$F = 5$: a) (?, ?, ?, ?, ?, ?, 1, 1), b) (?, ?, 0, 0, 0, 0, 1, 1), c) (0, 0, 0, 0, 0, 0, 1, 1)

ROC curves



Chromosome length $L=1.5$

- a) $\eta = 2, 3, 4, 5, 6, \rho = 2, d = 1$
- b) $\eta = 3, \rho = 1.5, 2, 2.5, 3, d = 1$
- c) $\eta = 5, \rho = 1.5, 2, 2.5, 3, d = 1$
- d) $\eta = 5, \rho = 2, d = 0.9, 1, 1.1$

Generalizations

- Different family types in the same data set. (Hössjer, 2005)
- Corrections for non-Gaussianity. (Ängquist and Hössjer, 2005)
- Corrections for incomplete marker data.
- Multilocus models (several disease genes on same chromosome).

References

- Aldous, D. (1989). *Probability approximations via the Poisson clumping heuristic*. Springer, New York.
- Ängquist, L. and Hössjer, O. (2005). Improving the calculation of statistical significance in genome-wide scans. To appear in *Biostatistics*.
- Diaconis, P. (1988). *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, California.
- Donnelly, P. (1983). The probability that some related individuals share some section of the genome identical by descent. *Theoret. Population Biol.*, **23**, 34-64.
- 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.
- 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. (2005). Spectral decomposition of score functions in linkage analysis. To appear in *Bernoulli*.
- Kruglyak, L. and Lander, E. (1998). Faster multipoint linkage analysis using Fourier transforms, *J. Comp. Biol.*, **5**(1), 1-7.
- Lander, E.L. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241-247.
- Leadbetter, R, Lindgren, G. and Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer, Berlin.
- Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *Ann. Statist.* **14**, 361-404.