Stationary Distribution of the Linkage Disequilibrium Coefficient r^2

Wei Zhang, Jing Liu, Rachel Fewster and Jesse Goodman

Department of Statistics, The University of Auckland

December 1, 2015

Overview

Introduction

1. Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

Introduction

1. Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

LD can be used to find genetic markers that might be linked to genes for diseases and understand the evolutionary history.

Introduction

1. Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

2. r^2 is a quantitative measure of LD and we are aiming to find its stationary distribution under models for genetic drift.

Introduction

1. Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

2. r^2 is a quantitative measure of LD and we are aiming to find its stationary distribution under models for genetic drift.

3. Given some **moments** of the unknown distribution, the maximum entropy (Maxent) principle can be used to approximate the density function of r^2.

Introduction

1. Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

2. r^2 is a quantitative measure of LD and we are aiming to find its stationary distribution under models for genetic drift.

3. Given some **moments** of the unknown distribution, the maximum entropy (Maxent) principle can be used to approximate the density function of r^2.

4. The diffusion approximation is a powerful tool to compute certain **expectations** at stationarity.

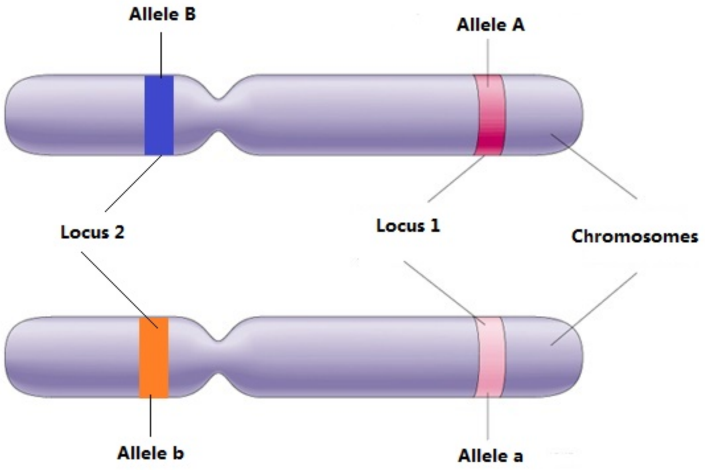
Background

- The TLD model (**Liu, 2012**)
- Diffusion approximation
- Maximum entropy principle

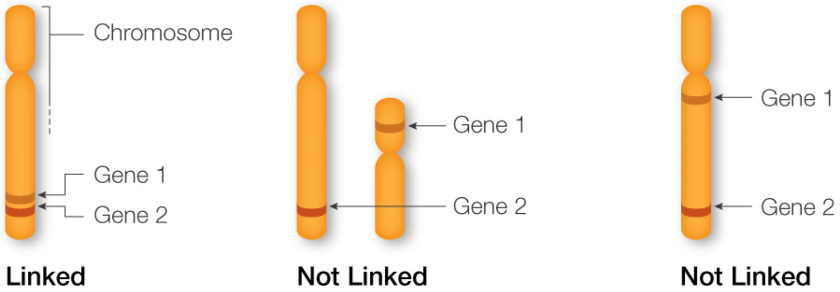
Some terminologies in genetics

1. One **locus** is a position of a gene or significant DNA sequence on a chromosome.
2. An **allele** is a variant form of a gene.
3. **Diploid** describes a cell or an organism that has paired chromosomes, one from each parent.
4. A **mutation** is a permanent change in the DNA sequence.
5. **Recombination** is the production of offspring with combinations of traits that differ from those found in either parent.

Some terminologies in genetics



Some terminologies in genetics



The TLD model

TLD is short for 'two-locus diallelic model with mutation and recombination'.

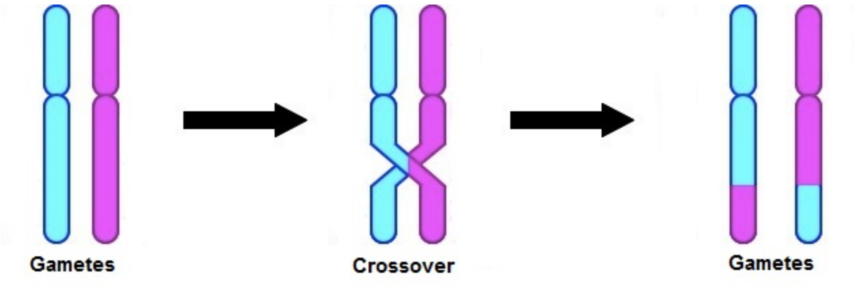
The TLD model

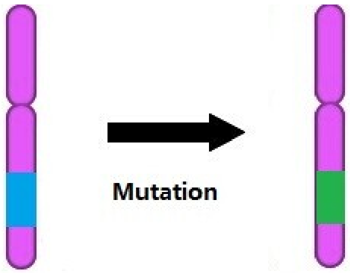
TLD is short for 'two-locus diallelic model with mutation and recombination'.

Model assumptions and notations:

- The two possible alleles on each of the two loci are assumed to be A_1, A_2 and B_1, B_2, thus the four possible types of gamete are A_1B_1, A_1B_2, A_2B_1 and A_2B_2.
- Recombination rate C: $A_iB_j + A_mB_n \Rightarrow A_iB_n/A_mB_j$.
- Equal mutation rate μ for both loci: $A_1 \rightleftharpoons A_2$ and $B_1 \rightleftharpoons B_2$.

The TLD model





The TLD model

Table 1: The proportions of gametes in generation T and the expected proportions in generation $T+1$ in the population. N is the population size.

Generation	Gamete			
	A_1B_1	A_1B_2	A_2B_1	A_2B_2
T	$\frac{x_1}{2N}$	$\frac{x_2}{2N}$	$\frac{x_3}{2N}$	$\frac{x_4}{2N}$
$T+1$	ϕ_1	ϕ_2	ϕ_3	ϕ_4

The TLD model

Suppose

$$D(\mathbf{x}) = \frac{x_1}{2N}\frac{x_4}{2N} - \frac{x_2}{2N}\frac{x_3}{2N}.$$

We have

$$\phi_1(\mathbf{x}) = \frac{x_1}{2N}(1-\mu)^2 + \left(\frac{x_2}{2N} + \frac{x_3}{2N}\right)\mu(1-\mu) + \frac{x_4}{2N}\mu^2 - CD(\mathbf{x})(1-2\mu)^2$$

$$\phi_2(\mathbf{x}) = \frac{x_2}{2N}(1-\mu)^2 + \left(\frac{x_1}{2N} + \frac{x_4}{2N}\right)\mu(1-\mu) + \frac{x_3}{2N}\mu^2 + CD(\mathbf{x})(1-2\mu)^2$$

$$\phi_3(\mathbf{x}) = \frac{x_3}{2N}(1-\mu)^2 + \left(\frac{x_1}{2N} + \frac{x_4}{2N}\right)\mu(1-\mu) + \frac{x_2}{2N}\mu^2 + CD(\mathbf{x})(1-2\mu)^2$$

$$\phi_4(\mathbf{x}) = \frac{x_4}{2N}(1-\mu)^2 + \left(\frac{x_2}{2N} + \frac{x_3}{2N}\right)\mu(1-\mu) + \frac{x_1}{2N}\mu^2 - CD(\mathbf{x})(1-2\mu)^2.$$

The TLD model

The transition probability of going from $\mathbf{x} = (x_1, x_2, x_3, x_4)$ to $\mathbf{y} = (y_1, y_2, y_3, y_4)$ is:

$$\begin{aligned} p_{\mathbf{xy}} &= \mathbb{P}\left(\mathbf{y}|\mathbf{x}\right) \\ &= \frac{(2N)!}{y_1!y_2!y_3!y_4!}\left(\phi_1\left(\mathbf{x}\right)\right)^{y_1}\left(\phi_2\left(\mathbf{x}\right)\right)^{y_2}\left(\phi_3\left(\mathbf{x}\right)\right)^{y_3}\left(\phi_4\left(\mathbf{x}\right)\right)^{y_4}. \end{aligned}$$

The TLD model

The transition probability of going from $\mathbf{x} = (x_1, x_2, x_3, x_4)$ to $\mathbf{y} = (y_1, y_2, y_3, y_4)$ is:

$$\begin{aligned} p_{\mathbf{xy}} &= \mathbb{P}(\mathbf{y}|\mathbf{x}) \\ &= \frac{(2N)!}{y_1!y_2!y_3!y_4!} (\phi_1(\mathbf{x}))^{y_1} (\phi_2(\mathbf{x}))^{y_2} (\phi_3(\mathbf{x}))^{y_3} (\phi_4(\mathbf{x}))^{y_4}. \end{aligned}$$

The TLD model is an irreducible aperiodic Markov chain, thus there exists a unique stationary distribution.

Diffusion approximation

The main idea is to rescale the discrete state space and time space by a factor related to the population size N, so that the gap between two successive states in the new space is infinitesimal when N is large enough.

Diffusion approximation

The main idea is to rescale the discrete state space and time space by a factor related to the population size N, so that the gap between two successive states in the new space is infinitesimal when N is large enough.

Example

$$\text{State space } \{0, 1, 2, \cdots, 2N\} \xrightarrow{(2N)^{-1}} \{0, \tfrac{1}{2N}, \tfrac{2}{2N}, \cdots, 1\}.$$

$$\text{Time space } \{0, 1, 2, \cdots\} \xrightarrow{(2N)^{-1}} \{0\delta t, 1\delta t, 2\delta t, \cdots\}, \text{ where } \delta t = \tfrac{1}{2N}.$$

Diffusion approximation

When N is large enough, the new chain is approximately continuous. In the derivation process, Taylor series expansion and the definition of derivative are used to get two important results for the TLD model.

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

The diffusion operator and master equation

The diffusion operator for the TLD model is

$$\mathcal{L} = \frac{1}{2}p\left(1-p\right)\frac{\partial^2}{\partial p^2} + \frac{1}{2}q\left(1-q\right)\frac{\partial^2}{\partial q^2} + \frac{1}{2}\left\{p\left(1-p\right)q\left(1-q\right) + D\left(1-2p\right)\left(1-2q\right) - D^2\right\}\frac{\partial^2}{\partial D^2}$$
$$+ D\frac{\partial^2}{\partial p \partial q} + D\left(1-2p\right)\frac{\partial^2}{\partial p \partial D} + D\left(1-2q\right)\frac{\partial^2}{\partial q \partial D} + \frac{\theta}{4}\left(1-2p\right)\frac{\partial}{\partial p} + \frac{\theta}{4}\left(1-2q\right)\frac{\partial}{\partial q}$$
$$- D\left(1 + \frac{\rho}{2} + \theta\right)\frac{\partial}{\partial D}$$

and the master equation at stationarity is $\mathbb{E}\left\{\mathcal{L}f\left(p,q,D\right)\right\} = 0$.

The master equation means the expected evolution over time of any nice function of p, q and D is zero at stationarity.

The diffusion operator and master equation

The diffusion operator for the TLD model is

$$\mathcal{L} =\frac{1}{2}p\left(1-p\right)\frac{\partial^2}{\partial p^2}+\frac{1}{2}q\left(1-q\right)\frac{\partial^2}{\partial q^2}+\frac{1}{2}\left\{p\left(1-p\right)q\left(1-q\right)+D\left(1-2p\right)\left(1-2q\right)-D^2\right\}\frac{\partial^2}{\partial D^2}$$
$$+D\frac{\partial^2}{\partial p\partial q}+D\left(1-2p\right)\frac{\partial^2}{\partial p\partial D}+D\left(1-2q\right)\frac{\partial^2}{\partial q\partial D}+\frac{\theta}{4}\left(1-2p\right)\frac{\partial}{\partial p}+\frac{\theta}{4}\left(1-2q\right)\frac{\partial}{\partial q}$$
$$-D\left(1+\frac{\rho}{2}+\theta\right)\frac{\partial}{\partial D}$$

and the master equation at stationarity is $\mathbb{E}\left\{\mathcal{L}f\left(p,q,D\right)\right\}=0$.

Here p and q are the frequencies of A_1 and B_1, $D=f_{11}-pq$, f_{11} is the frequency of gamete A_1B_1, $\rho=2NC$, $\theta=2N\mu$ and f is any twice continuously differentiable function with compact support.

A simple example

If letting $f(p,q,D) = D$, we can get that

$$\mathcal{L}f(p,q,D) = -D\left(1 + \frac{\rho}{2} + \theta\right)$$

and

$$\mathbb{E}\left\{\mathcal{L}f(p,q,D)\right\} = -\left(1 + \frac{\rho}{2} + \theta\right)\mathbb{E}(D) = 0$$

so

$$\mathbb{E}(D) = 0.$$

Maximum entropy principle

Definition

Entropy is the quantitative measure of disorder in a system.

Suppose a random variable Z has K possible outcomes with probabilities $p_1, p_2, p_3, \cdots, p_K$, the entropy is:

$$I(Z) = -\sum_{i=1}^{K} p_i \log_K (p_i).$$

Maximum entropy principle

Definition

Entropy is the quantitative measure of disorder in a system.

Suppose a random variable Z has K possible outcomes with probabilities $p_1, p_2, p_3, \cdots, p_K$, the entropy is:

$$I(Z) = -\sum_{i=1}^{K} p_i \log_K (p_i).$$

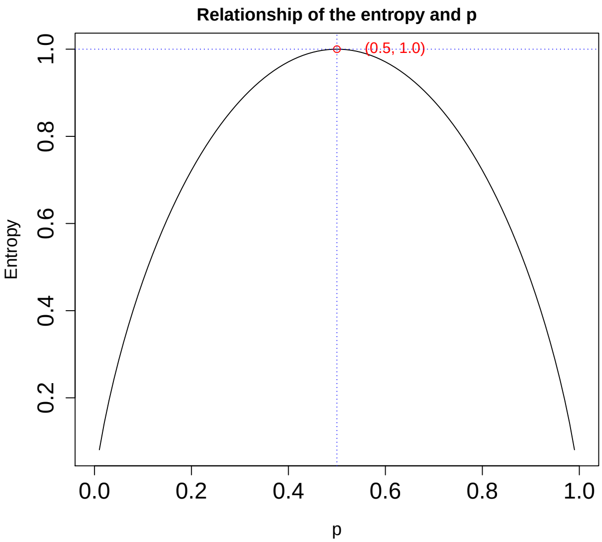
The maximum entropy principle states that the solution that maximises the entropy is the most honest one.

An example

In a coin toss experiment, suppose Pr(head)$=p$ and Pr(tail)$=1-p$, then the entropy is:

$$I = -p\log_2(p) - (1-p)\log_2(1-p).$$

An example


Relationship of the entropy and p
(0.5, 1.0)
Entropy
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
p

Maximum entropy principle

The Maxent solution of an unknown probability density function $\pi(p)$ given knowledge of n moments $m_i = \mathbb{E}\left(p^i\right), i = 1, \cdots, n$ is the solution $\widetilde{\pi}_n(p)$ that maximizes:

$$I = -\int_\Omega \widetilde{\pi}_n(p) \log\left\{\widetilde{\pi}_n(p)\right\} dp$$

subject to

$$m_i = \int_\Omega p^i \, \widetilde{\pi}_n(p) \, dp \quad \text{for} \quad i = 0, 1, \cdots, n$$

where Ω is the support of π.

Considering the Lagrange function and Euler-Lagrange equation, then the Maxent solution is:

$$\widetilde{\pi}_n(p) = \exp\left(\lambda_0 + \lambda_1 p + \lambda_2 p^2 + \cdots + \lambda_n p^n\right)$$

where $\lambda_i,\ i = 0, \cdots, n$ are the solutions of

$$\arg\min_{\lambda} \left\{ \int_{\Omega} \exp\left(\lambda_0 + \lambda_1 p + \lambda_2 p^2 + \cdots + \lambda_n p^n\right) dp - \sum_{i=0}^{n} \lambda_i m_i \right\}.$$

Linkage disequilibrium coefficient r^2

The definition of r^2 is:

$$r^2 = \frac{D^2}{p\,(1-p)\,q\,(1-q)}$$

where p and q are the frequencies of A_1 and B_1, $D = f_{11} - pq$ and f_{11} is the frequency of gamete $A_1 B_1$.

Reformulation of the problem

Let $u = 1 - 2p$ and $v = 1 - 2q$. Then the diffusion generator can be rewritten as

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\left(1-u^2\right)\frac{\partial^2}{\partial u^2} + \frac{1}{2}\left(1-v^2\right)\frac{\partial^2}{\partial v^2} + \frac{1}{2}\left\{\frac{1}{16}\left(1-u^2\right)\left(1-v^2\right) + Duv - D^2\right\}\frac{\partial^2}{\partial D^2} \\ &+ 4D\frac{\partial^2}{\partial u \partial v} - 2Du\frac{\partial^2}{\partial D \partial u} - 2Dv\frac{\partial^2}{\partial D \partial v} - \frac{1}{2}\theta u \frac{\partial}{\partial u} - \frac{1}{2}\theta v \frac{\partial}{\partial v} - D\left(1 + \frac{1}{2}\rho + \theta\right)\frac{\partial}{\partial D}.\end{aligned}$$

This reparameterization yields

$$r^2 = \frac{D^2}{p\left(1-p\right)q\left(1-q\right)} = \frac{16D^2}{\left(1-u^2\right)\left(1-v^2\right)}.$$

Note that when $0 \leq u^2, v^2 < 1$:

$$\frac{1}{1-u^2} = \sum_{k=0}^{\infty} u^{2k} \quad \text{and} \quad \frac{1}{1-v^2} = \sum_{l=0}^{\infty} v^{2l}.$$

Considering the new form of r^2, it follows that when $M = 1, 2 \cdots$

$$\mathbb{E}\left(r^{2M}\right) = 16^M \sum_{k_1=0}^{\infty} \cdots \sum_{k_M=0}^{\infty} \sum_{l_1=0}^{\infty} \cdots \sum_{l_M=0}^{\infty} \mathbb{E}\left\{D^{2M} u^{2(k_1+k_2+\cdots+k_M)} v^{2(l_1+l_2+\cdots+l_M)}\right\},$$

which can be simplified to

$$\mathbb{E}\left(r^{2M}\right) = 16^M \sum_{K=0}^{\infty} \sum_{L=0}^{\infty} \binom{K+M-1}{M-1} \binom{L+M-1}{M-1} \mathbb{E}\left(D^{2M} u^{2K} v^{2L}\right).$$

Analytic computation of the moments

Our problem now is how to compute $\mathbb{E}\left(D^{2M}u^{2K}v^{2L}\right)$ for all possible M, K and L.

Step 1:
Let f in the master equation be some specific forms of u^n, uv, u^2v and Du^n, we can get the results of $\mathbb{E}\left(u^n\right), \mathbb{E}\left(uv\right), \mathbb{E}\left(u^2v\right)$ etc.

Step 2:
Given a value of (m,n), apply the function $f = D^k u^{m+2-k} v^{n+2-k}$ into the master equation with $k \in \{0,1,2,\cdots,n+2\}$.

A system of $n+3$ linear equations is generated, whose solutions are:

$$\mathbb{E}\left(u^{m+2}v^{n+2}\right),\ \mathbb{E}\left(Du^{m+1}v^{n+1}\right),\ \mathbb{E}\left(D^2u^mv^n\right)\cdots$$

Analytic computation of the moments

Given a $M \in \mathbb{N}$ and a truncation level $\ell_{\text{max}} \in \mathbb{N}$, we use

$$\mathbb{E}\left(r^{2M}\right)_{\ell_{\text{max}}} = 16^M \left\{ \sum_{K,L \geq 0}^{2K+2L=\ell_{\text{max}}} \binom{K+M-1}{M-1} \binom{L+M-1}{M-1} \mathbb{E}\left(D^{2M} u^{2K} v^{2L}\right) \right\}$$

to approximate the M-th moment of r^2.

Variance of r^2

Table 2: $\mathbb{V}\left(r^2\right)$ computed by our method ($\ell_{\text{max}} = 700$)

ρ	θ						
	0.0125	0.0500	0.1000	0.2500	0.7500	1.2500	$\cdots$
0.00	0.006119	0.03109	0.03806	0.02433	0.00685	0.003239	$\cdots$
0.25	0.003412	0.01950	0.02614	0.01891	0.00608	0.002928	$\cdots$
0.50	0.002271	0.01372	0.01932	0.01517	0.00538	0.002739	$\cdots$
1.25	0.001062	0.00666	0.00991	0.00892	0.00403	0.002197	$\cdots$
2.50	0.000526	0.00327	0.00487	0.00476	0.00263	0.001590	$\cdots$
5.00	0.000244	0.00142	0.00206	0.00202	0.00141	0.000961	$\cdots$

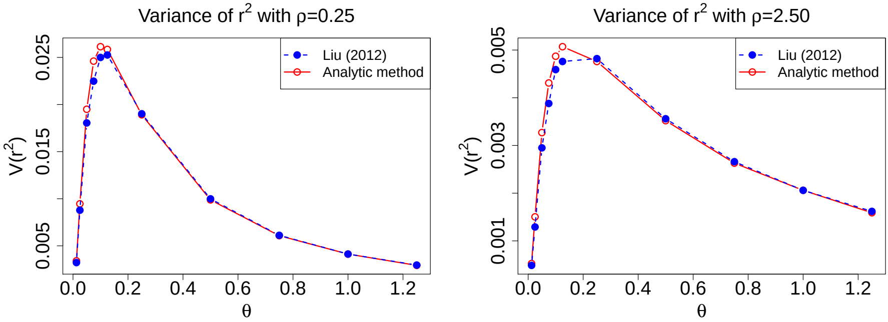
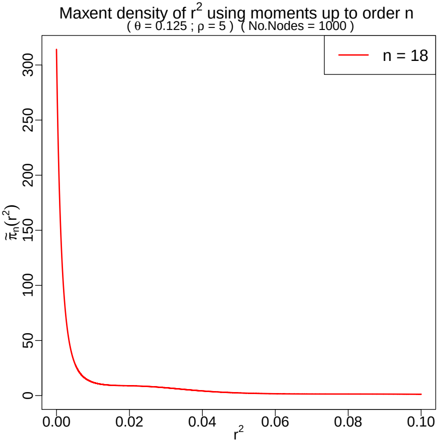


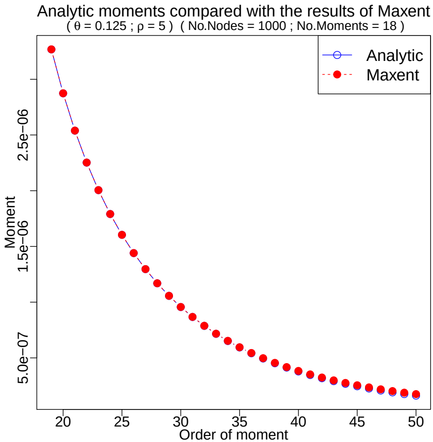
Figure 1: Comparison of $\mathbb{V}\left(r^2\right)$ between our analytic method and the method in **Liu(2012)**

Probability density function of r^2

We can compute 50 moments in 2.5 hours with $\ell_{\max} = 2000$ on a laptop.

- Use $n = 18$ moments to calculate Maxent $\tilde{\pi}\left(r^2\right)$.
- Compare moments $19, 20, \ldots, 50$ using $\tilde{\pi}\left(r^2\right)$ vs analytic method.





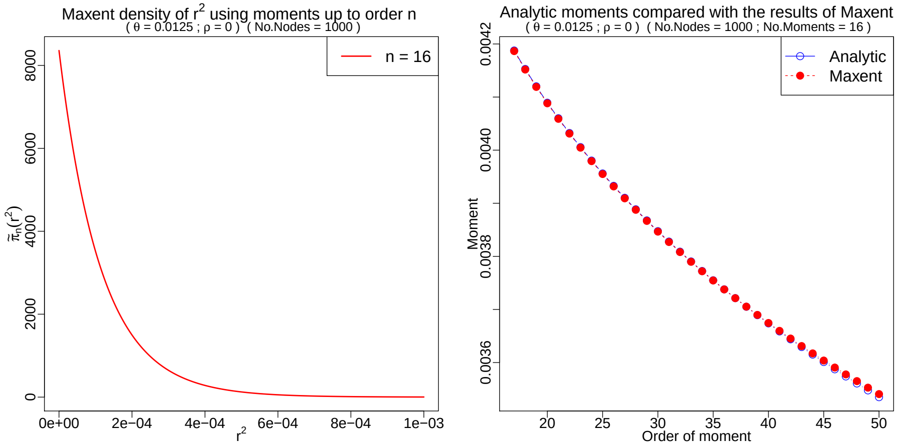


Figure 2: Stationary density functions of r^2 for two pairs of θ and ρ approximated by the numerical univariate Maxent method.

Acknowledgements

I would like to sincerely and gratefully thank my supervisor Rachel Fewster for her guidance, Dr. Jesse Goodman and Dr. Jing Liu for their useful ideas.

References

Y. S. Song and J. S. Song, "Analytic computation of the expectation of the linkage disequilibrium coefficient r^2," *Theoretical population biology*, vol. 71, no. 1, pp. 49–60, 2007.

J. Liu, *Reconstruction of probability distributions in population genetics*. PhD thesis, The University of Auckland, 2012.

X. Wu, "Calculation of maximum entropy densities with application to income distribution," *Journal of Econometrics*, vol. 115, no. 2, pp. 347–354, 2003.

M. Slatkin, "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008.

W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *Theoretical and Applied Genetics*, vol. 38, no. 6, pp. 226–231, 1968.

L. R. Mead and N. Papanicolaou, "Maximum entropy in the problem of moments," *Journal of Mathematical Physics*, vol. 25, no. 8, pp. 2404–2417, 1984.

Thanks!