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

$r^2$ is a quantitative measure of LD and we are aiming to find its stationary distribution under models for genetic drift.
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 $r^2$. 
Linkage disequilibrium (LD) indicates the statistical dependence of alleles at different loci in population genetics.

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

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

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

- Allele B
- Locus 2
- Allele a
- Locus 1
- Allele A
- Chromosomes
- Allele b
Some terminologies in genetics

- **Linked**
  - Chromosome
  - Gene 1
  - Gene 2

- **Not Linked**
  - Gene 1
  - Gene 2

- **Not Linked**
  - Gene 1
  - Gene 2
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 $\mu$ for both loci: $A_1 \leftrightharpoons A_2$ and $B_1 \leftrightharpoons B_2$. 
The TLD model

Gametes → Crossover → Gametes

Gametes → Mutation

Background  The TLD model  December 1, 2015
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.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Gamete</th>
<th>$A_1B_1$</th>
<th>$A_1B_2$</th>
<th>$A_2B_1$</th>
<th>$A_2B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td></td>
<td>$\frac{x_1}{2N}$</td>
<td>$\frac{x_2}{2N}$</td>
<td>$\frac{x_3}{2N}$</td>
<td>$\frac{x_4}{2N}$</td>
</tr>
<tr>
<td>$T + 1$</td>
<td>$\phi_1$</td>
<td>$\phi_2$</td>
<td>$\phi_3$</td>
<td>$\phi_4$</td>
<td></td>
</tr>
</tbody>
</table>
The TLD model

Suppose

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

We have

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

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

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

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

$$p_{xy} = \mathbb{P}(\mathbf{y} | \mathbf{x})$$

$$= \frac{(2N)!}{y_1! y_2! y_3! y_4!} (\phi_1 (\mathbf{x}))^{y_1} (\phi_2 (\mathbf{x}))^{y_2} (\phi_3 (\mathbf{x}))^{y_3} (\phi_4 (\mathbf{x}))^{y_4}.$$
The transition probability of going from $x = (x_1, x_2, x_3, x_4)$ to $y = (y_1, y_2, y_3, y_4)$ is:

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

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

State space $\{0, 1, 2, \cdots, 2N\}$ \(\frac{(2N)^{-1}}{(2N)^{-1}}\) $\rightarrow$ $\{0, \frac{1}{2N}, \frac{2}{2N}, \cdots, 1\}$.

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

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$
The diffusion operator for the TLD model is

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

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

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 for the TLD model is

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

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

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

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

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 $p_1, p_2, p_3, \ldots, p_K$, 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 $\Pr(\text{head}) = p$ and $\Pr(\text{tail}) = 1 - p$, then the entropy is:

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

Relationship of the entropy and \( p \)

![Graph showing the relationship between entropy and \( p \) with a peak at \((0.5, 1.0)\).]
The Maxent solution of an unknown probability density function $\pi(p)$ given knowledge of $n$ moments $m_i = \mathbb{E}(p^i), i = 1, \cdots, n$ is the solution $\tilde{\pi}_n(p)$ that maximizes:

$$I = -\int_{\Omega} \tilde{\pi}_n(p) \log \{\tilde{\pi}_n(p)\} \, dp$$

subject to

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

where $\Omega$ is the support of $\pi$. 
Maximum entropy principle

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

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

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

\[ \arg\min_{\lambda} \left\{ \int_{\Omega} \exp\left(\lambda_0 + \lambda_1 p + \lambda_2 p^2 + \cdots + \lambda_n p^n\right) dp - \sum_{i=0}^{n} \lambda_i m_i \right\}. \]
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 $A_1B_1$. 
Let $u = 1 - 2p$ and $v = 1 - 2q$. Then the diffusion generator can be rewritten as

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

This reparameterization yields

\[
r^2 = \frac{D^2}{p (1 - p) q (1 - q)} = \frac{16D^2}{(1 - u^2) (1 - v^2)}.\]
Analytic computation of the moments

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

$$\mathbb{E} \left( r^{2M} \right) = 16^M \sum_{K=0}^{\infty} \sum_{L=0}^{\infty} \left( \begin{array}{c} K + M - 1 \\ M - 1 \end{array} \right) \left( \begin{array}{c} L + M - 1 \\ M - 1 \end{array} \right) \mathbb{E} \left( D^{2M} u^{2K} v^{2L} \right).$$
Analytic computation of the moments

Our problem now is how to compute \( \mathbb{E}(D^{2M}u^{2K}v^{2L}) \) 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 \( \mathbb{E}(u^n), \mathbb{E}(uv), \mathbb{E}(u^2v) \) etc.

**Step 2:**
Given a value of \((m, n)\), apply the function \( f = D^ku^{m+2-k}v^{n+2-k} \) 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:

\[
\mathbb{E}(u^{m+2}v^{n+2}), \; \mathbb{E}(Du^{m+1}v^{n+1}), \; \mathbb{E}(D^2u^mv^n) \cdots
\]
Given a $M \in \mathbb{N}$ and a truncation level $\ell_{\text{max}} \in \mathbb{N}$, we use

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

to approximate the $M$-th moment of $r^2$. 
Table 2: $\nabla (r^2)$ computed by our method ($\ell_{\text{max}} = 700$)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>0.0125</th>
<th>0.0500</th>
<th>0.1000</th>
<th>0.2500</th>
<th>0.7500</th>
<th>1.2500</th>
<th>$\cdots$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.006119</td>
<td>0.03109</td>
<td>0.03806</td>
<td>0.02433</td>
<td>0.00685</td>
<td>0.003239</td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.003412</td>
<td>0.01950</td>
<td>0.02614</td>
<td>0.01891</td>
<td>0.00608</td>
<td>0.002928</td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.002271</td>
<td>0.01372</td>
<td>0.01932</td>
<td>0.01517</td>
<td>0.00538</td>
<td>0.002739</td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>0.001062</td>
<td>0.00666</td>
<td>0.00991</td>
<td>0.00892</td>
<td>0.00403</td>
<td>0.002197</td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>2.50</td>
<td>0.000526</td>
<td>0.00327</td>
<td>0.00487</td>
<td>0.00476</td>
<td>0.00263</td>
<td>0.001590</td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>0.000244</td>
<td>0.00142</td>
<td>0.00206</td>
<td>0.00202</td>
<td>0.00141</td>
<td>0.000961</td>
<td>$\cdots$</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Comparison of $\nabla (r^2)$ between our analytic method and the method in Liu(2012).
We can compute 50 moments in 2.5 hours with $\ell_{\text{max}} = 2000$ on a laptop.

- Use $n = 18$ moments to calculate Maxent $\tilde{\pi}(r^2)$.
- Compare moments $19, 20, \ldots, 50$ using $\tilde{\pi}(r^2)$ vs analytic method.
Figure 2: Stationary density functions of $r^2$ for two pairs of $\theta$ and $\rho$ approximated by the numerical univariate Maxent method.
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.


Thanks!