Stationary distribution of the linkage disequilibrium coefficient r^2

Wei Zhang[a,1,*], Jing Liu[a,b], Jesse Goodman[a], Bruce S. Weir[c],
Rachel M. Fewster[a]

[a] *Department of Statistics, University of Auckland,*
Private Bag 92019, Auckland, New Zealand
[b] *University of Michigan - Shanghai Jiao Tong University Joint Institute,*
Shanghai 200240, China
[c] *Department of Biostatistics, University of Washington,*
Box 357232, Seattle WA 98195-7232, USA

Abstract

The linkage disequilibrium coefficient r^2 is a measure of statistical dependence of the alleles possessed by an individual at different genetic loci. It is widely used in association studies to search for the locations of disease-causing genes on chromosomes. Most studies to date treat r^2 as a fixed property of two loci in a finite population, and investigate the sampling distribution of estimators due to the statistical sampling of individuals from the population. Here, we instead consider the distribution of r^2 itself under a process of genetic sampling through the generations. Using a classical two-locus model for genetic drift, mutation, and recombination, we investigate the probability density function of r^2 at stationarity. This density function provides a tool for inference on evolutionary parameters such as mutation and recombination rates. We reconstruct the approximate stationary density of r^2 by calculating a finite sequence of the distribution's moments and applying the maximum entropy principle. Our approach is based on the diffusion approximation, under which we demonstrate that for certain models in population genetics, moments of the stationary distribution can be obtained without knowing the probability distribution itself.

[*] Corresponding author
Email address: `wzhan592@uwo.ca` (Wei Zhang)
[1] Current address: Department of Statistical and Actuarial Sciences, University of Western Ontario, London ON N6A 5B7, Canada

To illustrate our approach, we show how the stationary probability density of r^2 can be used in a maximum likelihood framework to estimate mutation and recombination rates from sample data of r^2.

Keywords: linkage disequilibrium, r^2, stationary distribution, maximum entropy principle, diffusion approximation, maximum likelihood

1. Introduction

Linkage disequilibrium (LD) refers to the statistical dependence between alleles found at two genetic loci, taken across individuals in a population. It has various applications, for example, understanding the evolutionary history of a species, gene mapping in association studies, and detecting recombination hotspots (e.g. Pritchard & Przeworski, 2001; Slatkin, 2008). Many statistics for measuring LD can be found in the literature (Pritchard & Przeworski, 2001). A common and convenient measure of LD is the squared correlation coefficient r^2, defined below, which has been introduced and studied in many papers (e.g. Hill & Robertson, 1968; Mueller, 2004; Gupta et al., 2005).

Studies of the distribution of r^2 fall broadly into two categories. In the first case, r^2 is treated as a fixed property of two loci in a finite population of interest, and the aim is to investigate the distribution of the sample-based squared correlation coefficient from a sample of n individuals, which serves as an estimator of this quantity. These studies therefore investigate the distribution of estimators of r^2 due to statistical sampling within a population (e.g. Golding, 1984; Hudson, 1985; Hill & Weir, 1994; Abecasis et al., 2001; Service et al., 2006). In the second case, r^2 is treated as a random variable that evolves over generations due to genetic processes such as drift, mutation, recombination, and selection. Thus, variability arises from genetic sampling rather than from statistical sampling. The aim is to study the equilibrium distribution of this random variable, so that measurements of extant r^2 may be used to draw inference on historical genetic processes. In this paper we focus exclusively on this second case, building upon foundations established by Ohta & Kimura (1969a,b); Weir

& Hill (1980); Hill & Weir (1988); and Song & Song (2007).

To investigate the equilibrium distribution of r^2, we use the two-locus diallelic (TLD) model treated by previous authors. This model incorporates mutation and recombination, although not selection, and is a generalization of the classical Wright-Fisher model (Wright, 1931). Consider a population of N individuals, where N stays constant throughout all generations, and generations do not overlap. Further consider two diallelic loci with allele types A_1 or A_2 present at the first locus, and B_1 or B_2 at the second locus. The population of interest is diploid so there are four possible types of gamete: A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Counts of these four gamete types sum to $2N$ in any generation. The squared correlation coefficient r^2 in a particular generation is defined as

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

where p and q are the marginal frequencies of alleles A_1 and B_1 in that generation, $D = f_{11} - pq$, and f_{11} denotes the frequency of gamete A_1B_1. Following Song & Song (2007), we use $\theta = 8N\mu$ and $\rho = 4Nc$ to denote the scaled mutation and recombination rates, where μ is the equal mutation rate per gene per generation for both loci:

$$A_1 \underset{\mu}{\overset{\mu}{\leftrightharpoons}} A_2 \qquad B_1 \underset{\mu}{\overset{\mu}{\leftrightharpoons}} B_2\ , \tag{2}$$

and c is the recombination rate per gamete per generation between the two loci. The TLD model is an irreducible aperiodic discrete-time Markov chain over a finite discrete state space, which comprises the possible combinations of the four gamete counts summing to $2N$. It therefore converges to a unique stationary distribution. Our aim is to explore the distribution of r^2 with respect to this stationary distribution.

Exact characterization of the stationary distribution of r^2 under the TLD model is thought to be intractable, but several authors have explored properties of the distribution under various approximations. Most research has focused on the mean, $\mathrm{E}(r^2)$. In general, it is difficult to compute the expectation of a ratio, so initial work involved approximating this by the ratio of expectations:

$\mathrm{E}(r^2) \simeq \mathrm{E}(D^2)/\mathrm{E}\{p(1-p)q(1-q)\}$ (e.g. Ohta & Kimura, 1969a,b; Weir & Hill, 1980). More recently, Song & Song (2007) proposed an analytic method of computing $\mathrm{E}(r^2)$ under the diffusion approximation: we describe and build upon their method below. Further properties of the distribution of r^2 have been largely unexplored. Hill & Weir (1988) computed the variance of D^2, but commented that using it together with a Taylor series to compute the variance of r^2 would likely be inaccurate. Liu (2012) generated an approximate computation of $\mathrm{Var}(r^2)$, but his method is slow and requires considerable computing power. As far as we know, there is no existing tractable method for computing either $\mathrm{Var}(r^2)$ or the stationary probability density function, $\pi(r^2)$. Availability of a computable density function would establish an inferential framework in which researchers could perform maximum likelihood estimation or derive critical values for hypothesis tests. Our purpose in this paper is to derive computable approximations to the variance and higher moments of r^2, and thence to construct an approximate probability density function that can readily be used for inference.

Our computations build on the ideas of Song & Song (2007) (hereafter S2007). We first show how S2007's method for finding $\mathrm{E}(r^2)$ may be extended to compute a sequence of higher-order moments of r^2 at stationarity, from which the variance of r^2 can be obtained. As in S2007, our approach to computing stationary moments of r^2 depends on the technique of diffusion approximation, which requires the constant population size N to be sufficiently large. We then use the maximum entropy principle to approximate the stationary probability density of r^2 from the moments obtained. We show how the resulting approximate density of r^2 can be used for maximum likelihood estimation of mutation rate and recombination rate under the TLD model, using sample observations of r^2. We demonstrate the performance of our novel approach by simulation studies.

2. Materials and Methods

2.1. Diffusion approximation

Ohta & Kimura (1969a,b) showed that the computation of certain expectations at stationarity is greatly facilitated by the diffusion approximation for models in population genetics. S2007 extended the method of Ohta & Kimura (1969b) to compute the expectation of r^2 at stationarity under the TLD model. The diffusion approximation is a general technique for approximating a discrete process by a diffusion process that is continuous in both time and state. Associated with each diffusion process is a diffusion generator that consists of a set of partial differential operators, and is key to computing certain expectations.

Let $\boldsymbol{X}_t = (p, q, D)_t$ denote the diffusion process corresponding to the TLD model. In the discrete model, the three variables p, q, and D uniquely determine the state of the Markov chain in terms of the gamete counts at generation t. In the diffusion process, p, q, and D are treated as continuous variables which evolve continuously over time. The diffusion generator used by S2007 differs from that used by Ohta & Kimura (1969b) by a factor of 2, which does not affect the stationary distribution. Here we use the diffusion generator $\mathcal{L}$ of S2007:

$$\begin{aligned}
\mathcal{L} \;=\; & \frac{1}{2} p\,(1-p)\,\frac{\partial^2}{\partial p^2} + \frac{1}{2}\left\{p\,(1-p)\,q\,(1-q) + D\,(1-2p)\,(1-2q) - D^2\right\}\frac{\partial^2}{\partial D^2} \\
& + \frac{1}{2} q\,(1-q)\,\frac{\partial^2}{\partial q^2} + D\frac{\partial^2}{\partial p \partial q} + D\,(1-2p)\,\frac{\partial^2}{\partial p \partial D} + D\,(1-2q)\,\frac{\partial^2}{\partial q \partial D} \\
& + \frac{\theta}{4}\,(1-2p)\,\frac{\partial}{\partial p} + \frac{\theta}{4}\,(1-2q)\,\frac{\partial}{\partial q} - \left(1 + \frac{\rho}{2} + \theta\right) D\frac{\partial}{\partial D}\,.
\end{aligned} \tag{3}$$

The diffusion generator $\mathcal{L}$ for any diffusion process is defined such that

$$\frac{\partial}{\partial t}\mathrm{E}\left\{f\left(\boldsymbol{X}_t\right)\right\} = \mathrm{E}\left\{\mathcal{L} f\left(\boldsymbol{X}_t\right)\right\} \tag{4}$$

for any twice continuously differentiable function f with compact support. At stationarity, $\mathrm{E}\left\{f\left(\boldsymbol{X}_t\right)\right\}$ does not depend on the time parameter by definition, so its rate of change on the left-hand side of (4) is zero. Thus, for any suitable function f, and for $\boldsymbol{X}_t = (p, q, D)_t$ distributed over the continuous state space $[0,1]^2 \times [-1,1]$ according to the stationary distribution $\boldsymbol{\pi}$, we have

$$\mathrm{E}\left\{\mathcal{L}f\left(\boldsymbol{X}_t\right)\right\} = 0. \tag{5}$$

Equation (5) is called the master equation, and it provides considerable power for computing expectations of certain quantities at stationarity. The method described by (4) and (5) is general, but the specific form of $\mathcal{L}$ in (3) establishes stationary expectations under the TLD model. Equation (5) may be applied to any eligible function f. For example, given $f\left(p,q,D\right) = D$, we have $\mathcal{L}f\left(p,q,D\right) = -D\left(1+\rho/2+\theta\right)$ from (3), and therefore the master equation (5) gives $\mathrm{E}\left(D\right) = 0$ under the TLD model at stationarity.

2.2. The method of Song and Song (2007) for computing $E\left(r^2\right)$

The primary strategy developed by S2007 is to write r^2 as an infinite sum of monomials in terms of (p, q, D), for which stationary expectations can be computed using the master equation (5). We shall extend this method to compute higher-order stationary moments of r^2 in the following section. We first describe S2007's method in detail, as it is vital for the development that follows.

Because the infinite series $\sum_{k=0}^{\infty} y^k$ converges to $1/(1-y)$ for $0 \le y < 1$, we have $1/\left(1-p\right) = \sum_{k=0}^{\infty} p^k$ and $1/p = \sum_{k=0}^{\infty}\left(1-p\right)^k$ when $0 < p < 1$ for the TLD model. Similar results are obtained for $1/q$ and $1/\left(1-q\right)$. It follows that

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

$$= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\Big\{D^2p^mq^n + D^2(1-p)^mq^n + D^2p^m(1-q)^n + D^2(1-p)^m(1-q)^n\Big\}. \tag{6}$$

Under the TLD model, all alleles A_1, A_2, B_1, B_2 undergo identical genetic processes of mutation, recombination, and drift, so by symmetry the stationary expectations of $D^2p^mq^n$, $D^2\left(1-p\right)^m q^n$, $D^2p^m\left(1-q\right)^n$, and $D^2\left(1-p\right)^m\left(1-q\right)^n$ are all equal. Thus we have

$$\mathrm{E}\left(r^2\right) = 4\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathrm{E}\left(D^2p^mq^n\right). \tag{7}$$

The problem of computing $\mathrm{E}\left(r^2\right)$ therefore reduces to computing $\mathrm{E}\left(D^2p^mq^n\right)$ for all pairs of non-negative integers m and n.

Note that $\mathrm{E}\left(D^2p^mq^n\right) = \mathrm{E}\left(D^2p^nq^m\right)$ for the TLD model due to the symmetry of the two loci, so we only focus on expectations with $m \geq n$. For each pair of (m, n) where $m \geq n$, inserting $f(p, q, D) = D^kp^{m+2-k}q^{n+2-k}$ into the master equation (5) with $k = 0, \ldots, n+2$ generates a system of $n+3$ linear equations. This system includes more than $n+3$ unknown expectations, so it does not have a unique solution unless the computations are carried out in a particular order. S2007's innovative idea was to specify an order of computation such that the number of unknown expectations is reduced to $n+3$ at every iterative step. The computations are executed along an increasing level of $\ell = m+n$, while within each level the algorithm follows an increasing order of n. The computation order is as follows:

$$\begin{aligned}(m,n):\overbrace{(0,0)}^{\ell=0}\to\overbrace{(1,0)}^{\ell=1}\to\overbrace{(2,0)\to(1,1)}^{\ell=2}\to\overbrace{(3,0)\to(2,1)}^{\ell=3}&\\ \to\overbrace{(4,0)\to(3,1)\to(2,2)}^{\ell=4}\cdots.&\end{aligned} \tag{8}$$

When this order is followed, S2007 showed that each system of linear equations then includes exactly $n+3$ unknown expectations:

$$\mathrm{E}\left(p^{m+2}q^{n+2}\right),\ \mathrm{E}\left(Dp^{m+1}q^{n+1}\right),\ \mathrm{E}\left(D^2p^mq^n\right)\ ,\ \ldots,\ \mathrm{E}\left(D^{2+n}p^{m-n}\right), \tag{9}$$

one of which is the expectation $\mathrm{E}\left(D^2p^mq^n\right)$ needed for the computation of $\mathrm{E}\left(r^2\right)$ in (7). As byproducts, other expectations are obtained which are not needed for computing $\mathrm{E}\left(r^2\right)$, but we shall show in the following section that these can be used to compute higher-order moments of r^2.

In practice, we truncate the infinite sum (7) at some finite value to compute $\mathrm{E}\left(r^2\right)$. Given a value of ℓ, the truncated summation

$$\mathrm{E}\left(r^2\right)_\ell = 4\sum_{m,n\geq 0}^{m+n=\ell}\mathrm{E}\left(D^2p^mq^n\right) \tag{10}$$

serves as an approximation to $\mathrm{E}\left(r^2\right)$. As ℓ varies from 0 to ∞, we obtain a monotonically increasing sequence of partial sums $\left\{\mathrm{E}\left(r^2\right)_\ell\right\}_{\ell=0}^{\infty}$. Since $\mathrm{E}\left(r^2\right)$ is bounded, the convergence of this sequence is guaranteed. Practical results show that convergence is achieved fairly rapidly as ℓ increases, although the rate of convergence tends to be faster for smaller ρ and larger θ. S2007 demonstrated that for all settings of ρ and θ they explored, the truncation $\mathrm{E}\left(r^2\right)_{\ell_{\max}}$ served as a good approximation to $\mathrm{E}\left(r^2\right)$ when the truncation level was $\ell_{\max} = 700$.

2.3. Reformulation of the diffusion generator

To extend the method of S2007 to compute higher-order moments of r^2, we first change variables in the TLD model as follows. Let $u = 1 - 2p$ and $v = 1 - 2q$. By the fact that $0 < p, q < 1$, we have $-1 < u, v < 1$ and thus $0 \leq u^2, v^2 < 1$. This reformulation has the property that a relabeling of allele A_1 to A_2 has the effect of mapping u to $-u$. This is a more advantageous way of expressing the symmetry between the two alleles than the previous mapping of p to $1 - p$, and considerably simplifies the ongoing computations.

Since $p = (1-u)/2$ and $q = (1-v)/2$, we have $p\,(1-p) = \left(1-u^2\right)/4$ and $q\,(1-q) = \left(1-v^2\right)/4$. In addition, $\partial/\partial p = -2\partial/\partial u$ and $\partial/\partial q = -2\partial/\partial v$. Then the diffusion generator (3) of the TLD model in terms of the new variables (u, v, D) can be written as

$$\begin{aligned}\mathcal{L}^* = {} & \frac{1}{2}\left(1-v^2\right)\frac{\partial^2}{\partial v^2} + \frac{1}{2}\left\{\frac{1}{16}\left(1-u^2\right)\left(1-v^2\right) + Duv - D^2\right\}\frac{\partial^2}{\partial D^2} \\ & + \frac{1}{2}\left(1-u^2\right)\frac{\partial^2}{\partial u^2} + 4D\frac{\partial^2}{\partial u \partial v} - 2Du\frac{\partial^2}{\partial D \partial u} - 2Dv\frac{\partial^2}{\partial D \partial v} \\ & - \frac{1}{2}\theta u\frac{\partial}{\partial u} - \frac{1}{2}\theta v\frac{\partial}{\partial v} - \left(1 + \frac{1}{2}\rho + \theta\right)D\frac{\partial}{\partial D}\ .\end{aligned} \tag{11}$$

The definition of r^2 in terms of the new variables is

$$r^2 = \frac{D^2}{p\,(1-p)\,q\,(1-q)} = \frac{16D^2}{(1-u^2)\,(1-v^2)}\ . \tag{12}$$

For $0 \leq u^2, v^2 < 1$, we have $1/\left(1-u^2\right) = \sum_{k=0}^{\infty} u^{2k}$ and $1/\left(1-v^2\right) = \sum_{l=0}^{\infty} v^{2l}$. Substituting these into (12) yields

$$r^2 = 16\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} D^2 u^{2k} v^{2l}. \tag{13}$$

The advantage of the change in variables is seen by comparing (13) with the previous expression (6). The expression in (13) can be raised to higher powers while still producing only monomial terms, unlike the expression in (6) which would involve cross-terms of the form $p^k(1-p)^{k'}q^l(1-q)^{l'}$. Although such cross-terms could be reduced to monomials of the form $p^m q^n$ by binomial expansions, this would result in a mixture of positive and negative terms, which is problematic for numerical computation due to cancellation errors.

We now seek to find $\mathrm{E}\left(r^{2M}\right)$ for $M = 1, 2, \ldots$. Because $|u^2| < 1$ and $|v^2| < 1$, we can apply the negative binomial series to the factors of the denominator in (12), yielding

$$\left(1-u^2\right)^{-M} = \sum_{K=0}^{\infty} \binom{K+M-1}{K} u^{2K}, \quad \left(1-v^2\right)^{-M} = \sum_{L=0}^{\infty} \binom{L+M-1}{L} v^{2L}. \tag{14}$$

Combining equations (12) and (14) gives

$$\begin{aligned} \mathrm{E}\left(r^{2M}\right) &= \mathrm{E}\left\{16^M D^{2M} \left(1-u^2\right)^{-M} \left(1-v^2\right)^{-M}\right\} \\ &= 16^M \sum_{K=0}^{\infty} \sum_{L=0}^{\infty} \binom{K+M-1}{K} \binom{L+M-1}{L} \mathrm{E}\left(D^{2M} u^{2K} v^{2L}\right). \end{aligned} \tag{15}$$

Then the problem of computing the moment of r^2 of order M reduces to computing expectations of the form $\mathrm{E}\left(D^{2M} u^i v^j\right)$ for all combinations of non-negative even integers i and j. To compute these expectations, we follow the procedure of S2007's original algorithm described in the last section, but using variables u and v to replace p and q respectively, and using the new diffusion generator (11) to replace the original generator (3). For example, we can see from (9) that if m and n are sufficiently large positive even integers, the $n+3$ expectations contain expectation $\mathrm{E}\left(D^4 u^{m-2} v^{n-2}\right)$ needed for computing $\mathrm{E}\left(r^4\right)$, expectation $\mathrm{E}\left(D^6 u^{m-4} v^{n-4}\right)$ needed for computing $\mathrm{E}\left(r^6\right)$, and other expectations needed for computing higher-order moments of r^2. Similar to S2007, we set $\ell_{\max} = 2(K+L)$ to be the truncation level and use

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

as an approximation to $\mathrm{E}\left(r^{2M}\right)$.

2.4. Maximum entropy principle

Information entropy was first introduced by Shannon & Weaver (1948) as a measure of the probabilistic uncertainty inherent in a probability distribution. The maximum entropy principle states that, having satisfied all known constraints, the most reasonable probability distribution for representing a system is the solution with maximal entropy. The so-called Maxent principle is widely used in various scientific areas such as statistical mechanics, computer science, biology, economics and finance (e.g. Berger et al., 1996; Tagliani, 1999; Wu, 2003; Yeo & Burge, 2004), but as far as we know, its only application in population genetics to date is in Liu (2012). The Maxent principle is a powerful technique for representing an unknown distribution subject to some known constraints, such as the finite sequence of moments of the distribution as we have derived here.

Suppose we are seeking the true probability density function $\pi\left(x\right)$ of random variable X; however, we know nothing about the distribution except for some moments $m_i = \mathrm{E}\left(X^i\right), i = 1, \ldots, n$, where m_i is the ith moment of $\pi\left(x\right)$. The Maxent solution is $\widetilde{\pi}_n\left(x\right)$ that maximizes the expectation of the negative logarithm of the density function of X, while satisfying all the specified moment constraints. Let Ω denote the support of π, which here is $[0, 1]$ corresponding to the range of r^2. We therefore seek $\widetilde{\pi}_n$ that maximizes the entropy of the distribution:

$$I\left(\widetilde{\pi}_n\right) = -\int_{\Omega} \widetilde{\pi}_n\left(x\right) \log\left\{\widetilde{\pi}_n\left(x\right)\right\} dx \tag{17}$$

subject to

$$\int_{\Omega} x^i\, \widetilde{\pi}_n\left(x\right) dx = m_i \quad \text{for} \quad i = 0, 1, \ldots, n. \tag{18}$$

By application of the corresponding Lagrange function and the Euler-Lagrange equation, it is readily shown that the Maxent density function is

$$\widetilde{\pi}_n\left(x\right) = \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_n x^n\right), \tag{19}$$

where $\boldsymbol{\lambda} = \left(\lambda_0, \lambda_1, \ldots, \lambda_n\right)$ solves the following unconstrained minimization problem:

$$\arg\min_{\boldsymbol{\lambda}} \left\{ \int_{\Omega} \exp\left(\sum_{i=0}^{n} \lambda_i x^i \right) dx - \sum_{i=0}^{n} \lambda_i m_i \right\}. \tag{20}$$

The density $\widetilde{\pi}_n(x)$ in (19) is normalized due to the constraint for $i = 0$ in (18). Here, we apply (19) and (20) to obtain $\widetilde{\pi}_n(r^2)$, the Maxent solution based on n moments to the stationary probability density function π of the random variable r^2. Convergence of $\widetilde{\pi}_n$ to π is guaranteed as $n \to \infty$, because π is determined by the moment sequence $\{m_i\}_{i=0}^{\infty}$, and since $|r^2| < 1$ the error in omitting moments higher than n tends to zero as $n \to \infty$. We describe methods for selecting n in the next section.

There are two difficulties when solving the minimization problem (20) numerically. First, it is challenging to evaluate the definite integral in the objective function. We address this using Gaussian-Legendre quadrature nodes (Hildebrand, 1987). Second, computation of the objective in (20) is numerically unstable due to the power moments m_i, which become extremely small for large i. We therefore transform the power moments into corresponding Chebyshev moments, and then apply the Maxent principle to these shifted moments. The transformation to Chebyshev moments greatly improves the performance of numerical optimization algorithms for Maxent problems (Silver & Röder, 1997). After applying the Gaussian-Legendre nodes and Chebyshev moments, we employ the trust region optimization algorithm (Wright & Nocedal, 1999) to solve the resulting optimization problem. See Liu (2012) for more technical details about the computation. For more details about the Maxent approach, see Jaynes (1982) and Cover & Thomas (2012).

2.5. Sequential updating algorithm

Using our method based on (9) and (16), we can obtain a series of arbitrarily many stationary moments of r^2. We therefore need to decide upon an appropriate number of moments to use when constructing the Maxent density of r^2. In principle, we expect a larger number of moments to yield a more accurate approximation to the true density function π; however, it also becomes more difficult and time-consuming to compute the optimal solution to the minimization problem (20) because the dimension of the unknown $\boldsymbol{\lambda}$ increases. To address this problem, we propose a criterion based on the sequential updating method of Wu (2003). Instead of simultaneously using all the moments available, we incorporate the moments one by one from lower to higher order, and update the Maxent density of r^2 sequentially until some specified criterion is satisfied.

We describe the approach in a general setting following the last section. Suppose the first k moments, m_i for $i = 1, 2, \ldots, k$, are used to construct the Maxent density $\widetilde{\pi}_k\left(x\right)$. Using $\widetilde{\pi}_k\left(x\right)$, the moment of X of order $k+1$ can be predicted by

$$\widetilde{m}_{k+1} = \int_{\Omega} x^{k+1} \widetilde{\pi}_k\left(x\right) dx. \tag{21}$$

The difference between the predicted moment $\widetilde{m}_{k+1}$, based on the kth-order approximation $\widetilde{\pi}_k$, and the true moment m_{k+1}, from the analytic computation, serves as an indicator to decide whether more moments are needed. If $\widetilde{m}_{k+1}$ is very close to m_{k+1}, this suggests that most of the information contained in m_{k+1} may already be provided by the first k moments, in which case little may be lost by omitting the moment m_{k+1}. We use the percentage bias b_k defined below to measure the difference between m_{k+1} and $\widetilde{m}_{k+1}$:

$$b_k = \frac{\widetilde{m}_{k+1} - m_{k+1}}{m_{k+1}} \times 100\%. \tag{22}$$

Thus, b_k is an indicator of the performance of using moments of order up to k to construct the density of X. In general, we expect b_k to tend towards zero as k increases. In our computation, we first set a threshold for b_k at a small percentage, say 1%. Starting with $k = 1$ moment, we increment k at each step until b_k is found to be under the threshold. We then use the Maxent density $\widetilde{\pi}_k$ as our final estimate.

3. Results

3.1. Expectation and variance of r^2

To verify our proposed method, we computed a series of moments of r^2 for various values of ρ and θ following S2007. Results were obtained for all combinations of $\rho \in \{0, 1, 2, 5, 10, 20\}$ and $\theta \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 4.0, 6.0, 8.0, 10.0\}$.

Results of $\mathrm{E}\left(r^2\right)$ from our method using (16) were in all cases identical to those obtained by S2007 to the three decimal places supplied by S2007; see Table 1 in S2007 for details. This provides evidence of the correctness of our computations, since the two methods adopt different sets of model variables and diffusion generators.

In addition, by adopting the change of variables we have proposed and using (16) and (9), the variance and higher-order moments of r^2 can readily be evaluated. Using the same truncation level (say 700), both approaches require a similar amount of computation time (roughly two minutes on a 1.3 GHz laptop); however, S2007's method only gives $\mathrm{E}\left(r^2\right)$, whereas we obtain the first M moments of r^2 (say $M = 20$). The computation time for our method remains almost the same if more moments are computed while the truncation level remains constant.

Here we show the results of $\mathrm{Var}\left(r^2\right)$, the variance of r^2, which is of particular interest. This can be obtained by our method using the first two moments of r^2. As far as we know, $\mathrm{Var}\left(r^2\right)$ in the TLD model has previously only been obtained by a lengthy computation in Liu (2012); however, Liu (2012) indicated that his computation was not satisfactory for small θ due to limitations of computing power. Our method for $\mathrm{Var}\left(r^2\right)$, based on (16) and (9), is more reliable as it is analytic and does not involve any density approximation or numerical optimization or integration, only the solution of a system of linear equations. Both methods rely on the same assumption (in common with S2007) that the process is adequately approximated by the diffusion approximation.

Table 1 shows results for $\mathrm{Var}\left(r^2\right)$ from our computation for ρ, θ as above. Our results are almost identical to those of Liu (2012) for $\theta > 1$, although there are larger discrepancies for smaller θ, as anticipated. From Table 1 it can be seen that the variance of r^2 appears to be a decreasing function of ρ if θ is fixed. When ρ is fixed, $\mathrm{Var}\left(r^2\right)$ typically increases to a peak, and then declines again as θ increases.

Table 1: Computation of $\mathrm{Var}(r^2)$, the stationary variance of r^2, for various ρ, θ using truncation level $\ell_{\max} = 2000$

ρ	θ										
	0.1	0.2	0.4	0.6	0.8	1.0	2.0	4.0	6.0	8.0	10.0
0.0	0.006210	0.016406	0.03134	0.03878	0.03825	0.03664	0.02435	0.01174	0.00685	0.00450	0.003239
1.0	0.003477	0.009613	0.01972	0.02480	0.02629	0.02596	0.01894	0.00988	0.00608	0.00413	0.002928
2.0	0.002326	0.006571	0.01390	0.01789	0.01945	0.01965	0.01519	0.00856	0.00538	0.00374	0.002739
5.0	0.001102	0.003146	0.00679	0.00897	0.01001	0.01030	0.00893	0.00585	0.00403	0.00293	0.002197
10.0	0.000596	0.001574	0.00337	0.00440	0.00494	0.00512	0.00477	0.00352	0.00263	0.00206	0.001590
20.0	0.000265	0.000730	0.00149	0.00192	0.00211	0.00218	0.00202	0.00167	0.00141	0.00115	0.000961

Table 2: Maximum likelihood estimation results obtained by the Maxent approach from 100 simulations

(ρ, θ)	$\hat{\rho}$			$\hat{\theta}$		
	Mean	Median CI width	CI coverage	Mean	Median CI width	CI coverage
$(10, 1.00)$	9.96	1.06	0.97	1.00	0.06	0.97
$(10, 0.05)$	9.99	0.89	0.91	0.05	0.0012	0.93
$(5, 1.00)$	5.01	0.61	0.96	1.00	0.05	0.93
$(5, 0.10)$	5.00	0.50	0.96	0.10	0.0023	0.93

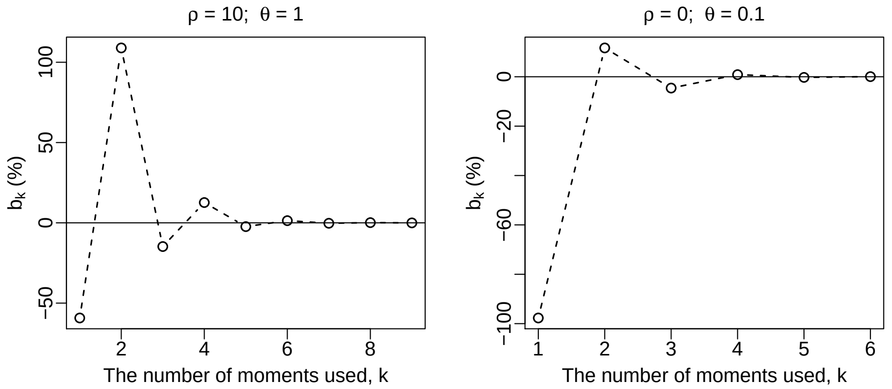

Figure 1: Relationship between the percentage bias b_k and the number of moments used, k, for $(\rho, \theta) = (10, 1)$ (left) and $(\rho, \theta) = (0, 0.1)$ (right).

3.2. Maxent density of r^2

Using the moments obtained from (9) and (16), we can also construct the Maxent density of r^2 for any parameters (ρ, θ). We use the sequential updating approach to select an appropriate number of moments for each (ρ, θ). Figure 1 shows convergence of the indicator b_k to zero as the number of moments k increases, for two settings of (ρ, θ). The rate of convergence depends on the values of ρ and θ. For example, b_4 is almost zero when $(\rho, \theta) = (0, 0.1)$, while for $(\rho, \theta) = (10, 1)$, it is clearly larger than zero.

Figure 2 presents Maxent densities of r^2 for two different choices of parameters (ρ, θ), using 1% as the threshold for terminating the sequential updating algorithm. Computation time for constructing the Maxent density of r^2 for one pair of (ρ, θ) is roughly two minutes on a 1.3 GHz laptop.

3.3. Maximum likelihood inference

We now demonstrate how the Maxent density of r^2 may be used to estimate ρ and θ from simulated data using maximum likelihood. Estimation of these parameters is of particular interest for comparing mutation and recombination

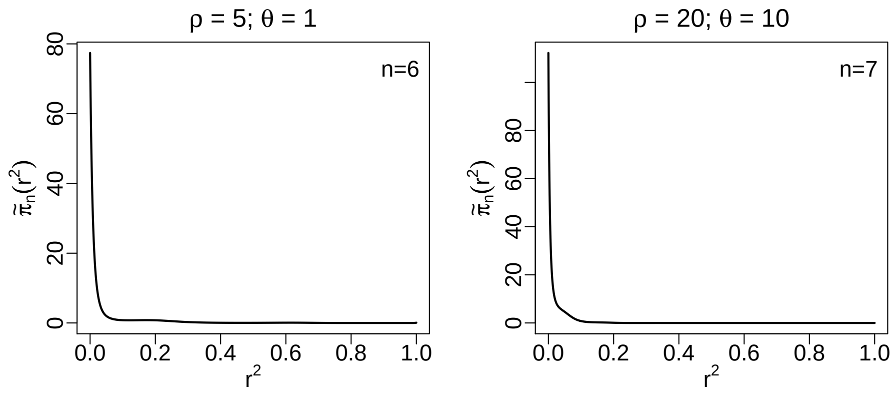

Figure 2: Maxent density functions of r^2 for two different settings of ρ and θ, constructed using moments of order up to n as selected by the sequential updating algorithm.

rates in different parts of the genome, or for detecting outliers that might suggest the influence of other genetic forces such as natural selection. In what follows, we consider how the parameters ρ and θ may be estimated from a sample of S values of r^2 obtained from S independent pairs of loci. For the present study, we disregard variability due to estimating r^2 from sample data at each pair of loci.

The Maxent density of r^2 we construct is based on the diffusion approximation, so the population size needs to be sufficiently large for this approximation to be reasonable. For this reason, it is difficult to simulate the discrete TLD model directly due to the enormous state space, which consists of all possible genotype counts $(x_1, x_2, x_3, x_4) \in \{0, 1, \ldots, 2N\}^4$ such that $x_1 + x_2 + x_3 + x_4 = 2N$. Adequate sampling of such a large state space at equilibrium is computationally impracticable. It is also unclear how to deal with the case where a locus is temporarily fixed for a single allele, because r^2 is then undefined for all generations until a mutation occurs. Simulations in S2007 imply that excluding these cases makes a substantial difference to the equilibrium distribution obtained. We are not aware of alternative methods for sampling directly from the

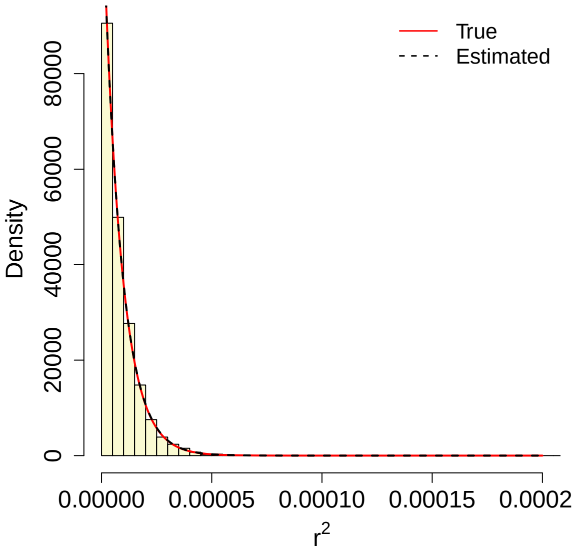

Figure 3: Histogram of 10,000 observations of r^2 generated by the rejection sampling method from the Maxent density of r^2 when $\rho_0 = 5$ and $\theta_0 = 0.01$. The red solid curve on the plot represents the true Maxent density. The dashed curve represents the Maxent density constructed using the MLEs $\left(\hat{\rho}, \hat{\theta}\right) = (5.10, 0.01)$ with 95% confidence intervals $(4.86, 5.35)$ for ρ and $(0.0099, 0.0102)$ for θ.

diffusion process corresponding to the TLD model. Since the Maxent density of r^2 for any parameters (ρ, θ) can be obtained, we instead generate samples of r^2 directly from this distribution using rejection sampling (Gilks & Wild, 1992).

As an example, we used $(\rho_0, \theta_0) = (10, 1)$ to generate sample observations of r^2. Based on a 5% threshold, the sequential updating algorithm terminated for these parameters at five moments, so the Maxent density of r^2 was taken to be $\widetilde{\pi}_5\left(r^2; \rho_0, \theta_0\right)$. We then generated a sample of 10,000 observations of r^2 from the density $\widetilde{\pi}_5\left(r^2; \rho_0, \theta_0\right)$ using the rejection sampling method, and then applied maximum likelihood estimation to obtain maximum likelihood estimates (MLEs) $\left(\hat{\rho}, \hat{\theta}\right)$ by numerically maximizing the log-likelihood based on reconstructing $\widetilde{\pi}_n\left(r^2; \rho, \theta\right)$ at every candidate pair (ρ, θ). Each likelihood evaluation involved constructing the moment sequence $m_1, m_2, \ldots,$ using (16) and (9), and reconstructing the density $\widetilde{\pi}_n$ using (19) and (20), where n is selected by the sequential updating algorithm at each iteration. Confidence intervals were constructed as Wald intervals based on the inverse Hessian at the maximum in the usual manner. From the sample data, the estimates of the parameters were $\left(\hat{\rho}, \hat{\theta}\right) = (9.86, 1.03)$ with 95% confidence intervals $(9.32, 10.40)$ for ρ and $(0.99, 1.06)$ for θ. The results have reasonably good precision for both parameters. On a 1.3 GHz laptop, the computation took approximately 25 minutes to generate the data and obtain the MLEs, so the approach is also feasible in terms of efficiency.

We repeated the procedure above 100 times. Inference results are shown in the first row of Table 2. Our procedure yields approximately unbiased estimation for both parameters with satisfactory confidence interval coverage for 95% confidence intervals. We also conducted simulation studies using other settings as shown in Table 2. The results indicate that the method has a similar performance in these cases.

For practical applications, smaller values of θ may be of interest, say 0.01. Simulating data by rejection sampling becomes rather slow for small θ, because the Maxent density of r^2 is very spiked close to zero. Once the data are obtained, maximum likelihood estimation is efficient as before. We show an example result in Figure 3, using data generated with $\rho_0 = 5$ and $\theta_0 = 0.01$.

An illustrative application of our maximum likelihood method to real data is given in Zhang (2017), using human genetic data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). Estimates of (ρ, θ) were obtained with good precision, with coefficients of variation typically around 2% to 8%. However, the estimates themselves do not have a straightforward biological interpretation, as loci were included in the study conditional on polymorphism with minor allele frequencies over 5%, thereby inflating the associated estimates of θ.

Various other inferential questions may be addressed using the framework we describe here. One application that might be of interest is the detection of recombination hotspots (Wall & Stevison, 2016) by estimating (ρ, θ) in different zones along a chromosome. Sample sizes may be enlarged by incorporating several different within-pair spacings, d, into a single analysis, for example by parametrizing ρ as a function of d and estimating the corresponding parameters using our method. Interest might then focus on changes in those parameter estimates along the chromosome. Another application is to use the Maxent density of r^2 to establish critical values for hypothesis tests, for example to test whether an observed value of r^2 is consistent with hypothesized values of ρ and θ.

4. Discussion

We have constructed the stationary probability density function of r^2 under the TLD model using a maximum entropy approach based on a series of the distribution's moments, which can be obtained under the diffusion approximation without knowledge of the distribution of r^2 itself. As a byproduct, we have created a fast, analytic computation of the variance of r^2, which was previously only available via an extremely lengthy computation derived by Liu (2012). We have shown how this Maxent density can be used to estimate the evolutionary parameters ρ and θ from sample observations of r^2. The performance of the method has been illustrated by simulation studies.

The Maxent principle is a powerful tool to represent the density of a probability distribution based on information known about the distribution, such as the sequence of moments we have used here. The Maxent density will converge to the true density for sufficiently many moments, but the rate of convergence is not known in general. Specifying the number of moments, n, therefore becomes an important part of a Maxent analysis. Our calculation in (16) delivers long moment sequences with little additional computational effort; but to convert n moments into a Maxent density, we optimize over the $(n+1)$-dimensional variable $\boldsymbol{\lambda}$ in (20), which does become more time-consuming as n increases. In this work we selected n using the percentage bias in forward-prediction of the next moment as a stopping criterion, and we found that the magnitude of the bias decreased rapidly and monotonically towards zero as n increased. However, a monotonic improvement cannot be guaranteed in all contexts, and empirical assessment such as that shown in Figure 1 is needed to verify the stopping criterion in practice. More stringent stopping criteria, for example involving forward-prediction of multiple moments instead of just one, are readily implemented.

The diffusion approximation provides a convenient and relatively fast way of computing the moments of r^2; however, we do not necessarily need to rely upon it. The Maxent approach can also be applied to other sources of moments, for example, the moments from accurate coalescent simulations, or the analytic moments of the original discrete process if these can be found. The diffusion approximation adds an extra layer of approximation error beyond the approximation of using Maxent procedures to represent a distribution from its moments. Thus the performance of our Maxent computation of the true density function at stationarity of the original discrete model can only be as good as the diffusion approximation itself.

S2007 mentioned that although mutation is assumed to be symmetric and recurrent for the TLD model, their method of calculating $\mathrm{E}\left(r^2\right)$ can be generalized to models with different mutation structures. They also commented that their method might also be generalized to genetic models with natural selection. Our approach to computing the moments of r^2 is based on the same ideas as S2007, so it is likely that it too has wider applicability than the TLD model alone. For models with different mutation structures, the diffusion generator in (11) is replaced by the appropriate alternative, but there is no impact on equations (12) to (16). Thus, although the system of linear equations will involve different values of $\mathrm{E}(D^{2M}u^{2K}v^{2L})$ under such models, we expect the solving algorithm to remain applicable. Investigating these alternative models, and deriving the corresponding diffusion generators and systems of linear equations similar to those presented in (9), is a suitable focus for future research.

Acknowledgements

We thank the editor and two referees for their constructive and insightful comments, and Y. S. Song for providing the C code for computing the expectation of r^2. Funding for WZ was provided by the China Scholarship Council and by the Royal Society of New Zealand through Marsden grant UOA1418.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., & Carey, A. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *The American Journal of Human Genetics*, *68*, 191–197.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*, 39–71.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, *41*, 337–348.

Golding, G. (1984). The sampling distribution of linkage disequilibrium. *Genetics*, *108*, 257–274.

Gupta, P. K., Rustgi, S., & Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology*, *57*, 461–485.

Hildebrand, F. B. (1987). *Introduction to numerical analysis*. Courier Corporation.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*, 226–231.

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, *33*, 54–78.

Hill, W. G., & Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, *54*, 705.

Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, *109*, 611–631.

Jaynes, E. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*, 939–952.

Liu, J. (2012). *Reconstruction of probability distributions in population genetics*. Ph.D. thesis The University of Auckland.

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, *5*, 355–364.

Ohta, T., & Kimura, M. (1969a). Linkage disequilibrium due to random genetic drift. *Genetics Research*, *13*, 47–55.

Ohta, T., & Kimura, M. (1969b). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, *63*, 229–238.

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, *69*, 1–14.

Service, S., DeYoung, J., Karayiorgou, M., Roos, J. L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J. A., & Heutink, P. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genetics*, *38*, 556–560.

Shannon, C., & Weaver, W. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Silver, R. N., & Röder, H. (1997). Calculation of densities of states and spectral functions by Chebyshev recursion and maximum entropy. *Physical Review E*, *56*, 4822–4829.

Slatkin, M. (2008). Linkage disequilibrium−understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*, 477–485.

Song, Y. S., & Song, J. S. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. *Theoretical Population Biology*, *71*, 49–60.

Tagliani, A. (1999). Hausdorff moment problem and maximum entropy: a unified approach. *Applied Mathematics and Computation*, *105*, 291–305.

Wall, J. D., & Stevison, L. S. (2016). Detecting recombination hotspots from patterns of linkage disequilibrium. *G3: Genes, Genomes, Genetics*, *6*, 2265–2271.

Weir, B. S., & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, *95*, 477–488.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

Wright, S., & Nocedal, J. (1999). *Numerical optimization*. Springer.

Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, *115*, 347–354.

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, *11*, 377–394.

Zhang, W. (2017). *Probability density approximation methods with applications in ecology and population genetics*. Ph.D. thesis The University of Auckland.