Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 65 (2004) 249–261
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Brownian models and coalescent structures

Michael G.B. Blum,[a,b,*] Christophe Damerval,[b] Stephanie Manel,[c] and Olivier François[b]

[a] Laboratoire Ecologie, Systematique et Evolution, Université Paris Sud, Bâtiment 360, F91405 Orsay, France
[b] TIMC-TIMB, Faculté de Médecine, F38706 La Tronche, France
[c] Laboratoire d'Ecologie Alpine, Université Joseph Fourier, BP 53 F38041 Grenoble, France

Received 29 April 2003

Abstract

Brownian motions on coalescent structures have a biological relevance, either as an approximation of the stepwise mutation model for microsatellites, or as a model of spatial evolution considering the locations of individuals at successive generations. We discuss estimation procedures for the dispersal parameter of a Brownian motion defined on coalescent trees. First, we consider the mean square distance unbiased estimator and compute its variance. In a second approach, we introduce a phylogenetic estimator. Given the UPGMA topology, the likelihood of the parameter is computed thanks to a new dynamical programming method. By a proper correction, an unbiased estimator is derived from the pseudomaximum of the likelihood. The last approach consists of computing the likelihood by a Markov chain Monte Carlo sampling method. In the one-dimensional Brownian motion, this method seems less reliable than pseudomaximum-likelihood.

Keywords: Coalescent theory; Brownian motion; Pairwise statistic; Pseudomaximum-likelihood; Microsatellite evolution; Spatial dispersal

1. Introduction

In this article we discuss estimation procedures for the parameter of a Brownian motion model defined on coalescent trees. Let us consider n random variables $X_1, X_2, \ldots, X_n$, resulting at the tips of a binary rooted tree from one-dimensional Brownian random walks on the branches of this tree. We assume that the trees are randomly sampled according to Kingman's model of coalescence (Kingman, 1982). Therefore, branch length corresponds to the time elapsed since the divergence of lineages in a neutral evolution. The random walk starts at the root of the tree, and splits into independent copies when it goes through a node. We assume that the two copies are conditionally independent given the common value at the split node. The second-order structure of Brownian motions (B_t) is specified as follows:

$$E[B_t^2] = \theta t, \quad t > 0$$

for some parameter $\theta > 0$ which is the object of the estimation procedures.

Brownian motion on coalescent trees were introduced as an approximation to the ladder model of microsatellite evolution and then implemented in a computer program by Beerli (2002). This approximation replaces the discrete stepwise mutation model of Kimura and Otha (1972) with a continuous model, and assumes that the changes in microsatellite length could be approximated by a Gaussian distribution. The approximation proved useful in the context of Markov chain Monte Carlo methods, because computations could be made many times faster. Beerli (2002) reported that it appeared to work well except when genealogies have very short branches (and gave the example of those associated with very small population sizes) on which it showed a significant upward bias.

Random walks were also introduced in the context of models of *isolation by distance* in continuous populations (Wright, 1943; Malécot, 1967) where spatial dispersal is often localized in space. This approach includes a parameter σ^2 that represents the rate of dispersal, i.e., the averaged squared distance between parents and offspring. Many theoretical attempts have been made in order to describe levels of genetic differentiation in terms of this parameter (e.g., Cox and Durrett, 2002). However these analyses often relied

*Corresponding author. Laboratoire Ecologie, Systematique et Evolution, Université Paris Sud, Bâtiment 360, F91405 Orsay, France.
E-mail address: michael.blum@imag.fr (M.G.B. Blum).

0040-5809/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2003.11.002

upon a discrete model, namely the stepping stone model of Kimura (1953). Parameter estimation methods based on the stepping stone model are discussed by Rousset (2003). Here we reexamine the inference problem from another point of view. We base parameter estimation on the separation of the spatial data (i.e., the locations of individuals) and the genetic data, considering the spatial data as *non-genetic* inherited characters. More specifically, Brownian motions on coalescent trees are regarded as a model for the evolution of spatial data in a large one-dimensional habitat modulo a correct rescaling of time. In fact, our approach involves the estimation of the product of the spatial dispersal rate σ^2 times the effective population size N_e

$$\theta = \sigma^2 N_e.$$

Hence, we base the estimation of θ on spatial data only. Nevertheless, estimating σ^2 yet requires genetic data because these data are usually necessary for estimating the effective population size N_e (Beaumont, 2003).

The paper is structured as follows. At the beginning of Section 2, we present the basic assumptions on which our model is based. At the end of Section 2, we describe Brownian motion on coalescent trees as the limit of discrete stepwise models on such genealogies. Two kinds of estimation methods are studied: the first based on a pairwise statistic (Section 3) and the others based on likelihoods (Section 4). Both are relevant to traditional approaches in statistical genetics. Estimation based on likelihoods is the most recent approach, and warrants optimal properties of estimators for large sample sizes. In our context, computing likelihood is a difficult issue because this function is expressed as a high-dimensional integral

$$L(\theta) = \int p(D|G)p(G|\theta)\, dG, \tag{1}$$

where $p(D/G)$ is the conditional distribution of the data given the genealogy of the sample G, and $p(G/\theta)$ is the distribution of such genealogies (see Stephens, 2003). The summation over all possible genealogies cannot be performed analytically unless the sample size remains very small. Section 4.2 presents a fast computational method for estimating θ from a pseudomaximum-likelihood approximation. Section 4.4 deals with Markov chain Monte Carlo approximations.

2. Models

2.1. Coalescent trees

Kingman's coalescent genealogies (Kingman, 1982) are large-size limits of genealogies under the assumption that populations reproduce according to an idealized neutral *Wright–Fisher* model. Given a sample of n individuals, the ancestral process can be defined as a continuous-time Markov chain for which the jumps correspond to the times of coalescence of ancestral lineages. Let T_{n-1} be the time since the most recent common ancestor (MRCA) in the sample, T_{n-2} the time since two distinct ancestors in the sample, $T_0 = 0$. Kingman's theory states that the durations separating coalescence events $Z_k = T_{n-k+1} - T_{n-k}$, $k = 2, \ldots, n$, are independent exponentially distributed random variables of rates $\lambda_k = k(k-1)/2$. Under the assumption that time is measured in units of N_e generations, the probability distribution of genealogies G can be described as

$$p(G) = \prod_{k=2}^{n} \exp(-\lambda_k z_k). \tag{2}$$

An alternative way of measuring time is by rescaling as follows:

$$t \equiv \theta t$$

for some $\theta > 0$. Under this transformation, the distribution of genealogies depends on the parameter θ as follows:

$$p(G|\theta) = \frac{1}{\theta^{n-1}} \exp\left(-\sum_{k=2}^{n} \lambda_k \frac{z_k}{\theta} \right). \tag{3}$$

According to Felsenstein et al. (1999), the approximation of discrete genealogies is valid when $n^2 \ll N_e$, and was observed as being extraordinary accurate in practice.

2.2. Random walks

The model of Kimura and Otha (1972) is a random walk model that has been applied to the evolution of microsatellites. Microsatellites are genetic markers where a given motif of DNA is repeated several times. These data are particularly useful for population genetics studies as they are abundant and widely dispersed in eukaryotic genomes, and have high mutation rates. The number of repetitions is called the length of the microsatellite.

In the discrete ladder model of Kimura and Otha, the population at generation ℓ consists of N_e diploid individuals. At generation $\ell + 1$, N_e offspring are created by sampling with replacement from the parental population and the parent allele can mutate with probability μ. Mutations randomly decrease or increase the length of the sequence. It is standard practice to set $\theta = 4N_e\mu$. However, we use the notation that $\theta = 2N_e\mu$ in order to obtain results homogenous with the spatial applications of Brownian motions. In the large-size limit, mutations occur according to independent Poisson processes of rate θ on each branch of the genealogy.

In a spatial model, we consider a haploid population. Each individual gives birth to a Poisson random number of offspring at rate $\lambda > 0$. Conditional to the fact that the population size is constant, this description is independent of λ, and equivalent to the Wright–Fisher haploid neutral model (Tavaré, 2001). We assume that the offspring locations are random independent variables centered around the parent location with a variance equal to σ^2. Whatever the genetic structure of the population, spatial data are therefore inherited in the same way that neutral markers could be. However, the term *neutral* is misleading as far as spatial locations are concerned. In this context, this term merely indicates the absence of density regulation. The important property is that coalescent approximation apply to this framework.

With time t measured in units of N_e generations, we set

$$\ell = \lfloor N_e t \rfloor.$$

The displacement of the offspring from an ancestor ℓ generations ago is

$$X_t = \xi_1 + \cdots + \xi_{\lfloor N_e t \rfloor},$$

where ξ_i corresponds to the displacement of the offspring from the parent in a single generation. In this situation, we take

$$\theta = \sigma^2 N_e,$$

and X_t has expectation $E[X_t] = 0$ and variance $\mathrm{Var}[X_t] = \theta t$. Considering the change of variable $t \equiv \theta t$ will be useful in likelihood computations. Under this transformation, the spatial diffusion rescales to the standard Brownian process independent of θ. This change of variable is only used in Section 4.

2.3. Brownian motion as an approximate model of stepwise mutation

In this section, we present informal arguments that motivate the use of Brownian models as limits of stepwise mutation models. We refer the reader to Appendix A for a more rigorous proof that Brownian models arise as the limit of sequences of compound Poisson processes which include the ladder model. Beerli (2002) reports that this kind of approximation breaks down when the effective population size is small ($\theta < 5$).

The stepwise mutation process is usually defined as follows. Let (M_t) count the number of mutations of the sequence before the time t. Mathematically, this process is defined as a homogeneous Poisson process of rate $\theta > 0$. The length variation of microsatellite markers at time t is given by

$$X_t^1 = \sum_{i=1}^{M_t} \xi_i,$$

where the ξ_i are independent identically distributed discrete random variables such that

$$E[\xi_i] = 0,$$

and

$$\mathrm{Var}[\xi_i] = v^2, \quad v > 0.$$

The basic idea that underpins the Brownian approximation is that the process (X_t^1) has the same covariance structure as the Brownian motion (B_t). A classical example of the stepwise mutation model is the symmetric random walk model

$$P(\xi_i = 1) = P(\xi_i = -1) = \tfrac{1}{2}$$

for which $v = 1$. For this model, let $s < t$, and compute

$$k(s,t) = \mathrm{cov}(X_s^1, X_t^1) = E\left[\sum_{i=1}^{M_s} \xi_i \sum_{j=1}^{M_t} \xi_j\right].$$

Using the fact that the Poisson process has independent increments, we obtain that

$$k(s,t) = E\left[\left(\sum_{i=1}^{M_s} \xi_i\right)^2\right] + E\left[\sum_{i=1}^{M_s} \xi_i\right] E\left[\sum_{j=M_s+1}^{M_t - M_s} \xi_j\right],$$

and

$$k(s,t) = E[M_s] = \theta s.$$

For large t, M_t is equivalent to θt almost surely. According to the Central Limit Theorem, X_t^1 behaves as a Gaussian random variable $\mathcal{N}(0, \theta t)$. In addition, we have

$$\mathrm{cov}(B_s, B_t) = \theta \min(s,t).$$

Since B_t is a Gaussian process, these equations tell that X_t^1 could be approximated by B_t.

Nevertheless, (X_t^1) is a continuous-time jump process that proceeds with discrete jumps. In contrast, (B_t) has continuous trajectories. In order to make rigorous statements, we need to rescale the processes (M_t) and (X_t^1) so that the mutations occur according to a Poisson process of rate θp. In addition, the basic jump of the rescaled process should be $\pm 1/\sqrt{p}$ instead of ± 1.

In this situation, a basic step ± 1 is the result of several steps of magnitude $\pm 1/\sqrt{p}$ which occur at rate p. When p goes to infinity, the rescaled process (X_t^p) converges to B_t where (B_t) is $\sqrt{\theta}$ times the standard Brownian motion.

3. Estimation based on pairwise statistics

In this section, we investigate the properties of an estimator of θ based on pairwise statistics. Consider a data set $D = X_1, \ldots, X_n$. The estimator is based on squared distance, as proposed by Slatkin (1995) and Goldstein et al. (1995) in the case of the stepwise mutation model. The idea behind such an estimator

relies on the fact that the squared distance increases linearly with time when going forward in the genealogy.

3.1. Basic results about X_n

The data set D is made of exchangeable variables, i.e., the X_i are identically distributed and the distribution of D is unchanged under arbitrary permutations of the variables. Because we assume that Brownian motions start from zero, the mean value of X_n is

$$E[X_n] = 0.$$

This can actually be shifted to any other value $E[X_n] = m$ by modifying the ancestral position from 0 to m. This may be important to do so in order to avoid negative values, in particular when microsatellite evolution is studied. The variance of X_n can be computed without difficulties

$$E[X_n^2] = \int_0^\infty E[B_t^2] f_{T_{n-1}}(t)\, dt = \theta E[T_{n-1}] = 2\theta(1 - 1/n)$$

and we see that an upper bound is 2θ (the averaged squared distance between clumps stay bounded away).

3.2. Squared distances

The distance between X_i and X_j is defined as follows:

$$d_{ij} = |X_i - X_j|,$$

and we study the pairwise squared distance statistics

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 - (\bar{X}_n)^2\right).$$

3.2.1. Moments

In this paragraph, we describe the moments of the difference $X_1 - X_2$ for two randomly chosen variables X_1 and X_2 in D. Let T be the time until the most recent common ancestor of the two individuals. Given $T = t$, $X_1 - X_2$ follows a Gaussian distribution of mean 0 and variance $2\theta t$. So, we have

$$E[X_1 - X_2|T] = E[(X_1 - X_2)^3|T] = 0,$$
$$E[(X_1 - X_2)^2|T] = E[d_{12}^2|T] = 2\theta T,$$
$$E[(X_1 - X_2)^4|T] = E[d_{12}^4|T] = 12\theta^2 T^2.$$

Because T follows an exponential distribution of parameter 1, we have

$$\begin{aligned} E[d_{12}^2] &= \int_0^\infty E[S_n^2|T = t] f_T(t)\, dt = 2\theta \int_0^\infty t f_T(t)\, dt \\ &= 2\theta E[T] \\ &= 2\theta. \end{aligned} \tag{4}$$

The fourth-order moment of d_{12} can be computed as follows:

$$\begin{aligned} E[d_{12}^4] &= \int_0^\infty E[d_{12}^4|T = t] f_T(t)\, dt = 12\theta^2 \int_0^\infty t^2 f_T(t)\, dt \\ &= 12\theta^2 E[T^2] \\ &= 24\theta^2. \end{aligned} \tag{5}$$

Then, the variance of the squared distance is

$$\begin{aligned} \mathrm{Var}[d_{12}^2] &= E[d_{12}^4] - E[d_{12}^2]^2 \\ &= 24\theta^2 - 4\theta^2 \\ &= 20\theta^2. \end{aligned} \tag{6}$$

Pritchard and Feldman (1996) obtained similar equations in the case of the stepwise mutation model. Nevertheless, their result regarding the fourth-order moment $E[d_{12}^4]$ was different and involved an additional 2θ. To conclude this paragraph, we remark that the distribution of d_{12} has a very simple expression

$$\begin{aligned} P(d_{12} > t) &= \int_{s=0}^\infty P(|B_{2s}| > t) e^{-s}\, ds \\ &= \exp(-t/\sqrt{\theta}), \quad t > 0, \end{aligned}$$

from which the moments could be deduced again.

3.2.2. Bias

The result is that S_n^2 is an unbiased estimator of θ. Indeed, we have

$$S_n^2 = \frac{1}{(n-1)n}\sum_{i<j} d_{ij}^2,$$

and since the X_i's are exchangeable, we find

$$E[S_n^2] = \tfrac{1}{2} E[d_{12}^2] = \theta.$$

3.2.3. Variance

In order to find the variance of S_n^2, we follow the same lines of proof as Pritchard and Feldman (1996). Their computations were based on the second and fourth moments of d_{12}, for which we obtained explicit expressions in a previous paragraph. The variance of S_n^2 is

$$\mathrm{Var}[S_n^2] = \frac{2\theta^2(1 + 3n + 2n^2)}{3(n^2 - n)}.$$

Note that S_n^2 is not a consistent estimator of θ

$$\mathrm{Var}[S_n^2] \to \frac{4}{3}\theta^2, \quad \text{as } n \to \infty.$$

Remark 1. Define $(X_1^p, X_2^p, \ldots, X_n^p)$ as being a data set obtained at the leaves of a coalescent tree from the dynamics described in Section 2 and in Appendix A. Assume the convergence of the moments of $(X_1^p, X_2^p, \ldots, X_n^p)$ to those of $(X_1, X_2, \ldots, X_n)$. Let $d_{ij}^{(p)}$

be the difference between X_i^p and X_j^p. Define

$$S_n^2(p) = \frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} d_{ij}^{(p)2} = \frac{1}{pn(n-1)} \sum_{1 \leqslant i < j \leqslant n} (\sqrt{p}\, d_{ij}^{(p)})^2.$$

For Bernoulli random walks, the random variable $\sqrt{p}\, d_{ij}^{(p)}$ has the same distribution as the difference in repeat number between two individuals in a single step mutation model where the mutation rate is equal to $p\theta$. Applying the result of Pritchard and Feldman (1996), we have

$$\mathrm{Var}[S_n^2(p)] = \frac{1}{p^2} \frac{\theta p[n(n+1)] + 2\theta^2 p^2[1 + 3n + 2n^2]}{3(n^2 - n)}.$$

Thanks to the convergence of moments, the variance of S_n^2 is

$$\mathrm{Var}[S_n^2] = \lim_{p \to \infty} \mathrm{Var}[S_n^2(p)] = \frac{2\theta^2(1 + 3n + 2n^2)}{3(n^2 - n)}$$

and this establishes a direct proof of the result.

4. Estimation based on likelihood

4.1. A peeling algorithm

Given the set of data $D = x_1, \ldots, x_n$, likelihoods can be computed as the integral of $p(D/G) \times p(G/\theta)$ over all possible neutral genealogies G according to equation (1). Kingman's formula gives the distribution of genealogies $p(G/\theta)$ (see Section 2). In this Section, we describe a peeling algorithm that enables computing the conditional distribution $p(D/G)$ given the genealogy analytically. This procedure is based on the explicit calculation of the integrals that arise at each internal node of the tree when applying Felsenstein's likelihood method (Felsenstein, 1981).

As usual in phylogenetic likelihood algorithms, we associate trees with $(n-1) \times 4$ arrays as follows

row i	node 1	node 2	ancestor	coalescence time

$(i = 1, \ldots, n-1)$ where n is the sample size. In addition, we assume that the coalescence times are ranked in increasing order, i.e., $t_1 < t_2 < \cdots < t_{n-1}$. The leaves are labelled from 1 to n and the other nodes (corresponding to the ancestors) are labelled from $n+1$ to $2n-1$. For example, consider the tree with 4 leaves given by

$$G = \begin{array}{cccc} 1 & 2 & 5 & t1 \\ 3 & 5 & 6 & t2 \\ 4 & 6 & 7 & t3 \end{array}$$

For $D = x_1, x_2, x_3, x_4$, Felsenstein's method computes $p(D|G)$ in the following way:

$$p(D|G) = \int \pi(x_7) p(x_4|x_7) \times \int p(x_6|x_7) p(x_3|x_6) \times \int p(x_5|x_6) p(x_1|x_5) p(x_2|x_5)\, dx_5\, dx_6\, dx_7, \quad (7)$$

where π is the distribution of $X_7 \equiv X_{MRCA}$ the value taken by the MRCA. Here, we consider a degenerate distribution where

$$X_{MRCA} = m,$$

for some m in $\mathbb{R}$. The procedure could be extended to arbitrary Gaussian distributions without difficulties. The computation of integrals of the type described above can be performed using the following technical result. Let $\alpha > 0$ and β, γ arbitrary real numbers, we have,

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + 2\beta x - \gamma}\, dx = \sqrt{\frac{\pi}{\alpha}} \exp(-\gamma + \beta^2/\alpha).$$

The distributions at each node can be characterized by four parameters A, α, β, γ, i.e.,

$$p(x|A, \alpha, \beta, \gamma) = A \exp(-\alpha x^2 + 2\beta x - \gamma).$$

We are now ready to describe the successive steps of the peeling algorithm.

(1) The peeling algorithm is initialized at the leaves of the tree as follows. If node n_1 corresponds to a leaf, then we take

$$A = \frac{1}{\sqrt{2\pi z}}, \quad \alpha = \frac{1}{2z}, \quad \beta = x_{n_1}\alpha, \quad \gamma = x_{n_1}^2 \alpha,$$

where z is the time until the most recent common ancestor with another node n_2 (z can be read in the fourth column of the tree structure). This computation corresponds for instance to the parameters that define $p(x_1|x_5)$ and $p(x_2|x_5)$ in the example genealogy G.

(2) Parameters at internal nodes, e.g., corresponding to the integral

$$I = \int p(x_5|x_6) p(x_1|x_5) p(x_2|x_5)\, dx_5,$$

are computed recursively thanks to the following induction formula:

$$A = A_1 A_2 / \sqrt{2z(\alpha_1 + \alpha_2) + 1},$$

$$\alpha = \frac{\alpha_1 + \alpha_2}{2z(\alpha_1 + \alpha_2) + 1},$$

$$\beta = \frac{\beta_1 + \beta_2}{2z(\alpha_1 + \alpha_2) + 1},$$

$$\gamma = \gamma_1 + \gamma_2 - \frac{2z(\beta_1 + \beta_2)^2}{2z(\alpha_1 + \alpha_2) + 1},$$

where $(A_1, \alpha_1, \beta_1, \gamma_1)$ and $(A_2, \alpha_2, \beta_2, \gamma_2)$ are the parameters at offspring nodes n_1 and n_2 and z is the time since the ancestor of the internal node.

(3) The algorithm terminates with an integral at the root of the tree, for which we use the same formula with $z = 0$. Finally, it returns

$$p(D|G) = A \exp(-\alpha m^2 + 2\beta m - \gamma),$$

where m is the mean of the sample.

A straightforward modification of this algorithm allows computing the log of distributions $\log P(D|G)$ instead of $P(D|G)$. The set of recursively computed parameters is then $\log A$, α, β and γ. In practice, we set $m = \bar{x}$ (the mean of the observed data). If a Gaussian distribution is assumed at the root of the tree, the final step of the algorithm could also use

$$z = \bar{s}^2(1 - 1/n)$$

with $\bar{s}^2$ the empirical variance of the data. In simulation experiments, we assumed a deterministic value at the root of the genealogy.

4.2. Pseudomaximum-likelihood algorithm

When computing $L(\theta)$, the genealogy of the data is unknown. Nevertheless, this genealogy may be viewed as an hidden or latent random variable over which an average must be performed. A reasonable guess of what the hidden variable looks like can help computing efficient numerical approximations of the likelihood. Techniques that employ such approximations are often called pseudomaximum-likelihood methods (Seo et al., 2002).

In this section, we build a pseudomaximum-likelihood method for estimating the parameter θ based on a specific genealogy. This tree is built from a phylogenetic reconstruction method that uses squared Euclidean distances between taxa (Slatkin, 1995). Because we make the hypothesis of constant evolution along the lineages, the UPGMA (Unweighted Pair Group Method using Averages) method is a natural mean to construct the topology of the tree (Nei, 1987).

Given the UPGMA topology, the lengths of the branches are taken equal to the average intercoalescence times, $z_n = 2\theta/n(n-1), \ldots, z_3 = 3\theta$, $z_2 = \theta$. Therefore, likelihoods are computed according to the peeling algorithm of Section 4.1 within an $O(n)$ time. The pseudomaximum-likelihood parameter $\hat{\theta}$ is then estimated thanks to a dichotomic search method. A limitation is that the method leads to a strong downward bias for large θ. This bias is due to the fact that squared distances do not account for the large deviations of Brownian motions.

In a first stage, we investigate the way of correcting the bias of the estimator using a Monte Carlo study for different parameter settings. In these experiments, data sets are created using coalescent simulations. For a given true parameter θ, a genealogy is sampled. This genealogy is then used to evolve the node variables according to the Brownian motion model. The data resulting at the tips of the simulated tree are then exploited in order to study the properties of the estimator. The experimental design consists of 1000 repetitions for each parameter setting (θ, n). The parameter θ ranges from 0.1 to 50, and the sample size ranges from $n = 10$ to 1000.

4.2.1. Bias

The main result of the simulation study is that the pseudomaximum-likelihood estimator $\hat{\theta}_n$ exhibits a constant multiplicative bias

$$E[\hat{\theta}_n] \approx b_n \theta.$$

In this relationship, the parameter b_n depends on the sample size n and is independent on θ (Table 1). This observation is consistent with the scaling property of Brownian motions. The values of the coefficient b_n are estimated as the correlation coefficient of a linear regression. In addition, a second regression analysis shows that the equation

$$E[\hat{\theta}] = \exp(-0.611 + 0.005n - 0.248\sqrt{n})\theta$$

fits the relationship between the average value of $\hat{\theta}$, θ and n extremely well ($R^2 \approx 0.99$). This formula provides a systematic way of correcting the estimation bias.

4.2.2. Variance

In a second stage, we investigate the quality of the pseudomaximum-likelihood estimator after bias correction. The corrected estimator is computed as

$$\tilde{\theta}_n = \frac{\hat{\theta}_n}{b_n} = e^{0.611 - 0.005n + 0.248\sqrt{n}}\hat{\theta}_n.$$

Table 1
Linear regression results for the multiplicative bias of $\hat{\theta}$, $E[\hat{\theta}_n] = b_n\theta + a_n$. The parameter θ ranges from 0.1 to 50. The linear coefficients are computed for different sample sizes. The intercepts a_n are not significant

n	b_n	Std. error	t-value	$\Pr(>\|t\|)$	R^2
10	0.277	0.0049	56.6	1.34e − 29	0.99
20	0.202	0.0034	58.1	6.58e − 30	0.99
30	0.165	0.0018	87.2	1.22e − 34	0.99
40	0.138	0.0019	70.8	3.24e − 32	0.99
50	0.119	0.0015	74.8	7.62e − 33	0.99
100	0.074	0.0007	101.4	2.09e − 36	0.99
150	0.057	0.0010	53.1	7.38e − 29	0.99
200	0.046	0.0005	82.6	5.23e − 34	0.99
250	0.040	0.0003	115.4	6.38e − 38	0.99
300	0.035	0.0004	78.8	1.84e − 33	0.99
350	0.031	0.0004	70.7	3.44e − 32	0.99
400	0.033	0.0012	26.1	1.07e − 20	0.96
450	0.028	0.0005	54.1	2.86e − 28	0.99
500	0.028	0.0010	27.8	1.91e − 21	0.96
1000	0.014	0.0001	95.5	1.20e − 34	0.99

Table 2
Bias and variance of pseudomaximum-likelihood estimator after correction. Parameter settings vary from $\theta = 0.1$ to 50, and sample sizes vary from $n = 20$ to 1000. The last column reports the standard deviations of the unbiaised pairwise estimator

θ	n	Mean($\tilde{\theta}$)	sd($\tilde{\theta}$)	sd(s_n^2)	θ	n	Mean($\tilde{\theta}$)	sd($\tilde{\theta}$)	sd(s_n^2)
0.1	20	0.104	0.080	0.122	0.5	20	0.522	0.401	0.614
	50	0.103	0.080	0.118		50	0.515	0.400	0.591
	100	0.095	0.067	0.116		100	0.476	0.337	0.584
	200	0.100	0.046	0.116		200	0.500	0.231	0.580
	300	0.091	0.047	0.115		300	0.459	0.237	0.579
	500	0.086	0.050	0.115		500	0.444	0.260	0.578
	1000	0.105	0.099	0.115		1000	0.467	0.273	0.578
1	20	1.086	0.850	1.229	5	20	4.738	4.251	6.145
	50	0.945	0.658	1.183		50	4.799	3.526	5.919
	100	1.078	0.680	1.169		100	4.832	3.443	5.846
	200	0.989	0.758	1.161		200	5.824	3.595	5.809
	300	1.038	0.778	1.159		300	4.592	2.548	5.797
	500	0.737	0.480	1.157		500	6.500	3.337	5.787
	1000	1.272	0.684	1.156		1000	4.887	2.794	5.780
10	20	11.154	7.502	12.295	50	20	48.541	38.142	61.451
	50	10.446	9.079	11.839		50	49.162	36.753	59.195
	100	10.980	8.883	11.692		100	50.951	34.421	58.460
	200	10.950	8.731	11.619		200	49.428	35.876	58.096
	300	9.020	4.851	11.595		300	51.542	26.653	57.976
	500	11.589	4.888	11.575		500	52.012	24.651	57.879
	1000	10.01	5.251	11.561		1000	48.013	29.651	57.807

Table 3
Regression results for the variance of the pseudomaximum-likelihood estimator $\mathrm{Var}[\tilde{\theta}] = c_n\theta + d_n\theta^2$. The linear coefficients c_n are nonsignificant

n	c_n	Std. error	t-value	Pr($>\|t\|$)	d_n	Std. error	t-value	Pr($>\|t\|$)
20	−0.276	0.223	−1.237	0.233	0.230	0.038	5.960	1.55e − 05
30	0.179	0.107	1.660	0.115	0.075	0.0186	4.041	8.4e − 04
40	0.068	0.222	0.308	0.762	0.058	0.038	1.536	0.143
100	−0.038	0.042	−0.911	0.375	0.041	0.007	5.604	3.16e − 05
150	0.0008	0.046	−0.018	0.985	0.025	0.007	3.201	0.005
200	0.013	0.024	0.527	0.605	0.014	0.004	3.387	0.003

For sample sizes below $n = 300$, a quadratic relationship between θ and the variance of $\tilde{\theta}$ can be identified thanks to a regression method

$$\mathrm{Var}(\tilde{\theta}_n) = d_n\theta^2.$$

Table 3 reports the values of the quadratic coefficient d_n and the significance levels of both coefficients c_n and d_n in the regression model $\mathrm{Var}(\tilde{\theta}_n) = c_n\theta + d_n\theta^2$. In Table 2, we report the average values of $\tilde{\theta}$ and the standard deviations of this estimator. The last column in Table 2 gives the standard deviations of the unbiaised pairwise estimator s_n^2. For small values of θ ($\theta \leqslant 5$), and samples of intermediate size (about 100–500) individuals the pseudomaximum-likelihood estimator is significantly better than the pairwise statistic. This observation remains true for larger θ, but the relative benefit is slighlty lower.

4.3. Lower bound of the variance of estimators of θ

Consider the random genealogy associated with the sample of data $X_1, \ldots, X_n$. There are exactly k branch segments in the tree during the time that separates the kth and the $(k-1)$th internal nodes.

As Fu and Li (1993), we measure time in generations and not in unit of N generations (in this subsection only). Let us denote $d_{k,i}$ $(i = 1, \ldots, k)$, the algebraic distance covered by the Brownian motions during z_k generations along the ith branch. The distribution of $d_{k,i}$ is Gaussian with mean 0 and variance $\sigma^2 z_k$, where we use the notation $\theta = \sigma^2 N$.

To establish a lower bound of the variance of estimators of θ, we follow the same lines of proofs as Fu and Li (1993). This approach consists of assuming that all evolutionary events are observable. More

precisely, we consider that the true topology of the genealogy is known as well as the number of generations between coalescent events $\{z_k, k = 2\ldots n\}$, and the algebraic distances covered by Brownian motions along each branch $\{d_{k,i},\ k = 2\ldots n, i = 1\ldots k\}$. Such a statistical model is called the complete case by Klein et al. (1999) in the context of the infinitely many-sites model. This model contains more informations than the incomplete model where the observations are the resulting values of the Brownian motions at the tips of the tree. Because the distribution of the topology is independent on N and σ (Kingman, 1982), we can restrict ourselves to the subset of data

$$S = \{z_k, d_{k,i}; k = 2\ldots n; i = 1\ldots k\}.$$

Given S, the likelihood can be computed as follows:

$$\begin{aligned} L(\sigma^2, N; S) &= \prod_{k=2}^{n}\left(\prod_{i=1}^{k}\frac{1}{\sqrt{2\pi\sigma^2 z_k}}e^{-\frac{d_{k,i}^2}{2\sigma^2 z_k}}\right)\frac{k(k-1)}{2N}e^{-\frac{k(k-1)z_k}{2N}} \\ &= \prod_{k=2}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2 z_k}}\right)^k e^{-\frac{d_k}{2\sigma^2 z_k}}\frac{k(k-1)}{2N} \\ &\quad \times e^{-\frac{k(k-1)z_k}{2N}}, \end{aligned} \tag{8}$$

where

$$d_k = \sum_{i=1}^{k} d_{k,i}^2.$$

The loglikelihood is

$$\begin{aligned} \log L = C &- \frac{(n-1)(n+2)}{4}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=2}^{n}\frac{d_k}{z_k} \\ &- (n-1)\log N - \sum_{k=2}^{n}\frac{k(k-1)z_k}{2N}, \end{aligned} \tag{9}$$

where C is independent of N and σ. Fisher's information matrix can be deduced from Eq. (9) by computing the second-order derivatives and taking the opposite of their expected values

$$I(N, \sigma^2) = \begin{pmatrix} \frac{n-1}{N^2} & 0 \\ 0 & \frac{(n-1)(n+2)}{4\sigma^4} \end{pmatrix}.$$

Theorem 2. *In the complete case and in the incomplete case, the variance of all unbiased estimator $\check{\theta}_n$ of θ is bounded below by*

$$\mathrm{Var}[\check{\theta}_n] \geqslant \frac{(n+6)}{(n-1)(n+2)}\theta^2. \tag{10}$$

This bound is asymptotically proportional to $1/n$ which is the typical variance of an estimator build according to independent observations. It can be compared to a similar bound found by Fu and Li (1993, Eq. (23)) in the infinite many sites model asymptotically equal to $1/\log n$. The difference between the two results is a consequence of the statistical properties of the Poisson process. In the infinitely many-sites model, the quantity of information depends linearly on the time elapsed since the root of the tree. Thus, the quantity of information in the complete case is proportional to the length of the tree which is asymptotically equal to $1/\log n$. In the Brownian model, the quantity of information is constant on any branch segment because of the rescaling property of Brownian motions. Thus, the quantity of information brought by the Brownian motions along the $O(n^2)$ branch segments is proportional to n^2. The information contained in the intercoalescence times is proportional to n and gives the major contribution to the Cramer–Rao estimate.

4.4. Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are computationally intensive statistical methods that have proven successful in estimating parameters of population genetics models. For instance, these methods have recently been used in estimating mutation rates (Kuhner et al., 1995), gene flow parameters (Beerli and Felsenstein, 2001), or the distribution of the time since the most recent common ancestor of a population (Tavaré, 2001; Griffiths and Tavaré, 1994).

4.4.1. Method description

Our approach is similar to the one described in Stephens (2003). In order to compute the likelihood of θ, an importance sampling method is applied. We use the Metropolis–Hastings (MH) algorithm for drawing samples of genealogies $G^1, G^2, \ldots$ from the importance distribution

$$q(G) = p(G|D) \propto p(D|G)p(G|\theta_0),$$

where θ_0 is an initial value. Our approach to importance sampling is based on the conditional coalescent. At this stage, we use a *Gibbs sampler* implementation (Robert and Casella, 1999). The strategy consists of removing a single internal node of the tree G at each proposal step, and then simulating the conditional coalescence time of this node given that the other times remain unchanged. The removed node is therefore randomly reintroduced into the tree with its new coalescence time.

Given M trees, we compute relative likelihoods as follows:

$$\frac{L(\theta)}{L(\theta_0)} \approx \frac{1}{M}\sum_i \frac{p(G^i|\theta)}{p(G^i|\theta_0)}.$$

In order to suppress the dependence on θ_0, this initial setting is kept only during a preliminary sampling. Then

a maximum likelihood value θ_1 based on this sample is found. A second Markov chain starts from θ_1, and a new maximum likelihood value θ_2 is found, etc. As suggested by Kuhner et al. (1995), we run 10 short chains and two longer ones at the end of the run.

4.4.2. Results

Data sets were created according to the same procedure as the one used in Section 5.2. Table 4 reports simulation results regarding the convergence of the MCMC estimator $\bar{\theta}$ for sample sizes $n = 20, 50, 100, 200$ and parameters $\theta = 1, 2.5, 5$. Biases and standard deviations are reported. In these experiments, the starting value was set to $\theta_0 = 3$. For small sample sizes ($n \leqslant 50$), the bias appears to be low. This can be explained as the set of genealogies is correctly explored, and the Monte Carlo Markov chain reached stationarity. For $n \geqslant 100$, the algorithm gets stuck in local optima more frequently. The standard deviations decrease as the sample sizes increase from $n = 20$ to 50. For large sample sizes, small variances may indicate that the algorithm has difficulties of escaping from the initial settings.

In a second series of experiments, the algorithm was run several times on two different simulated data sets ($n = 100, \theta = 50$). In the first data set, two different subpopulations are emerging whereas there are no distinct clusters in the second one (Fig. 1). In order to reconstruct the deep branches of the tree, more information is present in the first data set than in the second data set. The algorithm behaves differently in the two cases. While the estimation of θ is quite good for the first data set, it is inaccurate for the second data set (Table 5). This indicates that branches close to the root may have a strong influence on the variance of the estimator.

Table 4
Properties of the MCMC estimator $\bar{\theta}$. For each value of θ and n, mean and standard deviations are evaluated from 50 simulated data sets

θ	n	Mean($\bar{\theta}$)	s.d.($\bar{\theta}$)	5th percentile	95th percentile
1	20	0.99	0.29	0.60	1.63
	50	1.07	0.074	0.57	1.63
	100	1.44	0.34	0.78	2.15
	200	2.95	1.06	1.31	5.04
2.5	20	2.7	1.4	1.23	6.04
	50	2.36	0.45	1.52	3.29
	100	3.16	0.97	1.73	5.62
	200	3.03	0.66	1.89	4.29
5	20	4.2	1.16	1.70	5.88
	50	3.89	1.93	1.55	7.90
	100	3.38	0.97	1.97	5.36
	200	3.25	0.61	2.35	4.47

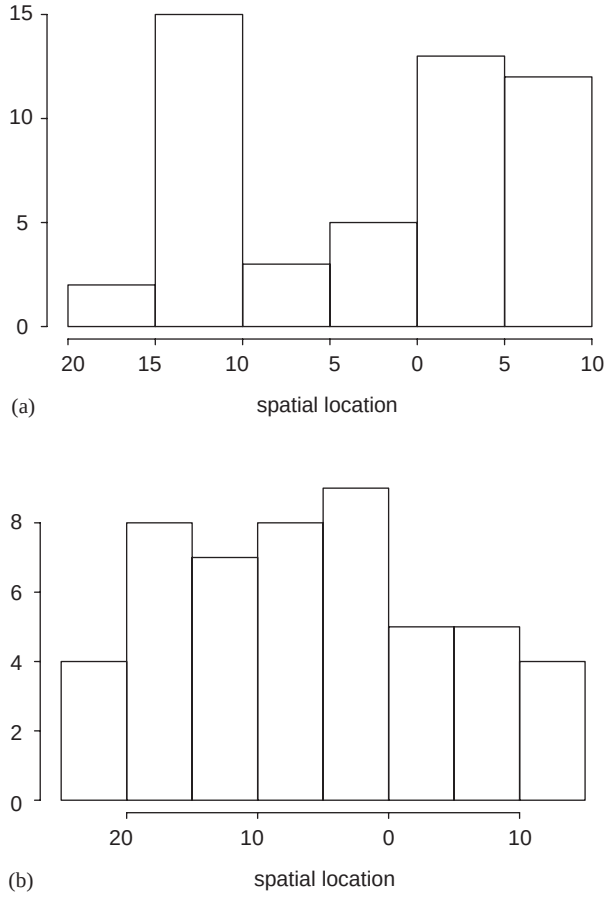

Fig. 1. Histograms of 2 samples simulated according to the Brownian model on coalescent trees. The first data set (a) exhibits two distinct clusters whereas the spatial distribution of the second one (b) is more uniform. There are 50 resulting individuals and $\theta = 50$ in both simulations.

Table 5
Properties of the MCMC estimator $\bar{\theta}$, 50 estimations have been run on each data set. In data set 1, two clusters are emerging whereas no distinct clusters are seen in data set 2. The estimator is more accurate on data set 1. The true value is $\theta = 50$, the initial value θ_0 has been set up to $\theta_0 = 70$

	Data set 1	Data set 2
Mean	59.7	93.3
Std. dev.	22.7	30.9
Median	55	87
5th percentile	33.4	49.35
95th percentile	92.6	149.5

5. Spatial dispersal: application to a biological data set

Estimating the amount of spatial dispersion of a species, σ^2, is crucial for many ecological studies. Two very different approaches to estimating this parameter can be used: (1) direct methods using direct observations of moving individuals (e.g., via mark and recapture), and (2) indirect methods using genetic data from samples of individuals. Direct methods can help to

determine the spatial pattern of dispersion during the study, and can deliver information about very recent history. However, there are also obvious shortcommings: The movements of individuals may be artefacts of the study, the accuracy of the parameter estimates may be small, and small dispersal rates may be undetectable.

Several indirect methods were devoted to the estimation of σ^2 based on molecular data. Slatkin (1993) showed that for allele frequency data, under a variety of dispersal models, there is an approximately linear log–log relationship between the product of the effective size and migration rates $N_e m_{ij}$ between a pair of populations i and j and their geographic distance, D_{ij}, namely

$$\log N_e m_{ij} = a + b \log D_{ij}.$$

A statistically significant negative regression coefficient, b, indicates that migration between populations becomes lower as their geographic separation increases, due to isolation by distance. This indicates that the regression coefficient contained information about the parameter σ^2. Rousset (1997) introduced a method based on the computation of F-statistics which exploit Slatkin's idea. He obtained an approximately linear relationship between F_{st} and the log of the geographic distance. This gave a practical method of estimating $d\sigma^2$, where d is the population density per surface unit. Rousset (2003) provides a recent survey of other methods available for estimating σ^2, and discusses the limitations of each method. For instance, the above method has the drawback of assuming well recognizable demes of several individuals, and estimators based on F_{st} may have high variance.

In Brownian models, the dispersal parameter is defined as being the standard deviation in a model where the locations of offspring followed from a probabilistic distribution centered around the parent. Felsenstein (1975) reported the existence of clumps in similar continuous models of isolation by distance, and some regulation of the density of individuals might be added in view of realistic applications (Barton et al., 2002). Brownian motions arise naturally in continuous limits from Kimura's stepwise migration models (Barton et al., 2002; Nagylaki, 2002). In these works, the authors usually study the conditional coalescence time given the spatial locations of two or more individuals. The typical conditional distribution has infinite mean, and hence is very different from the unconditional coalescence times used in the present article. However, although Brownian models might represent a rough approximation of the biological reality, it is useful because a number of theoretical insights are available.

As an example, we illustrated our approach with a sample of spatial locations of female brown bears in Scandinavia. The Scandinavian brown bear population has a strong phylopatry of females. Hence, the spatial locations of females can be thought as being maternally inherited like a haploid character. The Scandinavian brown bear population is subdivided in two distinct populations located at the South and North of an area covering both Sweden and Norway (Taberlet et al., 1994). These two subpopulations are isolated by the distance and regulated by male migration (Waits et al., 2000). We analyzed a sample of 64 female bears locations in the South area. These data were recorded as the latitude and the longitude of individuals at the instant of capture and then converted in kilometers (km). For this data set, the Mantel test of isolation by distance was non-significant, and GENEPOP gives an estimate of $\sigma = 2$ km (Raymond and Rousset, 1995), which is not in agreement with the knowledge of this population (Waits et al., 2000). The estimation of the effective size N_e is a critical step if one is interested in estimating σ^2 from θ. In this step, molecular data play an important role. We used as an estimate $N_e \approx 50$ (see Waits et al., 2000), confirmed by a multilocus microsatellite study that gave an expected homozygosity about 0.5 in this geographical area (the mutation rate can be taken as $\mu \approx 10e-2$ (Paetkau et al., 1998)). We obtain $\theta_x = \sigma^2 N_e \approx 4350$ and $\theta_y \approx 3034$ which means that $\sigma \approx 9$–10 km which is more consistent with field observations (Eva Bellemain, private comm.).

6. Discussion

In this article, Brownian motions were considered as models of evolution of genetic data (microsatellite) as well as models for the inheritance of non-genetic features (spatial locations). We proposed three estimators for the parameter of such models. The first estimator was based on mean pairwise square distance. The second estimator was a phylogenetic estimator relying on the UPGMA topology and mean coalescence times. The third estimator was based on approximate maximum likelihood using MCMC methods.

We found the exact variance of the mean pairwise estimator. Regarding the phylogenetic estimator, we found a systematic way of correcting the biases. After the correction, the quality of the estimation improves significantly. In addition, this approach has the merit of being very fast (few milliseconds runtimes for sizes of several hundred data). The MCMC method does not lead to improved estimation. In addition, the practical implementation of the MCMC method raises a number of questions that are specific to this family of algorithms. For instance, Wilson and Balding (1998) also reported biases for MCMC estimators in the context of microsatellite data when a single locus is used and evolution is modeled as a discrete random walk. This is in agreement with our results which show that the most likely genealogies can hardly be sampled

using the information contained in a one-dimensional random walk.

Regarding the convergence issue of MCMC, some difficult problems remain to be solved, where the relevance of the transition kernel is of primary importance. Even if our transition kernel had all the theoretical properties required, many other kernels should be tested and the choice for an optimal one is an open question. MCMC is time consuming. Theoretically it is asymptotically unbiaised, but our simulations shows the difficulty to tune the several internal parameters of the MCMC algorithm.

Some methods based on coalescent theory enable the estimation of the effective size N_e in recently isolated genetically diverging populations (O'Ryan et al., 1998). These approaches require the knowledge of an additional event: the time since the (usually two) populations have been isolated from each other. In the same spirit as (O'Ryan et al., 1998), our approach also provides an estimator of N_e given an estimator of σ^2 based on spatial data. For instance, indirect estimation of σ^2 using DNA fingerprinting (Bossart and Prowell, 1998) aims to exploit the recent shared genetic history between parents and offspring. Rousset's approach could be utilized as well, although the interpretation of the estimators should be different (Rousset, 1997). Nevertheless, the relevance of our method for estimating N_e could be a promising application, although its primary objective was estimating σ^2.

Acknowledgments

We would like to thank Professor Swenson for providing the data set on the bear locations.

Appendix A

Let us prove that Brownian motion can be obtained as limits of compound Poisson process which includes the stepwise mutation model.

Let $(\xi_i)_{i\geqslant 1}$ be independent and identically distributed random variables such that

$$E[\xi_i]=0 \quad \text{and} \quad \mathrm{Var}[\xi_i]=v^2, \quad v>0.$$

Let $\theta>0$. Consider a family of homogenous Poisson process (M_t^p) of rate θp $(p\geqslant 1)$ where M_t^p is the number of occurrences at time t. Define the *compound Poisson process* (X_t^p) as follows:

$$X_t^p=\frac{1}{v\sqrt{p}}\sum_{i=1}^{M_t^p}\xi_i. \tag{A.1}$$

Consider Bernoulli random variables

$$P(\xi_i=1)=P(\xi_i=-1)=1/2.$$

In this situation, we have $v^2=1$. For $p=1$, M_t^1 corresponds to the number of mutations at time t in the stepwise mutation model, and X_t^1 is the number of differences between the ancestral allelic state and the current state.

We consider more general mutation models than the ladder model, and establish the weak convergence of X_t^p to B_t.

Theorem A.1. *Let $(X_\cdot^p)$ be the stochastic process defined in Eq.* (A.1) *and $(B_\cdot)$ be a one-dimensional standard Brownian motion times $\sqrt{\theta}$. We have*

$$X_\cdot^p \overset{\mathscr{D}}{\Rightarrow} B_\cdot, \quad \text{as } p\to\infty,$$

where $\mathscr{D}$ denotes the weak convergence in $\mathscr{D}_{\mathscr{R}}(0,\infty)$, the set of càd-làg functions defined on $(0,\infty)$.

Proof. For all $t>0$ and $p\geqslant 1$, the first and second moments of X_t^p are

$$E[X_t^p]=0$$

and

$$\begin{aligned}\mathrm{Var}[X_t^p]&=E[(X_t^p)^2]=E[E[(X_t^p)^2|M_t^p]]\\&=\frac{1}{v^2p}E\left[E\left[\sum_{i=1}^{M_t^p}\xi_i^2|M_t^p\right]\right]=\frac{1}{v^2p}E[M_t^pv^2]\\&=\theta t.\end{aligned}$$

According to the Theorem 7.8 in Chapter 3 of Ethier and Kurtz (1986), the result follows from the convergence of the finite dimensional distributions and the relative compactness of $(X_\cdot^p)$ which are demonstrated below. □

A.1. Convergence of the finite-dimensional distributions

Let

$$Y_t^p=\frac{1}{v\sqrt{p}}\sum_{i=1}^{\lfloor\theta pt\rfloor}\xi_i. \tag{A.2}$$

First, we show that the random variables X_t^p converge weakly toward B_t for fixed $t>0$

$$X_t^p \overset{\mathscr{D}}{\Rightarrow} B_t.$$

Lemma A.1. *Let (X_t^p) and (Y_t^p) be defined in Eqs.* (A.1) *and* (A.2). *Then, $(Y_t^p-X_t^p)$ converges in probability to* 0 *as $p\to\infty$.*

Proof. Let $p\geqslant 1$ and $\varepsilon>0$. By the stationarity of the ξ_i's, we have

$$\begin{aligned}P(|X_t^p-Y_t^p|\geqslant\varepsilon)&=P\left(\frac{1}{v\sqrt{p}}\left|\sum_{i=\lfloor\theta pt\rfloor+1}^{M_t^p}\xi_i\right|\geqslant\varepsilon\right)\\&=P\left(\frac{1}{v\sqrt{p}}\left|\sum_{i=1}^{|M_t^p-\lfloor pt\rfloor|}\xi_i\right|\geqslant\varepsilon\right).\end{aligned}$$

Conditioning on M_t^p, this probability is equal to

$$\sum_{n=0}^{\infty} P\left(\left|\sum_{i=1}^{|n-\lfloor \theta pt \rfloor|} \xi_i\right| \geqslant v\sqrt{p}\,\varepsilon\right) P(M_t^p = n)$$
$$\leqslant \sum_{n=0}^{\infty} \frac{1}{v^2 p\,\varepsilon^2} \mathrm{Var}\left[\sum_{i=1}^{|n-\lfloor \theta pt \rfloor|} \xi_i\right] \times P(M_t^p = n)$$

and the upper bound follows from Chebyshev's inequality. Then we have

$$P(|X_t^p - Y_t^p| \geqslant \varepsilon) \leqslant \sum_{n=0}^{+\infty} \frac{|n - \lfloor \theta pt \rfloor|}{p\varepsilon^2} P(M_t^p = n)$$
$$\leqslant \frac{E[|M_t^p - \lfloor \theta pt \rfloor|]}{p\varepsilon^2}$$
$$\leqslant \frac{E[|(M_t^p - \theta pt) + (\theta pt - \lfloor \theta pt \rfloor)|]}{p\varepsilon^2}$$
$$\leqslant \frac{E[|M_t^p - \theta pt|] + E[|\theta pt - \lfloor \theta pt \rfloor|]}{p\varepsilon^2},$$

where we use the triangle inequality. We finish with the Cauchy–Schwarz inequality

$$P(|X_t^p - Y_t^p| \geqslant \varepsilon) \leqslant \frac{E[|M_t^p - \theta pt|] + 1}{p\varepsilon^2}$$
$$\leqslant \frac{\sqrt{E[(M_t^p - \theta pt)^2]} + 1}{p\varepsilon^2}$$
$$\leqslant \frac{\sqrt{\mathrm{Var}[M_t^p]} + 1}{p\varepsilon^2}$$
$$\leqslant \frac{\sqrt{\theta pt} + 1}{p\varepsilon^2}.$$

Since

$$\lim_{p \to \infty} \frac{\sqrt{\theta pt} + 1}{p\varepsilon^2} = 0,$$

the convergence is established. □

According to the central limit theorem and to the fact that $\lim_{p \to +\infty} \frac{\lfloor p\theta t \rfloor}{\theta p} = t$, we have

$$Y_t^p \xrightarrow{\mathscr{D}} B_t.$$

By Lemma 4, Slutsky's Theorem (Ethier and Kurtz, 1986) ensures the convergence in law of X_t^p to B_t. Now, fix $t_n > \cdots > t_1 > 0$. Showing

$$(X_{t_1}^p, \ldots, X_{t_n}^p) \xRightarrow{\mathscr{D}} (B_{t_1}, \ldots, B_{t_n})$$

amounts to prove that

$$(X_{t_1}^p, \ldots, X_{t_n}^p - X_{t_{n-1}}^p) \xRightarrow{\mathscr{D}} (B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}).$$

This result can be easily checked from the independence of increments in the compound Poisson process.

A.2. Relative compactness of $(X_\cdot^p)_{p \geqslant 1}$

Let $\mathscr{K}$ be the set of compact subsets of $\mathscr{R}$. From Theorems 8.6 and 8.8 in Chapter 3 of Ethier and Kurtz (1986), the following three conditions imply the relative compactness of $(X_\cdot^p)$.

- Condition (i)

  $\forall \eta > 0,\ \forall t \in \mathscr{Q}_+^*, \exists \Gamma_{\eta,t} \in \mathscr{K},\ \inf_{p \geqslant 1} P(X_t^p \in \Gamma_{\eta,t}) \geqslant 1 - \eta.$

- Condition (ii)

  $\exists C > 0,\ \forall T > 0,\ \forall t \in [0, T+1],\ \forall h \in [0, t],$
  $E[|X_{t+h}^p - X_t^p|^2 | X_t^p - X_{t-h}^p|^2] \leqslant Ch^2.$

- Condition (iii)

  $\lim_{\delta \to 0} \sup_{p \geqslant 1} E[|X_\delta^p - X_0^p|]^2 = 0.$

To check these three conditions, the formula of the variance of X_t^p (Eq. (A.2)) is useful. Let us prove (i). $(X_\cdot^p)$ is a $\mathbb{R}$-value process, so (i) is equivalent to

$$\forall \eta > 0,\ \forall t \in \mathscr{Q}_+^*, \exists a_{\eta,t} \in \mathscr{R}^+,\ \inf_{p \geqslant 1} P(X_t^p > a_{\eta,t}) \leqslant \eta.$$

Taking $a_{\eta,t} = \sqrt{\frac{t}{\eta}}$, the above property comes from Tchebychev's inequality.

We now prove (ii). Let $T > 0$ and $0 \leqslant t \leqslant T$:

$$E[|X_{t+h}^p - X_t^p|^2 | X_t^p - X_{t-h}^p|^2]$$
$$= E[|X_{t+h}^p - X_t^p|^2] E[|X_t^p - X_{t-h}^p|^2] = h^2.$$

The second inequality comes from the independence of increments in the compound Poisson process.

Let us prove (iii). We have

$$\lim_{\delta \to 0} \sup_{p \geqslant 1} E[||X_\delta^p - X_0^p|]^2 = \lim_{\delta \to 0} \sup_{p \geqslant 1} \delta = \lim_{\delta \to 0} \delta = 0.$$

References

Barton, N.H., Depaulis, F., Etheridge, A.M., 2002. Neutral evolution in spatially continuous populations. Theor. Popul. Biol. 61, 31–48.

Beaumont, M.A., 2003. Conservation genetics. In: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 779–812.

Beerli, P., 2002. MIGRATE: documentation and program, part of LAMARC. Version 1.5. Revised August 7, 2002. Distributed over the Internet, http://evolution.genetics.washington.edu/lamarc.html

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA 98 (8), 4563–4568.

Bossart, J.L., Prowell, D.P., 1998. Genetic estimates of population structure and gene flow: limitations, lessons, and new directions. Trends Ecol. Evol. 13, 171–212.

Cox, J.T., Durrett, R., 2002. The stepping stone model: new formulas expose old myths. Ann. Appl. Prob. 12, 1348–1377.

Ethier, S.N., Kurtz, T.G., 1986. Markov Processes, Characterization and Convergence. Wiley, New York.

Felsenstein, J., 1975. A pain in the torus: some difficulties with models of isolation by distance. Am. Nat. 109, 359–368.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., Kuhner, M.K., Yamato, J., Beerli, P., 1999. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. Statistics in Molecular Biology, IMS Lecture Notes-Monograph Series, Vol. 33, pp. 163–185.

Fu, Y.-X., Li, W.-H., 1993. Maximum likelihood estimation of population parameters. Genetics 134, 1261–1270.

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L., Feldman, M.W., 1995. An evaluation of genetic distances for use microsatellite loci. Genetics 139, 463–471.

Griffiths, R.C., Tavaré, S., 1994. Ancestral inference in population genetics. Statist. Sci. 9, 307–319.

Kimura, M., 1953. Stepping-stone model of population. Ann. Rep. Natl. Inst. Genet., Japan 3, 62–63.

Kimura, M., Otha, T., 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. 2, 87–90.

Kingman, J.F.C., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Klein, E.K., Austerlitz, F., Larédo, C., 1999. Some statistical improvements for estimating population size and mutation rate from segregating sites in DNA sequences. Theor. Popul. Biol. 55, 235–247.

Kuhner, M.K., Yamato, J., Felsenstein, J., 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140, 1421–1430.

Malécot, G., 1967. Identical loci and relationship. Proceedings of the Fifth Berkeley Symposium on Math. Stat. Prob. Vol. 4, University of California Press, Berkeley, pp. 317–332.

Nagylaki, T., 2002. When and where was the most recent common ancestor? J. Math. Biol. 44, 253–275.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

O'Ryan, C., Bruford, M., Beaumont, W.M., Wayne, R.K., Cherry, M.I., Harley, E.H., 1998. Genetics of fragmented populations of African buffalo (Syncerus caffer) in South Africa. Anim. Conservat. 1, 85–94.

Paetkau, D., Waits, L.P., Craighead, L., Clarkson, P., Strobeck, C., 1998. Dramatic variation in genetic diversity across the range of North America brown bears. Conservat. Biol. 12, 418–429.

Pritchard, J.K., Feldman, M.W., 1996. Statistics for microsatellite variation based on coalescence. Theor. Popul. Biol. 50, 325–344.

Raymond, M., Rousset, F., 1995. GENEPOP: a population genetics software for exact tests and ecumenicism. J. Hered. 86, 248–249.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods, Springer Series in Statistics. Springer, New York.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow using *F*-statistics under isolation by distance. Genetics 145, 1219–1228.

Rousset, F., 2003. Inferences from spatial population genetics. In: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 239–270.

Seo, T.K., Thorne, J.L., Hasegawa, M., Kishino, H., 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics 160, 1283–1293.

Slatkin, M., 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47, 264–279.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139, 457–462.

Stephens, M., 2003. Inference under the coalescent. In: Balding, D.J., Bishop, M.J., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 213–238.

Taberlet, P., Bouvet, J., 1994. Mitochondrial DNA polymorphism, phylogeography, and conservation genetics of the brown bear (Ursus arctos) in Europe. Proc. Roy. Soc. London B Biol. 255, 195–200.

Tavaré, S., 2001. Lecture Notes St Flour. Springer, Berlin.

Waits, L., Taberlet, P., Swenson, J.E., Sandegren, F., Franzen, R., 2000. Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (Ursus arctos). Mol. Evol. 9, 610–621.

Wilson, I.J., Balding, D.J., 1998. Genealogical inference from microsatellite data. Genetics 150, 499–510.

Wright, S., 1943. Isolation by distance. Genetics 28, 114–138.