THEORETICAL POPULATION BIOLOGY 15, 108–113 (1979)

Sojourn Times for Conditioned Markov Chains in Genetics

S. Tavaré

Department of Probability and Statistics, University of Sheffield, Sheffield, England

Received October 19, 1977

The use of diffusion methods to predict the distribution of the number of visits to a particular gene frequency in a class of finite population models is discussed. The models for which such approximations are accurate are found, and several results unified by this approach. Some difficulties in the application of diffusion methods to such sojourn times are highlighted.

1. Introduction

It is often the case in population genetics that diffusion approximations are used to model many properties of Markov chains which describe the evolution of the genetic composition of a population. It is therefore of interest to discover how accurate such approximations are.

Specifically, consider a single locus in a population of fixed size M, at which there are two possible alleles, denoted A and B. $\{X_n, n \geqslant 0\}$ is a Markov chain which gives the number of A alleles in the population. If there are no mutation pressures, then states 0 and M are absorbing. Recently, Ewens (1973) introduced the method of conditioned diffusion equations to study approximations to Markov chains $\{X_n^*\}$ which are derived by conditioning the X process on eventual fixation ($X = M$). Pollak and Arnold (1975) (henceforth referred to as PA) used diffusion methods to approximate several properties of the sojourn time distributions of the conditioned version of the Wright model. In this paper we look at a class of processes, discussed by Cannings (1974), which includes the Wright model. These processes have the same diffusion approximation. We determine for which models in this class the diffusion process predicts accurately the behavior of the sojourn time distributions. This unifies and extends the work of PA, and Maruyama (1972, 1973).

2. The Markov Chains, and Approximation to Mean Sojourn Times

Let $\mathscr{C}$ be the class of Markov chains X, with state space $\{0, 1,..., M\}$, whose transition matrix $P = (p_{ij})$ satisfies

$$\pi_i = iM^{-1} \tag{1}$$

$$p_{ij} = p_{M-i,M-j} \tag{2}$$

0040-5809/79/010108-06$02.00/0

where π_i is the probability that X is absorbed at M, given $X_0 = i$, and which, after suitable time and state space scaling we may approximate by a diffusion $X(t)$, with state space $[0, 1]$, which has drift coefficient $m(x)$ and diffusion coefficient $v(x)$ given by

$$m(x) \equiv 0, \qquad v(x) = x(1-x). \tag{3}$$

Condition (2) is a consequence of the exchangeability approach to a wide variety of Markov chains of use in genetics (Cannings (1974)). Two chains in $\mathscr{C}$ are the Wright model, for which

$$p_{ij} = \binom{M}{j}\left(\frac{i}{M}\right)^j\left(1-\frac{i}{M}\right)^{M-j} \tag{4}$$

and the Moran model (e.g., Moran (1962)), for which

$$\begin{aligned} p_{ij} &= \frac{i(M-i)}{M^2} = p_i\,; && j = i-1, i+1 \\ &= 1 - 2p_i\,; && j = i \\ &= 0; && |i-j| > 1. \end{aligned} \tag{5}$$

For $1 \leqslant i,\ j \leqslant M-1$, let T_{ij} be the number of visits to j, given $X_0 = i$, and let $n_{ij} = ET_{ij}$, and $N = (n_{ij})$. Denote by T_{ij}^*, n_{ij}^*, N^*, p_{ij}^*, etc. the corresponding conditioned quantities. From (3), and Ewens (1973), the elements of n_{ij}^* are approximated by (rescaled versions of)

$$\begin{aligned} t_p^*(x) &= \frac{2x(1-p)}{p(1-x)} && 0 < x \leqslant p \\ &= 2 && p \leqslant x < 1, \end{aligned} \tag{6}$$

where $t_p^*(x)$ is the mean of the time $T_p^*(x)$ spent at x for the conditioned diffusion process with $X^*(0) = p$. Equation (6) suggests that, for some constant $c > 0$, we should have

$$n_{ij}^* = c \qquad 1 \leqslant i \leqslant j \leqslant M-1. \tag{7}$$

The approximation holds on the diagonal by continuity of $t_p^*(x)$ at p. The constant c depends on the time and state-space scaling used to derive (3). For example, Wright's model gives $c = 2$, Moran's model $c = M$. The distribution of T_{ij}^* is determined by N^* (cf., PA). We therefore want to know which chains in $\mathscr{C}$ have n_{ij}^* given by (7).

Let $Q^* = (q_{ij}^*)$ be the matrix of transition probabilities between the transient states $\{1,\ldots, M-1\}$.

Then

$$N^* = (I - Q^*)^{-1}$$

and

$$|N^*| = c(c - n_{21}^*) \cdots (c - n_{M-1,M-2}^*) \neq 0. \tag{8}$$

THEOREM 1. *Let $\{X_n\}$ be any chain in $\mathscr{C}$, for which N^* satisfies* (7). *Then*

(i) $$p_{jM}^* = 0, \qquad j = 1, 2, \ldots, M-2$$
$$p_{M-1,M}^* = 1/c$$

(ii) $$q_{ij}^* = 0, \qquad 1 \leqslant i \leqslant j - 2 \leqslant M - 3.$$

Proof. Since $(I - Q^*) N^* = I$, premultiplying the last column of N^* by $(I - Q^*)$ gives

$$c - c\left(\sum_{k=1}^{M-1} q_{jk}^*\right) = \delta_{j,M-1}$$

where $\delta_{ij} = 0$, $i \neq j$; $\delta_{ij} = 1$, $i = j$. It follows that $c(1 - \sum_{k=1}^{M-1} q_{jk}^*) = \delta_{j,M-1}$. Hence $p_{jM}^* = 0$, $j = 1, \ldots, M-2$, and $p_{M-1,M}^* = 1/c$, so (i) is proved. Next, premultiplying the $(M-2)$th column of N^* by $(I - Q^*)$ gives, for $1 \leqslant i \leqslant M-2$,

$$c\left(1 - \sum_{k=1}^{M-2} q_{ik}^*\right) - q_{i,M-1}^* n_{M-1,M-2}^* = \delta_{i,M-2}$$

so that

$$c(q_{i,M-1}^* + p_{i,M}^*) - q_{i,M-1}^* n_{M-1,n-2}^* = \delta_{i,M-2}\,.$$

Rearranging and using (i) gives

$$q_{i,M-1}^*(c - n_{M-1,M-2}^*) = \delta_{i,M-2}$$

so that, from (8), $q_{i,M-1}^* = 0$, $1 \leqslant i \leqslant M-3$. Continuing in this way completes the proof.

We now use conditions (1) and (2) to reconstruct P from the form of P^* given by Theorem 1. It is, of course, possible to construct other matrices P^* which satisfy (7). However, if we do not know the absorption probabilities of the unconditioned chain, we cannot reconstruct P, and so deduce the form of the underlying process $\{X_n\}$. However, (1) and (2) yield

COROLLARY 1. *If $\{X_n\} \in \mathscr{C}$, and N^* satisfies* (7), *then P must be tridiagonal. Hence P is the transition matrix of a* (*possibly state dependent*) *random walk, which, by* (1), *must be symmetric.*

The symmetric random walk has transition matrix determined by

$$\begin{aligned} p_{ij} &= 0; \qquad |i-j| > 1 \\ p_{i,i+1} &= p_{i,i-1} = p_i\,, \qquad 0 < p_i \leqslant \tfrac{1}{2} \\ p_{i,i} &= 1 - 2p_i\,. \end{aligned} \tag{9}$$

LEMMA 2. *For a random walk $\{X_n\}$ with absorbing barriers at* 0 *and M, and transition matrix as* (9),

$$\begin{aligned} n_{ij} &= \frac{(M-i)j}{Mp_j} \qquad 1 \leqslant j \leqslant i \\ &= \frac{i(M-j)}{Mp_j} \qquad i \leqslant j \leqslant M-1. \end{aligned}$$

Proof. This is a special case of a result of Ewens (1964, p. 146).

Lemma 2 enables us to find precisely those $\{X_n\}$ in $\mathscr{C}$ for which (7) holds. We know from Corollary 1 that such a model must be a random walk. From the lemma, equation (1), and Ewens (1973, p. 23, Eq. (6)) we find that the conditioned random walk satisfies

$$\begin{aligned} n_{ij}^* &= \frac{(M-i)j^2}{Mip_j} \qquad 1 \leqslant j \leqslant i \\ &= \frac{j(M-j)}{Mp_j} \qquad i \leqslant j \leqslant M-1 \end{aligned} \tag{10}$$

so that if (7) holds, we must have $p_j = j(M-j)/Mc$, and so we have proved

THEOREM 3. *Let $\{X_n\} \in \mathscr{C}$. Then N^* satisfies* (7) *if, and only if, $\{X_n\}$ is a symmetric random walk, for which $p_j = j(M-j)/Mc$.*

In particular, Moran's model is such a random walk, and so satisfies (7) (a result which can be verified from (10)). However, Wright's model is not, and so we cannot expect the diffusion approximation to give the correct form for the elements of N^*. It follows that the diffusion process which approximates this class of Markov chains is really a continuous state-space version of Moran's model, and so diffusion approximations to sojourn time distributions are just (rescaled) multiples of the known results for the Moran model. In particular, for other models in the class $\mathscr{C}$, it is necessary to check the adequacy of the diffusion approximation to N^* in each particular case. PA have obtained some improved approximations for the case of the Wright model. They notice that the diagonal elements n_{jj}^* as predicted by the diffusion result are too small. The reason for this is not that the initial position, $X_0 = j$, is not counted when computing n_{jj}^*, but that the diffusion process is only "recreating" Moran's model.

3. Approximation to Higher Moments of T_{ij}^*.

It is well known from the theory of Markov chains (cf., PA) that the variance of T_{ij}^* is given by

$$\mathrm{Var}(T_{ij}^*) = n_{ij}^*(2n_{jj}^* - n_{ij}^* - 1) \tag{11}$$

while for the diffusion process, Nagylaki (1974) has shown that

$$\mathrm{Var}(T_p^*(x)) = t_p^*(x)(2t_x^*(x) - t_p^*(x)). \tag{12}$$

It is apparent from (11) and (12) that we have two methods for estimating $\mathrm{Var}(T_{ij}^*)$ from the diffusion approximation. Method 1 is to rescale $t_p^*(x)$ to approximate n_{ij}^*, and then use (12) as an approximation to $\mathrm{Var}(T_{ij}^*)$. For Wright's model, we obtain

$$\mathrm{Var}(T_{ij}^*) = 4, \qquad i \leqslant j. \tag{13}$$

This is essentially the method used by Maruyama (1972), and (13) agrees with his result. However, another method is to use the approximation to n_{ij}^*, and apply (11). Wright's model yields

$$\mathrm{Var}(T_{ij}^*) = 2, \qquad i \leqslant j. \tag{14}$$

Applied to Moran's model, we get

$$\mathrm{Var}(T_{ij}^*) = M^2, \qquad i \leqslant j \tag{15}$$

$$\mathrm{Var}(T_{ij}^*) = M^2 - M, \qquad i \leqslant j \tag{16}$$

for the first and second methods respectively. We notice that (15) and (16) are asymptotically equivalent, but that (13) and (14) are not. The reason for this discrepancy is accounted for by Theorem 3. We cannot expect asymptotic equivalence because the approximation really only applies to Moran's model. We note that (16) is the exact result for Moran's model. The approximations to $\mathrm{Var}(T_{jj}^*)$ for Wright's model which are derived in PA use the second method, but differ from (14) because they use an improved approximation to the diagonal elements, n_{jj}^*.

Conclusions

We have shown that certain results derived from the diffusion equation approach are continuous analogues of explicitly known properties of Moran's model. Largely for reasons of mathematical tractability, it is usual to use diffusion approximations to processes which are essentially discrete in nature.

The discrete models discussed in the introduction differ considerably in the probabilistic mechanism involved in the reproduction of individuals. However, all models in this class are approximated by the same diffusion process. This arises because the parameters that determine the diffusion are based only on the mean and variance of local changes in gene frequency, so that the class looks, locally at least, the same. Thus the net effect of the approximation is to remove any differences that we have modelled in reproductive strategy. Next, using the diffusion process, we try to recreate properties of the underlying models; the problem that now arises is the assessment of how well we can do this. In the context of sojourn time distributions, the results of this note show that the diffusion process accurately models only one chain in the class, and furthermore that results for this model are know explicitly (so there is no need for approximation anyway). Since we want to use the diffusion approximation to give qualitative and quantitative estimates for the other models, it is apparent that we must check their adequacy in each case. It is usually the case that such approximations are qualitatively reasonable, but quantitatively they may be quite inaccurate.

Acknowledgments

The author would like to thank Dr. C. Cannings for discussions on this subject, and the referee, whose comments have greatly improved the clarity of the paper.

References

Cannings, C. 1974. The latent roots of certain Markov chains arising in genetics. A new approach. I. Haploid models, *Advan. Appl. Prob.* **6**, 260–290.

Ewens, W. J. 1964. The pseudo-transient distribution and its uses in genetics, *J. Appl. Prob.* **1**, 141–156.

Ewens, W. J. 1973. Conditional diffusion processes in population genetics, *Theor. Pop. Biol.* **4**, 21–30.

Kemeny, J. G., and Snell, J. L. 1960. "Finite Markov Chains," Van Nostrand, New York.

Maruyama, T. 1973. The variance of the number of loci having a given gene frequency, *Genetics* **73**, 361–366.

Maruyama, T. 1972. The average number and variance of number of generations at a particular gene frequency in the course of fixation of a mutant gene in a finite population, *Genet. Res. Camb.* **19**, 109–113.

Moran, P. A. P. 1962. "The Statistical Processes of Evolutionary Theory," Oxford Univ. Press (Clarendon), London.

Nagylaki, T. 1974. The moments of stochastic integrals and the distribution of sojourn times, *Proc. Nat. Acad. Sci.* **71**, 746–749.

Pollak, E., and Arnold, B. C. 1975. On the sojourn times at particular gene frequencies, *Genet. Res. Camb.* **25**, 89–94.

Tavaré, S. 1976. Sojourn times for conditioned Markov chains in genetics (abstract), *Advan. Appl. Prob.* **8**, 645–648.