Chapter 4
More than Two Types

4.1 Multi-Dimensional Diffusion Models

So far we have considered only a very special case in which our population is classified into just two types. The frequencies are then characterised by a one-dimensional diffusion and one dimensional diffusions are, at least in principle, relatively straightforward to study. More generally, suppose that our population is classified into K different types. We're not going to develop the general theory of multidimensional diffusions, but let's see what happens in a special case.

Our starting point is a K-allele version of the Wright–Fisher model. The population configuration at any time can be described by a vector $(X_1, X_2, \ldots, X_K)$ where X_i is the number of genes of allelic type A_i and we assume that $X_1 + \cdots + X_K = N$. (Although only $K-1$ components are necessary to specify the vector $(X_1, X_2, \ldots, X_K)$, it is sometimes convenient to retain all K.)

In the simplest case when all the alleles are selectively neutral and there is no mutation, we have

$$\begin{aligned}\mathbb{P}[Y_i \text{ genes of type } A_i \text{ at } t+1, i=1,\ldots,K | X_j \text{ genes of type } A_j \text{ at } t,\, j=1,\ldots,K]\\ = \frac{N!}{Y_1!Y_2!\cdots Y_K!}\psi_1^{Y_1}\psi_2^{Y_2}\cdots\psi_K^{Y_K}\end{aligned}$$

where $\psi_i = \frac{X_i}{N}$ and $\sum_{i=1}^K Y_i = N$ (the probability is zero if this last condition is not satisfied).

We write $p_i(t) = X_i(t)/N$ and consider the increment $\delta p_i = p_i(t+1) - p_i(t)$. By 'pooling' all the alleles A_j for $j \neq i$ into a single class ('not A_i'), we recover the Wright–Fisher model for two alleles for which we already checked that

$$\mathbb{E}[\delta p_i] = 0, \quad \texttt{var}(\delta p_i) = \frac{1}{N}p_i(1-p_i) \quad \text{and} \quad \mathbb{E}[(\delta p_i)^k] = \mathscr{O}(\frac{1}{N^2}) \quad \forall k \geq 3.$$

To complete the picture we need the *covariances*, that is we must calculate, for $i \neq j$,

$$\begin{aligned}\mathbb{E}[\delta p_i \delta p_j] &= \frac{1}{N^2}\mathbb{E}\left[(X_i(t+1)-X_i(t))(X_j(t+1)-X_j(t))\middle|\mathscr{F}_t\right]\\ &= \frac{1}{N^2}\mathbb{E}\left[X_i(t+1)X_j(t+1)\middle|\mathscr{F}_t\right] - p_i(t)p_j(t).\end{aligned} \tag{4.1}$$

A. Etheridge, *Some Mathematical Models from Population Genetics*, Lecture Notes in Mathematics 2012, DOI 10.1007/978-3-642-16632-7_4,

Here $\mathscr{F}_t$ denotes all information about the process up until time t (more formally, it is the natural σ-field associated with $\{X_s\}_{s\leq t}$). In the calculations below we write $\mathbb{E}_t$ for the corresponding conditional expectation. Now

$$\begin{aligned}&\mathbb{E}_t[X_i(t+1)X_j(t+1)]\\&\qquad=\frac{1}{2}\left\{\mathbb{E}_t[(X_i(t+1)+X_j(t+1))^2]-\mathbb{E}_t[X_i(t+1)^2]-\mathbb{E}_t[X_j(t+1)^2]\right\}\end{aligned}$$

and again ‘pooling’ genes of type A_i and A_j we reduce to a calculation for the binomial distribution, $Bin(N,p)$, for which we know

$$\mathbb{E}[X^2]=Np(1-p)+N^2p^2.$$

This gives

$$\begin{aligned}\mathbb{E}_t[X_i(t+1)X_j(t+1)]&=\frac{1}{2}\left\{N(p_i+p_j)(1-(p_i+p_j))+N^2(p_i+p_j)^2\right.\\&\qquad\left.-Np_i(1-p_i)-N^2p_i^2-Np_j(1-p_j)-N^2p_j^2\right\}\\&=-Np_ip_j+N^2p_ip_j\end{aligned}$$

and substituting in (4.1) we obtain

$$\mathbb{E}[\delta p_i\delta p_j]=-\frac{1}{N}p_ip_j.$$

Now, just as in the two allele case, if we consider functions $f(p_1,\ldots,p_{K-1})$ (note that $p_K=1-\sum_{j=1}^{K-1}p_j$) and rescale time so that the inter-generation time is $1/N$ and let $N\to\infty$, then we can use Taylor's Theorem to identify the generator of the limiting model. This gives

$$\mathscr{L}f(p_1,\ldots,p_{K-1})=\frac{1}{2}\sum_{i=1}^{K-1}p_i(1-p_i)\frac{\partial^2 f}{\partial p_i^2}-\sum_{1\leq i<j\leq K-1}p_ip_j\frac{\partial^2 f}{\partial p_i\partial p_j}.$$

We can also add a mutation step. We did not do this explicitly in the two-allele Wright–Fisher model so let's be more explicit here. The idea is that for each offspring in each generation there is a small probability that it will not inherit the type of its parent, but rather it will mutate to another type. Suppose that with probability u_{ij} the offspring of a type A_i individual will (independently of one another) be type A_j. We will now have

$$\mathbb{E}[\delta p_i]=-\sum_{j\neq i}u_{ij}p_i+\sum_{j\neq i}u_{ji}p_j.$$

If we assume, as we did before, that mutation rates are very low (on the order of inverse population size), then writing $\beta_{ij}=Nu_{ij}$ we have

$$\mathbb{E}[\delta p_i] = \frac{1}{N}\left\{-p_i\sum_j \beta_{ij} + \sum_j p_j\beta_{ji}\right\}.$$

The correction to $\mathbb{E}[(\delta p_i)^2]$ is of $\mathscr{O}(1/N^2)$. This gives the following lemma.

Lemma 4.1 (Multi-allele Wright–Fisher diffusion with mutation). *The generator of the K-allele Wright–Fisher diffusion with mutation is*

$$\frac{1}{2}\sum_{i=1}^{K-1} p_i(1-p_i)\frac{\partial^2 f}{\partial p_i^2} - \sum_{1\le i<j\le K-1} p_ip_j\frac{\partial^2 f}{\partial p_i\partial p_j} + \sum_{i=1}^{K-1}\left(-p_i\sum_j\beta_{ij} + \sum_j p_j\beta_{ji}\right)\frac{\partial f}{\partial p_i}.$$

If each $u_{ij} > 0$ for $i \neq j$ then the joint frequency of $A_1,\ldots,A_{K-1}$ has a stationary distribution, but in general no closed form is known. It *is* known in the special case of symmetric parent-independent mutation.

Lemma 4.2. *Suppose that*

$$u_{ij} = \frac{u}{K-1}$$

so that the total mutation probability per gene per generation is u and it is equally likely to be a mutation to each of the other types. Then the corresponding K-allele Wright–Fisher diffusion with mutation has a stationary distribution with density

$$\psi(p_1,\ldots,p_{K-1}) = \frac{\Gamma(K\varepsilon)}{(\Gamma(\varepsilon))^K}(p_1\cdots p_K)^{\varepsilon-1} \tag{4.2}$$

where $\varepsilon = \frac{2Nu}{K-1}$ *and* $p_K = 1 - p_1 - \ldots - p_{K-1}$.

Proof. First note that in this case

$$-p_i\sum_j\beta_{ij} + \sum_j p_j\beta_{ji} = \frac{Nu}{K-1}(1-Kp_i).$$

Writing $\psi(p_1,\ldots,p_{K-1})$ for the density of the stationary distribution and integrating by parts, exactly as we did to obtain (3.13) for the two-allele case, we find

$$\frac{1}{2}\sum_{i=1}^{K-1}\frac{\partial^2}{\partial p_i^2}(p_i(1-p_i)\psi(p_1,\ldots p_{K-1})) - \sum_{1\le i<j\le K-1}\frac{\partial^2}{\partial p_i\partial p_j}(p_ip_j\psi(p_1,\ldots,p_{K-1}))$$
$$-\sum_{i=1}^{K-1}\frac{\partial}{\partial p_i}\left(\frac{Nu}{K-1}(1-Kp_i)\psi(p_1,\ldots,p_{K-1})\right)=0. \tag{4.3}$$

It is elementary (if tedious) to check that the expression in (4.2) solves this equation. □

Notice that when $K = 2$, (4.3) becomes

$$0 = -\frac{1}{2}\frac{d}{dp}\left(\mu(1-2p)f(p)\right) + \frac{1}{2}\frac{d^2}{dp^2}\left(p(1-p)f(p)\right)$$

where $\mu = 2Nu$ and (4.2) becomes

$$f(p) = \frac{\Gamma(2\mu)}{(\Gamma(\mu))^2}(p(1-p))^{\mu-1}$$

which is precisely the solution we found in (3.15), since in this notation $2\nu_1 = 2\nu_2 = \mu$.

In the 2-allele case we calculated the heterozygosity

$$H = \frac{4\nu_1\nu_2}{(\nu_1+\nu_2)(2(\nu_1+\nu_2)+1)}.$$

Substituting $\nu_1 = \nu_2 = Nu$ gives

$$H = \frac{2Nu}{4Nu+1}.$$

Writing $\theta = 2Nu$ this gives

$$H = \frac{\theta}{2\theta+1}.$$

Remark 4.3. This is the magic θ of Remark 2.18 but with 2 in place of 4 here because we have taken limits in a haploid population. To recover our previous θ we set $N = 2N_e$.

The expected *homozygosity*, F, which is the chance that a random sample of two genes is of the same allelic type is

$$F = 1 - H = \frac{\theta+1}{2\theta+1}.$$

For the K-allele model,

$$\begin{aligned} F &= \sum_{i=1}^{K}\mathbb{E}[p_i^2] = K\int_0^1 p^2 p^{\varepsilon-1}(1-p)^{(K-1)\varepsilon-1}\frac{\Gamma(K\varepsilon)}{\Gamma(\varepsilon)\Gamma((K-1)\varepsilon)}dp \\ &= K\frac{\Gamma((K-1)\varepsilon)\Gamma(\varepsilon+2)}{\Gamma(K\varepsilon+2)}\frac{\Gamma(K\varepsilon)}{\Gamma(\varepsilon)\Gamma((K-1)\varepsilon)} \\ &= \frac{K\Gamma(\varepsilon+2)}{\Gamma(\varepsilon)}\frac{\Gamma(K\varepsilon)}{\Gamma(K\varepsilon+2)} = \frac{\varepsilon+1}{K\varepsilon+1} \end{aligned}$$

and substituting $\varepsilon = \theta/(K-1)$ gives

$$F = \frac{\theta + K - 1}{K\theta + K - 1}. \tag{4.4}$$

Definition 4.4. The density (4.2) is called the *Dirichlet distribution*. It is usual to rearrange it and consider the sequence of allele frequencies in *decreasing* order:

$$p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(K)} \geq 0,$$

that is we look at the order statistics of $p_1, \ldots, p_K$. Their joint distribution is

$$f(p_{(1)}, \ldots, p_{(K)}) = \frac{K!\Gamma(K\varepsilon)}{\Gamma(\varepsilon)^K} \left(p_{(1)} \cdots p_{(K)}\right)^{\varepsilon - 1}.$$

Recall that the mutation model that led to this distribution was the symmetric parent-independent mutation model in which each individual mutates at the same rate to a type chosen uniformly from $\{1, \ldots, K\}$. If $K \to \infty$ this becomes the *infinitely many alleles model* (Definition 2.17). It is natural to ask whether we can find an analogue of the Dirichlet distribution for the stationary distribution of allele frequencies in the infinitely many alleles model. The answer, due to Kingman (1975, 1977), is yes.

4.2 The Poisson–Dirichlet and GEM Distributions

Kingman showed that, for every $j \geq 1$, the distribution of the first j order statistics of the Dirichlet distribution converges as $K \to \infty$ and called the corresponding limiting distribution the *Poisson–Dirichlet* distribution. In this section we shall try to understand why such a limit should exist. Direct manipulation of the Dirichlet distribution is difficult because of the linear dependence between the variables. However, it turns out that it can be represented in terms of *independent* Γ-random variables as follows.

Lemma 4.5. *Let $Y_1, \ldots, Y_K$ be independent positive random variables with probability density function*

$$g_\varepsilon(y) = \frac{y^{\varepsilon - 1} e^{-y}}{\Gamma(\varepsilon)}.$$

Then writing $Y = Y_1 + \cdots + Y_K$, the vector $\mathbf{p}$ *with components $p_i = \frac{Y_i}{Y}$ has the Dirichlet distribution and Y has a Γ-distribution with parameter $K\varepsilon$. Moreover,* $\mathbf{p}$ *is* independent *of Y.*

Proof. The proof of this claim is a simple change of variables,

$$(y_1, \ldots, y_K) \mapsto (p_1, \ldots, p_{K-1}, y).$$

In an obvious notation, $y_i = p_i y$ (and $p_K = 1 - \sum_{i=1}^{K-1} p_i$). Since the Y_i are independent,

$$\begin{aligned}
f_{(p_1,\dots,p_{K-1},Y)}(p_1,\dots,p_{K-1},y) &= f_{(Y_1,\dots,Y_K)}(p_1 y,\dots,p_k y)\left|\frac{\partial(y_1,\dots,y_k)}{\partial(p_1,\dots,p_{K-1},y)}\right| \\
&= \prod_{i=1}^{K}\frac{(p_i y)^{\varepsilon-1}e^{-p_i y}}{\Gamma(\varepsilon)}\begin{vmatrix} y & & \cdots & & p_1 \\ & y & & & p_2 \\ & & \ddots & & \\ & & & y & \\ -y & -y & \cdots & -y & p_K \end{vmatrix} \\
&= \frac{1}{\Gamma(\varepsilon)^K}(p_1\dots p_K)^{\varepsilon-1}y^{K\varepsilon-K}e^{-y}\times y^{K-1}\sum_{i=1}^{K}p_i \\
&= \frac{\Gamma(K\varepsilon)}{\Gamma(\varepsilon)^K}(p_1\cdots p_K)^{\varepsilon-1}\frac{1}{\Gamma(K\varepsilon)}e^{-y}y^{K\varepsilon-1}
\end{aligned}$$

as required. □

We now use this to find a representation of the Poisson–Dirichlet distribution (and in the process see why the name is natural). To do so we need a spatial analogue of the probability generating function of elementary probability.

Definition 4.6 (Probability generating functional). For a (possibly random) number of random points $\{Y_i\}_{i\in I}$ with each $Y_i \in (0,\infty)$ (say) we define the *probability generating functional* of $\{Y_i\}_{i\in I}$ by

$$G(\xi) = \mathbb{E}\Big[\prod_{i\in I}\xi(Y_i)\Big]$$

for any function $\xi : [0,\infty) \to \mathbb{R}$ for which the expectation exists.

If I is random, then we recover the probability generating function of I by choosing ξ to be constant.

Now choose the Y_i's to be independent Gamma random variables with parameter ε and consider the generating functional of $Y_1,\dots,Y_K$. By independence,

$$G_K(\xi) = \left[\int_0^\infty \xi(u)\frac{u^{\varepsilon-1}e^{-u}}{\Gamma(\varepsilon)}du\right]^K.$$

Recall from (4.2) that $\varepsilon = 2Nu/(K-1)$ and $\theta = 2Nu$ so that $K\varepsilon \to \theta$ as $K \to \infty$. Now rewrite the term in square brackets using

$$\int_0^\infty \frac{u^{\varepsilon-1}e^{-u}}{\Gamma(\varepsilon)}du = 1 \quad \text{and} \quad \frac{\varepsilon}{\Gamma(\varepsilon+1)} = \frac{1}{\Gamma(\varepsilon)}$$

to obtain that

$$
\begin{aligned}
G_K(\varepsilon) &= \left[1-\varepsilon\int_0^\infty (1-\xi(u))\frac{u^{\varepsilon-1}}{\Gamma(\varepsilon+1)}e^{-u}du\right]^K \\
&\rightarrow \exp\left(-\theta\int_0^\infty (1-\xi(u))u^{-1}e^{-u}du\right) \quad \text{as } K\rightarrow\infty.
\end{aligned}
$$

The right hand side is the probability generating functional of a Poisson point process with intensity $\theta e^{-u}/u$, so in the limit as $K\rightarrow\infty$ the number of points in each interval $(a,b)\subseteq(0,\infty)$ is Poisson distributed with mean $\int_a^b(\theta e^{-u}/u)du$.

Now write $Y_{(1)}\geq Y_{(2)}\geq\cdots$ for the *ordered* points and $Y=Y_{(1)}+Y_{(2)}+\cdots$. Since $K\varepsilon\rightarrow\theta$ as $K\rightarrow\infty$, Y has a Gamma distribution with parameter θ.

Definition 4.7. The points $p_{(i)}=Y_{(i)}/Y$ have the *Poisson–Dirichlet* distribution.

The finite dimensional distributions of the $p_{(i)}$ are complicated, but those of the $Y_{(i)}$ are relatively straightforward. The density function of $Y_{(i)}$ is

$$
\frac{\theta e^{-y}}{y}\frac{[\theta E_1(y)]^{i-1}}{(i-1)!}e^{-\theta E_1(y)}, \quad \text{for } y>0,
$$

where $E_1(y)=\int_y^\infty(e^{-u}/u)du$. Thus, for example, since $p_{(i)}$ is independent of Y (just as in Lemma 4.5)

$$
\mathbb{E}[Y_{(i)}]=\mathbb{E}[p_{(i)}Y]=\mathbb{E}[p_{(i)}]\mathbb{E}[Y]=\theta\mathbb{E}[p_{(i)}]
$$

gives

$$
\mathbb{E}[p_{(i)}]=\frac{\theta^{i-1}}{(i-1)!}\int_0^\infty e^{-y}[E_1(y)]^{i-1}e^{-\theta E_1(y)}dy
$$

which can be evaluated numerically.

In the Dirichlet distribution with K allelic types, the probability that there are alleles with frequencies in $(p_1,p_1+dp_1),\ldots,(p_r,p_r+dp_r)$ is

$$
\begin{aligned}
&\tbinom{K}{r}\tfrac{\Gamma(K\varepsilon)}{\Gamma(\varepsilon)^r\Gamma((K-r)\varepsilon)}(p_1\cdots p_r)^{\varepsilon-1}\left(1-\textstyle\sum_1^r p_i\right)^{\varepsilon(K-r)-1}dp_1\ldots dp_r \\
&\rightarrow \theta^r(p_1\cdots p_r)^{-1}\left(1-\textstyle\sum_1^r p_i\right)^{\theta-1}dp_1\ldots dp_r \quad \text{as } K\rightarrow\infty. \qquad (4.5)
\end{aligned}
$$

In particular, taking $r=1$, the probability that there is an allele with frequency in $(p,p+dp)$ under the limiting Poisson–Dirichlet distribution is $h(p)dp$ where

$$
h(p)=\theta p^{-1}(1-p)^{\theta-1}.
$$

Definition 4.8. The function $h(p)=\theta p^{-1}(1-p)^{\theta-1}$ is called the *frequency spectrum* of $\{p_{(i)}\}$.

The frequency spectrum allows us to calculate expressions of the form

$$\mathbb{E}\left[\sum_{1}^{\infty} f(p_{(i)})\right] = \int_0^1 f(p)h(p)dp$$

(provided this is finite). For example, taking $f(p_{(i)}) = p_{(i)}^2$ we calculate the *expected homozygosity*

$$F = \int_0^1 p^2 \theta p^{-1}(1-p)^{\theta-1} dp = \frac{1}{1+\theta}.$$

This is consistent with (4.4) as $K \longrightarrow \infty$. Similarly, the expected number of alleles with frequencies in (a,b) is

$$\mathbb{E}\left[\sum_{1}^{\infty} \mathbf{1}_{(a,b)}(p_{(i)})\right] = \int_a^b \theta p^{-1}(1-p)^{\theta-1} dp,$$

and so on. This is the same θ cropping up again and again in our calculations.

The Poisson–Dirichlet distribution is not all that user-friendly, but remarkably a distribution obtained from it by ‘size-biased’ sampling is extremely elegant.

Example 4.9. Suppose that a gene is sampled at random from the population. What is the distribution of the frequency of alleles of the same type as the sampled individual?

The probability that the sampled allele has frequency in $[p, p+dp)$ is the probability that there *is* an allele with frequency in $[p, p+dp)$ *and* we choose it which is just

$$p \cdot \frac{\theta}{p}(1-p)^{\theta-1} dp = \theta(1-p)^{\theta-1} dp.$$

Now suppose that we remove all the individuals of this type and sample again from the remaining population. The chance that our new allele is at *relative* frequency r is just calculated by Bayes’ rule. Let us write P_1 for the frequency of the first individual sampled and P_2 for the relative frequency of the second class. Then

$$\begin{aligned}
&\mathbb{P}\left[P_2 \in [r, r+dr) | P_1 \in [p, p+dp)\right] \\
&\quad = \frac{\mathbb{P}[\exists \text{ class with rel. freq} \in [r, r+dr) \text{ and sample from it and } P_1 \in [p, p+dp)]}{\mathbb{P}[P_1 \in [p, p+dp)]} \\
&\quad = \frac{\frac{\theta^2}{r(1-p)p}(1-p-r(1-p))^{\theta-1} pr\text{“}d(r(1-p))\text{”}dp}{\theta(1-p)^{\theta-1}dp} = \theta(1-r)^{\theta-1}dr,
\end{aligned}$$

where we have used (4.5) in the last line. In other words, P_2 has the same distribution as P_1. We can repeat this procedure and we find that the frequencies of the alleles picked in this way in our original population are

$$P_1, P_2(1-P_1), P_3(1-P_2)(1-P_1), \dots \tag{4.6}$$

where the P_i are independent identically distributed random variables with density $\theta(1-p)^{\theta-1}$, $0 < p < 1$.

Definition 4.10 (GEM distribution). The sequence of random variables in (4.6) are said to follow the *GEM distribution* after Griffiths, Engen and McCloskey (see Ewens (2004)).

All alleles in the population are eventually lost by the joint process of mutation and random drift and the probability that an allele at frequency p lives the longest of all current alleles is just p, so the GEM distribution can be thought of as allele frequencies when alleles are ordered according to their future persistence in the population. Reversibility arguments allow us to conclude that we have the same distribution if we order alleles by their age.

4.3 Ewens Sampling Formula

We now turn to one of the most famous results from mathematical population genetics. We continue to assume the infinitely many alleles model.

Definition 4.11 (Allele frequency spectrum). In a sample of size n, for $1 \le j \le n$ write $\alpha(j)$ for the number of alleles that occur exactly j times in our sample. The vector $(\alpha(1), \dots, \alpha(n))$ is called the *allele frequency spectrum*.

This is illustrated in an example in Fig. 4.1.

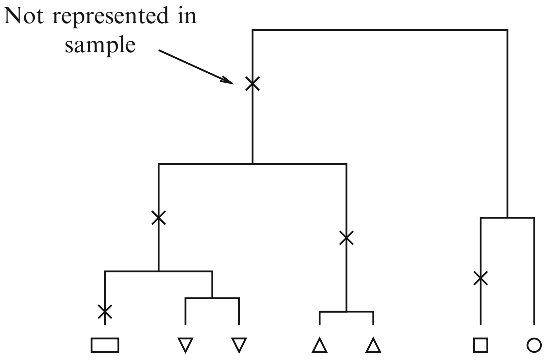

Fig. 4.1 The allele frequency spectrum. In this example, five mutations fall on the genealogical tree. Under the infinitely many alleles model, one of these mutations is not represented in the sample. The ancestral lineage of one individual has not experienced any mutations since the MRCA. In the notation of Definition 4.11, $\alpha(1) = 3$, $\alpha(2) = 2$, $\alpha(3) = \alpha(4) = \dots = \alpha(7) = 0$ and so the allele frequency spectrum is $(3, 2, 0, 0, 0, 0, 0)$

Theorem 4.12 (Ewens sampling formula, Ewens 1972). *If the genealogy of our sample is determined by Kingman's coalescent, then under the infinitely many alleles model with mutation rate $\theta/2$ (in the coalescent timescale)*

$$\mathbb{P}\left[(\alpha(j))_{1\leq j\leq n} = (a_j)_{1\leq j\leq n}\right] = \frac{n!\theta^k}{a_1!\cdots a_n! 1^{a_1}2^{a_2}\cdots n^{a_n}\theta_{(n)}}, \tag{4.7}$$

where $\sum_{j=1}^{n} a_j = k$ is the number of distinct alleles in the sample, $\sum_{j=1}^{n} j a_j = n$ and $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

Proof. There are many elegant derivations of this result (see, for example, Berestycki (2009) for one based on the 'Chinese Restaurant process'). Here we work directly with the Poisson–Dirichlet distribution.

First we show that if we take a sample of size n, then the probability that it falls into k distinct allelic types labelled $1,2,\ldots,k$, say, with n_i individuals of type i for each i is

$$\frac{n!\theta^k}{n_1\cdots n_k\theta(\theta+1)\cdots(\theta+n-1)}. \tag{4.8}$$

To see this, recall that under the Poisson–Dirichlet distribution, the probability that there are points in $(p_1, p_1+dp_1),\ldots,(p_k, p_k+dp_k)$ is

$$\theta^k(p_1\cdots p_k)^{-1}\left(1-\sum_1^k p_i\right)^{\theta-1} dp_1\ldots dp_k.$$

Given that such points exist, the probability that we see $n_1,\ldots,n_k$ copies of the corresponding alleles is the number of ways of assigning the n individuals in our sample to k classes of sizes $n_1,n_2,\ldots,n_k$ times the probability that the first n_1 are from class 1, the next n_2 from class 2 and so on. Combining these observations and integrating out over all choices of $p_1,\ldots,p_k$ gives

$$\begin{aligned}
&\frac{n!}{n_1!\cdots n_k!}\int_{\sum p_i\leq 1} p_1^{n_1}\cdots p_k^{n_k}\theta^k(p_1\cdots p_k)^{-1}\left(1-\sum_1^k p_i\right)^{\theta-1} dp_1\ldots dp_k\\
&= \frac{n!}{n_1!\cdots n_k!}\theta^k\int p_1^{n_1-1}\cdots p_k^{n_k-1}\left(1-\sum_1^k p_i\right)^{\theta-1} dp_1\ldots dp_k\\
&= \frac{n!}{n_1!\cdots n_k!}\theta^k\frac{\Gamma(n_1)\cdots\Gamma(n_k)\Gamma(\theta)}{\Gamma(n+\theta)}\\
&= \frac{n!\theta^k}{n_1\cdots n_k}\frac{1}{\theta(\theta+1)\cdots(\theta+n-1)}
\end{aligned}$$

as required.

Now to obtain the corresponding probability for the allelic partition $(\alpha(j))_{1\leq j\leq n}$, we evaluate this for a vector $(n_1,\ldots,n_k)$ for which $n_i = j$ exactly a_j times and multiply by

$$\frac{1}{a_1!\cdots a_n!},$$

that is the number of ways of assigning the class sizes to k types divided by $k!$ (because we don't care about the labels attached to types), to obtain

$$\begin{aligned}\mathbb{P}[(\alpha(j))_{1\le j\le n}=(a_j)_{1\le j\le n}] &= \frac{1}{a_1!\cdots a_n!}\frac{n!\theta^k}{n_1\cdots n_k}\frac{1}{\theta(\theta+1)\cdots(\theta+n-1)}\\ &= \frac{n!\theta^k}{a_1!\cdots a_n!}\frac{1}{\prod_{j=1}^k j^{a_j}}\frac{1}{\theta_{(n)}}\end{aligned}$$

where the last line follows because n_i is equal to j exactly a_j times. □

Remark 4.13. This formula has some remarkable properties. Most importantly, note that if we condition on the number of distinct alleles in the sample being k, the distribution of $(\alpha(1),\dots,\alpha(n))$ is *independent* of θ:

$$\begin{aligned}\mathbb{P}\left[(\alpha(j))_{1\le j\le n}=(a_j)_{1\le j\le n}\Big|\sum_{j=1}^n\alpha(j)=k\right] &= \frac{\dfrac{n!\theta^k}{a_1!\cdots a_n!}\dfrac{1}{\prod_{j=1}^n j^{a_j}\theta_{(n)}}}{\sum_{(b_j):\sum b_j=k}\dfrac{n!\theta^k}{b_1!\cdots b_n!}\dfrac{1}{\prod_{j=1}^n j^{b_j}\theta_{(n)}}}\\ &= \frac{1}{C_{n,j}}\frac{n!}{a_1!\cdots a_n!\prod j^{a_j}},\end{aligned}$$

where the constant $C_{n,j}$ depends only on n and j. Thus it is possible to test the neutral theory – so the goodness of fit of the Kingman coalescent – without making *any* assumptions about N_e or θ.

This remark tells us that the number K of distinct allelic classes in our sample of size n is a sufficient statistic for θ, so how can we use K to estimate θ?

Lemma 4.14. *For a sample of size n, let K be the number of distinct alleles. Then*

$$\mathbb{E}[K]=1+\theta\sum_{j=2}^n\frac{1}{j+\theta-1},\qquad \mathtt{var}(K)=\theta\sum_{j=2}^n\frac{j-1}{(j+\theta-1)^2}.$$

In particular,

$$\mathbb{E}[K]\sim\theta\log n,\quad \mathtt{var}(K)\sim\theta\log n\qquad \textit{as } n\to\infty.$$

Remark 4.15. This suggests

$$\hat{\theta}=\frac{K}{\log n}$$

as an estimator for θ, but like Watterson's estimator (which was based on the infinitely many sites model), because $\mathbb{E}[K]$ grows only like $\log n$, convergence of $\hat{\theta}$ to θ is extremely slow.

Proof of Lemma 4.14. Consider the coalescent with mutation. As we trace backwards in time we think of ancestral lineages as being lost by mutation or by coalescence and K is then the number lost by mutation. Write

$$X_j = \begin{cases} 1 \text{ if the } (n-j+1)\text{th lineage to be lost is lost by mutation} \\ 0 \text{ otherwise.} \end{cases}$$

With this convention, X_j records whether the transition from j to $j-1$ ancestral lineages is by mutation ($X_j = 1$) or coalescence ($X_j = 0$). Then

$$\mathbb{P}[X_j = 1] = \frac{j\theta/2}{\binom{j}{2} + j\theta/2} = \frac{\theta}{j+\theta-1},$$

and so

$$\mathbb{E}[K] = \sum_{j=2}^{n} \mathbb{E}[X_j] = \sum_{j=2}^{n} \frac{\theta}{j+\theta-1}$$

and (since the X_j are independent)

$$\texttt{var}(K) = \sum_{j=2}^{n} \texttt{var}(X_i) = \sum_{j=2}^{n} \frac{\theta(j-1)}{(j+\theta-1)^2}.$$

□