Genealogical Properties in the Moran Model with Large Sample Size

Arjun Biddanda

January 30, 2022

1 Derivations under the Moran Model

1.1 Definitions and Notation

There are many definitions for the Moran Model, but we will adopt the model according to the following definition:

1. Choose one individual to "die"
2. Choose one individual (it may be the same individual) to reproduce and have 2 offspring

Note that this differs from the properties of the Discrete-Time Wright-Fisher (DTWF) model as multiple-mergers are not allowed within a generation. However, all of the other relevant genealogical properties are available for interrogation.

1.2 Probability of Lineage Reduction

Initially we want to derive the probability that the number of lineages is reduced from n to m from the t^{th} to the $(t+1)^{\text{th}}$ generation :

$$p_{n,m}^{(t)} = \begin{cases} 1 - \frac{n(n-1)}{N_{t+1}^2} & \text{if } m = n, \\ \frac{n(n-1)}{N_{t+1}^2} & \text{if } m = n-1, \\ 0 & \text{else.} \end{cases} \tag{1}$$

where as t is increases we are moving farther back in time (e.g. the $(t+1)^{th}$ generation is gene generation immediately before the t^{th}). The rationale is that in order to reduce the number of lineages by 1, we must pick one of n individuals to then be "replaced" by one of $n-1$ individuals. This is a simple representation of the Moran model and can be extended to support a higher birth rate (which can potentially yield multiple mergers).

We can then extend this to $p_{n,m}^{(t,t')}$, or the probability that n samples in generation t have m ancestors in generation t'. Due to the construction of the Moran Model, we are required to bound the quantity m such that $n - t' \leq m \leq n$, otherwise the probability is 0. Using this bound, we obtain the following recursion :

$$\begin{aligned} p_{n,m}^{(t,t')} &= \left[\frac{n(n-1)}{N_{t+1}}\right] p_{n-1,m}^{(t,t'-1)} + \left[1 - \frac{n(n-1)}{N_{t+1}}\right] p_{n,m}^{(t,t'-1)} \\ p_{n,n}^{(t,t)} &= 1 \end{aligned} \tag{2}$$

The recursion holds when $t' \geq t$. This recursion allows us to look at lengths of time that are longer than a single generation and understand the long range time-scaling of the process needed to compare the Moran Model, DTWF, and the coalescent.

We expect that when the sample size gets larger, the proportion of lineages that are lost successively in the generations of the Moran model will be larger. See 2 for a direct comparison of the probability of m descendents under the Moran Model with time scaled to equal a generation in the DTWF model.

1.3 The Expected Number of Lineages as a Function of Time under the Moran Model

Now we can think of time t in much larger steps (e.g. in increments of t^*) such that we can now define the probability of having m ancestral lineages, abbreviated NLFT (conditional on n lineages at time $t = 0$). These results directly mirror those of Bhaskar et al Bhaskar et al. (2014), but now with the recursion derived for the Moran Model:

$$\mathbb{P}[A_n^M(t) = m] = \begin{cases} \sum_{k=m}^{n} p_{k,m}^{(t-t^*,t)} \mathbb{P}[A_n^M(t-t^*) = k], & \text{if } t > 0, \\ \delta_{n,m} & \text{if } t = 0 \end{cases}, \tag{3}$$

where δ is 1 when $n = m$ and 0 elsewhere. From this, we can directly calculate the expected NLFT in the Moran Model as:

$$\mathbb{E}[A_n^M(t)] = \sum_{m=1}^{n} m\mathbb{P}[A_n^M(t) = m] \tag{4}$$

1.4 Expected Number of Lineages Under the Coalescent

We use the moment-based method of Tavare (1984) to compute the expected NLFT (with no mutation) under the Kingman coalescent:

$$\begin{aligned} \mathbb{E}[A_n^C(t)] &= \sum_{i=1}^{n} e^{-\binom{i}{2}\Omega(t)}(2i-1)\frac{n_{(i)}}{n^{(i)}} \\ Var[A_n^C(t)] &= \sum_{i=1}^{n} e^{-\binom{i}{2}\Omega(t)}(2i-1)(i^2-i+1)\frac{n_{(i)}}{n^{(i)}} - \Bigg[\mathbb{E}[A_n^C(t)]\Bigg]^2 \end{aligned} \tag{5}$$

Where $n_{(i)}$ and $n^{(i)}$ are the falling and rising factorials respectively. Note that the t measured here is in “coalescent time” and will need to be rescaled appropriately when considering the discrete generations that we have for our model. We can define the following function to denote the rescaling of time (following Bhaskar et al. (2014)):

$$\Omega(t) = \begin{cases} \int_0^t \frac{N_0}{N_\tau} d\tau & \text{under the DTWF} \\ \int_0^t \frac{N_0^2/2}{N_\tau^2/2} d\tau & \text{under the Moran Model} \end{cases} \tag{6}$$

References

Anand Bhaskar, Andrew G Clark, and Yun S Song. Distortion of genealogical properties when the sample is very large. *Proceedings of the National Academy of Sciences*, 111(6):2385–90, 2014. doi: 10.1073/pnas.1322709111.

Simon Tavare. Line of descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164, 1984. doi: 10.1016/0040-5809(84)90027-3.

John Wakeley. *Coalescent Theory: An Introduction*. 2008. ISBN 0974707759.

2 Figures

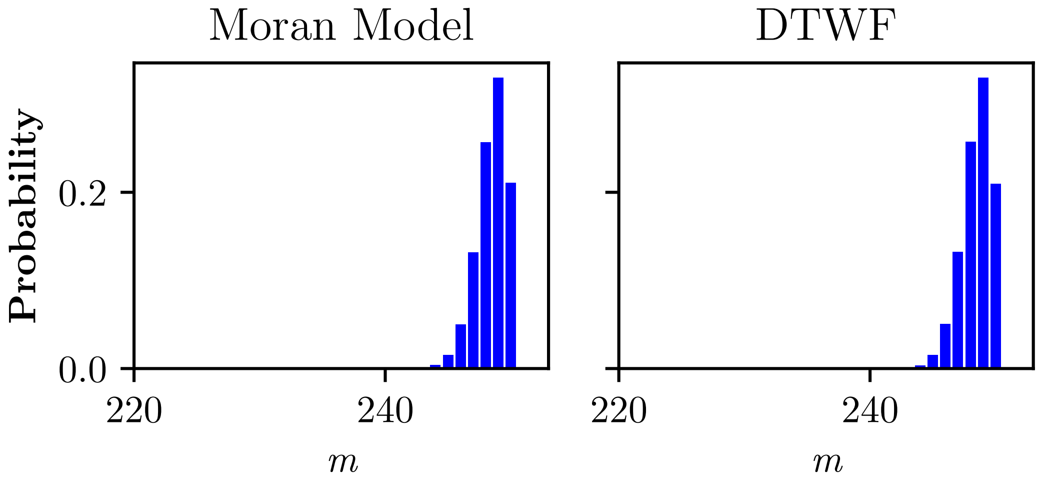

Figure 1: The distribution of parental lineages in the Moran Model and the Discrete-Time Wright-Fisher (DTWF) model. For both scenarios we maintained a constant population size of $N = 20000$ and a sample size of $n = 250$. The results in the right pane correspond exactly to Supplementary Figure 2 in Bhaskar et al. (2014). To compare the Moran model to the DTWF we only report after $N/2 = 20000/2 = 10000$ Moran generations based on the scaling result (Wakeley, 2008).

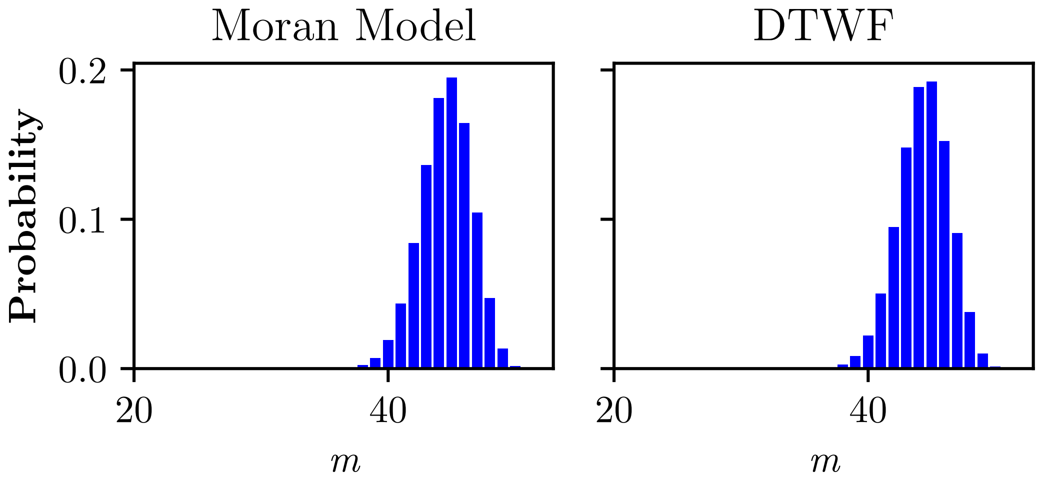

Figure 2: Comparing the distribution of parental lineages when the population size is relatively small ($N = 200$) with a sample size of $n = 50$ under the DTWF and Moran models. Under a smaller population size like this, we can start to observe differences in the ancestral DTWF model and the Moran model. Note that we have retained the same time scaling here as in the first figure as well.

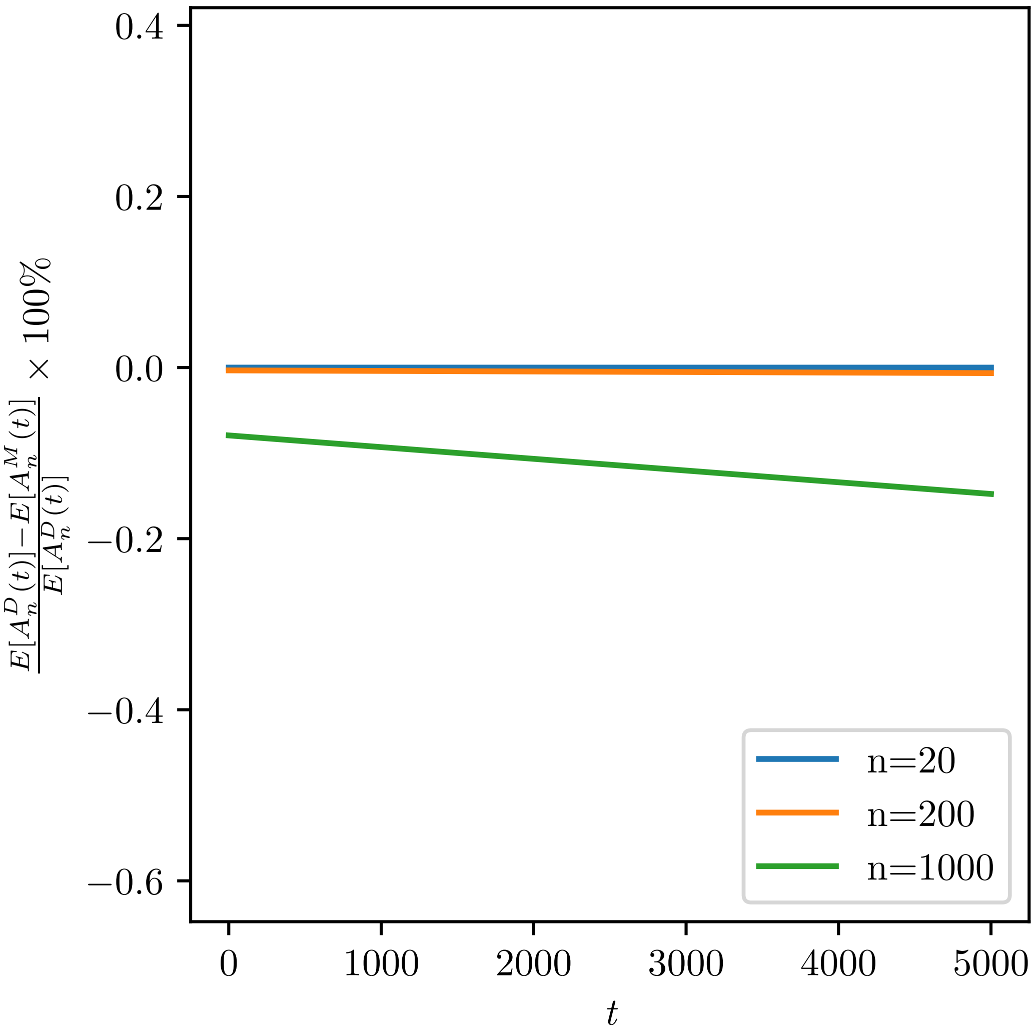

Figure 3: Difference in Expected NLFT between the DTWF and the Moran Model. Here we have scaled time appropriately since this is the constant sized case where $N = 20000$ (e.g. $\mathbb{E}[A_n^D(1)]$ vs. $\mathbb{E}[A_n^M(10000)]$) This suggests a consistent difference in the NLFT (and likely the expected SFS) between the Moran and the DTWF model when the sample size gets exceedingly large. This plot suggests that the DTWF model loses lineages at a faster rate than the Moran model (and therefore much more so than the Coalescent). The Moran model is also known to converge to the Kingman Coalescent at a rate of $O(N^{-2})$, where the DTWF model is known to converge to the coalescent at rate $O(N^{-1})$

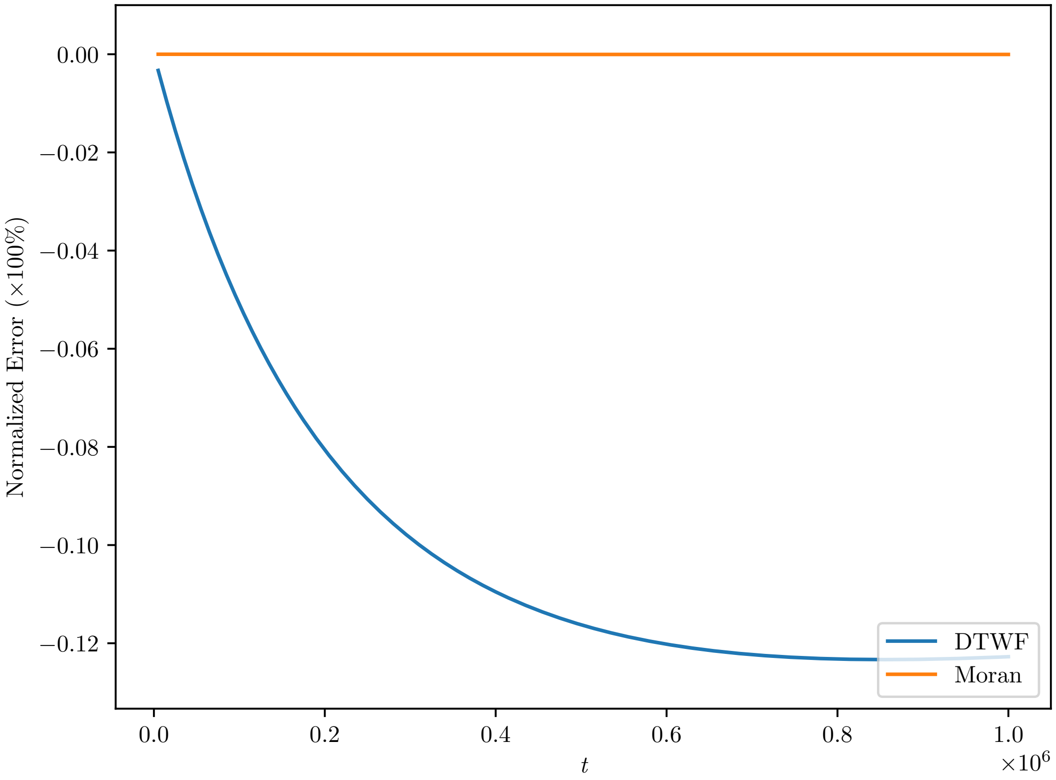

Figure 4: Normalized error in expected NLFT between the DTWF and the Moran Model. The expressions compared here are $\frac{\mathbb{E}(A^M)-\mathbb{E}(A^C)}{\mathbb{E}(A^M)}$ and $\frac{\mathbb{E}(A^D)-\mathbb{E}(A^C)}{\mathbb{E}(A^D)}$.