Likelihood of the model for the branching process

Author

October 2025

Supposed the process starts at time $T_0 = 0$ and the number of starting stem cell S_0. The interarrival time of the next event is exponentially distributed

$$\Delta_{T_i} = T_i - T_{i-1} \sim \text{Exp}(r \cdot S_{i-1}),\ i = 1, \cdots, n,$$

where r is the division rate. At the event time T_i, the triplet of random variable (X_i, Y_i, Z_i) has the following distribution

$$(X_i, Y_i, Z_i) = \begin{cases} (+1, 0, 0), & p_1(T_i) \\ (0, +1, 0), & p_2(T_i) \\ (-1, +2, 0), & p_3(T_i) \\ (0, 0, +1), & p_4(T_i). \end{cases} \tag{1}$$

Assume that there is no dud stem cells (i.e $p_4(t) = 0\,\forall t$), and we observe all the events (both the time of the events $T_0, T_1, \cdots, T_n$, and the number of stem cells $S_0, S_1, \cdots, S_n$. Since we observe every events, we can know how the cell changes (X_i, Y_i) at each event time.

The likelihood of observing the event times and division types

$$\mathbb{L}(T_1, T_2, \cdots, T_n, S_1, \cdots, S_i) = \prod_{i=1}^{n} \Big[p_1(T_i) I_{(X_i=1)} + p_2(T_i) I_{(X_i=0)} + p_3(T_i) I_{(X_i=-1)} \Big] \cdot r S_{i-1} e^{-r S_{i-1} \Delta_{T_i}}. \tag{2}$$

Given the division probability

$$\begin{aligned} P(X_i = 1 | T_i) &= \frac{p_1}{1 + c(T_i - m)^2} \\ P(X_i = 0 | T_i) &= \frac{p_2}{1 + c(T_i - m)^2} \\ P(X_i = -1 | T_i) &= 1 - \frac{p_1 + p_2}{1 + c(T_i - m)^2}, \end{aligned} \tag{3}$$

with $p_1, p_2, c, m > 0, p_1 + p_2 < 1$, the likelihood of observing the event is

$$\begin{aligned} &\mathbb{L}(T_1, T_2, \cdots, T_n, S_1, \cdots, S_i) \\ &= \prod_{i=1}^{n} \Big[\frac{p_1}{1 + c(T_i - m)^2} I_{(X_i=1)} + \frac{p_2}{1 + c(T_i - m)^2} I_{(X_i=0)} + \Big(1 - \frac{p_1 + p_2}{1 + c(T_i - m)^2} \Big) I_{(X_i=-1)} \Big] r S_{i-1} e^{-r S_{i-1} \Delta_{T_i}}. \end{aligned} \tag{4}$$

Let $f(t,c,m) = \frac{1}{1+c(t-m)^2}$, the log-likelihood is

$$
\begin{aligned}
&\ell(T_1, T_2, \cdots, T_n, S_1, \cdots, S_n) \\
&= \sum_{i=1}^{n} \Big[\log(p_1 \cdot f(T_i,c,m)) \cdot I_{(X_i=1)} + \log(p_2 \cdot f(T_i,c,m)) \cdot I_{(X_i=0)} + \log([1-(p_1+p_2)] \cdot f(T_i,c,m)) \cdot I_{(X_i=-1)} \Big] \\
&+ n\log(r) + \sum_{i=1}^{n} \log(S_{i-1}) - r\sum_{i=1}^{n} S_{i-1}\Delta_{T_i}.
\end{aligned} \tag{5}
$$

We take derivative of the log-likelihood with respect to each parameter r, p_1, p_2, c, m

$$
\begin{aligned}
\frac{\partial \ell}{\partial r} &= \frac{n}{r} - \sum_{i=1}^{n} S_{i-1}\Delta_{T_i}, \\
\frac{\partial \ell}{\partial p_1} &= \sum_{i=1}^{n} \frac{I_{(X_i=1)}}{p_1} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)} f(T_i,c,m)}{1-(p_1+p_2)f(T_i,c,m)}, \\
\frac{\partial \ell}{\partial p_2} &= \sum_{i=1}^{n} \frac{I_{(X_i=0)}}{p_2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)} f(T_i,c,m)}{1-(p_1+p_2)f(T_i,c,m)}, \\
\frac{\partial \ell}{\partial c} &= \sum_{i=1}^{n} \frac{I_{(X_i=1)} + I_{(X_i=0)}}{f(T_i,c,m)} [-(T_i-m)^2 f(T_i,c,m)] \\
&\quad - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}(p_1+p_2)}{1-(p_1+p_2)f(T_i,c,m)} [-(T_i-m)^2 f(T_i,c,m)], \\
\frac{\partial \ell}{\partial m} &= \sum_{i=1}^{n} \frac{I_{(X_i=1)} + I_{(X_i=0)}}{f(T_i,c,m)} 2c(T_i-m)[f(T_i,c,m)]^2 \\
&\quad - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}(p_1+p_2)}{1-(p_1+p_2)f(T_i,c,m)} 2c(T_i-m)[f(T_i,c,m)]^2.
\end{aligned} \tag{6}
$$

Setting $\frac{\partial \ell}{\partial r} = 0$, we have

$$\hat{r} = \frac{n}{\sum_{i=1}^{n} S_{i-1}\Delta_{T_i}}.$$

So we can get a closed-form solution for the MLE of parameter r.

MLE Estimates

I use the log-likelihood function in equation (5) and optimize it using the Nelder-Mead optimization in R with linear inequality constraints (function constrOptim) to estimates the parameters. I also include the gradients from equation (6) in the optimization. I simulate 100 replications using the parameters $S_0 = 200, r = 0.2, p_1 = 0.5, p_2 = 0.2, c = 0.005, m = 4$. Figure (1) shows the probability function given these parameters.

The estimates are stable at different starting points. Figure (2) and table (1) show estimates and their summary statistics across 100 replications with starting value $(p_1, p_2, c, m, r) = (0.1, 0.1, 5, 5, 1)$. Figure (3) and table (2) show estimates and their summary statistics across 100 replications with starting value $(p_1, p_2, c, m, r) = (0.2, 0.4, 10, 10, 1)$. Both starting values give really good estimates across all parameters.

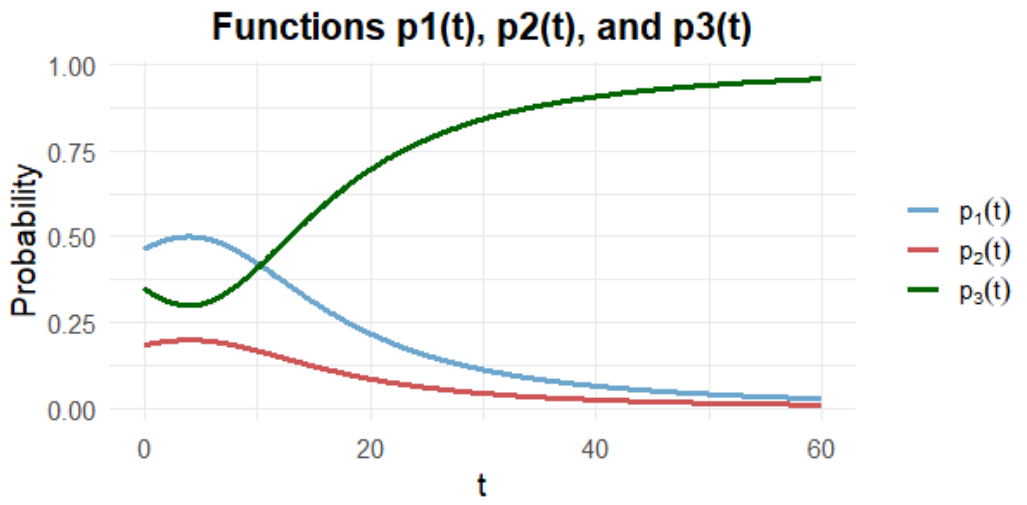

Figure 1: Functions $p_1(t), p_2(t), p_3(t)$ with parameters $p_1 = 0.5, p_2 = 0.2, c = 0.005, m = 4$.

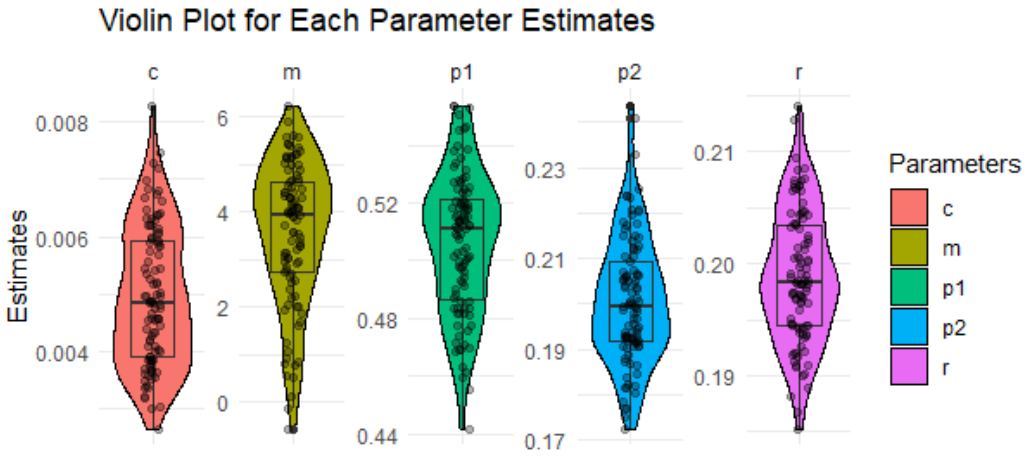

Figure 2: Violin plot for estimate results using starting values $(p_1, p_2, c, m, r) = (0.1, 0.1, 5, 5, 1)$ with 100 replications. The true parameters are $(p_1, p_2, c, m, r) = (0.5, 0.2, 0.005, 4, 0.2)$.

Parameter	p_1	p_2	c	m	r
Mean	0.506	0.201	0.00495	3.550	0.199
Median	0.511	0.199	0.00485	3.940	0.198
2.5 Percentile	0.461	0.177	0.00311	0.315	0.188
97.5 Percentile	0.552	0.229	0.00723	5.590	0.209

Table 1: Parameter estimate results using starting values $(p_1, p_2, c, m, r) = (0.1, 0.1, 5, 5, 1)$ with 100 replications. The true parameters are $(p_1, p_2, c, m, r) = (0.5, 0.2, 0.005, 4, 0.2)$.

Simulation to include tracking of each cell time of division

I'm also currently working on simulating data to track the time stem cells are created and undergo division. The simulation produces two datasets. The first one is the data that tracks cell counts after each division as

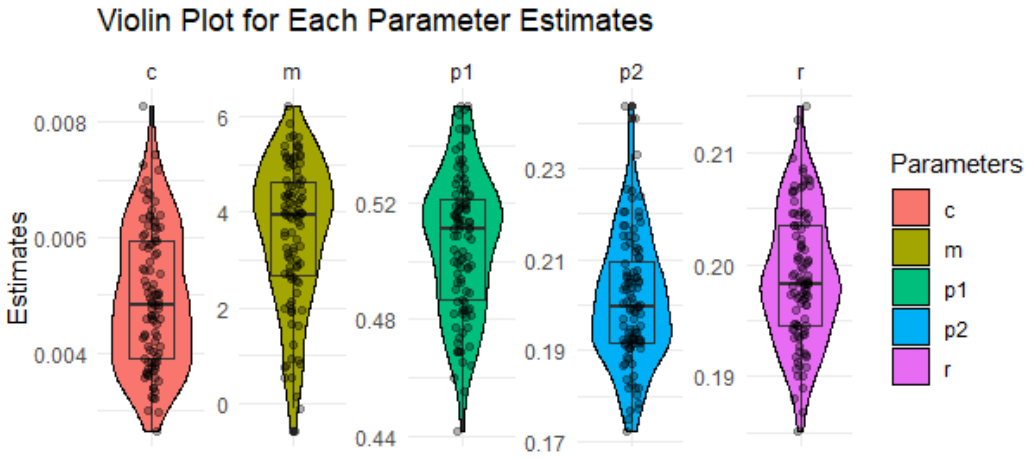

Figure 3: Violin plot for estimate results using starting values $(p_1, p_2, c, m, r) = (0.2, 0.4, 10, 10, 1)$ with 100 replications. The true parameters are $(p_1, p_2, c, m, r) = (0.5, 0.2, 0.005, 4, 0.2)$.

Parameter	p_1	p_2	c	m	r
Mean	0.506	0.201	0.00495	3.530	0.199
Median	0.511	0.200	0.00482	3.960	0.198
2.5 Percentile	0.462	0.177	0.00311	0.316	0.189
97.5 Percentile	0.552	0.229	0.00723	5.590	0.209

Table 2: Parameter estimate results using starting values $(p_1, p_2, c, m, r) = (0.1, 0.1, 5, 5, 1)$ with 100 replications. The true parameters are $(p_1, p_2, c, m, r) = (0.5, 0.2, 0.005, 4, 0.2)$.

we have before (figure (4). The second one is the data that tracks the parent cell of each cell and when each cell is created and ceased to exist (undergo division) (figure (5). Currently I have this tracking for viable stem cells only. I will add the tracking to the non-viable stem cells and differentiated cells as well. The goal is to used the information of the time cell is created to estimate whether it is a viable or non-viable stem cell.

reps	time.steps	sc.steps	ec.steps
1	0.00000000	200	0
1	0.04410613	199	2
1	0.08321523	199	3
1	0.10279252	199	4
1	0.10942688	200	4
1	0.23466071	201	4
1	0.23616411	202	4
1	0.26378449	201	6

Figure 4: Simulated data that tracks the cell counts after each division.

reps	id	parent	birth_time	death_time
1	487	378	5.344656	7.875954
1	488	50	5.344719	10.695169
1	489	50	5.344719	17.672922
1	490	145	5.356451	12.570040
1	491	17	5.395300	8.163755
1	492	306	5.434105	8.312166
1	493	306	5.434105	14.815076
1	494	465	5.446922	12.343150

Figure 5: Simulated data that tracks parent cells, birth time (when cell is created) and death time (when cell undergoes division and ceases to exists) of stem cell.

Add p_4 to simplest form of probability function with p_1, p_2, c, m

I tried to add p_4, probability of getting non-viable stem cells, to the simplest form of the probability function with parameters p_1, p_2, c, m

$$\begin{aligned} P((X_i, Z_i) = (1,0)|T_i) &= \frac{p_1}{1 + c(T_i - m)^2} \\ P((X_i, Z_i) = (0,0)|T_i) &= \frac{p_2}{1 + c(T_i - m)^2} \\ P((X_i, Z_i) = (-1,0)|T_i) &= 1 - p_4 - \frac{p_1 + p_2}{1 + c(T_i - m)^2} \\ P((X_i, Z_i) = (0,1)|T_i) &= p_4, \end{aligned} \tag{7}$$

When we observe viable and non-viable stem cells separately, the likelihood of observing the event times and division outcomes

$$\begin{aligned} &\mathbb{L}(T_1, T_2, \cdots, T_n, S_1, \cdots, S_i) \\ =&\prod_{i=1}^{n} \Big[p_1(T_i) I_{(X_i=1, Z_i=0)} + p_2(T_i) I_{(X_i=0, Z_i=0)} + p_3(T_i) I_{(X_i=-1, Z_i=0)} + p_4(T_i) I_{(X_i=0, Z_i=1)} \Big] r S_{i-1} e^{-r S_{i-1} \Delta_{T_i}}. \end{aligned} \tag{8}$$

I simulate 100 replications using the parameters $S_0 = 200, r = 0.2, p_1 = 0.5, p_2 = 0.2, p_4 = 0.05, c = 0.005, m = 4$. Table (3) and (4) show the estimate results with two different starting values. The estimates are very stable and accurate.

Parameter	p_1	p_2	p_4	c	m	r
Mean	0.502	0.202	0.049	0.00519	3.870	0.199
Median	0.500	0.200	0.050	0.00530	4.060	0.199
2.5 Percentile	0.450	0.174	0.037	0.00321	0.641	0.189
97.5 Percentile	0.544	0.241	0.060	0.00741	5.730	0.209

Table 3: Parameter estimate results for the simplest form of the probability function with non-viable stem cells. The true parameters are $(p_1, p_2, p_4, c, m) = (0.5, 0.2, 0.05, 0.005, 4)$. The starting values are $(p_1, p_2, p_4, c, m) = (0.1, 0.1, 0.1, 5, 5, 1)$.

Parameter	p_1	p_2	p_4	c	m	r
Mean	0.502	0.202	0.049	0.00519	3.880	0.199
Median	0.500	0.200	0.050	0.00529	4.080	0.199
2.5 Percentile	0.450	0.174	0.037	0.00320	0.641	0.189
97.5 Percentile	0.544	0.241	0.060	0.00741	5.730	0.209

Table 4: Parameter estimate results for the simplest form of the probability function with non-viable stem cells. The true parameters are $(p_1, p_2, p_4, c, m) = (0.5, 0.2, 0.05, 0.005, 4)$. The starting values are $(p_1, p_2, p_4, c, m) = (0.2, 0.4, 0.1, 20, 20, 1)$.

Variance of Cell Count in Branching Process

Derivation of the theoretical expectation and variance

Let $\Delta > 0$.

$$\begin{aligned} E[X(t+\Delta) - X(t)|X(t)] &= [p_1(t) - p_3(t) + \mathcal{O}(\Delta)] \cdot rX(t)\Delta + \wr(\Delta) \\ &= [p_1(t) - p_3(t)]rX(t) + \wr(\Delta). \end{aligned} \tag{9}$$

Taking expectation,

$$S(t+\Delta) - S(t) = [p_1(t) - p_3(t)]rS(t)\Delta + E[\xi] + \wr(\Delta). \tag{10}$$

Let $\Delta \to 0$,

$$\frac{dS(t)}{dt} = [p_1(t) - p_3(t)]rS(t). \tag{11}$$

Thus, the expected stem cell count of the branching process coincides with the differential equation.

Let $V(t)$ denote the theoretical variance of the stem cell in the branching process. Denote $S(t) = E[X(t)], M(t) = E[X(t)^2]$, then $V(t) = M(t) - S(t)^2$. Let $\Delta > 0$.

$$\begin{aligned} E[X(t+\Delta)^2 - X(t)^2|X(t)] &= [2X(t)(p_1(t) - p_3(t)) + (p_1(t) + p_3(t))]rX(t)\Delta \\ &= 2X(t)^2(p_1(t) - p_3(t))r\Delta + X(t)(p_1(t) + p_3(t))r\Delta. \end{aligned} \tag{12}$$

Taking expectation,

$$M(t+\Delta) - M(t) = 2[p_1(t) - p_3(t)]rM(t)\Delta + [p_1(t) + p_3(t)]rS(t)\Delta. \tag{13}$$

Let $\Delta \to 0$,

$$\begin{aligned} M'(t) &= 2[p_1(t) - p_3(t)]rM(t) + [p_1(t) + p_3(t)]rS(t), \\ V'(t) &= M'(t) - 2S(t)S'(t) \\ &= 2[p_1(t) - p_3(t)]rM(t) + [p_1(t) + p_3(t)]rS(t) - 2S(t)[p_1(t) - p_3(t)]rS(t) \\ &= 2[p_1(t) - p_3(t)]rV(t) + [p_1(t) + p_3(t)]rS(t). \end{aligned} \tag{14}$$

Verify the theoretical variance with simulation

Figure (6) displays the plots to compare variance from the branching process with variance from the compound nonhomogeneous Poisson process, using the form of the probability function

$$\begin{aligned} P(X_i = 1|T_i) &= \frac{p_1}{1 + c_1(T_i - m_1)^2}, \\ P(X_i = 0|T_i) &= \frac{p_2}{1 + c_2(T_i - m_2)^2}, \\ P(X_i = -1|T_i) &= 1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2}. \end{aligned} \tag{15}$$

The parameters used to construct these plots are $S_0 = 200, r = 0.2, p_1 = 0.5, c_1 = 0.005, m_1 = 4, p_2 = 0.2, c_2 = 0.1, m_2 = 12$. Figure (6) on the left is the plot of variances over time, and on right is the theoretical mean and the region within two standard deviations from the mean. In this set-up, the theoretical variances of the two processes start out similar in the beginning, however, as t gets large, the variance of the branching process converges to 0, whereas the variance of the compound nonhomogeneous Poisson process converges to a positive

number. It is also possible for the variance of the branching process to be greater than that of the compound Poisson process.

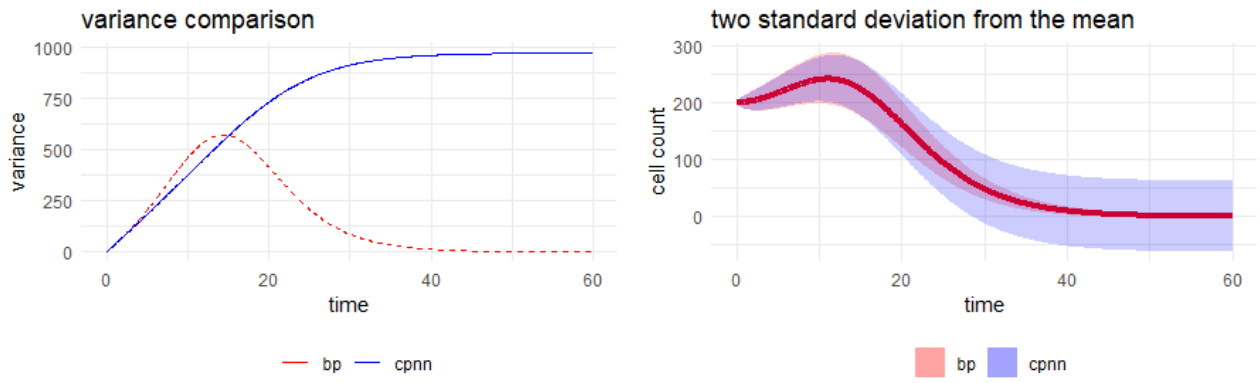

Figure 6: Comparing theoretical variances of the two process with parameters $S_0 = 200, r = 0.2, p_1 = 0.5, c_1 = 0.005, m_1 = 4, p_2 = 0.2, c_2 = 0.1, m_2 = 12$

Figure (7) compares the theoretical variances with the variability from the simulated data. 50 replications of the simulated data with the same parameters as in figure (6). The simulated cell count data is plotted, along with the theoretical mean and the region within 2 theoretical standard deviation from the mean.

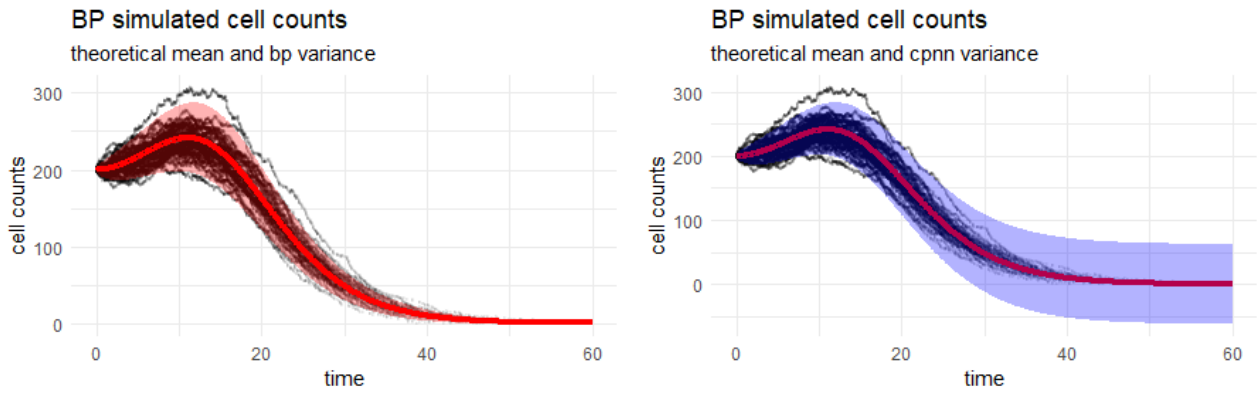

Figure 7: Compare theoretical variances with the simulated data with parameters $S_0 = 200, r = 0.2, p_1 = 0.5, c_1 = 0.005, m_1 = 4, p_2 = 0.2, c_2 = 0.1, m_2 = 12$.

Show variance of the stem cell is finite

$$V'(t) = 2[p_1(t) - p_3(t)]rV(t) + [p_1(t) + p_3(t)]rS(t),$$

$$
\begin{aligned}
\Rightarrow V(t) &= \exp\Big\{2r\int_0^t [p_1(u)-p_3(u)]du\Big\}\Big[\int_0^t [p_1(u)+p_3(u)]rS(u)\exp\Big\{-2r\int_0^u [p_1(v)-p_3(v)]dv\Big\}du + C\Big] \\
&= \exp\Big\{2r\int_0^t [p_1(u)-p_3(u)]du\Big\} \\
&\quad \Big[\int_0^t [p_1(u)+p_3(u)]rS_0\exp\Big\{r\int_0^u [p_1(v)-p_3(v)]dv\Big\}\exp\Big\{-2r\int_0^u [p_1(v)-p_3(v)]dv\Big\}du + C\Big] \\
&= S_0\cdot r\cdot \exp\Big\{2r\int_0^t [p_1(u)-p_3(u)]du\Big\}\Big[\int_0^t [p_1(u)+p_3(u)]\exp\Big\{-r\int_0^u [p_1(v)-p_3(v)]dv\Big\}du + C\Big].
\end{aligned} \tag{16}
$$

Using the initial condition $V(0)=0$, we can simplify the expression of $V(t)$ in terms of $S(t)$

$$V(t) = rS(t)^2\int_0^t \frac{p_1(u)+p_3(u)}{S(u)}du.$$

Let $P(t)=\int_0^t[p_1(u)-p_3(u)]du$. Since $p_1(u)+p_3(u)\le 1\ \forall u$ and r is a positive constant, to show $V(t)<\infty$ as $t\to\infty$, we can show

$$f(t) = \exp\Big\{P(t)\Big\}\Big[\int_0^t \exp\Big\{-P(u)\Big\}du\Big] < \infty$$

as $t\to\infty$. When $P(t)\to\infty$, $\exp\Big\{P(t)\Big\}\to\infty$ and $\int_0^t\exp\Big\{-P(u)\Big\}du<\infty$ as $t\to\infty$. Then $f(t)\to\infty$ as $t\to\infty$. Thus, we consider when $P(t)\to-\infty$ as $t\to\infty$.

We rewrite $f(t)$ as a quotient

$$f(t) = \frac{\int_0^t \exp\Big\{-P(u)\Big\}du}{\exp\Big\{-P(t)\Big\}}.$$

Both numerator and denominator go to $+\infty$ as $t\to\infty$, we can use the l'Hôpital's rule.

$$\lim_{t\to\infty} f(t) = \lim_{t\to\infty}\frac{\int_0^t \exp\Big\{-P(u)\Big\}du}{\exp\Big\{-P(t)\Big\}} = \lim_{t\to\infty} = \frac{\exp\{-P(t)\}}{-[p_1(t)-p_3(t)]\exp\{-P(t)\}} = \lim_{t\to\infty}\frac{-1}{p_1(t)-p_3(t)}.$$

Case 1. $\lim_{t\to\infty}[p_1(t)-p_3(t)]\to -p$ for some $p>0$. Then $\lim_{t\to\infty} f(t)$ is a finite constant.
Case 2. $\lim_{t\to\infty}[p_1(t)-p_3(t)]\to-\infty$. Then $\lim_{t\to\infty} f(t)\to 0$.
Case 3. $[p_1(t)-p_3(t)]$ does not converge but is negative on average

$$\limsup_{T\to\infty}\frac{1}{T}\int_0^T [p_1(u)-p_3(u)]du = \limsup_{T\to\infty}\frac{P(T)}{T} = \bar{p} < 0.$$

Then on average $P(t)$ decreases roughly like $\bar{p}t$ for large t with some remainder term $P(t)=\bar{p}t+R(t)$, that is bounded or growing slower than linear. Then

$$
\begin{aligned}
f(t) &= \exp\Big\{P(t)\Big\}\Big[\int_0^t \exp\Big\{-P(u)\Big\}du\Big] \\
&= \exp\{\bar{p}t+R(t)\}\int_0^t \exp\{-\bar{p}u-R(u)\}du \\
&= \exp\{\bar{p}t\}\exp\{R(t)\}\int_0^t \exp\{|\bar{p}|u\}\exp\{-R(u)\}du \text{ (since } \bar{p}<0).
\end{aligned} \tag{17}
$$

Since $R(t)$ oscillates and is bounded, $\exp\{-R(u)\}$ is bounded. Thus, $\int_0^t\exp\{|\bar{p}|u\}\exp\{-R(u)\}du \approx C_1\exp\{|\bar{p}|t\}$

for some $C > 0$. Then

$$f(t) \approx \exp\{\bar{p}t\} \exp\{R(t)\} C_1 \exp\{|\bar{p}|t\} = C_1 \exp\{R(t)\}.$$

Since $\exp\{R(u)\}$ is bounded, $f(t)$ is bounded. Thus, as $t \to \infty$ the variance $V(t)$ is finite when $\int_0^t [p_1(u) - p_3(u)]du \to -\infty$. $V(t) \to 0$ when $[p_1(t) - p_3(t)] \to -\infty$.

Add parameters for $p_2(t)$

Suppose we have the division probabilities

$$\begin{aligned} P(X_i = 1|T_i) &= \frac{p_1}{1 + c_1(T_i - m_1)^2}, \\ P(X_i = 0|T_i) &= \frac{p_2}{1 + c_2(T_i - m_2)^2}, \\ P(X_i = -1|T_i) &= 1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2}, \end{aligned} \tag{18}$$

with $p_1, p_2, c, m > 0, p_1 + p2 < 1$. The likelihood of observing the event is

$$\begin{aligned} \mathbb{L}(T_1, \cdots, T_n, S_0, \cdots, S_n) = \prod_{i=1}^{n} \Big[&\frac{p_1}{1 + c_1(T_i - m_1)^2} I_{(X_i=1)} + \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(X_i=0)} \\ &+ \Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(X_i=-1)}\Big)\Big] r S_{i-1} e^{-r S_{i-1} \Delta_{T_i}}. \end{aligned} \tag{19}$$

The log-likelihood is

$$\begin{aligned} \ell(T_1, \cdots, T_n, S_0, \cdots, S_n) = \sum_{i=1}^{n} \Big[&\log\Big(\frac{p_1}{1 + c_1(T_i - m_1)^2}\Big) I_{(X_i=1)} \\ &+ \log\Big(\frac{p_2}{1 + c_2(T_i - m_2)^2}\Big) I_{(X_i=0)} \\ &+ \log\Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2}\Big) I_{(X_i=-1)} \\ &+ \log r + \log S_{i-1} - r S_{i-1} \Delta_{T_1} \Big]. \end{aligned} \tag{20}$$

We take the derivative of the log-likelihood with respect to each parameter $r, p_1, p_2, c_1, c_2, m_1, m_2$.

$$\begin{aligned} \frac{\partial \ell}{\partial r} &= \frac{n}{r} - \sum_{i=1}^{n} S_{i-1} \Delta_{T_I} \\ \frac{\partial \ell}{\partial p_1} &= \sum_{i=1}^{n} \frac{I_{(X_i=1)}}{p_1} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \frac{1}{1 + c_1(T_i - m_1)^2} \\ \frac{\partial \ell}{\partial p_2} &= \sum_{i=1}^{n} \frac{I_{(X_i=0)}}{p_2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \frac{1}{1 + c_1(T_i - m_1)^2} \\ \frac{\partial \ell}{\partial c_1} &= -\sum_{i=1}^{n} \frac{I_{(X_i=1)}(T_i - m_1)^2}{1 + c_1(T_i - m_1)^2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \Big(- \frac{p_1(T_i - m_1)^2}{[1 + c_1(T_i - m_1)^2]^2} \Big) \\ \frac{\partial \ell}{\partial c_2} &= -\sum_{i=1}^{n} \frac{I_{(X_i=0)}(T_i - m_2)^2}{1 + c_2(T_i - m_2)^2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \Big(- \frac{p_2(T_i - m_2)^2}{[1 + c_2(T_i - m_2)^2]^2} \Big) \\ \frac{\partial \ell}{\partial m_1} &= \sum_{i=1}^{n} \frac{I_{(X_i=1)} 2c_1(T_i - m_1)}{1 + c_1(T_i - m_1)^2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \frac{2p_1 c_1(T_i - m_1)}{[1 + c_1(T_1 - m_1)^2]^2} \\ \frac{\partial \ell}{\partial m_2} &= \sum_{i=1}^{n} \frac{I_{(X_i=0)} 2c_2(T_i - m_2)}{1 + c_2(T_i - m_2)^2} - \sum_{i=1}^{n} \frac{I_{(X_i=-1)}}{1 - \frac{p_1}{1+c_1(T_i-m_1)^2} - \frac{p_2}{1+c_2(T_i-m_2)^2}} \frac{2p_2 c_2(T_i - m_2)}{[1 + c_2(T_i - m_2)^2]^2} \end{aligned} \tag{21}$$

I reviewed the code for the optimization of the log-likelihood function when we have the probability function as in (12) with different scale and location parameters (c_1, c_2, m_1, m_2) for $p_1(t)$ and $p_2(t)$. From the optimization results, the estimates for p_1 and p_2 are very stable, however, the estimates for c_1, c_2, m_1, m_2 are not. I ran the optimization multiple times with different starting points and chose the results with the highest log-likelihood. The starting points are shown in table (5). Since p_1 and p_2 are not sensitive to starting values, the starting values for these two parameters are the same, whereas starting values for c_1, c_2, m_1, m_2 describe different scenarios of where the peaks and how sharp the peaks are in functions $p_1(t)$ and $p_2(t)$.

Scenarios	p_1	p_2	c_1	c_2	m_1	m_2
Early peaks	0.3	0.3	0.05	0.05	10	10
Very early peaks	0.3	0.3	0.05	0.05	5	5
Late peaks	0.3	0.3	0.05	0.05	45	45
Sharp peak at center	0.3	0.3	0.2	0.2	25	25
Broad peak at center	0.3	0.3	0.001	0.001	25	25
Asymmetric time centers	0.3	0.3	0.02	0.03	35	15
Asymmetric time centers	0.3	0.3	0.1	0.05	20	30
Random start	unif(0.2,0.5)	unif(0.2,0.5)	unif(0.001,0.05)	unif(0.001,0.05)	unif(0,50)	unif(0,50)

Table 5: Multiple starting points for optimizations

Table (6) and (7) shows the estimate results for simulated data of two sets of parameters using the same multiple starting values. Both results are good and converge to the true parameters. I will try different optimization techniques to see if it would be less sensitive to starting points.

Parameter	p_1 (0.55)	p_2 (0.15)	c_1 (0.005)	c_2 (0.01)	m_1 (4)	m_2 (18)
Mean	0.560	0.152	0.00491	0.01120	3.340	17.600
Median	0.560	0.152	0.00479	0.01110	3.820	17.600
2.5 Percentile	0.503	0.107	0.00292	0.00374	0.0000416	14.200
97.5 Percentile	0.621	0.192	0.00759	0.02150	6.570	20.000

Table 6: Parameters estimate results using multiple starting values with 100 replications. The true parameters are $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.55, 0.15, 0.005, 0.01, 4, 18)$.

Parameter	p_1 (0.2)	p_2 (0.7)	c_1 (0.005)	c_2 (0.01)	m_1 (12)	m_2 (6)
Mean	0.200	0.699	0.00504	0.01030	10.700	6.090
Median	0.197	0.697	0.00454	0.01020	10.400	6.190
2.5 Percentile	0.163	0.639	0.00115	0.00679	5.590	4.540
97.5 Percentile	0.244	0.761	0.01110	0.01490	14.800	7.130

Table 7: Parameters estimate results using multiple starting values with 100 replications. The true parameters are $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.2, 0.6, 0.005, 0.1, 12, 6)$.

Stopping times

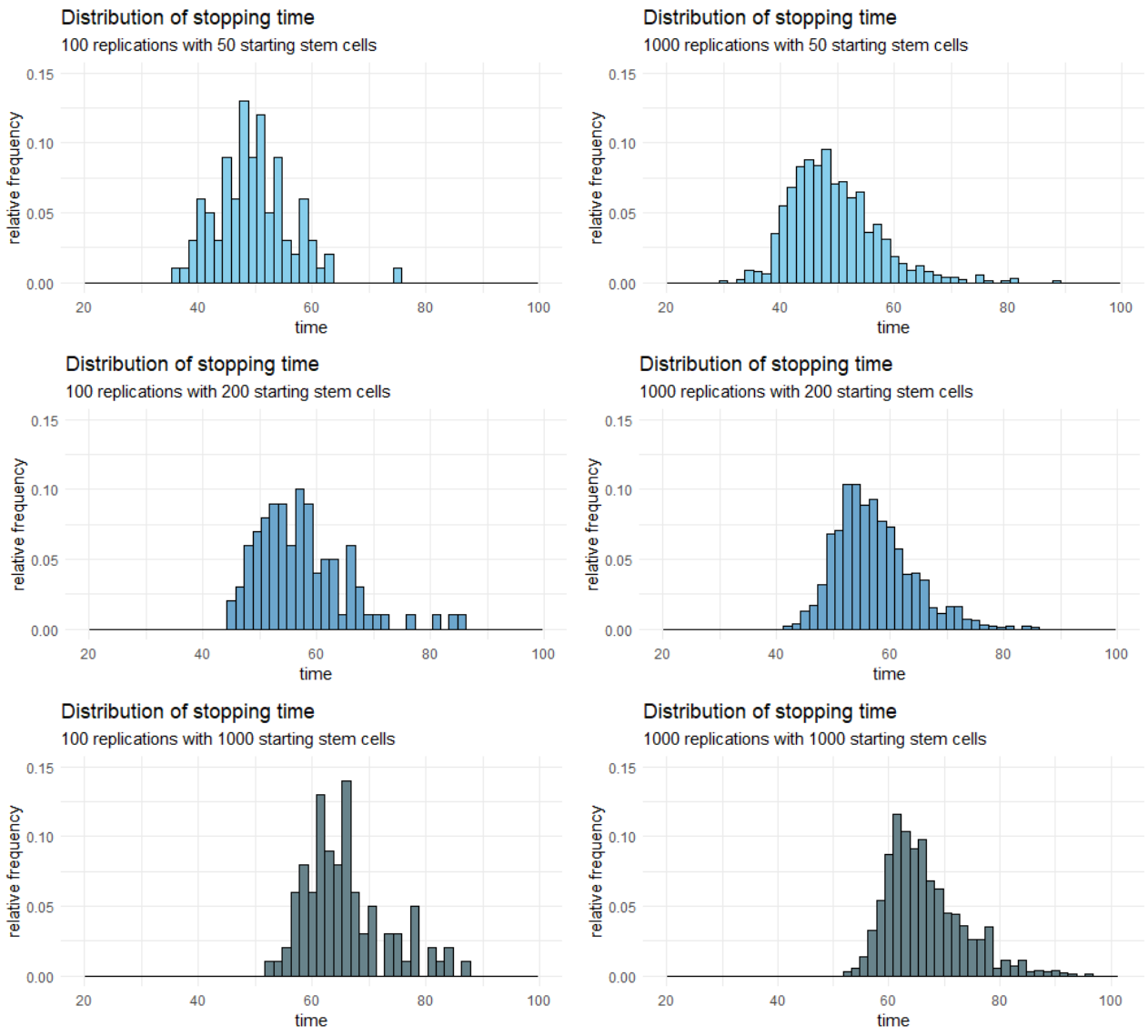

Figure 8: Distributions of stopping times from simulated data with varying numbers of initial stem cells and replication counts. The number of starting stem cells is 50 (top row), 200 (middle row), and 1000 (bottom row). Each column represents a different number of replications: 100 (left column) and 1000 (right column). The parameters for the probability function is $p_1 = 0.55, c_1 = 0.005, m_1 = 4, p_2 = 0.15, c_2 = 0.01, m_2 = 18$ and division rate is $r = 0.2$.

Fitting the inverse Gaussian distribution to the shifted stopping times

We first derive the minimum stopping times. Since the interarrival times between two consecutive division events is the a exponential distribution with rate proportional to the current viable stem cells. Then, the average interarrival time between two consecutive division events $T_{i+1} - T_i$ is $1/(rS_i)$. The minimum stopping times occurs when at each division event is a differentiation event, in which no new viable stem cell is created and the number of viable stem cells decreases by 1. Then the minimum stopping time is

$$\tau_{\min} = \frac{1}{rS_0} + \frac{1}{r(S_0 - 1)} + \cdots + \frac{1}{1}.$$

Using the harmonic sum approximation,

$$\tau_{\min} \approx r \log(S_0) + \gamma,$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. To fit the inverse Gaussian distribution, we shift the stopping times by $-\tau_{\min}$.

In the following, we fit the inverse Gaussian distributions to the stopping times of 1000 simulated datasets with the parameters $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.55, 0.15, 0.005, 0.01, 4, 18)$ for the division probabilities, division rate $r = 0.2$, and three different number of initial stem cells $S_0 = 50, 200, 1000$.

For each scenario of initial stem cells,

- The simulated stopping time by $-\tau_{\min}$.
- The mean and shape parameters of the inverse Gaussian distribution are estimated using the MLE (using *fitdist* function in the package *fitdistrplus*.
- Distributional tests (Kolmogorov-Smirnov and Anderson-Darling) to evaluate the shifted simulated stopping times for the inverse Gaussian distribution.
- The shifted simulated data are also fitted for the Gamma and Log-normal distribution. The goodness of fit are compared between the three distributions using the AIC and BIC criteria.

In scenario 1 and 2, with the initial stem cells being 50 and 200, respectively, the distributional tests indicate that the inverse Gaussian distribution is a good fit for the shifted simulated stopping times. Using the AIC and BIC criteria, the fit of the inverse Gaussian distribution and of the log-normal distribution are very compatible, both are good fit for shifted stopping time.

However, in scenario 3 with the initial stem cells being 1000, the distributional tests indicate a poor fit of the inverse Gaussian distribution for the shifted simulated stopping times. Neither the gamma or the log-normal distribution has better fit.

Scenario 1. $S_0 = 50$ When $S_0 = 50$, the minimum stopping times is $\tau_{\min} \approx r \log(50) + \gamma = 20.1373$.

	Estimates	Std. Error
Mean	29.1554	0.2356
Shape	446.1868	19.9544

Table 8: MLE for mean and shape parameters for the inverse Gaussian distribution with initial stem cells of 50 and 1000 replications.

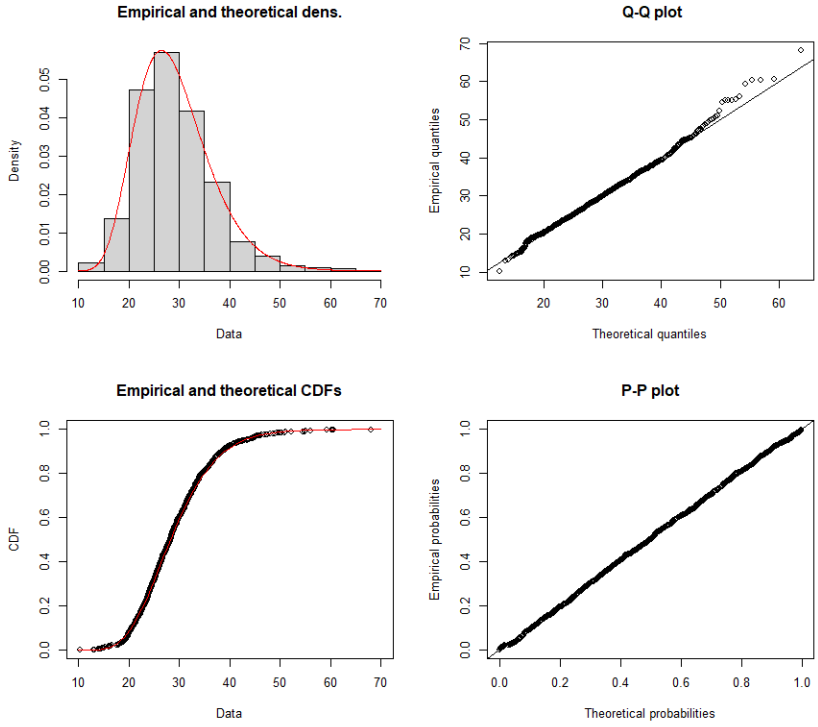

Figure 9: Diagnostic plots for the fit of shifted stopping times (with initial stem cells of 50) for inverse Gaussian distribution.

Test	Test Statistics	P-value
Kolmogorov-Smirnov	0.0189	0.8637
Anderson-Darling	0.5752	0.6717

Table 9: Distributional tests to evaluate the shifted stopping times (with initial stem cells of 50) for inverse Gaussian distribution.

	Inverse Gaussian	Gamma	Log Normal
AIC	6763.453	6778.186	6761.191
BIC	6773.269	6788.001	6771.006

Table 10: Comparing the fit of shifted stopping times (with initial stem cells of 50) with inverse Gaussian, Gamma, and Log-normal distributions.

Scenario 2. $S_0 = 200$ When $S_0 = 200$, the minimum stopping times is $\tau_{\min} \approx r \log(200) + \gamma = 27.0688$.

	Estimates	Std. Error
Mean	30.0482	0.2093
Shape	619.2376	27.6928

Table 11: MLE for mean and shape parameters for the inverse Gaussian distribution with initial stem cells of 200 and 1000 replications.

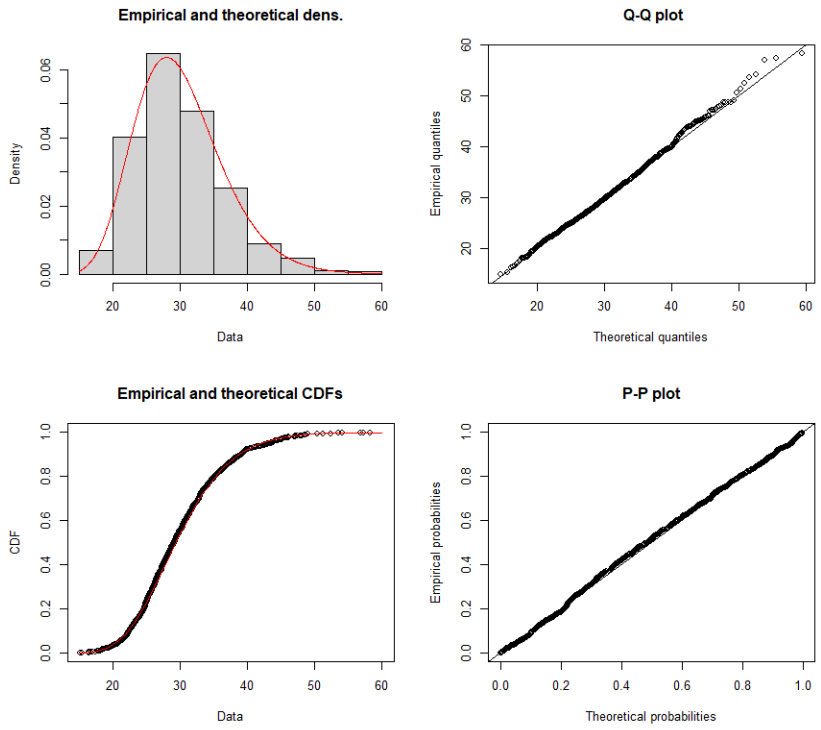

Figure 10: Diagnostic plots for the fit of shifted stopping times (with initial stem cells of 200) for inverse Gaussian distribution.

Test	Test Statistics	P-value
Kolmogorov-Smirnov	0.0233	0.6475
Anderson-Darling	0.8613	0.4388

Table 12: Distributional tests to evaluate the shifted stopping times (with initial stem cells of 200) for inverse Gaussian distribution.

	Inverse Gaussian	Gamma	Log Normal
AIC	6549.471	6568.875	6549.636
BIC	6559.286	6578.691	6559.452

Table 13: Comparing the fit of shifted stopping times (with initial stem cells of 200) with inverse Gaussian, Gamma, and Log-normal distributions.

Scenario 3. $S_0 = 1000$

When $S_0 = 1000$, the minimum stopping times is $\tau_{\min} \approx r \log(1000) + \gamma = 35.1159$.

	Estimates	Std. Error
Mean	31.3807	0.2140
Shape	674.5831	30.1662

Table 14: MLE for mean and shape parameters for the inverse Gaussian distribution with initial stem cells of 1000 and 1000 replications.

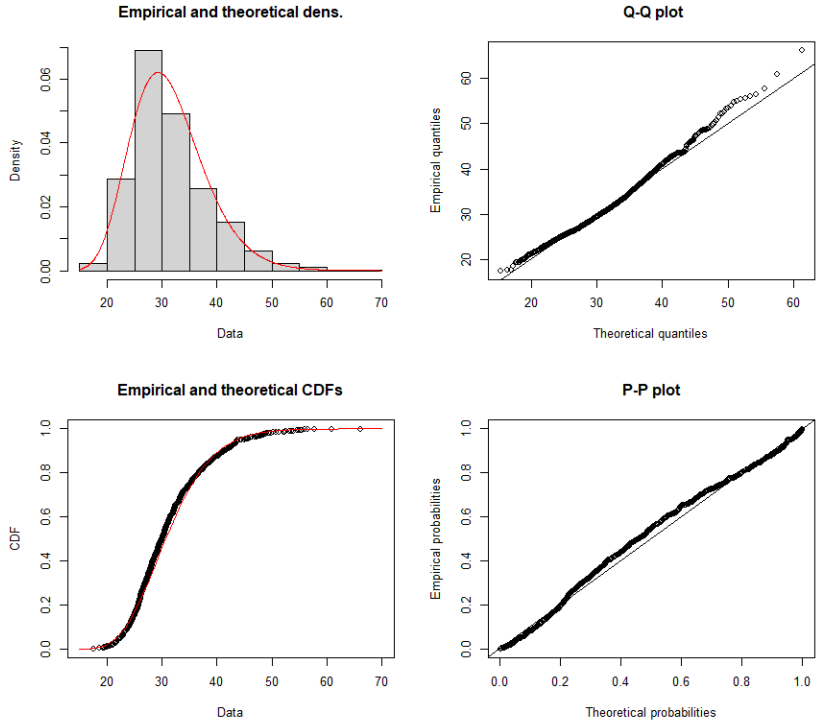

Figure 11: Diagnostic plots for the fit of shifted stopping times (with initial stem cells of 1000) for inverse Gaussian distribution.

Test	Test Statistics	P-value
Kolmogorov-Smirnov	0.0507	0.0114
Anderson-Darling	4.4769	0.0051

Table 15: Distributional tests to evaluate the shifted stopping times (with initial stem cells of 1000) for inverse Gaussian distribution.

	Inverse Gaussian	Gamma	Log Normal
AIC	6596.175	6634.945	6596.631
BIC	6605.991	6644.761	6606.447

Table 16: Comparing the fit of shifted stopping times (with initial stem cells of 1000) with inverse Gaussian, Gamma, and Log-normal distributions.

Verify the autocorrelation formula

For $t > u$,

$$\begin{aligned} E[X(t) \cdot X(u)] &= E[E[X(t) \cdot X(u)|X(u)]] \\ &= E[X(u) \cdot E[(X(t)|X(u)]] \\ &= E[X(u) \cdot X(u) \cdot \exp\{r[P(t) - P(u)]\}] \\ &= E[X(u)^2] \cdot \frac{S(t)}{S(u)}. \end{aligned} \tag{22}$$

From the variance derivation, we have

$$E[X(u)^2] = S(u)^2 \Big[1 + r \int_0^t \frac{p_1(v) + p_3(v)}{S(v)} dv\Big].$$

Then,

$$\begin{aligned} E[X(t) \cdot X(u)] - S(t) \cdot S(u) &= S(t) \cdot S(u) \cdot r \int_0^u \frac{p_1(v) + p_3(v)}{S(v)} dv \\ &= \frac{S(t)}{S(u)} \cdot V(u) \end{aligned} \tag{23}$$

To verify the theoretical result, the process is simulated for $N = 100$ replications with 200 initial stem cells and division rate 0.2 and parameters $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.5, 0.2, 0.005, 0.01, 4, 12)$ of the probability function. The stem cell counts are recorded at time points $t = \{5, 10, 15, 20, 25\}$. The empirical autocovariance are calculated

$$\widehat{\mathbb{E}}[X(t_i)X(t_j)] - \widehat{\mathbb{E}}[X(t_i)]\widehat{\mathbb{E}}[X(t_j)] = \frac{1}{N}\sum_{r=1}^{N} X_r(t_i)\, X_r(t_j) - \frac{1}{N}\sum_{r=1}^{N} X_r(t_i)\, \frac{1}{N}\sum_{r=1}^{N} X_r(t_j). \tag{24}$$

Normalized to autocorrelation, we have the theoretical autocorrelation

$$\mathrm{Corr}(X(t_i), X(t_j)) = \left[\begin{array}{c|ccccc} & 5 & 10 & 15 & 20 & 25 \\ \hline 5 & 1.000 & 0.738 & 0.617 & 0.525 & 0.434 \\ 10 & 0.738 & 1.000 & 0.837 & 0.711 & 0.588 \\ 15 & 0.617 & 0.837 & 1.000 & 0.849 & 0.703 \\ 20 & 0.525 & 0.711 & 0.849 & 1.000 & 0.827 \\ 25 & 0.434 & 0.588 & 0.703 & 0.827 & 1.000 \end{array}\right],$$

and the empirical autocorrelation

$$\widehat{\mathrm{Corr}}(X(t_i), X(t_j)) = \left[\begin{array}{c|ccccc} & 5 & 10 & 15 & 20 & 25 \\ \hline 5 & 1.000 & 0.714 & 0.500 & 0.366 & 0.286 \\ 10 & 0.714 & 1.000 & 0.800 & 0.663 & 0.562 \\ 15 & 0.500 & 0.800 & 1.000 & 0.841 & 0.726 \\ 20 & 0.366 & 0.663 & 0.841 & 1.000 & 0.870 \\ 25 & 0.286 & 0.562 & 0.726 & 0.870 & 1.000 \end{array}\right].$$

To further quantify the similarity, I computed the mean absolute difference

$$\text{MAD} = \frac{1}{T^2} \sum_{i,j} \left| \widehat{\text{Corr}}(X(t_i), X(t_j)) - \text{Corr}(X(t_i), X(t_j)) \right|, \tag{25}$$

and the Pearson correlation between the vectorized empirical and theoretical matrices. Both metrics indicate strong agreement (see Table 17).

Metric	Value
Mean Absolute Difference	0.050
Maximum Absolute Difference	0.159
Matrix Correlation	0.982

Table 17: Quantitative comparison between theoretical and empirical autocorrelation matrices.

Parameter estimates using differential evolution algorithm

Since the estimation result using the gradient-based method is highly sensitive to the starting values, I'm trying different optimization method to find the MLE. The method I'm trying is differential evolution optimization algorithm, which is a method used to find minimum or maximum of a function for when the function is nonlinear, non-differentiable, or has many local optima. It works by iteratively searching for a better solution candiate in comparison to the previous one:

- DE creates a new trial solution candidate by adding the weighted difference between two randomly chosen individuals to a third one.
- This new candidate is then mixed with the original solution candidate through crossover, and the better of the two is kept for the next generation.
- Through repeated mutation, crossover, and selection, the population gradually moves toward the optimal solution.

Table (18) and (19) shows the estimate results for simulated data of two sets of parameters using the differential evolution algorithm with random starting values. The differential evolution algorithm is implemented from the package *DEoptim* in R. The estimates are good in both cases.

The only issue that might arise with implementing this algorithm is the constraints as we need to provide a finite bound for the parameters. For example, in our case, we have the constraint for parameters c_1, c_2, m_1, m_2 to be greater than 0. However, the algorithm requires the parameters to have finite bound, i.e $a_1 < c_1, c_2 < a_2, b_1 < m_1, m_2 < b_2$ for some constants a_1, a_2, b_1, b_2. When implementing this algorithm to obtain the result in table (18) and (19), I use the bounds $0 < c_1, c_2 < 100$ and $0 < m_1, m_2 <$ max(observed time). I will try this optimization method with when we have $p_4 \neq 0$.

Parameter	p_1 (0.55)	p_2 (0.15)	c_1 (0.005)	c_2 (0.01)	m_1 (4)	m_2 (18)
Mean	0.558	0.153	0.00515	0.011	3.780	18.100
Median	0.557	0.152	0.00507	0.010	3.850	18.200
2.5 Percentile	0.504	0.110	0.00300	0.00377	0.377	15.700
97.5 Percentile	0.617	0.193	0.00762	0.021	6.450	20.100

Table 18: Parameters estimate results using differential evolution with 100 replications. The true parameters are $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.55, 0.15, 0.005, 0.01, 4, 18)$.

Parameter	p_1 (0.2)	p_2 (0.7)	c_1 (0.005)	c_2 (0.01)	m_1 (12)	m_2 (6)
Mean	0.202	0.701	0.00543	0.01030	11.200	6.030
Median	0.201	0.699	0.00490	0.01020	11.600	6.080
2.5 Percentile	0.166	0.640	0.000973	0.00675	5.760	4.730
97.5 Percentile	0.247	0.766	0.01150	0.01500	14.700	7.090

Table 19: Parameters estimate results using differential evolution with 100 replications. The true parameters are $(p_1, p_2, c_1, c_2, m_1, m_2) = (0.2, 0.7, 0.005, 0.1, 12, 6)$.

Add p_4 and use differential evolution for maximum likelihood estimates

Define the time-dependent probability $p_4(\cdot)$ similar to $p_1(\cdot)$ and $p_2(\cdot)$

$$p_4(t) = \frac{p_4}{1 + c_4(t - m_4)^2}.$$

Then the log-likelihood is

$$\begin{aligned}
&\ell(T_1, \cdots, T_n, S_0, \cdots, S_n, D_0, \cdots D_n) \\
&= \sum_{i=1}^{k} \Big[\log\Big(\frac{p_1}{1 + c_1(T_i - m_1)^2}\Big) I_{(X_i=1, Z_i=0)} \\
&+ \log\Big(\frac{p_2}{1 + c_2(T_i - m_2)^2}\Big) I_{(X_i=0, Z_i=0)} \\
&+ \log\Big(\frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(X_i=0, Z_i=1)} \\
&+ \log\Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} - \frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(X_i=-1, Z_i=0)} \\
&+ \log r + \log S_{i-1} - r S_{i-1} \Delta_{T_1} \Big].
\end{aligned} \tag{26}$$

Since the MLE of r has the closed form solution $\hat{r} = \frac{n}{\sum_{i=1}^{n} S_{i-1}\Delta_{T_i}}$, I applied the differential evolution optimization method to find the maximum likelihood estimates for the parameter of the probability functions $p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4$ by maximizing

$$\begin{aligned}
A = \sum_{i=1}^{k} &\Big[\log\Big(\frac{p_1}{1 + c_1(T_i - m_1)^2}\Big) I_{(X_i=1, Z_i=0)} \\
&+ \log\Big(\frac{p_2}{1 + c_2(T_i - m_2)^2}\Big) I_{(X_i=0, Z_i=0)} \\
&+ \log\Big(\frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(X_i=0, Z_i=1)} \\
&+ \log\Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} - \frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(X_i=-1, Z_i=0)}].
\end{aligned}$$

Table (20) and (21) show the estimate results for simulated data of two sets of parameters using differential evolution algorithm with random starting values. In both cases the estimates are good. However, compare to the when we only have parameters for $p_1(\cdot)$ and $p_2(\cdot)$ (i.e $p_4(t) = 0\ \forall t$), the number of iterations required is much larger and thus results in longer computational time. It took between 30-40 minutes to obtain the each result in table (20) and (21).

Parameter	p_1 (0.55)	p_2 (0.20)	p_4(0.15)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (4)	m_2 (18)	m_4 (28)
Mean	0.552	0.203	0.151	0.00528	0.0127	0.00876	4.070	17.900	28.200
Median	0.551	0.202	0.152	0.00497	0.0122	0.00842	4.120	17.900	28.000
2.5 Percentile	0.505	0.167	0.108	0.00333	0.00618	0.00302	0.914	16.200	25.100
97.5 Percentile	0.604	0.245	0.202	0.00834	0.0208	0.0156	6.450	19.700	32.300

Table 20: Parameters estimate results using differential evolution with 100 replications. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m4) = (0.55, 0.20, 0.15, 0.005, 0.012, 0.008, 4, 18, 28)$.

Parameter	p_1 (0.20)	p_2 (0.70)	p_4(0.10)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (12)	m_2 (6)	m_4 (28)
Mean	0.205	0.704	0.112	0.00594	0.0122	0.0122	11.800	5.960	27.600
Median	0.206	0.705	0.102	0.00541	0.0125	0.00979	11.800	5.970	26.700
2.5 Percentile	0.163	0.649	0.049	0.00195	0.00831	0.00265	8.690	4.860	20.000
97.5 Percentile	0.250	0.773	0.208	0.0120	0.0165	0.0348	14.800	7.050	39.800

Table 21: Parameters estimate results using differential evolution with 100 replications. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m4) = (0.20, 0.70, 0.10, 0.005, 0.012, 0.008, 12, 6, 28)$.

Estimation for observing data up until predetermined time t^*.

Suppose we observe the events (both event time and division types) up until time t^*. The likelihood of observing the data up until this time point is

$$\begin{aligned}
&L_{t^*}(\boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{Z}) \\
&= \prod_{i=1}^{n} \Big[\frac{p_1}{1 + c_1(T_i - m_1)^2} I_{(\Delta X_i = 1, \Delta Z_i = 0)} \\
&+ \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(\Delta X_i = 0, \Delta Z_i = 0)} \\
&+ \frac{p_4}{1 + c_4(T_i - m_4)^2} I_{(\Delta X_i = 0, \Delta Z_i = 1)} \\
&+ \Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} - \frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(\Delta X_i = -1, \Delta Z_i = 0)} \Big] \\
&\cdot r X_{i-1} e^{-r X_{i-1} \Delta T_i} \cdot e^{-r X_n (t^* - T_n)}.
\end{aligned} \tag{27}$$

Then the log-likelihood

$$\begin{aligned}
&\ell_{t^*}(\boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{Z}) \\
&= \sum_{i=1}^{n} \Big[\log \frac{p_1}{1 + c_1(T_i - m_1)^2} I_{(\Delta X_i = 1, \Delta Z_i = 0)} \\
&+ \log \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(\Delta X_i = 0, \Delta Z_i = 0)} \\
&+ \log \frac{p_4}{1 + c_4(T_i - m_4)^2} I_{(\Delta X_i = 0, \Delta Z_i = 1)} \\
&+ \log \Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} - \frac{p_4}{1 + c_4(T_i - m_4)^2}\Big) I_{(\Delta X_i = -1, \Delta Z_i = 0)} \Big] \\
&+ \log r + \log X_{i-1} - r X_{i-1} \Delta T_i - r X_n (t^* - T_n).
\end{aligned} \tag{28}$$

Then the MLE of $\hat{r}$ has the closed-form

$$\hat{r} = \frac{n}{\sum_{i=1}^{n} X_{i-1} \Delta T_i + X_n (t^* - T_n)}.$$

The estimates of $p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4$ are obtained by the maximimizing the quantity

$$\begin{aligned}
A = \sum_{i=1}^{n} \Big[&\log \frac{p_1}{1 + c_1(T_i - m_1)^2} I_{(\Delta X_i = 1, \Delta Z_i = 0)} \\
&+ \log \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(\Delta X_i = 0, \Delta Z_i = 0)} \\
&+ \log \frac{p_4}{1 + c_4(T_i - m_4)^2} I_{(\Delta X_i = 0, \Delta Z_i = 1)} \\
&+ \log \Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2}\Big) I_{(\Delta X_i = -1, \Delta Z_i = 0)} \Big].
\end{aligned}$$

Essentially, I'm doing the same optimization as before with less data (only using the data up until the time t^*).

Table (22) through (25) display the parameter estimate results of the same simulated data with different maximum observation times $t^* = 40, 30, 20, 10$. The estimates for r are good in all cases. When $t^* = 40$ and $t^* = 30$ (table (22) and (23, respectively), the estimates are still good. However, the estimates for the probability parameters get worse as the maximum observation time gets smaller (when $t^* = 20$ and $t^* = 10$ in

table (24) and (25), respectively).

It's also interesting to note that when $t^* = 20$, the estimates for $p_1, p_2, c_1, c_2, m_1, m_2$ are still good and it coincides with the fact that the true values $m_1, m_2 < t^*$ while the true value of $m_4 > t^*$ in this case. I suspect that with $t^* = 20$, there were not enough events for $p_4(\cdot)$ observed to estimates the parameters of $p_4(\cdot)$. Similarly, when $t^* = 10$, the estimates for $p_1(\cdot)$ is still pretty good but those for $p_2(\cdot)$ and $p_4(\cdot)$ are not, and when $t^* = 10$, the true value $m_1 < t^*$ and the true values of $m_2, m_4 > t^*$.

Parameter	p_1 (0.55)	p_2 (0.15)	p_4(0.20)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (4)	m_2 (18)	m_4 (28)
Mean	0.556	0.156	0.207	0.00535	0.0138	0.00849	3.960	18.000	28.600
Median	0.557	0.155	0.204	0.00527	0.0139	0.00785	4.140	18.000	28.100
2.5 Percentile	0.500	0.123	0.162	0.00331	0.00657	0.00313	0.425	16.100	25.500
97.5 Percentile	0.609	0.190	0.263	0.00805	0.0231	0.0149	6.460	20.300	33.900

Parameter	r (0.2)
Mean	0.199
Median	0.199
2.5 Percentile	0.189
97.5 Percentile	0.209

Table 22: Parameter estimates using data up until time point $t^* = 40$. The estimates of the probability parameters p_i, c_i, m_i are obtained by the optimize the quantity A with differential evolution and the estimates of r are obtained by the closed-form solution. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r) = (0.55, 0.15, 0.20, 0.005, 0.012, 0.008, 4, 18, 28, 0.2)$.

Parameter	p_1 (0.55)	p_2 (0.15)	p_4(0.20)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (4)	m_2 (18)	m_4 (28)
Mean	0.556	0.158	0.265	0.00536	0.0142	0.00959	3.962	17.900	29.900
Median	0.560	0.156	0.218	0.00534	0.0139	0.00899	4.340	17.800	27.300
2.5 Percentile	0.497	0.125	0.149	0.00332	0.00610	0.00381	0.296	15.800	24.200
97.5 Percentile	0.606	0.192	0.840	0.00840	0.0259	0.0171	6.470	20.400	47.800

Parameter	r (0.2)
Mean	0.199
Median	0.199
2.5 Percentile	0.188
97.5 Percentile	0.209

Table 23: Parameter estimates using data up until time point $t^* = 30$. The estimates of the probability parameters p_i, c_i, m_i are obtained by the optimize the quantity A with differential evolution and the estimates of r are obtained by the closed-form solution. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r) = (0.55, 0.15, 0.20, 0.005, 0.012, 0.008, 4, 18, 28, 0.2)$.

Parameter	p_1 (0.55)	p_2 (0.15)	p_4(0.20)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (4)	m_2 (18)	m_4 (28)
Mean	0.559	0.236	0.575	0.00579	0.0189	0.0259	4.120	19.400	28.700
Median	0.560	0.167	0.670	0.00563	0.0174	0.0258	4.230	17.300	29.900
2.5 Percentile	0.500	0.119	0.0736	0.00292	0.00527	0.00367	0.827	14.200	15.600
97.5 Percentile	0.610	0.908	0.990	0.00928	0.0416	0.0593	6.600	33.000	40.900

Parameter	r (0.2)
Mean	0.199
Median	0.199
2.5 Percentile	0.188
97.5 Percentile	0.211

Table 24: Parameter estimates using data up until time point $t^* = 20$. The estimates of the probability parameters p_i, c_i, m_i are obtained by the optimize the quantity A with differential evolution and the estimates of r are obtained by the closed-form solution. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r) = (0.55, 0.15, 0.20, 0.005, 0.012, 0.008, 4, 18, 28, 0.2)$.

Parameter	p_1 (0.55)	p_2 (0.15)	p_4(0.20)	c_1 (0.005)	c_2 (0.012)	c_4 (0.008)	m_1 (4)	m_2 (18)	m_4 (28)
Mean	0.572	0.490	0.450	0.00815	0.125	0.139	4.030	14.900	13.800
Median	0.572	0.537	0.465	0.00701	0.0914	0.0837	4.140	15.300	14.800
2.5 Percentile	0.502	0.0629	0.0470	0.00077	0.00544	0.00769	0.386	4.260	0.898
97.5 Percentile	0.644	0.983	0.975	0.0197	0.375	0.577	8.440	26.700	25.800

Parameter	r (0.2)
Mean	0.200
Median	0.201
2.5 Percentile	0.186
97.5 Percentile	0.214

Table 25: Parameter estimates using data up until time point $t^* = 10$. The estimates of the probability parameters p_i, c_i, m_i are obtained by the optimize the quantity A with differential evolution and the estimates of r are obtained by the closed-form solution. The true parameters are $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r) = (0.55, 0.15, 0.20, 0.005, 0.012, 0.008, 4, 18, 28, 0.2)$.

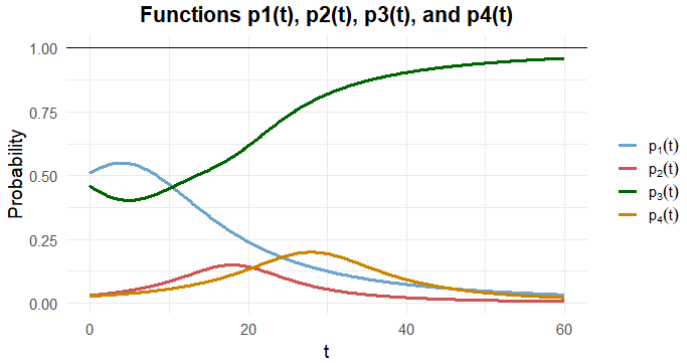

Figure 12: Functions $p_1(t), p_2(t), p_3(t), p_4(t)$ with parameters $(p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r) = (0.55, 0.15, 0.20, 0.005, 0.012, 0.008, 4, 18, 28, 0.2)$

Forward algorithm for when viable and non-viable stem cells are not observed

Let $(X_{t_i}, Y_{t_i} Z_{t_i})$ denote numbers of viable stem cells, differentiated cells, and nonviable stem cells. Assume that we can not differentiate between viable and non-viable stem cells, we have the hidden states

$$V_{t_i} = (X_{t_i}, Y_{t_i}, Z_{t_i})$$

we can only observe

$$M_{t_i} = X_{t_i} + Z_{t_i} \text{ and } Y_{t_i}.$$

Then, the probability of observing (M_{t_i}, Y_{t_i}) given the hidden states $V_{t_i} = (X_{t_i}, Y_{t_i}, Z_{t_i})$ (emission probability) is

$$P((M_{t_i}, Y_{t_i})|V_{t_i}) = I_{(M_{t_i}=X_{t_i}+Z_{t_i})}.$$

Each hidden state $v = (x, y, z)$ can transition to a new states $v' = (x', y', z')$ through one of the four events. The

Event k	Probability	ΔX	ΔY	ΔZ
1	$p_1(t_i)$	+1	0	0
2	$p_2(t_i)$	0	+1	0
3	$1 - p_1(t_i) - p_2(t_i) - p_4(t_i)$	-1	2	0
4	$p_4(t_i)$	0	0	+1

Table 26: Division event type and resulting change of cell counts.

transition probability of transitioning from hidden state $v = (x, y, z)$ to the new hidden state $v' = (x', y', z')$ is

$$P(V_{t_i} = v'|V_{t_{i-1}} = v) = rxe^{-rx\Delta t_i} \sum_{k=1}^{4} p_k(t_i) I_{(v'=v+\Delta V_k)}.$$

We want the likelihood of observed data sequence of cell counts and of even time

$$P((M_{T_1}, Y_{T_1}), (M_{T_2}, Y_{T_2}), \cdots, (M_{T_n}, Y_{T_n}), T_1, \cdots T_n).$$

We define the forward variable $\alpha_i(v)$ as the probability of being in hidden state $v = (x, y, z)$ after i-th event, and having seen all previous observation. Then, the forward recursion is

$$\alpha_i(v') = \sum_{v \in \mathcal{V}} \alpha_{i-1}(v) P(V_{t_i} = v'|V_{t_{i-1}} = v) P((M_{t_i}, Y_{t_i})|V_{t_i}). \tag{29}$$

Number of hidden states exploding

I'm trying to figure out how to address the issue of number of hidden states exploding as the number of division increases. For example,

- after the first division, there are three possible hidden states with number of stem cells as $X_{t_1} = \{S_0 - 1, S_0, S_0 + 1\}$
- after the second division, there are five possible hidden states with number of stem cells $X_{t_2} = \{S_0 - 2, S_0 - 1, S_0, S_0 + 1, S_0 + 2\}$.

However, since we can observe the number of differentiated cells Y_{t_i} and how it changes, we know some of these hidden states are not possible. For example, after the first division

- if $\Delta Y_{t_1} = 1$, then $X_{t_1} = S_0$,
- if $\Delta Y_{t_1} = 2$, then $X_{t_1} = S_0 + 2$,

- $\Delta Y_{t_1} = 0$, then $X_{t_1} = \{S_0, S_1\}$.

Maybe this could be used to eliminate some of the possible hidden states.

Another consideration is that the hidden states directly affect the interarrival times since the rate of arrival times depends on the number of viable stem cells, but the probability of event types given that we observe the interarrival times does not. In other word, given the event times T_i's, the conditional likelihood of observing the cell count change is

$$\begin{aligned} L(M_{t_i}, Y_{t_i}|T_i) = \prod_{i=1}^{n} \Big[\Big(& \frac{p_1}{1 + c_1(T_i - m_1)^2} + \frac{p_4}{1 + c_4(T_i - m_4)^2} \Big) I_{(\Delta M_i = 1, \Delta Y_i = 0)} \\ & + \frac{p_2}{1 + c_2(T_i - m_2)^2} I_{(\Delta M_i = 0, \Delta Y_i = 1)} \\ & + \Big(1 - \frac{p_1}{1 + c_1(T_i - m_1)^2} - \frac{p_2}{1 + c_2(T_i - m_2)^2} - \frac{p_4}{1 + c_4(T_i - m_4)^2} \Big) I_{(\Delta M_i = -1, \Delta Y_i = 2)} \Big]. \end{aligned}$$

I'm not quite sure if this could be useful to simplify the expression/calculation of the forward variable yet.

Forward algorithm

Let (X_i, Y_i, Z_i) denote viable stem cells, differentiated cells, and non-viable stem cells and T_i are event times. Let M_i denote the sum of the viable and non-viable stem cells. Note that $(M_i, Y_i, T_i),\ k = 1, 2, \cdots, n$ are observable and

$$T_0 = 0, M_0 = S_0, Y_0 = 0, Z_0 = 0.$$

Forward variables (for $k = 1, \cdots, n$)

$$\alpha_k(u) = \ell(T_1, \cdots, T_k, Y_1, \cdots Y_k, M_1, \cdots M_k, Z_k = u),$$

where u is an integer, $0 \le z \le M_k$. Then for $0 \le k \le n-1$,

$$\alpha_{k+1}(v) = \sum_{u \in \{v, v-1\}} \alpha_k(u) \cdot h_k(u, v),$$

where $0 \le v \le M_{k+1}$. Here $h_k(u, v)$ are transitional "probabilities, and for $i \le m \le 4$

$$h_k(u, v) = \begin{cases} h_{mk}(u, v), & \text{event } m \text{ at } T_{k+1} \\ 0, & \text{otherwise,} \end{cases}$$

$h_{mk}(u, v) = r(M_k - u)e^{-r(M_k - u)(T_{k+1} - T_k)} \cdot p_m(T_{k+1})$.

The events in terms of (Y, M, Z)

- Event 1: $Y_{k+1} = Y_k, M_{k+1} = M_k + 1, v = u$
- Event 2: $Y_{k+1} = Y_k + 1, M_{k+1} = M_k, v = u$
- Event 3: $Y_{k+1} = Y_k + 2, M_{k+1} = M_k - 1, v = u$
- Event 4: $Y_{k+1} = Y_k, M_{k+1} = M_k + 1, v = u + 1$

Finally, the likelihood is given by

$$L_n = \sum_{0 \le u \le M_n} \alpha_n(u).$$

Normalization

Let

$$L_k = \sum_{0 \le u \le M_k} \alpha_k(u), \quad k = 1, \cdots, n.$$

Consider $\bar{\alpha}_k(v) = \frac{\alpha_k(v)}{L_k}$ then

$$\begin{aligned} \bar{\alpha}_{k+1}(v) &= \frac{\alpha_{k+1}(v)}{L_{k+1}} \\ &= \frac{1}{L_{k+1}} \sum_{u = v-1, v} \alpha_k(u) h_k(u, v) \\ &= \frac{L_k}{L_{k+1}} \sum_{u = v-1, v} \bar{\alpha}_k h_k(u, v). \end{aligned} \tag{30}$$

Now for $0 \leq k \leq n-1$ let

$$\begin{aligned} d_{k+1} &= \frac{L_{k+1}}{L_k} \\ &= \frac{1}{L_k} \sum_{0 \leq v \leq M_{k+1}} \alpha_{k+1}(v) \\ &= \frac{1}{L_k} \sum_{0 \leq v \leq M_{k+1}} \sum_{u=v-1,v} \alpha_k(u) h_k(u,v) \\ &= \sum_{0 \leq v \leq M_{k+1}} \sum_{u=v-1,v} \bar{\alpha_k}(u) h_k(u,v) \end{aligned} \tag{31}$$

So, $\bar{\alpha}_0(0) = 1$ and $\bar{\alpha}_k \to d_{k+1} \to \bar{\alpha}_{k+1}$.

$$\log(L_n) = \sum_{k=0}^{n-1} \log(d_{k+1})$$

Code implementation

Algorithm 1 Forward Log-Likelihood Computation

Extract M, Y, t from data
Extract parameters $p_1, p_2, p_4, c_1, c_2, c_4, m_1, m_2, m_4, r$
Initialize hidden states: $H \leftarrow \{0\}$
$\alpha_{\text{prev}} \leftarrow \mathbf{1}$ (vector of ones)
$n \leftarrow \text{length}(t)$
scale[1] $\leftarrow 1$
for $k = 2$ to n **do**
  $\Delta T = t_k - t_{k-1}$
  $\Delta Y = Y_k - Y_{k-1}$
  $\Delta M = M_k - M_{k-1}$
  Compute time–varying probabilities:

$$p_{1,t} = \frac{p_1}{1 + c_1(t_k - m_1)^2}, \quad p_{2,t} = \frac{p_2}{1 + c_2(t_k - m_2)^2}, \quad p_{4,t} = \frac{p_4}{1 + c_4(t_k - m_4)^2}$$

  if $\Delta Y = 1$ and $\Delta M = 0$ **then**
    Construct diagonal transition matrix T
    for each state $h \in H$ **do**
      $\lambda = r(M_{k-1} - h)$
      $T_{hh} = f_{\exp}(\Delta T; \lambda) \cdot p_{2,t}$
    end for
    $\alpha_{\text{new}} = \alpha_{\text{prev}} T$
  else if $\Delta Y = 2$ and $\Delta M = -1$ **then**
    $p_{3,t} = \max(1 - p_{1,t} - p_{2,t} - p_{4,t}, \varepsilon)$
    Construct diagonal transition matrix T
    for each state $h \in H$ **do**
      $\lambda = r(M_{k-1} - h)$
      $T_{hh} = f_{\exp}(\Delta T; \lambda) \cdot p_{3,t}$
    end for
    $\alpha_{\text{new}} = \alpha_{\text{prev}} T$
    Remove states in H exceeding M_k
    Truncate α_{new} accordingly
  else if $\Delta Y = 0$ and $\Delta M = 1$ **then**
    Construct transition matrix T with one extra column
    for each state $h \in H$ **do**
      $\lambda = r(M_{k-1} - h)$
      $T_{h,h} = f_{\exp}(\Delta T; \lambda) \cdot p_{1,t}$
      $T_{h,h+1} = f_{\exp}(\Delta T; \lambda) \cdot p_{4,t}$
    end for
    $\alpha_{\text{new}} = \alpha_{\text{prev}} T$
    Add new state to H if $\leq M_k$
    Truncate α_{new} to match H
  end if
  $\alpha_\Sigma = \sum_i \alpha_{\text{new}}[i]$
  if $\alpha_\Sigma \leq \varepsilon$ **then**
    $\alpha_\Sigma \leftarrow \varepsilon$
  end if
  scale[k] = α_Σ/scale[$k-1$]
  $\alpha_{\text{prev}} = \alpha_{\text{new}}/\alpha_\Sigma$
end for
Compute log-likelihood:

$$\ell = \sum_{k=1}^{n} \log(\text{scale}[k])$$