# Statistical Inference in a Two-Compartment Model for Hematopoiesis by S. Catlin, J. Abkowitz, P. Guttorp, 2001

Jason Xu

**Abstract**

This report examines the estimating equations approach presented by Catlin et al. [5] for inference in a stochastic two compartment model for hematopoiesis. We review previous work on this class of models and other attempts at inference in this setting. We then detail the methods presented in this paper, and reproduce the authors' results applied to experimental data, with validation via simulation. We conclude by assessing contributions and shortcomings of the proposed method.

## 1 Introduction and Literature Review

Hematopoiesis is the process of specialization of stem cells into mature blood cells. Understanding the behavior of stem cells during this multi-stage process is biologically significant and has many clinical applications, including cancer and gene therapies. Hematopoietic stem cells (HSCs) may self-renew or differentiate to become more specialized progenitor cells, and while the further development of progenitor cells is well-studied, relatively little is known about early stage HSC behavior due to identifiability issues. HSCs are impossible to distinguish and observe directly, but an experimental method developed by Abkowitz [1] allows for analysis and inference of HSC behavior. The experimental design leads to inferential challenges, but by exploiting a binary marker on the cells and observing relative proportions in a two compartment model, we may indirectly study behavior in the first compartment through samples from the second.

### 1.1 Stochastic compartmental models

A stochastic compartmental model is a type of Markov population process with state space represented by a vector $\{X_1(t), \ldots, X_s(t)\}$, where $X_i(t)$ denotes the number of individuals in compartment $i$ at time $t$. By the Markov assumption, each compartment evolves in a memoryless way: formally, given a sequence of times $s_0 < s_1 \ldots < s_n < s$ and possible states $x_j$, for any compartment $i$, we have

$$P(X_i(t+s) = y | X_i(s) = x_s, X_i(s_n) = x_n, \ldots, X_i(s_0) = x_0) = P(X_i(t+s) = y | X_i(s) = x_s).$$

Individuals can move between compartments, reproduce within compartments, or "die" and leave the system. These compartmental models have been widely used to model different phenomena, particularly in the field of biology– one such example is the SIR model for describing the spread of disease.

Stochastic compartment models that include birth are also referred to as interconnected birth-death models or stepping-stone models in the literature. Puri studied systems of finitely many interconnected linear birth-death processes, and solved the backward equations for the probability generating function for several simple cases [9]. However, even the simplified settings yield complicated solutions, from which it is infeasible to derive transition probabilities of the process. Similarly, Renshaw solved a generalization that includes immigration from outside of the system [11]. The solution is rather unwieldy and results cannot be applied to the model in the present paper [5], as it does not allow immigration from outside of the system.

The vast majority of the literature on interconnected population processes concerns completely observed processes [4]. It is often of interest to obtain estimates of the rates of the process in a setting where observation of the complete process is not feasible. When the underlying process is Markovian, the partially observed process becomes a hidden Markov model: formally, if $X_t$ is a Markov process, and $Y_t$ is a deterministic or stochastic function of $X_t$, then $Y_t$ is a hidden Markov model.

### 1.2 A two compartment model for hematopoiesis

Various experimental techniques support that a stochastic model for hematopoiesis is biologically reasonable; Ogawa provides an overview [8]. The first application of a stochastic compartmental model to the study of hematopoiesis was proposed in 1964 by Till et al [10]. A series of modifications were introduced by Abkowitz et al. [1], Newton et al. [7], and refined to the two-compartment model proposed by Abkowitz et al. [2] used in this paper. The current model contains a reserve compartment where HSCs reside, and a contributing compartment where progenitor cells reside. The first compartment is a linear birth-death process: HSCs have birth rate $\lambda$ corresponding to self-renewal, or death rate $\nu$ corresponding to differentiation into the contributing compartment. Progenitors in the contributing compartment can only further differentiate at rate $\mu$, thereby leaving the system. The contributing compartment is thus a non-homogeneous immigration-death process. Because HSCs are not identifiable in the reserve compartment, observations are available only from the second compartment.

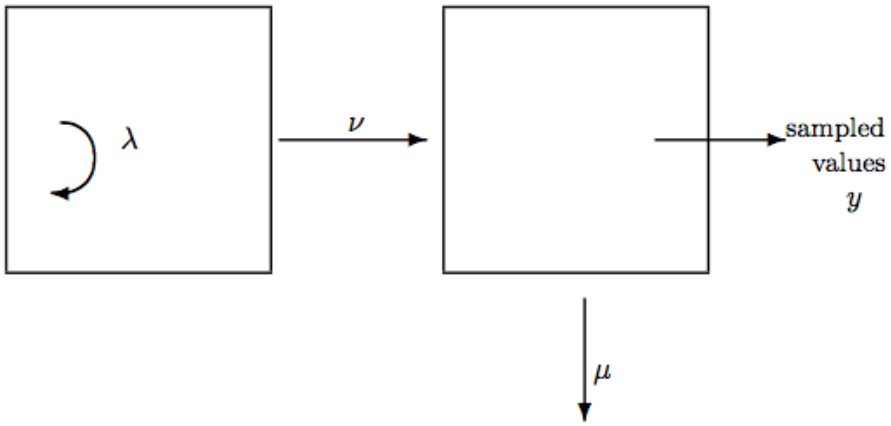


Figure 1: Diagram of the two-compartment model.

### 1.3 Experimental studies and inferential attempts

To study HSC behavior, Abkowitz et al. designed an experiment on female Safari cats, which have distinct phenotypes of the G6PD enzyme [2]. This phenotype is expressed as either $d$ or $G$ during embryogenesis, and is retained through the differentiation process. Because it is neutral to differentiation, it provides a binary marker on each cell and its clones. Samples are taken from the cats roughly every two to six weeks when possible, and the proportion of the label $d$ among progenitor cells in the contributing compartment is observed at these discrete times. Studying the relative number of progenitor cells marked $d$ will allow for inference of the migration of unobserved HSCs into the second compartment.

Previous studies on normal cats suggested that hematopoiesis is a fairly stable process, and insufficient variation in the proportion of the $d$ cells over time yielded little information about stem cell behavior [7]. This led researchers to believe that more information might be provided when the process is supported by fewer HSCs. In this dataset, transplant cats are initially irradiated to kill bone marrow where HSCs reside, and a small amount of their own marrow cells collected prior to radiation are transplanted back. Doing so allows for more variability over time in the percentage of marked $d$ cells.

Previous attempts to obtain transition probabilities for this model have been unsuccessful [1]. Approximation of the transition probabilities, such as normal approximations, are inaccurate, particularly at earlier time points when the population is small [4]. A simulation study by Abokwitz et al. searched the parameter space systematically, simulating realizations of the hidden process

and evaluating results compared to the observed data based on various criteria. This method becomes infeasible when population sizes become too large, and has been criticized for being rather informal with somewhat subjective criteria [2]. A later study by Golinelli et al. employs a Bayesian approach [6], using stochastic integration via RJMCMC to estimate the posterior distribution of the parameters. The Bayesian integration approach method makes more efficient use of the data at a large computational tradeoff. However, it does provide the most precise estimates, and can be applied even when data from only one realization are available.

# 2 Methods

In the current paper by Catlin et al. [5], a method of moments approach is proposed as an alternative to exact likelihood methods. Before detailing this approach, we discuss the difficulties of inference from a likelihood perspective.

## 2.1 Inference for partially observed processes

We may formally denote the process by a vector $\{R(t), C(t)\} = \{R_d(t), R_G(t), C_d(t), C_G(t)\}$, where $R(t)$ represents the size of the reserve compartment at time $t$ and $C(t)$ represents the size of the contributing compartment. Each is two-dimensional, where the subscripts denote the numbers of cells expressing each phenotype. Estimation of the parameters $\lambda, \nu, \mu$ in this model when it is continuously observed in a time interval $[0, T]$ is straightforward: the full likelihood can be written up to proportionality as

$$L(\lambda, \nu, \mu) \propto \lambda^{B_T} \nu^{M_T} \mu^{D_T} \exp\{-(\lambda + \nu) S_T^1 - \mu S_T^2\}$$

where $B_T, M_T, D_T$ denote the total number of births, migrations, and deaths in $[0, T]$, and $S_T^i$ is the total dwell time in compartment $i$. Then $(B_T, M_T, D_T, S_T^1, S_T^2)$ is the minimal sufficient statistic, and the maximum likelihood estimators $\hat{\lambda} = B_T/S_T^1$, $\hat{\nu} = M_T/S_T^1$, and $\hat{\mu} = D_T/S_T^2$ are readily available. Further, the MLE's have good asymptotic properties conditional upon non-extinction of the process, and estimation becomes a straightforward extension of continuous birth-death process theory; this is discussed in more detail by Catlin [4].

In the current experimental setting, the process is not observed continuously in time, and further, samples are only obtainable from the second compartment. Thus we have two sources of missing information. If $Y(t)$ is a random variable representing the total number of cells expressing $d$

at time $t$, then $Y(t)$ defines a hidden Markov model; specifically, observations are assumed binomial so that

$$[Y(t)|(R(t), C(t))] \sim Binom\left(n(t), \frac{C_d(t)}{C_d(t) + C_G(t)}\right)$$

where $n(t)$ is the total number of cells in the sample at time $t$. As mentioned earlier, analyzing this binomial proportion is mathematically difficult. Further, the hidden process is neither stationary nor irreducible and has a countably infinite state space, rendering standard methods in the literature such as Baum-Welch inapplicable. Determining the likelihood or posterior such as in the study by Golinelli et al. [6] requires a very large integration step: if $\theta = (\lambda, \nu, \mu)$,

$$L(\theta) = \pi(y|\theta) = \int_{w \in W} \pi(y|w, \theta)\pi(w|\theta)dw$$

where $W$ is all possible realizations in $[0, T]$. That is, we must sum over all possible sample paths between observation times and over possible sizes of the process at observation times.

### 2.2 Deriving the moments

We can write down the transition probabilities over an infinitesimally short time period $h$ for the hidden component of the process, given the state at time $t$, as follows:

$$\begin{aligned}
&P\{R(t+h) = r+1 | R(t) = r\} = \lambda r h + o(h),\\
&P\{R(t+h) = r-1, C(t+h) = c+1 | R(t) = r, C(t) = c\} = \nu r h + o(h),\\
&P\{C(t+h) = c-1 | C(t) = c\} = \mu c h + o(h).
\end{aligned}$$

Here we have dropped the subscripts $d$ and $G$ for notational convenience since they have identical expressions for these rates and are assumed to behave independently.

To calculate the means and variances $\{m_R(t), m_C(t), V_R(t), V_C(t)\}$, we follow a technique detailed by Bailey [3]. Bailey derives a general procedure to obtain the partial differential equation for a generating function of a continuous Markov jump process. In our case, we apply this "random-variable" technique to the cumulant generating function, since the first two cumulants correspond to the desired mean and variance.

On page 118, Bailey derives the general form of the forward equation PDE for the cumulant generating function $K(\theta_1, \theta_2, t)$, where $\theta_i$ are dummy variables corresponding to each compartment:

$$\frac{dK(\theta_1, \theta_2; t)}{dt} = \sum_{j,k} (e^{j\theta_1 + k\theta_2} - 1) f_{jk}\left(\frac{d}{d\theta_1}, \frac{d}{d\theta_2}\right) K(\theta_1, \theta_2; t)$$

The terms $f_{jk}$ are defined such that

$$P\{X(t+h)=j, Y(t+h)=k|X(t),Y(t)\} = f_{jk}(X,Y)h.$$

In our case, the only nonzero $f_{jk}$ are given by $f_{1,0}=\lambda x$, $f_{-1,1}=\nu x$, and $f_{0,1}=\mu y$. Plugging in, we arrive at the Kolmogorov forward equation for $K(\theta_1,\theta_2,t)$ given by

$$\frac{dK(\theta_1,\theta_2;t)}{dt} = [\lambda(e_1^{\theta}-1)+\nu(e^{-\theta_1+\theta_2}-1)]\frac{dK}{d\theta_1}+\mu(e^{-\theta_2}-1)\frac{dK}{d\theta_2} \tag{1}$$

To obtain the moments from this expression, we expand the cumulant generating function

$$K(\theta_1,\theta_2,t)=\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\frac{\kappa_{uv}(t)\theta_1^u\theta_2^v}{u!v!} \tag{2}$$

and take partial derivatives of this expansion as well as corresponding partial derivatives of (1). Equating coefficients of the $\theta_1^u\theta_2^v$ terms and evaluating at $t=0$ now yields a system of equations for the moments:

$$\begin{aligned}
dm_R(t)/dt &= (\lambda-\nu)m_R(t),\\
dm_C(t)/dt &= \nu m_R(t)-\mu m_C(t),\\
dV_R(t)/dt &= (\lambda+\nu)m_R(t)+2(\lambda-\nu)V_R(t),\\
dV_{RC}(t)/dt &= -\nu m_R(t)+\nu V_R(t)+(\lambda-\nu-\mu)V_{RC}(t),\\
dV_C(t)/dt &= \nu m_R(t)+\mu m_C(t)+2\nu V_{RC}(t)-2\mu V_C(t).
\end{aligned}$$

Thus, we have reduced to a system of first-order linear differential equations. Letting $r_0$ and $c_0$ denote the initial number of cells in the reserve and contributing compartments respectively, this system is solvable given initial conditions $m_R(0)=r_0, m_C(0)=c_0$, and $V_R(0)=V_C(0)=V_{RC}(0)=0$. Notice that the mean and variance of the reserve compartment are identical to the equations for a linear birth-death process, as we would expect. The relevant solutions for the contributing compartment are given by

$$m_C(t)=r_0\left\{\frac{\nu}{(\lambda-\nu+\mu)}(\exp\{(\lambda-\nu)t\}-\exp(-\mu t))+\frac{c_0}{r_0}\exp(-\mu t)\right\} \tag{3}$$

with a more complicated expression for the variance

$$\begin{aligned}
V_C(t)=r_0\{&k_1\exp\{(\lambda-\nu)t\}+k_2\exp\{2(\lambda-\nu)t\}+k_3\exp\{(\lambda-\nu-\mu)t\}\\
&+k_4\exp(-\mu t)+(k_1+k_2+k_3+k_4)\exp(-2\mu t)\},
\end{aligned} \tag{4}$$

where

$$k_1 = \frac{\nu}{\lambda - \nu + 2\mu}\left\{1 + \frac{\mu}{\lambda - \nu + \mu} - \frac{4\nu\lambda}{\mu(\lambda - \nu)}\right\},$$
$$k_2 = \frac{\lambda + \nu}{\lambda - \nu}\frac{\nu^2}{(\lambda - \nu + \mu)^2},$$
$$k_3 = \frac{2\nu^2}{(\lambda - \nu)(\lambda - \nu + \mu)}\left\{\frac{2\lambda}{\mu} - \frac{\lambda + \nu}{\lambda - \nu + \mu}\right\},$$
$$k_4 = \frac{c_0}{r_0} - \frac{\nu}{\lambda - \nu + \mu}.$$

We see that the last two terms in (4) become negligible as $t$ becomes large, and similarly the effect of $c_0$ becomes negligible in (3) as long as $\mu$ is not very small.

### 2.3 Observing a proportion in the two-compartment hidden model

By particle independence, we can consider each dimension of the process as the sum of $r_0$ independent processes starting with one cell. Thus, standard normal theory applies to analyze the asymptotic distribution of the true proportion $P^C(t) = \frac{C_d(t)}{C_d(t)+C_G(t)}$ as $r_0$ becomes large. Assuming that the process begins with $r_0$ individuals of each type in the reserve and that the ratio $c_0/r_0$ is finite, the multivariate CLT applies:

$$\begin{Bmatrix} \sqrt{r_0}(\frac{C^d(t)}{r_0} - m_{C_1^d}(t)) \\ \sqrt{r_0}(\frac{C^G(t)}{r_0} - m_{C_1^G}(t)) \end{Bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{C_1^d}(t) & 0 \\ 0 & V_{C_1^G}(t) \end{pmatrix}\right]$$

where each $C(t)/r_0$ term is equivalent to a sample average, and the subscripts on the mean and variance denote that they are evaluated for the case of one particle: i.e. $r_0 m_{C_1}(t) = m_C(t)$ and $r_0 V_{C_1}(t) = V_C(t)$.

Because we are interested in the distribution of the proportion $P^C(t)$, we additionally must apply the multivariate delta method with $g(X, Y) = \frac{X}{X+Y}$. Assuming equal initial sizes and intensities in each dimension, we have $g(m_{C_1^d}(t), m_{C_1^g}(t)) = 1/2$ and $g'(m_{C_1^d}(t), m_{C_1^d}(t)) = (\eta, -\eta)$, where

$$\eta = \left[\frac{4\nu}{\lambda - \nu + \mu}(\exp\{(\lambda - \nu)t\} - \exp(-\mu t))\right]^{-1};$$

thus the delta method yields

$$\sqrt{r_0}\{P^C(t) - 1/2\} \xrightarrow{d} N\left(0, \frac{(\lambda - \nu + \mu)^2}{8\nu^2(\exp\{(\lambda - \nu)t\} - \exp(-\mu t))^2} V_{C_1}(t)\right). \tag{5}$$

Next, recall that the samples are assumed to be distributed binomially $Y(t) \sim Binom\left(n(t), P^C(t)\right)$, so we may obtain the expectation and variance of the observed proportion by conditioning on the

true proportion and applying laws of iterated expectation and variance:

$$E(Y(t)/n(t)) = E(P^C(t)) = 1/2,$$

$$V(Y(t)/n(t)) = E(V\{Y(t)/n(t)|P^C(t)\}) + V(E\{Y(t)/n(t)|P^C(t)\}) = (1 - \frac{1}{n(t)})\sigma^2(t) + \frac{1}{4n(t)},$$

where

$$\sigma^2(t) = \frac{(\lambda - \nu + \mu)^2}{8r_0\nu^2(\exp\{(\lambda - \nu)t\} - \exp(-\mu t))^2} V_C(t). \tag{6}$$

Because we have multiple independent realizations of the process, we may carry out inference by essentially equating this variance function with the sample variance across realizations, given observations at fixed points in time: that is, across observations $(y_i, n_i)$ where $i = 1, \ldots, m_j$, and $m_j$ is the number of realizations available at time $t_j$. Since the sample sizes $n_i$ vary across realizations and times, we must create an estimating function rather than equating directly using a method of moments estimator. Define

$$g_t(\frac{y_i}{n_i}) = (\frac{y_i}{n_i} - \frac{1}{2})/\sqrt{(1 - \frac{1}{n_i})\sigma^2(t) + \frac{1}{4n_i}}$$

so that by construction, $g_t(\frac{Y_i}{n_i})$ has variance equal to 1 and mean 0. Then over $m$ realizations, we may set

$$\sum_{i=1}^{m} g_t^2(\frac{y_i}{n_i})/m = V(g_t(\frac{Y_i}{n_i})) = 1$$

and thus

$$\sum_{i=1}^{m} (\frac{y_i}{n_i} - \frac{1}{2})^2/\{(1 - \frac{1}{n_i})\sigma^2(t) + \frac{1}{4n_i}\} = m$$

The estimating function is now obtained by rearranging:

$$\psi_{j,m_j}(\theta) = \frac{1}{m_j} \sum_{i=1}^{m} \frac{(\frac{y_i}{n_i} - \frac{1}{2})^2}{\{(1 - \frac{1}{n_i})\sigma^2(t) + \frac{1}{4n_i}\}} - 1 = 0 \tag{7}$$

Here $\theta = (\lambda, \nu, \mu)$ so it follows that this system can be solved using a minimum of three time points to estimate the three parameters. As this is decreasing in $\sigma^2(t)$, a solution exists and is unique. Notice if we assume all $n_i = n$, the equation simplifies and is equivalent to equating the sample variance with the theoretical variance:

$$\sum_{i=1}^{m} (\frac{y_i}{n_i} - \frac{1}{2}^2)/m = (1 - \frac{1}{n(t)})\sigma^2(t) + \frac{1}{4n(t)}. \tag{8}$$

### 2.4 Standard errors of point estimates

To approximate the asymptotic variance of parameter estimates obtained using $\psi$, the authors apply an empirical sandwich estimator, detailed by Wakefield [13]; the authors reference van der Vaart where it is discussed in a general Banach space context [12]. The information $I(\theta)$ from the observations is given by the sandwich form

$$I(\theta) = A^T B^{-1} A \tag{9}$$

where $A$ is the $j$ by 3 Jacobian matrix of $\psi$ at $\theta$, where $j$ is the number of observation times, and the $j$th row of $A$ is given by

$$\dot{\psi_{j\cdot}}(\lambda, \nu, \mu) = [\frac{\partial \psi_{j\cdot}}{\partial \lambda}, \frac{\psi_{j\cdot}}{\partial \nu}, \frac{\partial \psi_{j\cdot}}{\partial \mu}]$$

It should be noted that while asymptotic normality of $\psi_{j,m_j}(\theta)$ is guaranteed by the Lindeberg-Feller central limit theorem as long as the $n_i$ are of the same magnitude, the size of the $m_j$ in our case (which is at most 11) is rather small for asymptotic methods. The inner term $B$ represents the variance of the estimating function, with $\psi_{j,m_j}(\theta)$ terms along the diagonal. $B$ is assumed to be a $j$ by $j$ diagonal matrix: this is reasonable if observation times are chosen such that they are far apart enough that correlation is realistically close to zero, justifying the zero entries off the diagonal. We estimate $\theta$ using observations from 3 time points, so $A$ and $B$ are both square 3 by 3 matrices, and standard error approximations are empirical: $\hat{A}$ is evaluated by numerically differentiating at the estimated $\hat{\theta}$, and the entries in $\hat{B}$ are also evaluated as empirical variances of $\psi_{j,\cdot}(\hat{\theta})$.

The authors provide point estimates in terms of $p = \lambda/(\lambda+\nu)$ and $g = \lambda - \nu$ with $\mu$ unchanged, so we must apply a further delta method step to obtain standard errors for $(p, g, \mu)$. The Jacobian matrix for this transformation is given by

$$J = \begin{bmatrix} \frac{\nu}{(\lambda+\nu)^2} & \frac{-\lambda}{(\lambda+\nu)^2} & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

so that standard errors for $(p, g, \mu)$ are obtained by taking the square roots of the diagonal entries of $J(\theta)I(\theta)^{-1}J(\theta)^T$. The parameter $p$ can be interpreted as the probability that a given event in the reserve compartment is a self-renewal, and $g$ represents the overall intensity of growth in the reserve compartment.

# 3 Results

We apply these methods to the experimental safari cat data. Data for 11 transplant cats were provided by Peter Guttorp: we display the observed percentages of $d$ G6PD phenotype for each cat in Figure 2. Notice that there is more variability in $\%d$ at earlier times that settles to approximate binomial variability by the end. We also notice that observations are often sparse, and data from several cats become unavailable at an early time.

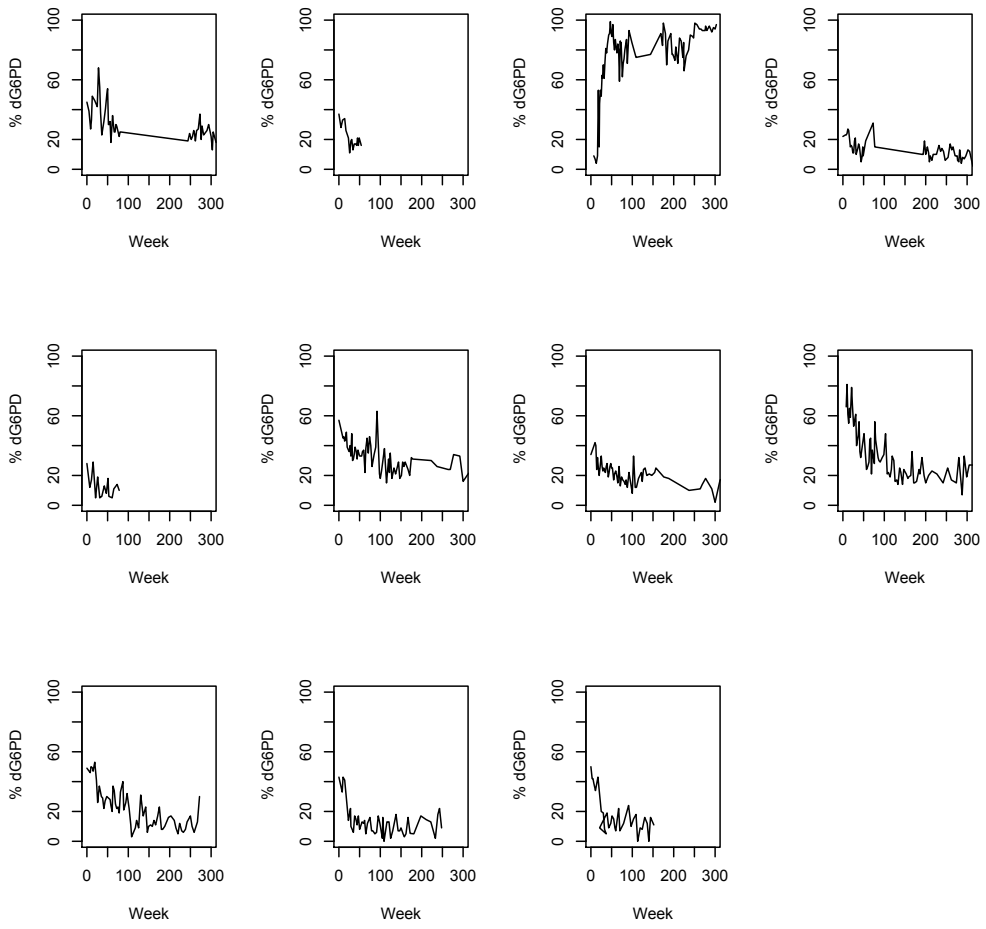


Figure 2: Observed percentages of $d$ G6PD of 11 transplant cats; week zero represents date of transplant.

### 3.1 Three time points

To estimate the parameters using three time points, the authors choose weeks $15, 51$, and 267. These time points are chosen for being spread out enough relative to the growth rate that they are representative of the data as a whole, and for providing data from sufficiently many cats. At each of these times, observations within $\pm 3$ weeks are pooled together to provide a reasonable number of observations: doing so, 11, 11, and 6 animals are available at $t = 15, 51, 267$, respectively. The authors assert that pooling of nearby observations in time is biologically justified, but do not specify how they are chosen: for instance, if data for a particular cat were not available at week 51, but available at both weeks 50 and 52, it is not clear which point is chosen for use with the estimating equation. The ambiguity in grouping observations may explain minor discrepancies from our estimates and those provided in the paper: this effect is illustrated in Figures 3 and 4, comparing point estimates based on always choosing the earlier of two equidistant points from a given time $t_j$ and always choosing the latter.

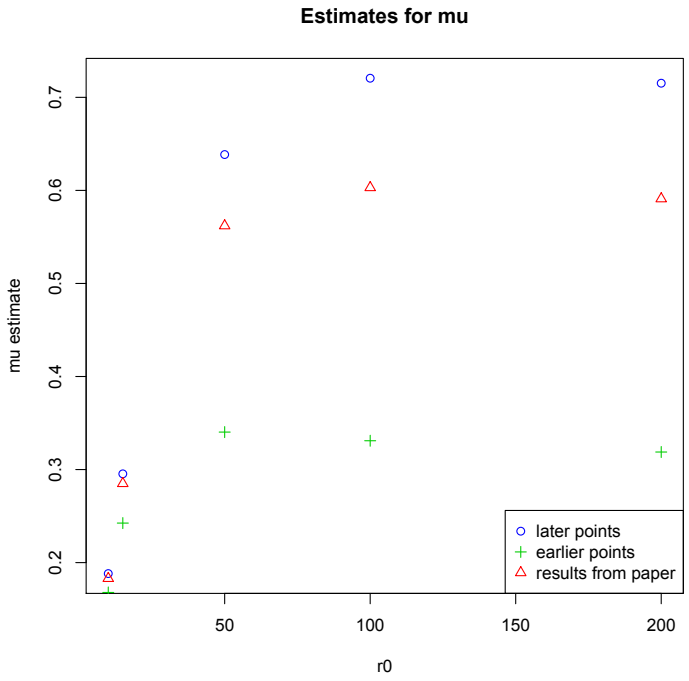


Figure 3: Point estimates for $\hat{\mu}$: here we see a noticeable difference in choice of observations, with later points yielding closer results to the paper.

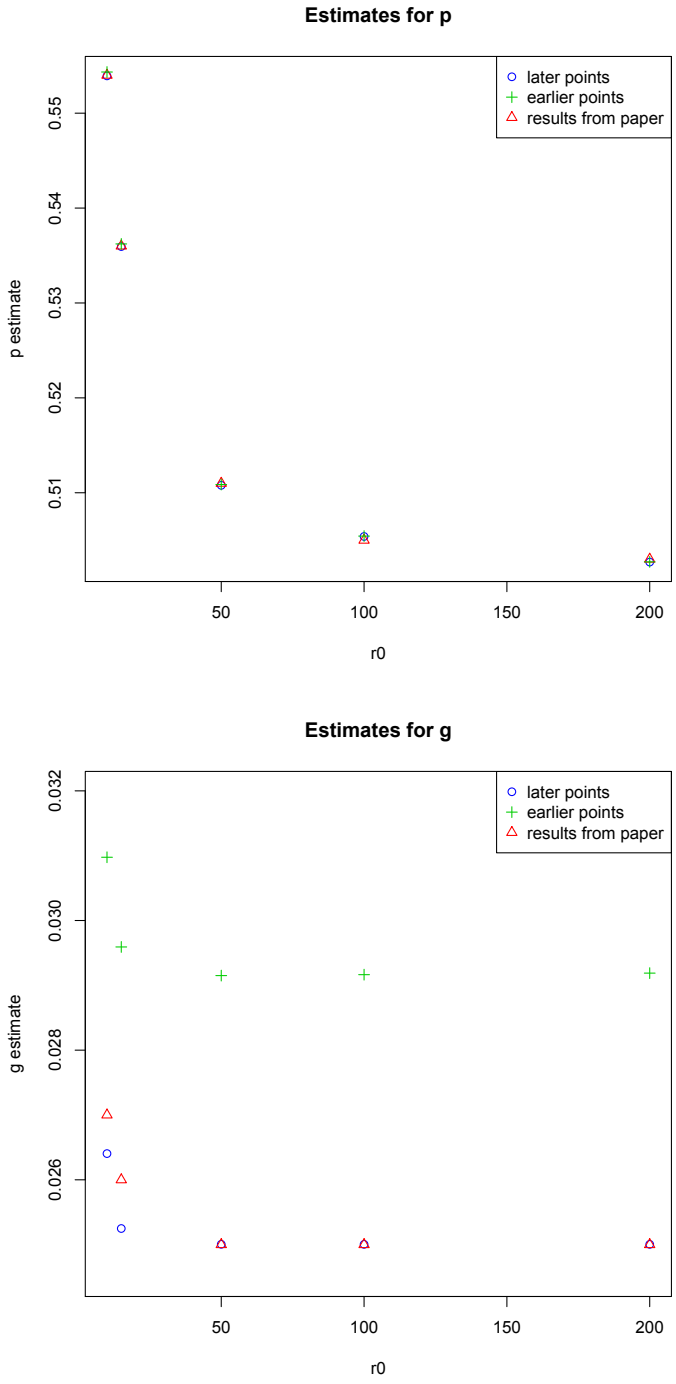


Figure 4: Point estimate comparisons for $\hat{p}$ and $\hat{g}$. While all three methods are close for estimating $\hat{p}$, we see a noticeable difference between choices of pooled observations for $\hat{g}$: always choosing the later point yields results closer to those reported in the paper.

Complete parameter estimates and corresponding sandwich standard errors using equations (7) and (9) are summarized in Table 1, based on consistently choosing the later observation when presented with a choice of pooling equidistant data points. Estimates are reported over a range of possible values $r_0$ because the starting size is unidentifiable, and are very similar to those reported in the paper. Our estimates are computed in three ways: the zeros of $\psi(t_j)$, $j = 15, 51, 267$ are found numerically in R using multiroot, based on a Newton-Raphson method, and with BBsolve, which alternates between the derivative-free spectral approach with multiple step lengths and Nelder-Mead algorithm for robustness. The third method is to minimize the sum of squares $(||\psi(t_j)||_2)^2$ via optim. All methods produced similar results within three significant figures, and while optim is slower than the others, it is least sensitive to initial guesses.

| $r_0$ | $\hat{p}$ | SE | $\hat{g}$ | SE | $\hat{\mu}$ | SE |
|---|---|---|---|---|---|---|
| 10 | 0.554 | 0.025 | 0.026 | 0.035 | 0.188 | 0.155 |
| | 0.554 | 0.023 | 0.027 | 0.040 | 0.183 | 0.187 |
| 15 | 0.536 | 0.017 | 0.025 | 0.038 | 0.295 | 0.397 |
| | 0.536 | 0.015 | 0.026 | 0.039 | 0.285 | 0.477 |
| 50 | 0.511 | 0.005 | 0.025 | 0.038 | 0.640 | 3.54 |
| | 0.511 | 0.005 | 0.025 | 0.038 | 0.603 | 3.84 |
| 100 | 0.505 | 0.002 | 0.025 | 0.037 | 0.721 | 5.97 |
| | 0.505 | 0.002 | 0.025 | 0.038 | 0.603 | 5.61 |
| 200 | 0.503 | 0.001 | 0.025 | 0.037 | 0.716 | 6.53 |
| | 0.503 | 0.001 | 0.025 | 0.038 | 0.591 | 5.82 |

Table 1: Parameter estimates obtained from weeks $15, 51, 267$. Estimates obtained by authors of paper in blue.

### 3.2 Full data

In Table 2, estimates are reported based on data from all time points. Estimates are obtained using nonlinear least squares via optim, assuming an equal sample size by setting $n_i = n = 67$ (the average sample size) to allow for fitting a single function (8) as described in the methods section. The authors do not include theoretical standard errors for this approach, but discuss measures of error in the following section via simulation.

| $r_0$ | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| 10 | 0.551 | 0.018 | 0.319 |
| | 0.551 | 0.015 | 0.304 |
| 15 | 0.533 | 0.014 | 0.599 |
| | 0.534 | 0.014 | 0.610 |
| 50 | 0.510 | 0.014 | 5.670 |
| | 0.510 | 0.014 | 5.893 |
| 100 | 0.505 | 0.014 | 22.090 |
| | 0.505 | 0.014 | 22.973 |
| 200 | 0.502 | 0.014 | 87.235 |
| | 0.503 | 0.014 | 90.811 |

Table 2: Point estimates using all data between $t = 0$ and $t = 330$, assuming all $n_i = 67$. Again, estimates from paper in blue.

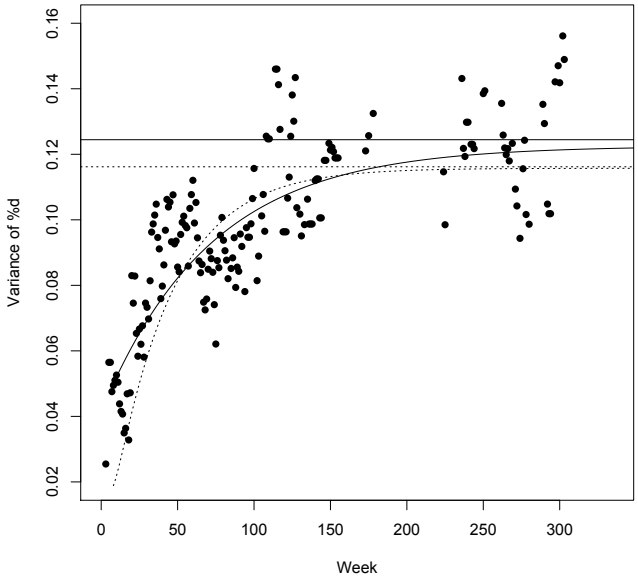


Figure 5: Sample variances at time points with at least six observations within ±3 weeks. The curves represent the variance of observed proportions based on estimates using the full data (solid) or three time points (dotted), and the lines represent the long-run variances.

Using the parameter estimates for the $r_0 = 15$ case, we reproduce the variance plot that compares sample variances, variance functions based on estimates from full data and three time points, and long-run variances in Figure 5.

### 3.3 Simulation and model validation

We reproduce the Monte Carlo procedure employed by the authors to validate estimates above and provide measures of error for the full data point estimates. One thousand sets of 11 simulated "cats" are generated, beginning with the estimated rates using either approach. New parameter estimates are calculated using the same method, observation times, and number of realizations as the true data for each set of 11 simulations. From this set of 1000 sets of estimates, we calculate means, medians, standard deviations, and median absolute deviations (MADs).

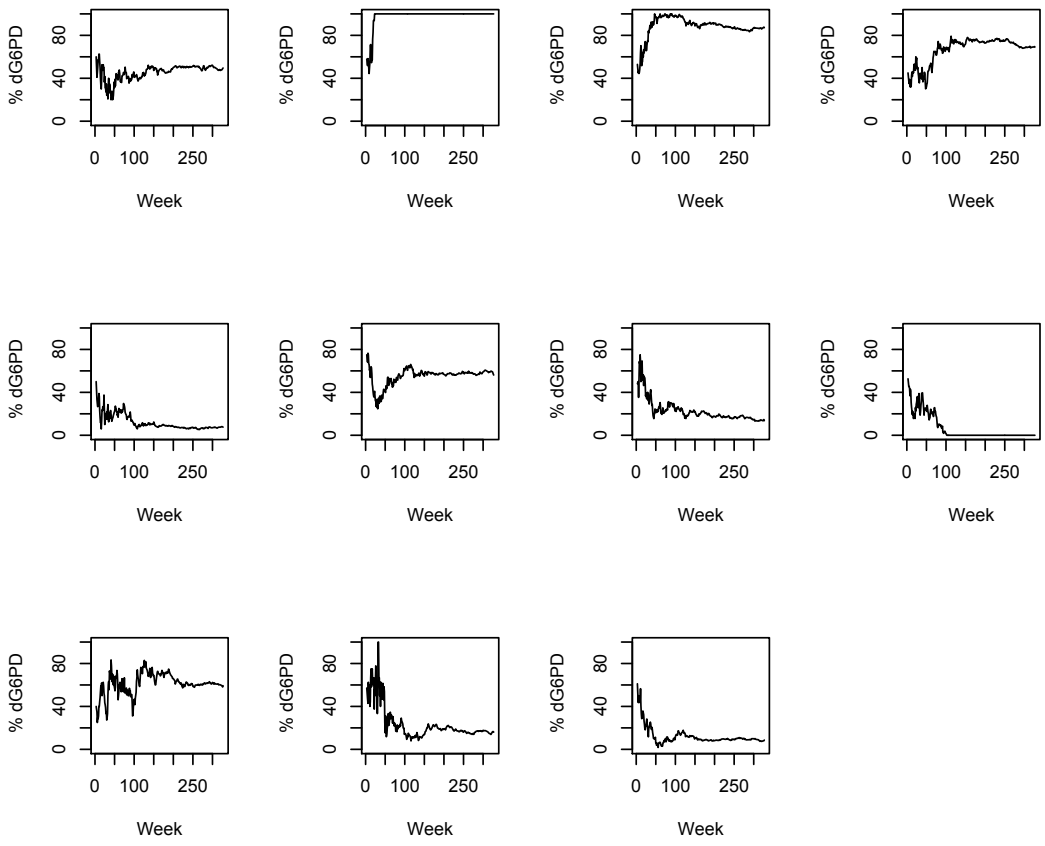


Figure 6: Example of 11 simulated cats with $r_0 = 15$. Here $\%d$ is the true proportion in the contributing compartment, prior to binomial sampling.

Each simulation begins with $r_0$ cells in the reserve for each $d$ and $G$ dimension. While $c_0$ is set to 0 in calculations, the authors do not specify the starting size used in the actual simulations. The choice of $c_0$ is negligible in the long run, but impacts variability at early observation times. We choose to initiate with $c_0 = \lceil r_0 \cdot \frac{\lambda}{\lambda-\nu+\mu} \rceil$ scaling $r_0$ by the equilibrium proportion of $c_0$; a similar choice is used in an earlier study [2]. Further details of simulations can be found in the appendix.

Sample simulation paths in Figure 6 are visually similar to the true data, with high levels of variability near the beginning of the observation period settling down as $t$ increases. There is slightly less variability in these graphs since we are graphing the true proportion of $d$ cells in the simulated contributing compartment. Because the experimental data are from samples from the cats, we sample binomially from these plotted true proportions prior to estimation. Notice the event of extinction of one phenotype population, evident when $\%d$ becomes constant 0 or 100 in some simulation paths, which was not observed in the true data. This is a general problem in the study of population processes, and the authors remark that there is no clear way to incorporate this non-extinction information in our methods.

### 3.3.1 Simulation results

The authors present simulation results for the case of $r_0 = 15$ for the full data and three time point methods, but only report MADs for the other values of $r_0$. Medians of parameter estimates obtained from simulated data at three time points are included in Table 3. Our final estimates are calculated using optim, which proved to be most stable after some experimentation. Estimates using multiroot were highly unstable and thus not included, but earlier simulation estimates using BBsolve are included for comparison.

An upper limit $K$ is placed on the size of the reserve compartment for computational reasons. Such a bound creates a competitive effect at the boundary that causes one phenotype to eventually become extinct; however Catlin [4] justifies its inclusion if $K$ is large enough that its effect on extinction in our given time frame is small. Catlin chooses $K = 750$ in her study; we choose $K = 9000$ when $r_0 = 15$ and include simulation results with $K = 2000$ for comparison. Indeed there is a small but noticeable difference in estimates when $K$ is increased.

Tables 3 and 4 illustrate that while medians are similar, there is a noticeable difference in measures of error between optimization methods. BBsolve reports unsuccessful convergence for many sets of simulations, and providing multiple initial guesses via Multistart did not noticeably remedy this problem. Even when holding the choice of optimization routine constant, there is

quite a large difference in standard deviation between two sets of simulations with $K = 2000$ and $K = 9000$, using optim in both cases. These illustrate both a sensitivity to choice of optimization routine and some inherent numerical instability to the problem. Nonetheless, results with $K = 9000$ using optim are very close to those reported by the authors.

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Starting parameters | 0.536 | 0.026 | 0.285 |
| Authors | 0.536 | 0.025 | 0.118 |
| $K = 9000$: Optim | 0.536 | 0.026 | 0.115 |
| $K = 2000$: Optim | 0.556 | 0.028 | 0.101 |
| BBsolve | 0.556 | 0.028 | 0.105 |

Table 3: Medians of parameter estimates using three time points, $r_0 = 15$.

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Theoretical SE | 0.015 | 0.039 | 0.477 |
| SD: Authors | 0.018 | 0.054 | 2.99 |
| $K = 9000$: Optim | 0.018 | 0.053 | 0.492 |
| $K = 2000$: Optim | 0.034 | 0.135 | 4.453 |
| $K = 2000$: BBsolve | 0.097 | 0.086 | 0.724 |
| MAD: Authors | 0.013 | 0.017 | 0.057 |
| $K = 9000$: Optim | 0.012 | 0.018 | 0.063 |
| $K = 2000$: Optim | 0.025 | 0.028 | 0.054 |
| $K = 2000$: BBsolve | 0.022 | 0.022 | 0.077 |

Table 4: Corresponding error comparisons

Means of estimates are higher than the corresponding medians and suggest a positive skew: when $r_0 = 15$ for the three time point approach, the authors obtain means $(0.540, 0.036, 0.324)$ compared to $(0.540, 0.038, 0.236)$ from our simulations. Using the full data, their estimated means are $(0.536, 0.023, 14.43)$ compared to $(0.555, 0.026, 0.622)$ from our calculations. Notice their mean for $\hat{\mu}$ is much larger than the true value, and the corresponding standard deviation for $\hat{\mu}$ estimates is extremely large, suggesting the potential presence of extreme outliers and finite sample bias in

the reported estimates. This again reiterates the difficulty in estimation for $\mu$ due to a very wide range of possible values and numerical sensitivity.

Complete results from simulations beginning with the full data estimates are included in Table 5: while not quite as close as the three time point approach, estimates are again similar to those in the paper. The more pronounced competitive effect induced by $K$ contributes to this discrepancy: since we have a longer observation interval ($t = 330$ compared to $t = 267$ in the three point approach) and positive growth rate $g$ of the reserve, achieving the limiting size $K$ and in turn extinction of one phenotype is more likely. While this effect is difficult to quantify mathematically, recall our inferential procedure is conditional upon non-extinction, so increased extinction events may bias the simulation results slightly further from the initial input parameters. The errors are harder yet to assess, as was the case in the three time point approach. Theoretical standard errors from a delta method approximation rely on an asymptotic assumption that is not fully justified with a small sample size, and do not take into account that estimates are bounded below by zero, leading to discrepancies between simulation results and sandwich errors. The large differences between standard deviations and MADs in both our findings as well as the authors' illustrates a difficulty in choosing an appropriate estimator for error comparison.

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Starting parameters: | 0.534 | 0.014 | 0.610 |
| Median: Authors | 0.533 | 0.012 | 0.419 |
| $K = 9000$ | 0.542 | 0.018 | 0.628 |
| $K = 2000$ | 0.545 | 0.020 | 0.628 |
| SD: Authors | 0.022 | 0.040 | 405.59 |
| $K = 9000$ | 0.032 | 0.022 | 0.032 |
| $K = 2000$ | 0.043 | 0.021 | 0.037 |
| MAD: Authors | 0.015 | 0.014 | 0.346 |
| $K = 9000$ | 0.026 | 0.016 | 0.021 |
| $K = 2000$ | 0.040 | 0.019 | 0.022 |

Table 5: Medians and errors of estimates from full data, $r_0 = 15$.

Results for cases $r_0 = 10, 50$ are included in the appendix, and results based on a reduced set of 200 estimates are provided for the $r_0 = 100$ case as well. Simulation for the two largest values

$r_0 = 100, 200$ became infeasible toward the end of the observation period on my computer; the authors mention similar computational difficulties for these two cases. Attempts to simulate data using the HematopoiesisSimulator developed by Abkowitz and Guttorp were also unsuccessful for these cases: estimated rates from the observed data are higher than the maximum input initial parameters compatible with the software, and inputting the largest possible $\mu$ value still led to many errors.

### 3.4 Discussion

The estimates obtained by solving the estimating equations using either the three time point or full data approach are similar to rates in other studies. Abkowitz et al. [2] report $\lambda = 1/10, \nu = 1/12.5, \mu = 1/6.7$ when $r_0 = 15$, which corresponds to $(p, g, \mu) = (.556, .020, .149)$, and the Golinelli study [6] reports $(p, g, \mu) = (.546, .021, .147)$ under gamma priors. The most noticeable discrepancy between these results and estimates using the moments method is in the value of $\hat{\mu}$, but with such large standard errors, their findings are well within any sensible confidence interval.

Even though point estimates obtained here are similar to results from other studies, it is hard to justify the underlying asymptotic assumptions *a priori* with such a small sample, especially with a sensitivity to the choice of time points used with the estimating equation. That appropriate choice of observation times depends on the process itself is worrisome– the method is somewhat ad hoc in this sense, and this problem is reflected in the size of standard errors. In her thesis, Catlin obtains a range of parameter estimates when varying the three time points, and remarks that lower estimates for $g$ when including later observation weeks raises the possibility that overall growth rate, which we have assumed constant, may decrease over time [4]. However, with such large standard errors, statements such as these cannot be made with any kind of certainty. On the other hand, standard deviations of estimates available using a Bayesian integration approach [6] are much smaller, leaving something to be desired from this technique.

Despite its shortcomings, the proposed estimating equation approach provides a more elegant and computationally efficient solution in this problem of inference. Its most notable contribution lies in enabling consideration of large population sizes when a simulation study becomes infeasible. Many of the shortcomings of this method arise from its inefficiency– it is not quite satisfactory in this application with such a small number of realizations available from the safari cat data. It should be noted that this method does better with a large number of realizations, precisely when computationally intensive techniques such as stochastic integration or simulation become less

tractable. Thus the proposed method is nicely complementary in this sense, and may even be used in conjunction with existing methods: if one were hesitant to rely solely on estimates obtained from this approach, it can provide a large reduction in the parameter space to be searched with a following simulation study. Further, the proposed technique is valid whenever expressions for the mean and variance of the process are available and additive, and thus can be applied generally to a large class of stochastic compartmental models.

# 4 Conclusion

The estimating equation approach presented by Catlin et al. provides a computationally inexpensive inferential procedure in the complex process of hematopoiesis via a stochastic two-compartment model, where exact likelihood methods are unavailable. Parameter estimates using well-chosen data points in the safari cat study support results obtained in other studies, although standard errors are quite large due to the heavy dependence on the amount of realizations available. Because of its inefficiency and sensitivity to choice of observation times, this technique falls short of a completely adequate solution. Nonetheless, its contribution in enabling inference in a stochastic compartmental model when large population sizes render simulation intractable is a considerable advancement over previous studies. It has been shown in later studies that Bayesian integration techniques make much more efficient use of the data, providing similar point estimates with significantly smaller variance estimates, and are thus preferable when feasible. These still are not quite satisfactory due to their large computational cost, and it may be worthwhile to apply analytical derivations such as those leading to the estimating equation to reduce the integration step necessary in such techniques. Perhaps combining these attempts may allow for inference in this model that is at once accurate, precise, and computationally efficient.

## References


[1] J. L. Abkowitz, M. Linenberger, M. A. Newton, G. H . Shelton, R. L. Ott, P. Guttorp, *Evidence for maintenance of hematopoiesis in a large animal by the sequential activation of stem cell clones*, Proceedings of the National Academy of Science, **87**, 9062-9066, 1990

[2] J. L. Abkowitz, S. N. Catlin, P. Guttorp, *Evidence that hematopoiesis may be a stochastic process in vivo* , Nature Medicine **2(2)**, 190-197, 1996

[3] N. Bailey, *The elements of stochastic processes with applications to the natural sciences*, New York: Wily, 1964

[4] S. N. Catlin, *Statistical inference for partially observed Markov population processes*, Ph.D. dissertation, University of Washington, Seattle, 1997

[5] S. N. Catlin, J. L. Abkowitz, P. Guttorp, *Statistical Inference in a Two-Compartment Model for Hematopoiesis*, Biometrics, **57(2)**, 546-553, 2001

[6] D. Golinelli, P. Guttorp, J. L. Abkowitz, *Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis*, Math Med Bio **23**, 153-172, 2006

[7] M. A. Newton, P. Guttorp, S. Catlin, R. Assuncao, J. Abkowitz, *Stochastic modeling of early hematopoiesis*, Journal of the Statistical Association of America **90**, 1146-1155, 1995

[8] M. Ogawa, *Differentiation and proliferation of hematopoietic stem cells*, Blood **81**, 2844-2853, 1993

[9] P. S. Puri, *Interconnected birth and death processes*, Journal of Applied Probability **5**, 334-349, 1968

[10] J. E. Till, E. A. McCulloch, L. A. Siminovitch, *A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells*, Proceedings of the National Academy of Science, U.S.A. **51**, 29-36, 1964

[11] E. Renshaw, *Spatial population processes*, Ph.D Dissertation, University of Edinburgh, 1976

[12] A. W. van der Vaart, J. A. Wellner, *Weak Convergence and Empirical Processes*, New York: Springer-Verlag, 1996

[13] J. Wakefield, *Bayesian and Frequentist Regression Methods*, New York: Springer, 2013

# 5 Appendix

## 5.1 Simulation details

Both the $d$ and $G$ population in the reserve compartment are initiated with $r_0$ cells, and similarly each dimension in the contributing compartment begins at size $c_0 = \lceil r_0 \cdot \frac{\lambda}{\lambda-\nu+\mu} \rceil$. Each event has exponential waiting time, and thus the time until next event is distributed exponentially, with rate $(\lambda + \nu)[R_d(t) + R_G(t)] + \mu[C_d(t) + C_G(t)]$, the sum of all the individual rates. The type of event that occurs is then determined by the ratio of its corresponding rate to the total rate: for instance, a self-renewal of $d$ type cell in the reserve compartment occurs with probability $\frac{\lambda R_d(t)}{(\lambda+\nu)[R_d(t)+R_G(t)]+\mu[C_d(t)+C_G(t)]}$. Population sizes are adjusted according to the event: in the case of self-renewal of $d$ in the reserve, for example, we set $R_d(t+h) = R_d(t)+1$, where $h$ is the waiting time given by the value of the simulated exponential variate. When a self-renewal event occurs but the total population of the reserve has already reached the limiting size $K$, it is ignored. This procedure of simulating exponential random variables to generate events is repeated until a specified end time. Notice that the rates of these exponentials grow with the population size, and the rapid increase in frequency of events when the population becomes too large renders computation infeasible in some cases.

Simulations in which one phenotype becomes extinct are used in estimation, while we must discard cases in which the entire process dies out, as the proportion we observe is undefined after extinction. We return the proportion of $d$ cells in the contributing compartment at specified time points identical to those used to analyze the safari cat data. Next, we sample binomially with this proportion, with sample size equal to those from the data at the chosen time points. This artificially inflates the variance appropriately to mirror the uncertainty from experimental sampling, since we can observe the true proportions in the contributing compartment in simulated data. Using these proportions, we estimate rates by solving the estimating equation analogously as we did with the observed data.

## 5.2 Additional Simulation Results

In the simulations when $r_0 \neq 15$, the authors only report MAD's which are included for comparison. Here we report results for $r_0 = 10, 50, 100$: our medians are similar to the true parameters, and MAD's are similar to the authors' estimates with the exception of a noticeable difference for $\mu$.

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Initial Values | 0.551 | 0.015 | 0.304 |
| Mean | 0.555 | 0.025 | 0.644 |
| Median | 0.539 | 0.017 | 0.631 |
| SD | 0.040 | 0.018 | 0.097 |
| MAD | 0.022 | 0.011 | 0.035 |
| MAD (authors) | 0.020 | 0.014 | 0.192 |

Table 6: Simulation estimates when $r_0 = 10$, $K = 20,000$, using data from all time points

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Initial Values | 0.510 | 0.014 | 5.893 |
| Mean | 0.522 | 0.048 | 5.865 |
| Median | 0.512 | 0.018 | 5.907 |
| SD | 0.051 | 0.090 | 0.604 |
| MAD | 0.010 | 0.013 | 0.301 |
| MAD (authors) | 0.005 | 0.017 | 3.447 |

Table 7: Estimates when $r_0 = 50$, $K = 20,000$, using data from all time points

| | $\hat{p}$ | $\hat{g}$ | $\hat{\mu}$ |
|---|---|---|---|
| Initial Values | 0.505 | 0.014 | 22.973 |
| Mean | 0.509 | 0.034 | 24.011 |
| Median | 0.506 | 0.016 | 23.993 |
| SD | 0.009 | 0.043 | 0.771 |
| MAD | 0.004 | 0.014 | 1.195 |
| MAD (authors) | 0.002 | 0.017 | 12.504 |

Table 8: Estimates when $r_0 = 100$, $K = 20,000$, using data from all time points with only 200 sets of estimates.