Statistical Inference in a Two-Compartment Model for Hematopoiesis

Sandra N. Catlin, Janis L. Abkowitz, Peter Guttorp

STAT 518 Jason Xu

May 28, 2013

What is Hematopoiesis?

Hematopoeisis: Process of specialization of stem cells into mature blood cells

- HSCs differentiate (specialize) into progenitor cells: multi-stage process
- Progenitor cells further differentiate to white/red blood cells, platelets, etc. This is well-studied.
- Little is known about early stages: unidentifiability of HSCs

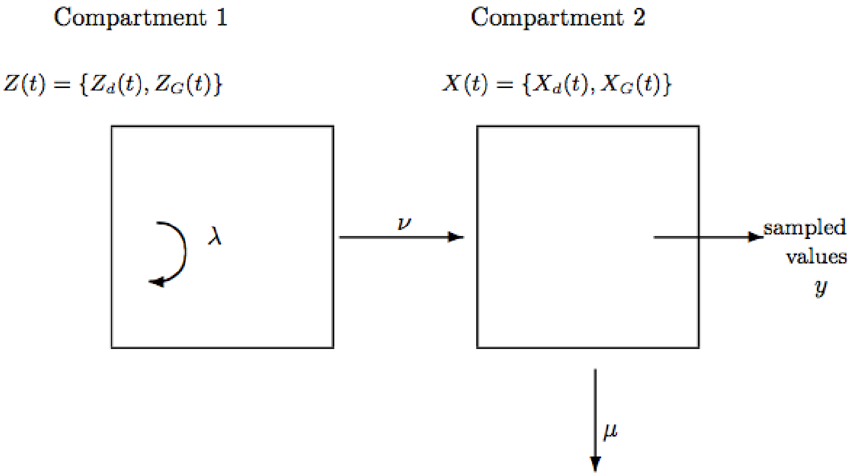
Experimental design

Female safari cat study

- Distinct G6PD phenotype expressed as d or G
- Retained after replication/differentiation; neutral
- Provides binary marker of each cell and its clones

Observing proportion of, say d, allows us to "track" HSC behavior

The model



Compartment 1
Compartment 2
$Z(t) = \{Z_d(t), Z_G(t)\}$
$X(t) = \{X_d(t), X_G(t)\}$
λ
ν
sampled values y
μ

The model

Simple continuous time, discrete state process:

- Compartment 1 is a linear birth-death (BD) process
- Compartment 2 is a non-homogeneous immigration-death process
- Inference: rates λ, ν, μ

Likelihood: $L(\lambda, \nu, \mu) \propto \lambda^{B_T} \nu^{E_T} \mu^{D_T} \exp(-(\lambda + \nu) S_T^z - \mu S_T^x)$

- B_T = births, E_T = emigrations, D_T = deaths, S_T^i = total time in i
- MLEs available: $\hat{\lambda} = B_T / S_T^z$, $\hat{\nu} = E_T / S_T^Z$, $\hat{\mu} = D_T / S_T^x$; nice asymptotic properties

Difficulty: Partial Observations

We only have sampled values from the second compartment: $Y(t)$, the total cells marked d, is a hidden Markov process

$$[Y(t)|(x(t), z(t))] \sim Binom(N_t, \frac{x_d(t)}{x_d(t) + x_G(t)})$$

- Distribution of this binomial proportion mathematically difficult
- Exact likelihood methods infeasible
- No successful attempts in obtaining transition probabilities

Current Method

Outline of current approach

- Calculate moments of process by solving Kolmogorov forward equation
- Create estimating function relating these expressions and data
- Solve numerically using full data and three time points
- Simulate process from estimated parameters

The Kolmogorov Forward Equation

From Bailey (1964), we can obtain a PDE for CGF of multi-dimensional Markov processes as

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

In our case, the f_{jk} are simple rates: $f_{1,0} = \lambda x$, $f_{-1,1} = \nu x$, and $f_{0,1} = \mu y$. Thus,

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

Getting the moments via cumulants

- Since CGF $=$ log(MGF), the first and second cumulants κ_1, κ_2 yield mean, variance
- We can obtain a system of ODE's for cumulants by expanding the CGF, taking partial derivatives, and equating coefficients of products of θ_i
- Successively solving yields desired moments

Deriving the estimating equation: setup

- Particle independence: treat the process beginning with r_0 cells as a sum of r_0 independent processes beginning with 1 cell: justifies application of CLT.
- Aymptotics of true proportion $P(t) := \frac{x_d(t)}{x_d(t)+x_G(t)}$ obtained using the moments calculated and applying delta method:

$$\sqrt{(r_0)}(P(t) - 1/2) \to N[0, \sigma^2_{P_1}(t)]$$

where the asymptotic variance is a nonlinear function of λ, ν, μ:

$$\sigma^2_{P_1}(t) = \frac{(\lambda - \nu + \mu)^2}{8\nu^2(\exp\{(\lambda - \nu)t\} - \exp(-\mu t))^2} V_{C_1}(t)$$

Deriving the estimating equation: expectation and variance

Remember, we observe $Y(t)/n(t)$, which is binomial with proportion $P(t)$: using iterated expectations/variances by conditioning on $P(t)$,

- $E(\frac{Y(t)}{n(t)}) = 1/2$
- $Var(\frac{Y(t)}{n(t)}) = (1 - \frac{1}{n(t)})\sigma^2_{P(t)} + \frac{1}{4n(t)}$
- Across realizations given a time t, inference can be based on the sample variance for (y_i, n_i) at realizations (cats) $i = 1, \ldots, m$.

Ta-da!

Thus, we come up with a function

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

constructed to have variance equal to 1.

Setting $\sum_{i=1}^{m} g_t^2(\frac{y_i}{n_i})/m = 1$ and rearranging, we arrive at the estimating function

$$\Psi_{j,m_j}(\theta) = \frac{1}{m_j}\sum_{i=1}^{m_j} \frac{(\frac{y_i}{n_i} - \frac{1}{2})^2}{(1 - \frac{1}{n_i})\sigma^2_{P(t_j)} + \frac{1}{4n_i}} - 1 = 0$$

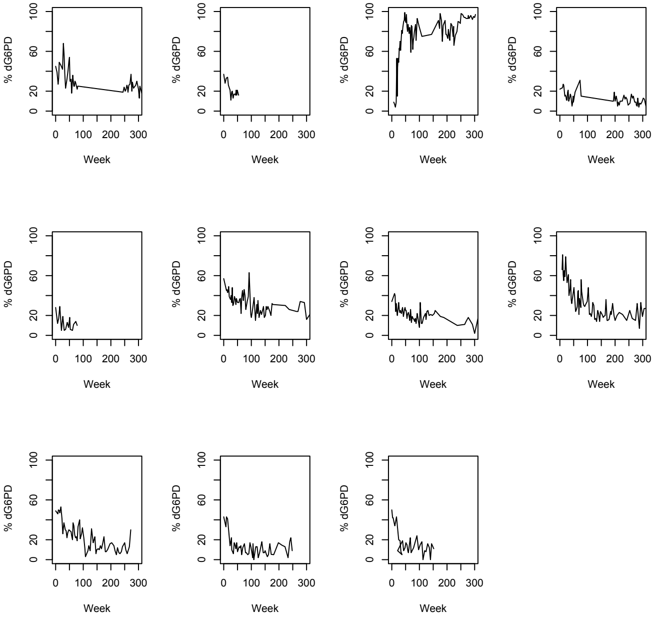
where $\theta = (\lambda, \nu, \mu)$

Solving the equation

- Observations from at least three times t_j allows us to solve for the three unknowns.
- Nonlinear system: numerical solution
- Asymptotic variance of estimates: Huber M Theorem/Sandwich estimates + delta method
- Let's try it out on the experimental data (Abkowitz)

The Data

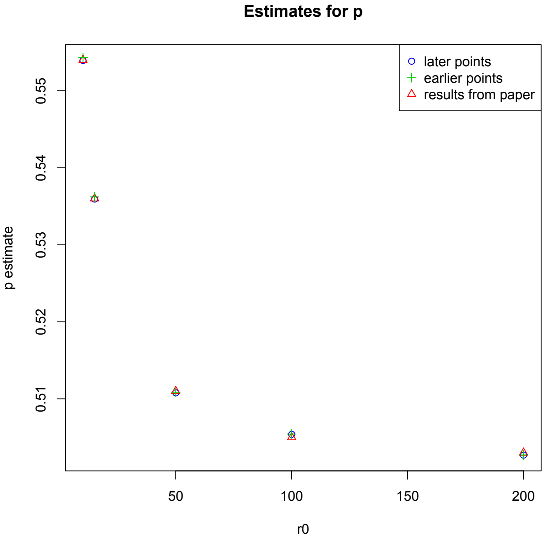
The Data



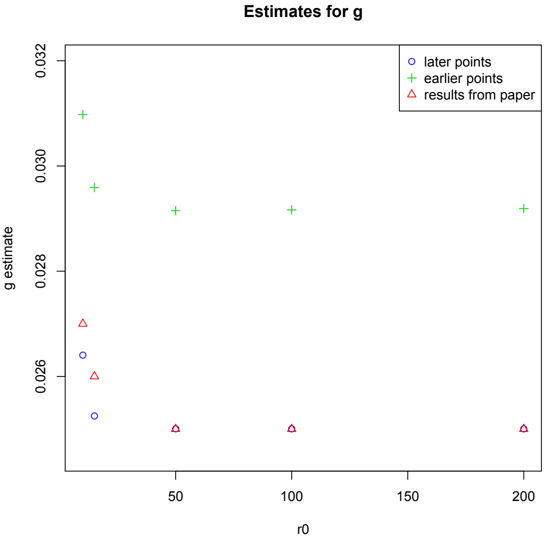
Solving the equation in R

- Weeks 15, 51, and 267 are used, grouping observations within 3 week intervals
- Similar estimates using optim, rootSolve, BB packages; sensitive to initial guess and choice of observations
- Point estimates reported in terms of $p = \frac{\lambda}{\lambda+\nu}$ and $g = \lambda - \nu$, over range of r_0 values
- Interpretation: p is probability that a decision in reserve is self-renewal, g is the intensity of growth in reserve

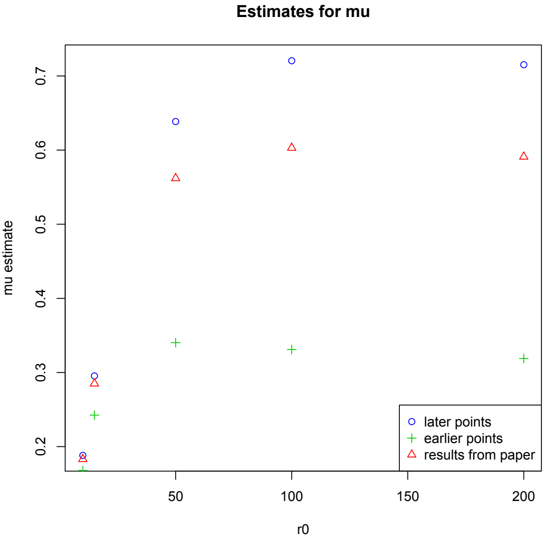
Point estimates: p



Point estimates: g



Point estimates: μ



Comparison of estimates and SE: 3 time points

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: Here we use the point estimates based on keeping the later points using optim(). Estimates obtained by authors of paper in gray

Comparison of point estimates: full data

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

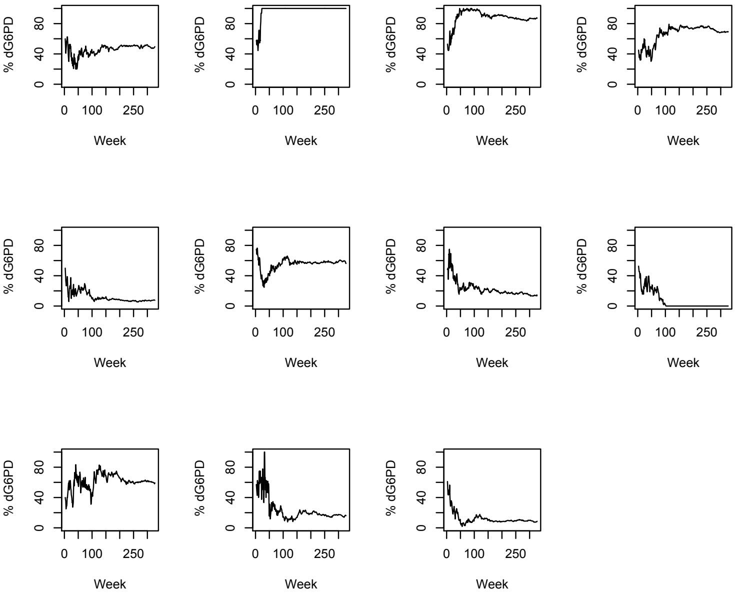
Simulation and model validation

1000 sets of 11 realizations/"cats" are generated, starting with the estimated rates and specified r_0 sizes.

- Using simulated data, 1000 new sets of estimates are calculated
- Evaluate using same time points and sample sizes; binomially sample
- Calculate empirical means, medians, SD, MAD

Simulation plots: $r_0 = 15$, upper limit 5000

Simulation plots: $r_0 = 15$, upper limit 5000



Simulation Estimates: 3 Time Points

	$\hat{p}$	$\hat{g}$	$\hat{\mu}$
True parameters	0.536	0.026	0.285
Authors	0.536	0.025	0.118
Optim	0.556	0.027	0.101
BBsolve	0.555	0.028	0.105

Table: Medians of parameter estimates from simulated data, evaluated at three time points

Error Comparison: 3 Time Points

	$\hat{p}$	$\hat{g}$	$\hat{\mu}$
Theoretical SE	0.015	0.039	0.477
SD: Authors	0.018	0.054	2.99
Optim	0.034	0.135	4.453
BBsolve	0.097	0.086	0.724
MAD: Authors	0.013	0.017	0.057
Optim	0.025	0.028	0.054
BBsolve	0.022	0.022	0.077

Table: Theoretical standard errors compared to standard deviations and MADs from simulation estimates

Simulation Estimates: Full Data

	$\hat{p}$	$\hat{g}$	$\hat{\mu}$
Median: True	0.534	0.014	0.610
Authors	0.533	0.012	0.419
Me	0.542	0.018	0.628
SD: Authors	0.022	0.040	405.59
Me	0.032	0.022	0.032
MAD: Authors	0.015	0.014	0.346
Me	0.036	0.016	0.021

Table: Comparison of preliminary results using full data when $r_0 = 15$. Upper limit of 2000 for the reserve

“Illustrates the difficulty in finding an appropriate estimator for comparison”

Problems and ambiguities

- Possible numerical instability of solvers
- Simulation infeasible for large r_0
- Uncertain of authors' initial sizes, upper limits, extinction events
- "No clear way to incorporate information that neither dimension in any observed processes became extinct"

Concluding Remarks

- Accurate point estimates similar to other studies, using "elegant" solution at low computational cost
- Enables estimation for large populations when simulation approach infeasible
- Huge standard errors, questionable asymptotic assumptions
- Sensitive numerical solutions
- Simplified model: biological limitations
- Significantly less efficient use of data than stochastic integration methods