Statistical Inference in a Two-Compartment Model for Hematopoiesis

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

STAT 518 Jason Xu

April 29, 2013
What is Hematopoiesis?

Hematopoiesis: 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
A Stochastic Model

- First birth-death model for hematopoeisis: Till et al, 1963
- Experimentally justified, refined over several studies
- Current paper analyzes hidden two-compartment model
A Stochastic Model

Goal: Develop inferential tools for this problem, and for a class of stochastic population processes

Statistical motivation: Tools for inference in a useful class of models. Hidden compartmental processes include

- SIR models
- Spread of malaria in human host (Gravenor 1998)

Application: Clinical and biological importance

- Cancer therapy: stem cell transplantation
- Gene therapy
- "A remarkable cell renewal process"; close to 1 trillion cells per day supported by HSCs (M. Ogawa)
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

\[ Z(t) = \{Z_d(t), Z_G(t)\} \]

Compartment 2

\[ X(t) = \{X_d(t), X_G(t)\} \]

\[ \lambda \]

\[ \nu \]

\[ \mu \]

sampled values \[ y \]
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 $\lambda, \nu, \mu$

Likelihood: $L(\lambda, \nu, \mu) \propto \lambda^{B_T} \nu^{E_T} \mu^{D_T} \exp\left(- (\lambda + \nu) S^z_T - \mu S^x_T \right)$

- $B_T = \text{births, } E_T = \text{emigrations, } D_T = \text{deaths, } S^i_T = \text{total time in } i$
- MLEs available: $\hat{\lambda} = B_T / S^z_T, \hat{\nu} = E_T / S^z_T, \hat{\mu} = D_T / S^x_T$; nice asymptotic properties
We only have sampled values from the second compartment: \( Y(t) \), the total cells marked \( d \), is a hidden Markov process

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

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

• Abkowitz (1996): vary parameters and simulate realizations: compare simulations to true data
• Catlin (1997): normal approximation of transition probabilities
• Golinelli (2006): Bayesian inference via RJMCMC
  • Integrate over paths between discrete observations
  • Most precise estimates and effective use of data at computational cost
Other Approaches
Current Method

Estimating equation approach

• Calculate moments of process by solving Kolmogorov forward equation
• Create estimating function relating these expressions and data
  • Method of moments cannot be used directly: differing population sizes over realizations at given time
• Solve using nonlinear least squares
Discussion

- Simulations starting with estimated rates close to observed data
- Parameter estimates very similar to results from other studies
- Minor discrepancies: theoretical and simulated errors
- **Advantages:** not restricted to large population sizes
  - Accurate parameter estimates without much computational cost
  - Provides standard error estimates
- **Drawbacks:** does not utilize all data efficiently
  - Dependent on number of realizations
  - Biological shortcomings
Closing remarks: While not able to make as efficient use of data as stochastic integration methods, provides a more “elegant” solution that is accurate and applicable when MCMC methods become infeasible.

Studying hematopoiesis via two-compartment stochastic model has provided much insight to understanding the complex behavior of HSCs.
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,

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

• Since CGF = log(MGF), the first and second cumulants \( \kappa_1, \kappa_2 \) 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 \( \theta_i \).

• Successively solving yields desired moments.
Getting the cumulants: example

Consider the simple case of a linear birth-death process:

\[
\frac{dK}{dt} = [\lambda(e^\theta - 1) + \mu(e^{-\theta} - 1)] \frac{dK}{d\theta}
\]

The cumulant generating function

\[
K(\theta) = \kappa_1 \theta + \frac{\kappa_2 \theta^2}{2!} + \frac{\kappa_3 \theta^3}{3!} + \ldots
\]

Differentiating this with respect to \(\theta\) and \(t\) yields

\[
\frac{d^2 K}{dt d\theta} = \frac{d\kappa_1}{dt} + \theta \frac{d\kappa_2}{dt} + \ldots
\]

To get \(\kappa_1\)…
Getting the cumulants: example

- Differentiate forward equation with respect to $\theta$:

$$\frac{d^2 K}{dt d\theta} = (\lambda e^\theta - \mu e^{-\theta}) \frac{dK}{d\theta} + [\lambda(e^\theta - 1) + \mu(e^{-\theta} - 1)] \frac{d^2 K}{d\theta^2}$$

- Evaluate at $\theta = 0$ in both expressions and equate:

$$\frac{d\kappa_1}{dt} = (\lambda - \mu) \kappa_1$$

- We arrive at an ODE! In this case, it is easily solvable:

$$\kappa_1 = e^{(\lambda-\mu)t}$$
Getting the cumulants: example

Similarly, $\kappa_2$ is obtained by taking $\frac{d^2}{d\theta^2}$: we obtain

$$\frac{d\kappa_2}{dt} = (\lambda + \mu)\kappa_1 + 2(\lambda - \mu)\kappa_2,$$

which has solution

$$\kappa_2 = \frac{\lambda + \mu}{\lambda - \mu} e^{(\lambda - \mu)t} \left( e^{(\lambda - \mu)t} - 1 \right)$$

• These solutions actually relevant: recall, reserve compartment is a linear birth-death process

• Analogous expansion of our bivariate CGF: system of five ODE’s; closed forms for means and variances available
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.

- Asymptotics of observed proportion $P(t) := \frac{x_d(t)}{x_d(t) + x_G(t)}$ obtained using the moments calculated and applying delta method:

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

- $\sigma^2_{P_1(t)}$ is a nasty expression: it is important that it is a nonlinear function of three variables $(\lambda, \nu, \mu)$
Deriving the estimating equation: expectation and variance

Remember, we observe the proportion \( P(t) \) in the second compartment to estimate the true proportion \( Y(t)/n(t) \): using iterated expectations/variances by conditioning on \( P(t) \),

- \( E\left( \frac{Y(t)}{n(t)} \right) = \frac{1}{2} \)
- \( \text{Var}\left( \frac{Y(t)}{n(t)} \right) = (1 - \frac{1}{n(t)})\sigma^2 P(t) + \frac{1}{4n(t)} \)
- Across realizations given a time \( t \), inference can be based on the sample variance for \( (y_i, n_i) \) at realizations (cats) \( i = 1, \ldots, m \).
Thus, we cook up a function

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

constructed to have variance equal to 1.

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

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

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: use modified Huber’s M Theorem (maybe next time...)
- Next, let’s try it out on some data
Missing data (seriously)
Missing data (seriously)

% dG6PD

Week

% dG6PD

Week

% dG6PD

Week

% dG6PD

Week

% dG6PD

Week
Solving the equation in R

- Because observations are sparse, we choose weeks 15, 51, and 267, and group together observations within 3 week intervals
- 11, 11, and 6 cats were available, respectively
- Solution using rootSolve, BB packages: similar results, sensitive to initial guess
Point estimates

Given in terms of $p = \frac{\lambda}{\lambda + \nu}$ and $g = \lambda - \nu$, over range of $r_0$

- Interpretation: $p$ is probability that a given decision in reserve is self-renewal, $g$ is the intensity of the reserve
- Similar to results from paper, but not identical
- Could be due to differing dataset, or choice of observations
Point estimates: $p$

Estimates for $p$

- later points
- earlier points
- results from paper
Point estimates: $g$

Estimates for $g$

- $g_0$
- $g$ estimate
- later points
- earlier points
- results from paper

![Graph of Estimates for $g$](image-url)
Point estimates: $\mu$

Estimates for $\mu$

earlier points

later points

results from paper
What’s next

- Solve using data from **all time points** using non-linear least squares
- Understand and compute standard errors
- Simulation and validation starting with point estimates
- Investigate transition probability calculations: re-derive Kolmogorov equation for pseudo-generating functions