STOCHASTIC MODELING OF EARLY HEMATOPOIESIS

Michael A. Newton
University of Wisconsin
Departments of Biostatistics and Statistics

Renato Assuncao
Peter Guttorp
University of Washington
Department of Statistics

Janis L. Abkowitz
University of Washington
Department of Medicine

MADISON, WISCONSIN
Stochastic Modeling of Early Hematopoiesis

Michael A. Newton, Renato Assuncao, Peter Guttorp, Janis L. Abkowitz

October 27, 1992

Abstract

Hematopoiesis is the body's way of making the cellular constituents of blood. Oxygen transport, response to infections, and control of bleeding are among the properties of different mature blood cells. These specific functions are acquired as cells mature in the bone marrow. Stem cells are the "master cells" at the top of this pedigree, having within them the capacity to reconstitute the entire system. While the latter stages of hematopoiesis are fairly well understood, the functioning of stem cells and other multi-potential cells is currently a matter of intense research.

One of us, J. Abkowitz, and colleagues at the University of Washington have developed an experimental method for studying the kinetics of early hematopoiesis in a large animal. The essence of the method is to analyze G6PD (glucose-6-phosphate-dehydrogenase), an enzyme linked to the X-chromosome. Somatic cells in female heterozygotes express either the maternal or the paternal G6PD type, but not both. It therefore acts as a binary marker of cells. Importantly, the G6PD type of any cell is passed down to all its descendant cells. Therefore, by monitoring over time the distribution of G6PD types in cells from the bone marrow, information is obtained about the number and lifetime of unobservable stem cells.

Preliminary analysis of the observed proportions in the experimental circumstance of autologous transplantation indicates that the proportion of cells with domestic-type G6PD in the marrow is not constant over time. The pattern of observed fluctuations reflects changes in the unobservable stem cell populations of interest. A simple stochastic model can be used to quantify the relationship between observed proportions and unobserved stem cell populations. The model has a two-stage, hidden-Markov structure. The unobserved stem cell population is modeled by a Markov process. This stem cell population produces, through differentiation and proliferation, the population of blood-cell precursor cells in the marrow. This latter population is sampled to produce the observed proportions, yielding the second stage of the two-stage model.

Statistical inference on the parameters of the hidden Markov process (describing the properties of the stem cell and precursor cell populations) is not at all straightforward. For the simple model in which \( N \), the number of active clones, is fixed, a recursive updating algorithm allows computation of the multi-modal likelihood function on a grid of parameter values. A similar algorithm produces estimates of the realized Markov process. The nature of transition probabilities makes these computations very difficult for large \( N \). Frequency calibration of likelihood-based confidence sets is done using the bootstrap, since Wilks' theorem does not apply. The use of Markov chain Monte Carlo as an alternative method of likelihood and Bayesian inference is reviewed.
1 Hematopoiesis: the Production of Blood Cells

The bone marrow of humans and other vertebrates contains a relatively small number of very remarkable cells, the hematopoietic stem cells. Spangrude et al. (1991) and Golde (1991) are two recent review articles. These cells produce by replication and differentiation all the different kinds of blood cells: red cells which transport oxygen throughout the body; white cells which form the immune defense; and platelets which initiate clotting. A human creates 3-10 billion platelets, red cells, and neutrophils (one kind of white cells) every hour, and in an emergency the rate of production can increase by an order of magnitude (Golde, 1991). Neutrophils are active for eight hours in humans, while platelets and red cells circulate for seven days and four months, respectively.

When an uncommitted progenitor cell divides it may replicate itself and/or differentiate and become the head of a sequence of cell divisions leading to mature blood cells. We assume that the regulation of the stem cell is similar. As differentiation proceeds, cells become committed to particular developmental pathways. This complex process of specialization of stem cells into mature blood cells is called hematopoiesis. Near the beginning of these pathways we find totipotent cells, i.e., cells that can produce all the different types of mature blood cells. Further along the path a partially committed cell may still be able to produce several different types of mature cells. It appears that the capacity to differentiate along individual pathways is lost randomly, until only one lineage remains. It is impossible to distinguish morphologically stem cells from progenitor cells.

There are two different theories regarding the kinetics of early hematopoiesis. In both, a large supply of stem cells is postulated. These stem cells represent the only source for mature blood cells, and must last the lifetime of the animal. One theory, which we will call the standard theory, is that the entire supply of stem cells is proliferating (actively dividing), perhaps with a very slow rate of division. Thus, at any time all stem cells contribute to hematopoiesis. The other, Kay’s (1965) theory of clonal succession, hypothesizes that most stem cells are inactive, and that at any time a small number are proliferating. This theory postulates that the stem cells have a finite lifespan, at the end of which they are replaced by previously dormant stem cells (see Brecher et al. (1986) for a discussion). Evidence of clonal succession may have important implications for research on cancer treatments, bone marrow transplantation, and gene transfer methods. However, these two theories may be indistinguishable in normal animals, since blood production, as most biological systems, has severe redundancy. We would therefore only see the effects of these theories in severely stressed animals, where one can study the physiological backup processes.

In the next section we describe an experiment that enables us to study these backup processes. Section 3 contains some preliminary data analytic questions, while section 4 consists of a brief description of a stochastic model for early hematopoiesis. There are many interesting and difficult questions of statistical inference for this stochastic model, and our approach to handling these questions is outlined in section 5. Finally, in section 6 we discuss some possible avenues towards relaxing some of the assumptions made in the model of section 4, as well as some alternatives to that model.
2 An Experimental Window to Early Hematopoiesis

Much of the work on stem cells, particularly the development of the standard theory of hematopoiesis, has been done using experiments on mice. Since the lifetime of mice is relatively short, it is possible that this evidence may be spurious. Single cells from one mouse can maintain hematopoiesis in another throughout its life time. Retrovirally marked stem cells can contribute to hematopoiesis through 2-3 serial transplants. However, this does not by itself prove that stem cells have very long life times, only that these life times are long compared to the life time of a mouse. For this reason, Abkowitz et al. (1988,1990,1992) embarked on a series of experiments using female Safari cats: a cross between the domestic cat (Felis catus) and the South American Geoffroy wild cat (Leopardus geoffroyi).

Although stem cells cannot be isolated in the laboratory, it is possible to analyze different types of progenitor cells. Colonies can be derived from both erythroid progenitors (red cell precursors) and granulocyte/macrophage progenitors (precursors of two kinds of white cells). In female cells only one of the X-chromosomes, chosen at random early in embryogenesis, remains active. Hence each cell can be classified as having active X-chromosome deriving from either the domestic or Geoffroy parent. The X-linked enzyme glucose 6-phosphate dehydrogenase (G6PD) forms a genetic marker for each cell as domestic and Geoffroy G6PD-types are distinguishable by electrophoresis. The label is conserved through replication and differentiation, so that the G6PD type is the same for a stem cell and all its progeny (the clone of the stem cell). As we did not see significant difference in the distribution of G6PD phenotypes between the two progenitor types (erythroid and granulocyte/macrophage), the data were combined. This observation is consistent with a stochastic mechanism for early differentiation and commitment. Measurements of the proportion of domestic cells were done among colonies grown from progenitor cells sampled from the marrow of the experimental animals. Typically there were 50-100 such colonies grown, so that the G6PD type of 50-100 progenitors could be determined.

Preliminary work on normal cats indicated that G6PD is a neutral marker. However, these cats provided no evidence for or against the clonal succession hypothesis, since our model indicated that a relatively large number (more than 40) of stem cells were operating. This is similar to recent work by O'Quigley (1992) using data from Nash et al. (1988) on human bone marrow transplant recipients which yielded estimates of about 100 active stem cells post transplant with large numbers (over $2 \times 10^8$ per kg) allogeneric cells. In order to obtain a more focused view of the system in operation, autologous bone marrow transplantations were done. Some bone marrow cells were extracted, the subject was irradiated to kill all remaining marrow, and only very few (approximately $1 \times 10^6$ per kg) of the reserved cells were returned. One cat (number 40005), inadvertently received excess marrow cells, and was retransplanted after about a year. As before the transplantations, percentages of domestic type committed progenitor cells were recorded every 2-3 weeks following the recovery of normal blood cell counts. Hematopoiesis was operating as before, but G6PD analysis suggested that much fewer cells were responsible for maintaining normal blood production. The data for three experimental and five control cats are given in Guttorp et al. (1990).
3 Preliminary Data Analysis

For each cat in the experiment, raw data form a time-series of counts \( \{Y_i\} \) of domestic-type colonies in samples of size \( \{n_i\} \) at sampling times \( \{t_i\} \), for \( i = 1, 2, \ldots, m \). The nature of sampling suggests that

\[
(Y_i | p_i) \sim \text{Binomial} (n_i, p_i)
\]

where \( p_i \) is the proportion of domestic-type cells in the progenitor pool at time \( t_i \). From considerations of cell kinetics, models can be developed for the proportions \( \{p_i\} \). But before we attempt sophisticated modeling, it is necessary to study the gross features of the time-series. In fact, two important questions must be addressed:

1. Is there evidence that the \( p_i \) are not equal?
2. If they are not equal, is there evidence that they follow some pattern? For example, do they exhibit positive autocorrelation?

If the answer to question 1 is no, then there is no need to explore models for fluctuating \( \{p_i\} \). If the answer to question 2 is yes, then there is structure in the time series which may provide insight into the workings of early hematopoiesis.

The observed variance of the \( Y_i \) can be used to test the null hypothesis that the answer to question 1 is no. Letting \( p \) equal the common value of the \( \{p_i\} \), it is convenient to introduce a standardized count:

\[
Z_i = \sqrt{n_i} \left( \frac{Y_i}{n_i} - \hat{p} \right)
\]

where \( \hat{p} = \sum_i Y_i / \sum_i n_i \) is the MLE of \( p \) under this null hypothesis. For large \( n_i \), which we have in this experiment, the \( Z_i \) are well approximated by a normal distribution with mean 0 and variance \( p(1 - p) \), under the null hypothesis. By a standard approximation, the test statistic is

\[
\sum_{i=1}^{m} Z_i^2 / (\hat{p}(1 - \hat{p})) \sim \chi_{m-1}^2
\]

Significant extra-binomial variation existed in three of the four post-transplant series published in Guttrop et al., 1990.

To address the second question, we inspect lag-1 autocorrelations of the \( Z_i \). These autocorrelations were .43, .47 and .91 for the three experimental series exhibiting extra-binomial variation. None of the series from the control cats had significant extra-binomial variation, although their autocorrelations ranged from .15 to .68.

4 Modeling Hematopoiesis

4.1 General Considerations

The complex processes which control the differentiation and amplification of hematopoietic cells can not be determined from the G6PD data alone. Any model designed to learn something from
Figure 1: An idealization of the precursor-cell compartment: Each number corresponds to a progenitor cell. Cells with the same number are in the same clone. Boxes represent cells having the domestic-type G6PD, and circles cells having Geoffroy-type G6PD.

these data must necessarily ignore some details, while at the same time be rich enough to account for most of the structure of the cell kinetics of early hematopoiesis. The first simplifying measure is to consider the cells of interest as falling into one of two compartments: a stem-cell compartment, denoted $C_1$, and a precursor-cell compartment, denoted $C_2$.

The essential modeling problems are well illustrated in Figure 1 which is an idealized view of the precursor-cell compartment at two time points. (Note that in a real system, many thousands of cells inhabit $C_2$, although only about 30 or so are shown in this idealization.) At time 1, $C_2$ is composed of three clones; that is the offspring from three stem cells released into $C_2$ from $C_1$. In this schematic, clone 2 dominates the pool, having more cells than clones 1 and 3 combined. For the cells sampled at time 1, the G6PD type (□ = domestic and ○ = Geoffroy) is recorded. Later, at time 2, the composition of $C_2$ has changed—this evolution being driven by several factors:

- expansion of a clone (e.g., clones 1 and 3)
- terminal differentiation (e.g., most of clone 2)
- release of new stem cells (e.g., clone 4)
We refer to the lifetime of a clone as the time from release of its stem-cell ancestor until terminal differentiation of its constituent cells.

One limitation of the data is that the clone of a cell cannot be identified. For example, cells from clone 2 cannot be distinguished from cells of clone 3, given the binary nature of the marker.

If clonal succession explains cellular development, then we would expect the clones to be relatively few in number, and to last a relatively short amount of time. If the standard theory is more appropriate, then clones would have a relatively longer lifetime.

A simple stochastic model is given below. It describes, in probabilistic terms, both the transfer of cells from $C_1$ to $C_2$ and the changes in $C_2$ driven by the three factors above. The model has a hidden-Markov structure: an unobservable Markov process describes the evolution of $C_1$ and $C_2$, then observed counts have a distribution determined by the unobserved state. To date, statistical inference has been limited to this very simple model, however the computational and analytic problems are non-trivial. In section 6, several other models sharing this hidden-Markov structure, but bound by fewer assumptions, are described.

### 4.2 A simple model

If the first compartment, $C_1$, has a large number of dormant or self-replicating cells, it is reasonable to assume that the proportion of domestic-type cells in this compartment is constant over time. We denote this proportion by $p_d$.

As a dynamic process, the number of clones composing $C_2$ may fluctuate because of terminal differentiation and new stem-cell release. A balance is expected if the process is stable. If the fluctuations are not too large, it is reasonable to assume (at least as a first approximation) that the total number of clones, denoted $N$, stays constant over time. A number $X(t)$ of these $N$ clones are of the domestic type at time $t$. The fraction $X(t)/N$ fluctuates between 0 and 1 because a depleted clone may be replaced by a clone having a different G6PD type. Strictly speaking in this model, a new stem cell is released if and only if an existing clone dies; that is, terminally differentiates.

If it is further assumed that the lifetimes of different clones are independent, and also that these lifetimes follow an exponential distribution with rate parameter $\lambda$, then $X(t)$ is a continuous-time, finite-state, Markov process. This state $X(t)$ stays constant except when a clone dies. At such a time, the state can either go up by one, or drop by one, or stay the same, depending on the G6PD type of the clone and the newly released stem cell. The transition intensities of such changes are

$$x \rightarrow x + 1 \text{ with intensity } (N - x)\lambda p_d$$

$$x \rightarrow x - 1 \text{ with intensity } x \lambda (1 - p_d)$$

This continuous-time process induces a finite Markov chain $\{X_1, X_2, \ldots, X_n\}$ by restriction to the sampling times. Because sampling times are unequal, the transition probabilities for this chain are not stationary. They can be written

$$P(X_i = k|X_{i-1} = j) = \sum_{l=0}^{N} \alpha_{ij} \alpha_{lj} \exp \{-l \lambda (t_i - t_{i-1})\} \tag{1}$$

6
where the coefficients $\alpha_{ij}$ are somewhat unwieldy:

$$
\alpha_{ij} = (-1)^i(1-p_d)^{N/2} \sum_{\nu=0}^{(N-i)/2} \left( -\frac{p_d}{1-p_d} \right)^\nu \binom{i}{i-j+\nu} \binom{N-i}{N-i-\nu}.
$$

The development is given in more detail in Guttrop et al., 1990.

The number $X(t)$ of domestic-type clones in $C_2$ influences the number of domestic-type cells in this progenitor pool. The proportion $p_t$ of domestic-type cells in $C_2$ differs from $X(t)/N$ because all the clones do not have exactly the same number of cells at every point in time. Intuitively, we expect $p_t$ to equal $X(t)/N$ on average, and in fact for certain models the fluctuations of $p_t$ can be quantified. A first approximation is to assume that these fluctuations are negligible, and under this assumption the observation distribution is:

$$(Y_t | X(t), N) \sim \text{Binomial} \left( n_t, \frac{X(t)}{N} \right),$$

where $Y_t$, $n_t$, and $t$ are as defined in section 3.

A simple analogy, described in Newton et al., 1992, relates this model to a coin-tossing experiment. Consider a room in which each of $N$ people are independently flipping coins at random exponential time intervals. If the probability of heads for each coin is $p$, and the mean time between flips for each person is $1/\lambda$, then $X(t) =$number of facing heads at time $t$ has the same probability structure as the number of active, domestic-type clones described above. At each of $m$ sampling times, a sample with replacement from the faces of the $N$ coins is shown to an observer, who records the total number $n_t$ and the number of heads $Y_t$. From the observed series, the observer is asked to estimate $N$, $p$, and $\lambda$. The difficulties involved with inference in such a model are described in more detail in the next section.

Two important assumptions used to derive the simple model above are that $N$ is constant and that there is a constant number of equally-contributing clones. Models which relax these two assumptions are described in section 6.

5 Inference

The likelihood function $L$ forms the basis of all our inference procedures, however, evaluation of $L$ at a given parameter triple $\theta = (N, \lambda, p)$ is a nontrivial exercise because the model probabilities are specified in terms of an unobservable process. As is somewhat standard for hidden Markov models, a recursive algorithm for evaluating $L$ can be designed which takes advantage of the Markov structure and the conditional independence of observations given the state. This recursive algorithm is reviewed in the next section.

Whereas a likelihoodist might be satisfied with simple inspection of the likelihood function, most statisticians demand some type of probabilistic calibration. A frequentist might want to know where to slice a level set to get an approximate 95% confidence set, or a Bayesian might want to know how to compute a 95% HPD (highest posterior density) region. In standard problems, the frequentist would appeal to the famous theorem of Wilks, 1938, which ensures that loglikelihood ratios have
asymptotic chi-squared distributions. The regularity conditions for Wilks’ theorem do not hold for this model. We discuss this issue, and methods of bootstrap calibration in section 5.2. The use of bootstrapping to assess goodness-of-fit is also described. Bayesian inference for this model has been studied in some detail in Newton et al., 1992, and this material is reviewed in section 5.3.

5.1 Recursive updating

The likelihood function is

\[ L(\theta) = P_\theta (Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) \]
\[ = P_\theta (Y = y). \]  

(3)

where \( y = (y_1, y_2, \ldots, y_m) \) are the observed counts. Because the model is specified in two stages, it is natural to rewrite the likelihood

\[ L(\theta) = \sum_{x \in S} P_\theta (Y = y | X = x) P_\theta (X = x) \]
\[ = \sum_{x \in S} \prod_{i=1}^m \{ P_\theta (Y_i = y_i | X_i = x_i) P_\theta (X_i = x_i | X_{i-1} = x_{i-1}) \} \]

(4)

where the summation is over the set of possible state values at the sampling times:

\[ S = \{ x = (x_1, x_2, \ldots, x_m) : 0 \leq x_i \leq N \ \forall i \} . \]

The cardinality of \( S \) is \((N + 1)^m\), making an algorithm based on equation 4 unviable. A recursive algorithm can be developed by writing the likelihood

\[ L(\theta) = P_\theta (Y_1 = y_1) \prod_{i=2}^m P_\theta (Y_i = y_i | Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}) \]

and noting that each factor in this product can be expanded into a sum over the \( N + 1 \) possible levels of the state at that time. Expansion of the first factor uses the marginal binomial distribution of \( X_1 \). The latter factors can be expressed, writing \( y_i^k \) for \((y_1 \ldots y_k)\), as

\[ P_\theta (Y_i = y_i | Y_i^{i-1} = y_i^{i-1}) = \sum_{j=0}^N P_\theta (Y_i = y_i | X_i = j) P (X_i = j | Y_i^{i-1} = y_i^{i-1}) \]
\[ =: \sum_{j=0}^N u_i(j) v_i(j) . \]

Because the hidden state is Markov,

\[ v_i(j) \propto \sum_{k=0}^N P_\theta (X_i = j | X_{i-1} = k) u_{i-1}(k) v_{i-1}(k) . \]

From the observation distribution of section 4, the \( u_i \)'s are binomial probabilities. Importantly, the \( v_i \)'s can be computed recursively for increasing \( i \) using the past \( u_j \)'s and \( v_j \)'s. Thus the likelihood
function can be evaluated recursively. This recursion allows computation of the likelihood at a given parameter value. The entire likelihood function can be approximated by evaluating $L$ on a large grid. Figure 2 shows the contours of $L$ for the parameters $N$ and $1/\lambda$ from one experimental cat (40005 after the second transplant). The third parameter, $p_d$, is concentrated out using an estimate from pre-transplant data. The likelihood is multimodal, and has a fairly unusual shape. Contours are associated with approximate coverage probabilities determined from a bootstrap calibration described in the next section. For data from all the experimental cats, the likelihoods are maximized at relatively small values of the mean lifetime $(1/\lambda)$ parameter. This provides support for the clonal succession hypothesis.

In the recursion to compute the likelihood, the vector of $v_t(j)$'s is computed using the transition probabilities (1) of the state. Given the expression in (2), these transition probabilities are subject to significant rounding errors when $N$ is large. Even with the application of standard rescaling tricks, the computations are still quite tricky, and we have not successfully computed these transitions for large $N$.

If instead of parameter estimation, we want to reconstruct the hidden state $X$ based on the observed data, a similar recursive algorithm can be devised. Whereas the algorithm above has one recursion (for increasing $i$), the smoothing algorithm has both a forward and a backward recursion. This algorithm allows computation of the probabilities

$$P_\theta(X_i = j | Y_1 = y_1, \ldots, Y_m = y_m)$$

for all $i$ and $j$, and for a given value of $\theta$. Baum, 1972, derived the appropriate recursions in early applications of hidden Markov models (see more recently Devijver, 1985). The precise recursions are also given in Guttrop et al., 1990. With a maximum likelihood estimate of the parameter $\theta$ plugged into 5, a reconstruction of the state is the sequence $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_m)$ for which $\hat{x}_i$ is the mode of the distribution in 5. Figure 3(d) shows such a reconstruction for data from cat 40005, compared to simulations of the hidden state used in a bootstrap algorithm. The reconstruction shows us the extent and time scale of fluctuations in the underlying precursor-cell population.

5.2 Calibrating level sets of the likelihood

With a way to compute the likelihood on a grid, the following question arises. How far down from the MLE do you go in likelihood units to have an approximate $100(1 - \alpha)$% confidence set, for some small number $\alpha$? This is a question of frequency calibration of the likelihood-based confidence set, which we address in this section. The set of parameters $\theta$ whose likelihood is at least $100r\%$ of the maximum $L(\hat{\theta})$ is called a $100r\%$ likelihood region for $\theta$ (see Kalbfleisch, 1979, for example). Such a set is

$$LR = \left\{ \theta : R = \frac{L(\theta)}{L(\hat{\theta})} \geq r \right\}.$$  

Nothing in this definition relates to coverage probability for the presumed true parameter value $\theta_0$,

$$P_{\theta_0}(LR \text{ includes } \theta_0).$$
Figure 2: Calibrated contours of the joint likelihood of $N$ and $1/\lambda$ for the data from cat 40005.
Figure 3: The bottom right panel shows the reconstruction of the hidden state for cat 40005 after the second transplant. The first three panels show simulated hidden states from the fitted model.
The cutoff \( r \) of the likelihood region determines that region's coverage probability, and so it is important to know how to choose \( r \) to achieve approximate 95% coverage, say. In general, the coverage also depends on \( \theta_0 \), and is not knowable for any particular procedure for choosing \( r \). Hence we aim for approximate coverage.

In models where \( R \) is a smooth function of real-valued parameters \( \theta \), Wilks' theorem (Wilks, 1938), ensures that \(-2 \log R\) is asymptotically \( \chi^2 \) on \( \dim(\theta) \) degrees of freedom. This translates to dropping about 3 units of log-likelihood to define a 95% likelihood-based confidence set in a two-parameter problem. Since \( N \) is an integer-valued parameter, the regularity conditions for Wilks' theorem do not apply. It is not clear that the theorem should apply even approximately, since \( N \) determines the support of the underlying Markov process. Without theoretical results on the distribution of \( R \), we can perform approximate calibration using the bootstrap.

To apply the bootstrap, it is convenient to modify our notation slightly. Since the likelihood depends on both data and parameters, we use the notation \( L(\theta; y) \) for the likelihood determined by an observed time series \( y = (y_1, \ldots, y_n) \). Introduce

\[
T(y; \theta) = -2 \log \frac{L(\theta; y)}{L(\tilde{\theta}(y); y)}
\]

as the random variable whose distribution we want to determine using the bootstrap. The bootstrap algorithm starts with a model fit, that is \( \theta \) is estimated from the data \( y \) by the maximum likelihood estimate \( \tilde{\theta}(y) \). (This involves the recursive updating procedure and a grid search.) The next step is to mimic the sampling process via simulation. On the computer, generate \( n_{\text{boot}} \) time series \( z_1, z_2, \ldots, z_{n_{\text{boot}}} \); each \( z_j \) is a series like \( y \). Generation of each \( z_j \) takes two steps. First a hidden state \( X^* \) is generated by running continuous time Markov process with parameters determined by the fitted model. Then, restriction to the fixed sampling times gives us the binomial success probabilities of the observation distribution which are used to get \( z_j \). This two-stage sampling procedure is illustrated in Figures 3 and 4. Figure 3 shows three realized \( X^*/N \)'s simulated from the model fit to the data for cat 40005, compared to the reconstruction of \( X/N \) using the smoothing algorithm of section 5.1. Figure 4 shows three \( z_j \)'s derived by adding binomial noise to the simulated \( X^*/N \)'s of Figure 3. The fourth panel shows the actual observed series for cat 40005. The vertical bars at the bottom of each graph represent the sample sizes \( n_i \) at the sampling times \( t_i \). These sizes are the same for every bootstrap replication, equalling about 100 on average. The largest is 227 at \( t_4 \). Note that the simulated series in Figure 4 are qualitatively similar to the observed series, suggesting that the model fits the data reasonably well.

For each bootstrapped series \( z_j \), the loglikelihood ratio statistic is computed;

\[
T_j := T(z_j; \tilde{\theta}(y)).
\]

Note that the entire maximum likelihood computation has to be redone for every bootstrapped series \( z_j \). The empirical distribution of the \( T_j \)'s converges to the bootstrap distribution of \( T \) as \( n_{\text{boot}} \) increases. This bootstrap distribution is used to determine the cutoff \( r \) for the likelihood-based confidence set.
Figure 4: The bottom right panel shows the observed proportions for cat 40005 after the second transplant. The first three panels show simulated data sets used in the bootstrap calibration of the likelihood.
Figure 5 compares the cdf of a $\chi^2$ random variable to the empirical distribution function of 240 $T_j$'s computed using the bootstrap algorithm described above. The bootstrapped $T_j$ is stochastically larger than a $\chi^2$ random variable, which means that confidence sets produced by bootstrap calibration are larger than those produced appealing to a Wilks' theorem argument. In Newton, 1991, the accuracy of the bootstrap intervals is checked using a double bootstrap procedure. This computationally intensive diagnostic showed that the bootstrap calibration was accurate in this case.

As mentioned above, the simulation of $z_j$'s used in bootstrap calibration can also be used to assess the goodness-of-fit of the model. In the spirit of Monte Carlo testing, we select a test statistic, like the autocorrelation of the standardized counts, and simulate the distribution of that test statistic under the fitted model. No significant departures in autocorrelation were found in the experimental data, showing that the simple model can explain the observed autocorrelation in the time series. Such tests probably do not have much power against other similar models.

5.3 Bayesian inference

Likelihood calculations for the simple model rely on expressions for the transition probabilities of the state from one sampling time to another. If we perturb the simple model slightly, then such expressions may not be available, and so the recursive updating technique will be impossible to apply. Also if the hidden state has more than a finite number of possible levels, the same problem arises. Furthermore, without theoretical results about the shape of the likelihood in the simple model, the grid evaluations alone do not guarantee that we actually have the MLE or that we have found all the modes of the likelihood. All these arguments suggest that from a computational standpoint, it is important to find other ways to derive the likelihood. One such way is conditional simulation within the framework of Bayesian inference. Besides computational arguments, there are of course philosophical arguments which say that a Bayesian approach is appropriate. For instance, if $N$ and $p$ are nuisance parameters, they can be conveniently integrated out to yield a marginal posterior distribution on the parameter of interest.

Bayesian inference is based on the posterior distribution of the $\theta$ given some particular prior. Because the likelihood is complex, straightforward use of Bayes rule does not make things any easier. The advent of Markov chain Monte Carlo methods like the Gibbs sampler, Metropolis algorithm, and Hastings algorithm (see Tierney, 1991) has revolutionized Bayesian inference. In Newton et al., 1992, a particular Markov chain is constructed to move through the space of $(\theta, X)$ in such a way that the stationary distribution of this chain equals the conditional distribution of $(\theta, X)$ given the data $y$. Here, $X$ is the hidden state at all times within the sampling window. The algorithm is not fully operational, at this time, however initial work shows that it will not provide an efficient way to compute the likelihood in the simple model. In fact, this model provides an interesting test case for the Markov chain simulation methods.
Figure 5: Estimated distributions of -2 times the loglikelihood ratio statistic: The solid curve is the cdf of a $\chi^2$ random variable, and the dots determine the empirical distribution function of 240 bootstrapped statistics.
6 Other Models

While the work reported above appears to lend support to the clonal succession theory of early hematopoiesis, there are several assumptions that may be oversimplifications.

- Each stem cell clone contributes equally to the progenitor pool
- The number of active stem cells is constant
- The stem cell life times are exponentially distributed
- Post-transplant hematopoiesis exhibits qualitatively similar kinetics to that of controls

In this section we discuss how the model may be modified to relax the first two of these assumptions.

6.1 Contributions to the Progenitor Pool

A conditionally binomial observation distribution is correct if all active stem cells contribute equally to the progenitor pool, so that a randomly chosen cell has probability $X_i/N$ to be of domestic type. We present here two more complex models, yielding conditionally superbinomial variability.

We may assume that stem cells give rise to offspring cascades at the times of a Poisson process of constant rate. The description in Brecher et al. (1986) indicates that these cells undergo binary splitting up to $K$ ($\approx 7$) times in the population of cells under study. Further cell division would imply diversification, which is a different population. We say that a cascade is active if the splitting process has not reached the $K + 1^{st}$ division. We further assume that cell division takes place at independent and identically distributed times. The proportion $p_k$ of domestic type cells in the population at time $t_k$ can be written

$$p_k = \frac{\sum_{i=1}^{X_k} m_i}{\sum_{j=1}^{N} m_j}$$

where $m_i$ is the number of cells originating from stem cell $i$, and we number the stem cells so that the first $X_k$ are of domestic type. We can further write $m_i = \sum_{A_i} m_{j,i}$ where $A_i$ is the number of active cascades of stem cell $i$, and $m_{j,i}$ is the number of offspring to the $j^{th}$ active cascade of stem cell $i$. Each $m_{j,i}$ is the population size of an age-dependent branching process (see Harris, 1963, Ch. VI), frozen when reaching the $K^{th}$ generation and started at the time of cloning. Given the (positive) number of cascades in the active set, these times are uniformly distributed. From these assumptions it follows that the $m_i$ are iid, and therefore by symmetry

$$E(p_k|X_k) = \frac{X_k}{N}.$$ 

Moreover $\text{Var}(p_k|X_k) > 0$, so the counts $Y_k$ are overdispersed relative to a binomial model.

A different model for $p_k$ can be obtained in the following manner. Assume again that division occurs at the times of a point process, and that each clone gives rise to a Bienaymé-Galton-Watson process (see Harris, 1963, Ch. I) with positive geometric offspring distribution. In this model we describe the cell division by looking at the number of offspring at discrete times, and allowing for
features such as dormant cells which have a delayed cell division process. For simplicity we assume that only one cascade is active from each stem cell at the time of observation. If more than one are active, the offspring from the earlier cascade would typically dominate the later ones. The age (generation-number) of the \( i \)th stem cell cascade at the time of observation is a random variable \( I_i \). If the underlying point process is homogeneous Poisson, \( I_i \) has the distribution of the minimum of a random number of uniform random variables. From branching process theory we know that the size of a population of age \( l \) is asymptotic to \( \mu^l \zeta \), where \( \mu \) is the mean of the offspring distribution and \( \zeta \) has a standard exponential distribution. This approximation is good even for relatively small \( l \), and suggests approximating \( p_k \) by \( \sum_1^{X_k} \mu^l \zeta / \sum_1^N \mu^l \zeta \). When the process is in steady state, we may be inclined to replace \( I_i \) by its expected value, in which case the conditional distribution of \( p_k \), given \( X_k \) and \( N \), is a beta distribution with parameters \( X_k \) and \( N - X_k \). This yields a beta-binomial distribution for \( Y_k \), given \( X_k \).

6.2 Non-constant \( N \)

It is conceivable that the system is adapting continuously to the situation post-transplant. For example, an increasing proportion of the dormant cell pool may be activated in order to reduce the production needed per active clone. A simple model for hematopoiesis in equilibrium would have \( N \) following a birth-and-death process with each birth being marked domestic with probability \( p \). In contrast to the simpler model on which we have based inference, this model decouples births and deaths of clones. Clones are born at a constant rate, and live independent exponential lifetimes. Since no finite Markov process is involved, the recursive updating technology does not apply, and likelihood theory may be quite difficult. Such a model is, however, very easy to simulate. A simulation study has indicated that cell lifetimes much longer than those estimated for cats 40004 and 40005 cannot produce fluctuations in the binomial success probability large enough (and at the right time scale) to account for the observed proportions. Therefore, even if \( N \) is not constant, it appears that stem cell lifetimes must be relatively short. Figure 6 shows a simulated sample path from this model where the parameters are set to make it comparable with the simpler fixed \( N \) model. The upper panel of Figure 6 shows the fluctuating number of clones, where the expected number of clones equals 8, and the mean lifetime of clones is 5 weeks (the MLE from Figure 2). The lower panel shows the fluctuating domestic-proportion in the precursor compartment. By comparison with Figure 4, we see that fluctuations of similar extent and on similar time scales occur in the two models.

6.3 Proliferative Potential

A simple measure of the proliferative potential (number of progeny per stem cell life span) of a stem cell is the ratio \( L/N \). For example, in cat 40004 an individual cell may support \( 1/7 \) of hematopoiesis for 26 weeks, with a proliferative potential of 3.7. In a study of cats subjected to chemotherapy instead of autologous transplantation (Abkowitz et al., 1992) the analysis of section 5 yielded a very different picture. The biology in these two settings is very different, since the cells in the chemotherapy experiment were affected by the treatment, while unmanipulated cells were returned.
Figure 6: A simulated state for a birth-death model allowing fluctuations in $N$, the number of clones in $C_2$. See section 6.2.
to the cat in the case of autologous bone marrow transplants. For one cat, followed for six years after first therapy, the estimated proliferative potential was less than 0.05. Separate analysis of the early and late parts of the data, using the model of section 4, indicated that as time goes on, the hematopoietic stem cell behavior becomes more and more abnormal. This suggests that initially cells with minimal damage supported hematopoiesis, and that as these cells were exhausted through terminal differentiation, more damaged cells supported blood cell production.

We did consider a model for transplanted animals in which the proliferative potential was held constant between cats, while $L$ and $N$ were varying proportionally. This model, however, did not explain the data as well as the model with independently varying parameters. As more data become available we will look into this type of modeling effort more carefully.

References


