Title
Estimation for the simple linear Boolean model

Journal
Methodology and Computing in Applied Probability, 8(4)

ISSN
1387-5841

Authors
Crespi, Catherine M
Lange, K

Publication Date
2006-12-01

Peer reviewed

Methodology and Computing in Applied Probability (2006) 8: 559-571

Estimation for the Simple Linear Boolean Model

Catherine M. Crespi

Department of Biostatistics, University of California, Los Angeles, Box 951772, Los Angeles, CA 90095-1772, ccrespi@ucla.edu

Kenneth Lange

Departments of Biomathematics and Human Genetics, University of California, Los Angeles, Box 951766, Los Angeles, CA 90095-1766, klange@ucla.edu

Abstract. We consider the simple linear Boolean model, a fundamental coverage process also known as the Markov/General/∞ queue. In the model, line segments of independent and identically distributed length are located at the points of a Poisson process. The segments may overlap, resulting in a pattern of "clumps" – regions of the line that are covered by one or more segments – alternating with uncovered regions or "spacings." Study and application of the model have been impeded by the difficulty of obtaining the distribution of clump length. We present explicit expressions for the clump length distribution and density functions. The expressions take the form of integral equations, and we develop a method of successive approximation to solve them numerically. Use of the fast Fourier transform greatly enhances the computational efficiency of the method. We further present inference procedures for the model using maximum likelihood techniques. Applications in engineering and biomedicine illustrate the methods.

Keywords: Boolean model; Coverage process; Markov/General/∞ queue; Type II counter

AMS 2000 Subject Classification: Primary 60-08; Secondary, 60K25, 62M09, 65R20

Corresponding author and reprint requests: Catherine M. Crespi, Ph.D., Department of Biostatistics, UCLA School of Public Health, CHS 51-236D, Los Angeles, CA 90095-1772. **Phone:** 310-267-1808 **Email:** ccrespi@ucla.edu

1. Introduction

Coverage processes are random set processes in which a random mechanism governs the position of random sets on a line, plane or other space. We study inference for one of the most fundamental coverage processes, the Poisson distribution of segments on a line. This process, termed the simple linear Boolean model by Hall (1988), has been broadly applied in diverse disciplines, including engineering (Hall, 1988; Takacs, 1962), physics (Bingham and Pitts, 1999), epidemiology (Parthasarathy, 1997), medicine (Crespi et al., 2005), and genetics (Arratia et al., 1991; Percus, 2002). In queueing theory, the model is equivalent to the Markov/General/∞ queue (Kleinrock, 1975).

Figure 1 illustrates the model. In the model, segments of independent and identically distributed (iid) length $\{S_i, i \geq 1\}$ are located at the points of a stationary Poisson process, $\{\xi_i, i \geq 1\}$. The Poisson process is taken to be independent of the segment lengths. The segments may overlap, resulting in a pattern of "clumps" – regions of the line that are covered by one or more segments – alternating with uncovered regions or "spacings."

The problem we consider is to estimate the intensity of the Poisson process and the segment length distribution from a sample of clump and spacing lengths. The challenge of this problem has been noted in the literature (Hall, 1988; Handley, 1999) and stems from the difficulty of obtaining the distribution of the length of a clump. Expressing the distribution of clump length has been uniquely challenging since each clump arises from a random number of line segments in a random configuration. Previous expressions for the clump length distribution have involved the degenerate case of fixed segment length (Hall, 1988), analytically intractable expressions such as an infinite sum of self-convolutions (Stadje, 1985), discrete approximation (Handley, 1999) and recursion (Daley, 2001). Thus tractable methods of obtaining the clump length distribution, preferably for an arbitrary segment length distribution, are needed.

We develop a procedure for obtaining the distribution and density of clump length in the simple linear Boolean model when segment lengths follow an arbitrary distribution. Beginning with the expression of Daley (2001), we derive explicit representations of the clump length distribution and density functions. Each representation takes the form of an integral equation, which can be solved numerically by constructing a sequence of functions that converges uniformly to its solution. The familiar method of successive approximations furnishes the functional iterates (Perko, 1991). We greatly reduce the computational complexity of each iteration by invoking the fast Fourier transform (Lange, 1999). Thus we are able to obtain the clump length distribution and density functions as

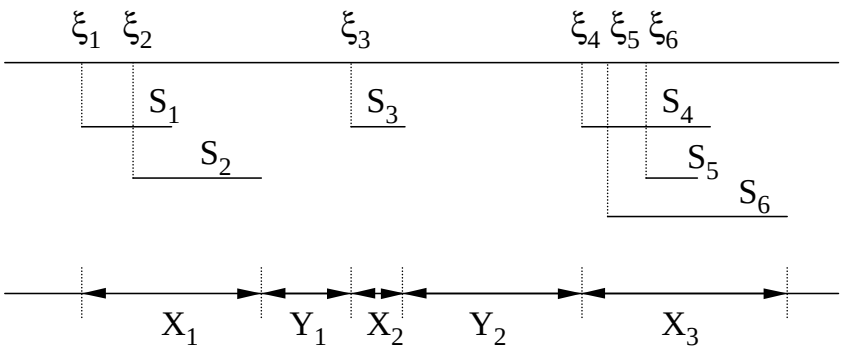

Figure 1. The simple linear Boolean model.

the limit of a suitably defined sequence of functions in a computationally efficient manner.

We further show how these methods can be applied to estimate the intensity of the Poisson process and the segment length distribution, when determined up to a set of unknown parameters, from a sample of clumps and spacings. We use maximum likelihood techniques and include a procedure for estimating the asymptotic variances of the estimators. We illustrate the methods with two examples: the measurement of particle mass flow and the estimation of the frequency and duration of a recurrent viral infection.

Previous work on estimation for the model includes Handley (2004), who uses a discrete Boolean model as an approximation to the continuous model and provides an algorithm to compute the discrete clump length density. We compare this algorithm with our method. We focus on parametric estimation procedures; nonparametric methods are discussed in Bingham and Pitts (1999) and Handley (1999).

2. Main Results

We begin with a statement of the model and its properties, drawing from Hall (1988). Suppose that points on the real line are produced by a stationary Poisson process with unknown constant intensity λ and that each point serves as the left-hand end of a line segment. Equivalently, we may consider placing the segment midpoints or right-hand ends at the points. Line segments are chosen independently of anchoring points and have independent lengths with common distribution $G(x)$, determined up to an unknown parameter vector $\boldsymbol{\theta}$. A connected set of segments not intersected by any other segments is called a clump, and the interval between successive clumps is called a spacing.

Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ denote the lengths of successive clumps and spacings, respectively. All of these random variables are independent (Hall, 1988). Each of the random variables $Y_1, Y_2, ...$ is exponentially distributed with mean λ^{-1} (Hall, 1988). The variables $X_1, X_2, ...$ are also identically distributed. Daley (2001) has shown that their common cumulative distribution function $Z(x)$ satisfies the integral equation

$$\bar{Z}(x) \quad = \quad \bar{G}(x) + \int_0^x \int_0^u \lambda \bar{Z}(x-v) \mathrm{e}^{-\lambda \int_0^v \bar{G}(t)dt} dv \, dG(u) \tag{1}$$

where $\bar{Z}(x) = 1 - Z(x)$ and $\bar{G}(x) = 1 - G(x)$. This equation uses the logic that for the clump length to exceed x, either (i) the length of the first segment exceeds x, or else (ii) the first segment has length $u < x$, and (iii) another segment begins at the point $v < u$ which starts (iv) a "pseudo-clump" with length at least $x - v$. A pseudo-clump is the clump that a segment would have started if the line had been vacant at the epoch of the segment's arrival. Pseudo-clumps have the same distribution $Z(x)$ as true clumps. The probabilities of these events are (i) $\bar{G}(x)$, (ii) $dG(u)$, (iii) λdv and (iv) $\bar{Z}(x-v)$. In order to capture the first segment with pseudo-clump length extending beyond x, we additionally require that there are no segments starting in $[0, v]$ that extend beyond v. If we let $Q(v)$ denote the probability of this event, it follows that

$$Q(v + \triangle v) \quad = \quad [1 - \lambda \triangle v + \lambda \triangle v G(v) + o(\triangle v)] Q(v).$$

Forming the difference quotient

$$\frac{Q(v + \triangle v) - Q(v)}{\triangle v} \quad = \quad -\lambda [1 - G(v)] Q(v) + \frac{o(\triangle v)}{\triangle v}$$

and sending $\triangle v$ to 0 produce the ordinary differential equation

$$Q'(v) \quad = \quad -\lambda [1 - G(v)] Q(v),$$

with explicit solution $Q(v) = \exp\{-\lambda \int_0^v [1 - G(t)] dt\}$ subject to the initial condition $Q(0) = 1$.

We note that equation (1) may be simplified via Fubini's Theorem to

$$\bar{Z}(x) \quad = \quad \bar{G}(x) + \int_0^x \bar{Z}(x-v)[G(x) - G(v)] h(v) dv, \tag{2}$$

where $h(v) = \lambda \mathrm{e}^{-\lambda \int_0^v \bar{G}(t)dt}$. This expression is a Volterra integral equation (Yosida, 1960) of the form

$$f(x) \quad = \quad b(x) + \int_0^x f(x-v)[G(x) - G(v)] h(v) dv, \tag{3}$$

which can be solved by functional iteration, as laid out in the following proposition.

Proposition 1. *Consider the sequence of functions defined by*

$$f_n(x) \;=\; b(x) + \int_0^x f_{n-1}(x-v)[G(x) - G(v)]h(v)dv \tag{4}$$

starting from $f_0(x) \equiv 0$. If the function $b(x)$ is bounded on a closed interval I containing 0, then a unique bounded solution $f(x)$ exists, and $f_n(x)$ converges to $f(x)$ uniformly on I. If $b(x)$ is continuous as well as bounded, then $f(x)$ is also continuous.

The proof of this and the subsequent proposition can be found in the Appendix.

To implement this result and compute the function $f(x)$ at m equally spaced points in an interval $[0, d]$, one may approximate the integral in (4) by a sum and compute, for $k = 0, ..., m$, the nth iterate

$$\begin{aligned} f_n\left(\frac{kd}{m}\right) \;=\; & b\left(\frac{kd}{m}\right) + G\left(\frac{kd}{m}\right)\frac{d}{m}\sum_{j=0}^{k} f_{n-1}\left[\frac{(k-j)d}{m}\right] h\left(\frac{jd}{m}\right) \\ & -\frac{d}{m}\sum_{j=0}^{k} f_{n-1}\left[\frac{(k-j)d}{m}\right] G\left(\frac{jd}{m}\right) h\left(\frac{jd}{m}\right). \end{aligned} \tag{5}$$

The sums in equation (5) are convolutions and may be computed efficiently using the fast Fourier transform (Brigham, 1988; Lange, 1999). This tactic decreases the computational complexity of each iteration from $O(m^2)$ to $O(m \ln m)$.

If we assume that the segment length distribution $G(x)$ possesses a bounded density $g(x)$, then it is natural to conjecture that the clump length distribution $Z(x)$ also possesses a density $z(x)$. This is indeed the case, and differentiation of equation (2) using Leibniz's rule for differentiation under the integral sign shows that $z(x)$ satisfies the integral equation in the next proposition.

Proposition 2. *The density of clump length exists and satisfies the equation*

$$z(x) \;=\; g(x)\left[1 - \int_0^x \bar{Z}(x-v)h(v)dv\right] + \int_0^x z(x-v)[G(x) - G(v)]h(v)dv. \tag{6}$$

This integral equation has the form of equation (3), and Proposition 1 may be used to solve it numerically. In particular, solutions for both the distribution and density of clump length for an arbitrary segment length distribution may be obtained by first computing $\bar{Z}(x)$ as the limit of the sequence

$$\bar{Z}_n(x) \;=\; \bar{G}(x) + \int_0^x \bar{Z}_{n-1}(x-v)[G(x) - G(v)]h(v)dv$$

starting from $\bar{Z}_0(x) \equiv 0$ and then using $\bar{Z}(x)$ to compute the limit $z(x)$ of the sequence

$$z_n(x) \;=\; g(x)\left[1 - \int_0^x \bar{Z}(x-v)h(v)dv\right] + \int_0^x z_{n-1}(x-v)[G(x) - G(v)]h(v)dv$$

starting from $z_0(x) \equiv 0$.

3. Estimation

The methods in Section 2 can be used to estimate the intensity of the Poisson process and the parameters of the segment length distribution from a sample of clump and spacings lengths. The first step is to express the likelihood. Imagine that a simple linear Boolean model is observed over an interval whose left and right endpoints each coincide with the end of a clump or spacing, yielding a sample S of spacing lengths, $\{y_i : i \in S\}$, and a sample C of clump lengths, $\{x_i : i \in C\}$. Collect unknown parameters into a vector $\boldsymbol{\psi}^t = (\lambda, \boldsymbol{\theta}^t)$ of length p. Due to the mutual independence of clump and spacing lengths, the likelihood function appropriate for maximum likelihood estimation of $\boldsymbol{\psi}$ is

$$\mathcal{L}_1(\boldsymbol{\psi}) \;=\; \prod_{i \in S} \lambda \mathrm{e}^{-\lambda y_i} \prod_{i \in C} z(x_i, \boldsymbol{\psi}). \tag{7}$$

In some cases, the endpoints of the observation interval may not coincide with the endpoints of clumps or spacings, so there may be remnants of clumps or spacings at the beginning or end of the interval. If the left endpoint of the interval occurs at a random point within a spacing, this spacing remnant contributes a factor $\lambda \mathrm{e}^{-\lambda y_i}$ to the likelihood. A spacing remnant at the right will contribute a factor $\mathrm{e}^{-\lambda y_i}$, and a clump remnant at the right will contribute a factor $\bar{Z}(x_i)$. The contribution of a clump remnant at the left of the interval is complicated and we neglect this possibility. Thus a more general expression for the likelihood is

$$\mathcal{L}_2(\boldsymbol{\psi}) \;=\; \prod_{i \in S \cup S_L} \lambda \mathrm{e}^{-\lambda y_i} \prod_{i \in S_R} \mathrm{e}^{-\lambda y_i} \prod_{i \in C} z(x_i, \boldsymbol{\psi}) \prod_{i \in C_R} \bar{Z}(x_i, \boldsymbol{\psi}) \tag{8}$$

where S_L and S_R denote the sets of remnant spacings at the left and right, respectively, and C_R denotes the set of remnant clumps at the right. The terms $z(x_i, \boldsymbol{\psi})$ and $\bar{Z}(x_i, \boldsymbol{\psi})$ in (8) may be obtained as described in Section 2.

The first partial derivatives $\frac{\partial}{\partial \psi_i} \ln \mathcal{L}(\boldsymbol{\psi})$ of the log likelihood provide the gradient for optimization purposes and the raw material for obtaining the asymptotic variances of the estimators. To compute one of these derivatives accurately, it clearly suffices to compute the components $\frac{\partial}{\partial \psi_i} \bar{Z}(x, \boldsymbol{\psi})$ and $\frac{\partial}{\partial \psi_i} z(x, \boldsymbol{\psi})$ accurately. Differentiating equation (2) shows that the former function

satisfies

$$\begin{aligned}
\frac{\partial}{\partial\psi_i}\bar{Z}(x,\boldsymbol{\psi}) &= \frac{\partial}{\partial\psi_i}\bar{G}(x,\boldsymbol{\psi}) + \int_0^x \frac{\partial}{\partial\psi_i}\bar{Z}(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv \\
&\quad + \int_0^x \bar{Z}(x-v,\boldsymbol{\psi})\frac{\partial}{\partial\psi_i}[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv \\
&\quad + \int_0^x \bar{Z}(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]\frac{\partial}{\partial\psi_i}h(v,\boldsymbol{\psi})dv \qquad (9) \\
&= b_1(x,\boldsymbol{\psi}) + \int_0^x \frac{\partial}{\partial\psi_i}\bar{Z}(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv,
\end{aligned}$$

and differentiating equation (6) shows that the latter function satisfies

$$\begin{aligned}
\frac{\partial}{\partial\psi_i}z(x,\boldsymbol{\psi}) &= \frac{\partial}{\partial\psi_i}g(x,\boldsymbol{\psi}) + \int_0^x \frac{\partial}{\partial\psi_i}z(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv \\
&\quad + \int_0^x z(x-v,\boldsymbol{\psi})\frac{\partial}{\partial\psi_i}[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv \\
&\quad + \int_0^x z(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]\frac{\partial}{\partial\psi_i}h(v,\boldsymbol{\psi})dv \\
&\quad - \frac{\partial}{\partial\psi_i}g(x,\boldsymbol{\psi})\int_0^x \bar{Z}(x-v,\boldsymbol{\psi})h(v,\boldsymbol{\psi})dv \qquad (10) \\
&\quad - g(x,\boldsymbol{\psi})\int_0^x \frac{\partial}{\partial\psi_i}\bar{Z}(x-v,\boldsymbol{\psi})h(v,\boldsymbol{\psi})dv \\
&\quad - g(x,\boldsymbol{\psi})\int_0^x \bar{Z}(x-v,\boldsymbol{\psi})\frac{\partial}{\partial\psi_i}h(v,\boldsymbol{\psi})dv \\
&= b_2(x,\boldsymbol{\psi}) + \int_0^x \frac{\partial}{\partial\psi_i}z(x-v,\boldsymbol{\psi})[G(x,\boldsymbol{\psi})-G(v,\boldsymbol{\psi})]h(v,\boldsymbol{\psi})dv.
\end{aligned}$$

These derivatives are integral equations of the form (3), and if $b_1(x,\boldsymbol{\psi})$ and $b_2(x,\boldsymbol{\psi})$ are bounded, then the partial derivatives can be obtained as the limits of sequences suitably defined by equation (4).

In principle, second partial derivatives can be computed by the same methods. In practice, numerical differentiation of the first partial derivatives is simpler to implement. Standard errors may be attached to the maximum likelihood estimates by inversion of the observed information matrix

$$\mathcal{I} = \left(\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ln\mathcal{L}(\psi)\right)$$

evaluated at the maximum likelihood estimate $\hat{\psi}$.

4. Comparison to Discrete Approximation

Handley (2004) uses the one-dimensional binomial germ-grain model as a discrete approximation to the continuous Boolean model and presents approximate likelihood procedures. The discrete

model involves a Bernoulli process that marks points on the discrete line with marking probability p. Marks are associated with line segments whose lengths have distribution $C(x)$. Using a recursive accounting of all possible events leading to a covered interval of length x, Handley and Dougherty (1996) show that clump length in their model has discrete probability density

$$\Pr(X = x) \;=\; \sum_{j=1}^{x} [F(x) - F(j-1)] \prod_{i=1}^{j-1} F(i-1) \Pr(X = x - j)$$

for $x = 1, 2, ...$, where $F(x) = 1 - p + pC(x)$ and $\Pr(X = 0) = (1-p)/p$. By discretizing an interval of length x into mx pieces of length $1/m$ and approximating a Poisson process by a Bernoulli process with $p = \lambda/m$, one obtains the recurrence

$$\Pr(X = x) \;=\; \sum_{j=1}^{[mx]} \left[F(x) - F\left(\frac{j-1}{m}\right) \right] \prod_{i=1}^{j-1} F\left(\frac{i-1}{m}\right) \Pr(X = x - j/m)$$

for $x = 1/m, 2/m, ...$, where $F(x) = 1 - \lambda/m + \lambda C(x)/m$. This expression can be used to compute the discrete density function on an array of m points starting at $\Pr(X = 1) = [F(1) - 1 + p](1-p)/p$.

The computational efficiency of Handley's algorithm can be increased by storing intermediate products and distribution values. However, because computing most values requires $O(m)$ operations, the total number of operations required to obtain values for all m points is $O(m^2)$. In contrast, our algorithm sequentially obtains $\bar{Z}(x)$ and $z(x)$ by functional iteration, and application of the fast Fourier transform reduces the computational complexity of each iteration to $O(m \ln m)$. In practice, convergence is typically achieved in 3-8 iterations, resulting in an overall computation complexity well below $O(m^2)$. In general, the number of iterations will depend on the rate of convergence. In equation (12) of the Appendix, we prove that convergence occurs at a geometric rate of $1 - e^{-\lambda\mu}$, where μ is the mean segment length. Note that $1 - e^{-\lambda\mu} = E(X)/[E(X) + E(Y)]$, the proportion of the real line expected to be covered by clumps (Kleinrock, 1975). Thus our algorithm converges rapidly when clumps are sparse and slowly when clumps are large relative to spacings. Even in extreme cases, the number of iterations is likely to be well below $m/\ln m$. However, statistical inference will be difficult using either method in this limiting case.

Our approach has the further advantage of providing accurate approximation to partial derivatives, which are essential to fast optimization. Solution of the integral equations for the various partial derivatives is apt to be more accurate than numerical differentiation, which by definition is prone to round-off error owing to cancellation of quantities of comparable magnitude. Functional iteration for the partial derivatives also benefits from the fast Fourier transform, producing even further gains in computational speed.

Handley (2004) constructs approximate $100(1-\alpha)\%$ confidence regions for maximum likelihood estimates using the likelihood ratio statistic, whereas we invert the observed Fisher information matrix to obtain standard errors. Both methods rely on asymptotic normality assumptions and are likely to produce similar results.

5. Examples

We illustrate our procedure with two examples: measurement of particle mass flow, and estimation of the frequency and duration of a recurrent viral infection. In both examples, the clump length distribution (2), density (6) and first partial derivatives (9) and (10) were computed at $m = 2000$ points in less than a second using the R software package on a Windows-based personal computer. The negative log likelihood was minimized using the nlm function in R, which uses a Newton-type algorithm.

5.1. Example: Particle Flow Measurement

A Type II counter (Hall, 1988; Takacs, 1962) is a basic type of device for measuring particle flow. Some electronic counters follow similar principles. In a Type II counter, flowing particles pass by a sensor, giving rise to a signal that is "on" as any particle passes the sensor and "off" otherwise. The passage times of particles may overlap; thus an "on" signal corresponds to a clump with an unknown number of particles. The "off" signals correspond to the spacings between clumps. When particle arrivals are Poisson, the clump and spacing lengths follow a simple linear Boolean model.

A fundamental problem with Type II counters is to estimate the total number of particles passing through the device. We illustrate how our methods can be used to solve this problem using data collected from the particle flow measurement device described by Grift (2003).

In an experiment with the device, 4000 identical steel spheres of diameter 4.45 millimeters (mm) were dispensed through a flow tube in a manner approximating a Poisson process. Optical sensors recorded the lengths of the clumps and spacings. Our objective was to use the clump and spacing lengths to estimate the particle flow rate (in particles per second), total number of particles flowing through the device, and the mean particle diameter.

Particle diameter, which corresponds to segment length in the Boolean model, is measured with approximately normal error by the device (Grift et al., 2001). Thus particle diameter was assumed to be normally distributed with unknown mean and standard deviation (μ, σ), and the object of inference was $\boldsymbol{\psi} = (\lambda, \mu, \sigma)$, where λ represents flow rate. The total number of particles N flowing

through the device was calculated as λL, where L is the length of the observation period, recorded as 8.5255 seconds. Preliminary analysis suggested a small excess of short clumps and spacings compared to a perfect Boolean model. Hence the shortest six percent of clumps and spacings were omitted, yielding a sample of 2781 clumps and 2781 spacings for the present analysis.

The likelihood (7) of the clump and spacing data was maximized to obtain $\widehat{\psi}$. Standard errors were obtained as described in Section 3. The estimated flow rate was $\hat{\lambda} = 478.7 \pm 7.9$ particles per second, corresponding to a total particle flow of $\hat{N} = 4081 \pm 67$, an estimate that compares favorably with the known total of 4000. Mean particle diameter was estimated as $\hat{\mu} = 4.445 \pm .005$ mm, which also compares favorably with the known value of 4.45 mm. The estimated standard deviation of particle diameter was $\hat{\sigma} = 0.187 \pm .004$ mm.

Figure 2 provides a histogram of clump lengths overlaid with an estimate of the clump length density, calculated using equation (6) and the methods in Section 2 with the maximum likelihood estimates as parameter values. The density features a peak at the mean particle diameter, suggesting clumps composed predominantly of single particles, a plateau spanning the range of one to two particle diameters, and then a monotone decline. These features reflect this particular model specification.

5.2. Example: Frequency and Duration of a Recurrent Viral Infection

Crespi et al. (2005) use the Markov/Markov/∞ queueing model, which is equivalent to the simple linear Boolean model with exponentially distributed segment lengths, to describe chronic infection with herpes simplex virus type 2 (HSV-2), the principal cause of genital herpes. Our methods provide a way of generalizing the model to other segment length distributions. Application of the model to HSV-2 is described in detail in Crespi et al. (2005). In brief, HSV-2 establishes a persistent infection in the nervous system and intermittently reactivates. Reactivations are associated with short periods of viral shedding from the mucosa. Successive reactivations may overlap in time. Thus individuals alternate between viral shedding periods, composed of one or more reactivations, and non-shedding periods. Referring to Figure 1, let $\xi_1, \xi_2, ...$ represent the reactivation times and let $S_1, S_2, ...$ represent the reactivation durations. Then $X_1, X_2, ...$ represent viral shedding periods, in which one or more reactivations are present, and $Y_1, Y_2, ...$ represent non-shedding periods.

During the clinical trial described by Wald et al. (1997), subjects received anti-HSV therapy for 10 weeks and placebo for 10 weeks, in random order. The time course of the subjects' shedding and non-shedding periods, during both treatment and placebo, was assayed by daily collection and analysis of mucosal secretions to detect viral shedding. Our objective was to use the durations of

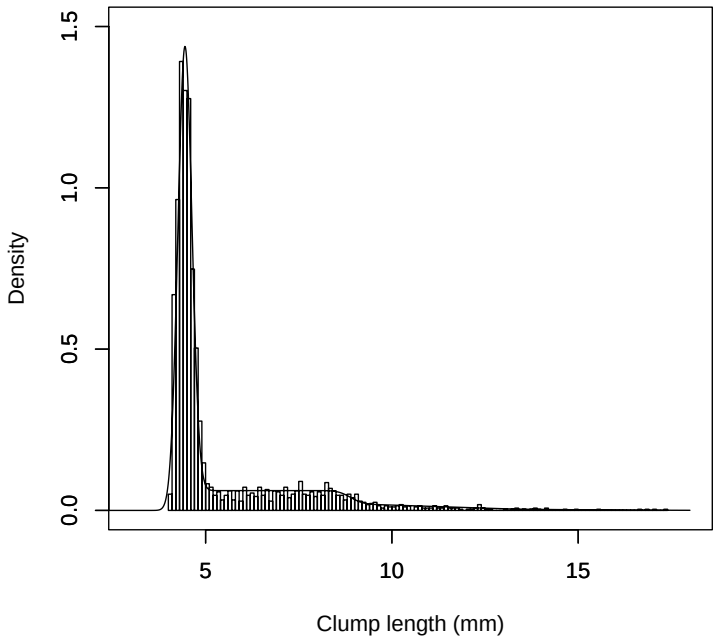

Figure 2. Histogram of clump lengths from a particle flow counter overlaid with the estimated clump length density, computed using the methods in Section 2.

the shedding and non-shedding periods to estimate the frequency and duration of the reactivations of the virus.

For simplicity, we consider data from a single subject. For both the placebo and treatment arms of the trial, observation of the subject began and ended during non-shedding periods. Thus the data include spacing remnants at the left and right endpoints. The data for the placebo arm are $S^{Pl} = \{15, 2, 4, 1, 3, 13\}$, $C^{Pl} = \{2, 1, 2, 7, 1, 1\}$ and $S_R^{Pl} = \{13\}$, and the data for the treatment arm are $S^{Tx} = \{38, 20\}$, $C^{Tx} = \{6, 1\}$ and $S_R^{Tx} = \{7\}$. We note that since the secretion samples were taken at discrete daily time points, these data are an approximation of the continuous-time process.

We modeled reactivation duration as a Weibull(α, β) distribution, where α is the shape parameter and β is the scale parameter. The Weibull is a flexible distribution covering a range of plausible biological assumptions, and includes the exponential distribution as a special case. The mean of the Weibull is $\beta\Gamma(1 + 1/\alpha)$.

The likelihood of the data was expressed using equation (8) without the fourth term and maximized to obtain $\widehat{\boldsymbol{\psi}} = (\hat{\lambda}, \hat{\alpha}, \hat{\beta})$. Standard errors were obtained as described in Section 3. The mean reactivation duration was estimated as $\hat{\beta}\Gamma(1+1/\hat{\alpha})$, and its standard error was approximated using

a first-order Taylor expansion. Based on this approach, the reactivation frequency, with standard error, was estimated to be 3.6 ± 1.5 reactivations per month under placebo and 0.9 ± 0.7 reactivations per month under treatment. The mean duration of a reactivation was estimated to be 2.1 ± 0.6 days under placebo and 3.3 ± 1.7 days under treatment. A likelihood ratio test of the null hypothesis $H_0 : \boldsymbol{\psi}^{Pl} = \boldsymbol{\psi}^{Tx}$ yielded a test statistic of 3.629, distributed as χ^2 with 3 degrees of freedom, corresponding to a p-value of 0.304. Although this is nonsignificant, the estimates hint that the treatment was efficacious. More data are needed to resolve the issue.

6. Discussion

We have derived an analytical expression for the clump length density in the simple linear Boolean model along with a numerical solution and procedures for conducting maximum likelihood estimation under the model. The analytical expression and method of solution are general and can be applied with an arbitrary segment length distribution. Implementation is straightforward, with minimal computational effort.

The simple linear Boolean model is the basis for more complex coverage processes, including two- and three-dimensional models (Hall, 1988; Handley, 1999). The model also has a close connection to certain queuing theory (Kleinrock, 1975) and spatial models (Molchanov, 1997). We hope that the results presented here may stimulate further study, application, and improvement of these models.

Appendix

Proof of Proposition 1. We demonstrate convergence of the telescoping series

$$f_n(x) \;=\; \sum_{m=1}^{n} [f_m(x) - f_{m-1}(x)]$$

under the sup norm $\|a\|_\infty = \sup_{x \in I} |a(x)|$ defined for bounded functions $a(x)$ on I. Because the terms of the series satisfy

$$f_m(x) - f_{m-1}(x) \;=\; \int_0^x [f_{m-1}(x-v) - f_{m-2}(x-v)][G(x) - G(v)]h(v)dv,$$

the fundamental theorem of calculus entails

$$\begin{aligned} \|f_m - f_{m-1}\|_\infty &\leq \|f_{m-1} - f_{m-2}\|_\infty \int_0^\infty \bar{G}(v)h(v)dv \\ &= \|f_{m-1} - f_{m-2}\|_\infty (1 - \mathrm{e}^{-\lambda\mu}), \end{aligned}$$

where $\mu = \int_0^\infty \bar{G}(t)dt < \infty$. If $b(x)$ is bounded, induction demonstrates that

$$\begin{aligned} \|f_m - f_{m-1}\|_\infty &\leq \|f_{m-2} - f_{m-3}\|_\infty (1 - \mathrm{e}^{-\lambda\mu})^2 \\ &\vdots \\ &\leq \|f_1 - f_0\|_\infty (1 - \mathrm{e}^{-\lambda\mu})^{m-1} \\ &= \|b\|_\infty (1 - \mathrm{e}^{-\lambda\mu})^{m-1}. \end{aligned}$$

For $n \geq m$, this inequality in turn implies that

$$\begin{aligned} \|f_n - f_m\|_\infty &\leq \|b\|_\infty \sum_{k=m}^{n-1} (1 - \mathrm{e}^{-\lambda\mu})^k \\ &\leq \|b\|_\infty \sum_{k=m}^{\infty} (1 - \mathrm{e}^{-\lambda\mu})^k \\ &= \|b\|_\infty \mathrm{e}^{\lambda\mu} (1 - \mathrm{e}^{-\lambda\mu})^m. \end{aligned} \tag{11}$$

Hence, the sequence $f_n(x)$ satisfies Cauchy's criterion and converges uniformly to a bounded limit $f(x)$ on I. If $b(x)$ is continuous as well as bounded, then each iterate $f_n(x)$ is also continuous, and the limit $f(x)$ must be continuous on I.

To see that the solution is unique, let $f(x)$ and $f^*(x)$ be two bounded solutions. Then the equation

$$f(x) - f^*(x) = \int_0^x [f(x-v) - f^*(x-v)][G(x) - G(v)]h(v)dv$$

implies the contraction property

$$\|f - f^*\|_\infty \leq (1 - \mathrm{e}^{-\lambda\mu})\|f - f^*\|_\infty,$$

which is untenable unless $\|f - f^*\|_\infty = 0$.

Finally, for the record, it is worth observing that inequality (11) carries over to

$$\|f - f_n\|_\infty \leq \|b\|_\infty \mathrm{e}^{\lambda\mu} (1 - \mathrm{e}^{-\lambda\mu})^n. \tag{12}$$

In other words, $f_n(x)$ converges to $f(x)$ at the geometric rate $1 - \mathrm{e}^{-\lambda\mu}$ in the sup norm. □

Proof of Proposition 2. If a sequence of differentiable functions $f_n(x)$ and its sequence of derivatives $f_n'(x)$ both converge uniformly on I, then the limit function $f(x) = \lim_{n\to\infty} f_n(x)$ is differentiable with derivative $f'(x) = \lim_{n\to\infty} f_n'(x)$ (Theorem 7.17, Rudin (1964)). We will apply

this principle to $f_n(x) = -\bar{Z}_n(x)$. Now induction and Leibniz's rule imply that $f_n(x)$ is differentiable with derivative satisfying the integral equation

$$f_n'(x) \;=\; g(x)\left[1 - \int_0^x \bar{Z}_{n-1}(x-v)h(v)dv\right] + \int_0^x f_{n-1}'(x-v)[G(x)-G(v)]h(v)dv$$

On the other hand, the sequence

$$z_n(x) \;=\; g(x)\left[1 - \int_0^x \bar{Z}(x-v)h(v)dv\right] + \int_0^x z_{n-1}(x-v)[G(x)-G(v)]h(v)dv$$

is known to converge uniformly to a limit $z(x)$. Our strategy therefore is to prove that $f_n'(x)$ also converges uniformly to $z(x)$. With this end in mind, we first note that the integral $c = \int_I h(v)dv$ is finite whenever I is finite. We put this fact to use in the bound

$$\|z_n - f_n'\|_\infty \;\leq\; c\|g\|_\infty\|\bar{Z} - \bar{Z}_{n-1}\|_\infty + \|z_{n-1} - f_{n-1}'\|_\infty(1 - \mathrm{e}^{-\lambda\mu}). \tag{13}$$

Inequality (12) with $b(x) = \bar{G}(x)$ implies that

$$\|\bar{Z} - \bar{Z}_{n-1}\|_\infty \;\leq\; \mathrm{e}^{\lambda\mu}(1 - \mathrm{e}^{-\lambda\mu})^{n-1}.$$

Substituting this bound in inequality (13), setting $a = c\|g\|_\infty e^{\lambda\mu}$ and $r = 1 - e^{-\lambda\mu}$, and iterating produce

$$\begin{aligned}
\|z_n - f_n'\|_\infty &\leq c\|g\|_\infty \mathrm{e}^{\lambda\mu}(1 - \mathrm{e}^{-\lambda\mu})^{n-1} + \|z_{n-1} - f_{n-1}'\|_\infty(1 - \mathrm{e}^{-\lambda\mu}) \\
&= ar^{n-1} + \|z_{n-1} - f_{n-1}'\|_\infty r \\
&\leq ar^{n-1} + ar^{n-1} + \|z_{n-2} - f_{n-2}'\|_\infty r^2 \\
&\;\;\vdots \\
&\leq nar^{n-1}.
\end{aligned}$$

Since $0 \leq r < 1$, the product nar^{n-1} tends to 0, and $f_n'(x)$ converges uniformly to $z(x)$ on I. □

Acknowledgments

The authors wish to acknowledge Tony Grift for providing the particle mass flow data and Larry Corey and Anna Wald for providing the herpes simplex virus data. Crespi was supported by National Institutes of Health grant T32 AI007370.

References

R. Arratia, E. S. Lander, S. Tavare, and M. S. Waterman, "Genomic mapping by anchoring random clones: a mathematical analysis," *Genomics* vol. 11 pp. 806–827, 1991.

N. H. Bingham and S. M. Pitts, "Non-parametric estimation for the M/G/∞ queue," *Annals of the Institute of Statistical Mathematics* vol. 51 pp. 71–97, 1999.

E. O. Brigham, *The Fast Fourier Transform and its Applications*, Prentice Hall, Upper Saddle River, NJ, 1988.

C. M. Crespi, W. G. Cumberland, and S. Blower, "A queueing model for chronic recurrent conditions under panel observation," *Biometrics* vol. 61 pp. 194–199, 2005.

D. J. Daley, "The busy periods of the M/GI/∞ queue," *Queueing Systems* vol. 38 pp. 195–204, 2001.

T. Grift, "Fundamental mass flow measurement of solid particles," *Particulate Science and Technology* vol. 21 pp. 177–193, 2003.

T. Grift, J. T. Walker, and J. W. Hofstee, "Mass flow measurement of granular materials in aerial application – Part 2: experimental model validation," *Transactions of the ASAE* vol. 44 pp. 27–34, 2001.

P. Hall, *Introduction to the Theory of Coverage Processes*, John Wiley and Sons, New York, 1988.

J. C. Handley, "Discrete approximation of the linear Boolean model of heterogeneous materials," *Physical Review E* vol. 60 pp. 6150–52, 1999.

J. C. Handley, "Computationally efficient approximate likelihood procedures for the Boolean model," *Computational Statistics and Data Analysis* vol. 45 pp. 125–136, 2004.

J. C. Handley and E. R. Dougherty, "Optimal nonlinear filter for signal-union-noise and runlength analysis in the directional one-dimensional discrete Boolean random set model," *Signal Processing* vol. 51 pp. 147–166, 1996.

L. Kleinrock, *Queueing Systems Volume I: Theory*, John Wiley and Sons, New York, 1975.

K. Lange, *Numerical Analysis for Statisticians*, Springer-Verlag, New York, 1999.

I. Molchanov, *Statistics of the Boolean Model for Practitioners and Mathematicians*, John Wiley and Sons, Chichester, England, 1997.

P. R. Parthasarathy, "The effect of superinfection on the distribution of the infectious period – a continued fraction approximation," *IMA Journal of Mathematics Applied in Medicine and Biology* vol. 14 pp. 113–123, 1997.

J. K. Percus, *Mathematics of Genome Analysis*, Cambridge University Press, Cambridge, 2002.

L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.

W. Rudin, *Principles of Mathematical Analysis*, John Wiley and Sons, McGraw-Hill, 2nd ed., 1964.

W. Stadje, "The busy period of the queueing system M/G/∞," *Journal of Applied Probability* vol. 22 pp. 697–704, 1985.

L. Takacs, *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.

A. Wald, L. Corey, R. Cone, A. Hobson, G. Davis, and J. Zeh, "Frequent genital herpes virus 2 shedding in immunocompetent women: effect of acyclovir treatment," *Journal of Clinical Investigation* vol. 99 pp. 1092–97, 1997.

K. Yosida, *Lectures on Differential and Integral Equations*, Interscience Publishers, Inc., New York, 1960.