A Three-State Biological Point Process Model and Its Parameter Estimation

G. Tong Zhou, *Member, IEEE*, William R. Schafer, and Ronald W. Schafer, *Fellow, IEEE*

***Abstract*—The Poisson random process is widely used to describe experiments involving discrete arrival data. However, for creating models of egg-laying behavior in recent neural biology studies on the nematode *C. elegans*, we have found that homogeneous Poisson processes are inadequate to capture the measured temporal patterns. We present here a novel three-state model that effectively represents the measured temporal patterns and that correlates well with the cellular and molecular mechanisms that are known to be responsible for the measured behavior. Although the model involves a combination of two Poisson processes, it is surprisingly tractable. We derive closed-form expressions for the probabilistic and statistical properties of the model and present a maximum likelihood method to estimate its parameters. Both simulated and experimental results are illustrated. The experiments with measured data show that the egg-laying patterns fit the three-state model very well. The model also may be applicable in quantifying the link between other neural processes and behaviors or in other situations where discrete events occur in clusters.**

***Index Terms*—Biological system modeling, parameter estimation, point process.**

I. Introduction

MANY biological events can be modeled as point processes [3], [4], [7], [9]. A random point process is a mathematical model for a physical phenomenon characterized by highly localized events distributed randomly in a continuum [7]. Recently, we have studied the egg-laying behavior of a nematode *Caenorhabditis elegans* and found that traditional point process models are inappropriate to describe the observed temporal pattern of behavior.

The nervous system is responsible for generating temporal patterns of muscle contraction, which are the basis for behavior (Fig. 1). Because the nervous system of *C. elegans* is extremely simple and well characterized, it is feasible to understand how patterns of behavior arise from the actions of specific molecular pathways in specific cells. The present study has two principal objectives. The first is to devise a mathematical model that accurately describes the temporal pattern of egg-laying exhibited by real animals. It is hoped that a quantitative model for the behavior will allow us to gain insight into the underlying biological basis for the egg-laying pattern. In addition, application of such a model will allow us to analyze quantitatively the egg-laying patterns of animals with specific nervous system defects and thus to ascertain the roles of specific neurons and genes in egg-laying behavior. In order to do this, it is necessary to extract model parameters from real data and to compare parameters extracted from different data sets (i.e., normal versus mutant animals). Thus, our second objective will be to derive algorithms for estimating model parameters from experimental data and to evaluate their accuracy and usefulness in analyzing real egg-laying data.

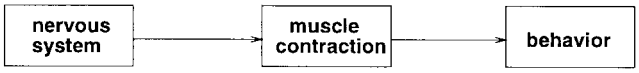

Fig. 1. Nervous system establishes muscle contraction which in turn induces behavior. In *C. elegans*, 302 neurons are responsible for behaviors such as movement, feeding, egg-laying, defecation, and mating. The egg-laying behavior involves the contraction of eight vulval muscles, which open the vulva, and eight uterine muscles, which contract the uterus to expel eggs.

The organization of the paper is as follows: In Section II, we build a biologically interpretable model to account for the temporal pattern of egg-laying behavior. In Section III, we derive a maximum likelihood algorithm to extract the model parameters. We show results from various biological experiments in Section IV, and we draw conclusions in Section V.

II. A Three-State Point Process Model

A. Model Development

Fig. 2(a) shows a partial record of egg laying in wild-type[1] *C. elegans*. The horizontal axis corresponds to time in seconds, and the vertical axis indicates whether an egg was laid at a specific instant. We observed that egg-laying events were often clustered, with events separated by an average of about 15 s. Between these clusters, long inactive periods averaging 20 min in duration were observed during which egg-laying did not occur. Unlike many other behavioral processes (such as feeding and defecation in nematodes) [1], neither the onset of egg-laying clusters nor individual egg-laying events within those clusters appeared to be periodic. This suggested that egg

Manuscript received June 11, 1997; revised April 3, 1998. Some results from this paper were presented at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 3–5, 1997. This work was supported in part by National Science Foundation Grant BES-9615565, by an award from the Arnold and Mabel Beckman Foundation, and by a grant from the National Alliance for Research on Schizophrenia and Depression. The associate editor coordinating the review of this paper and approving it for publication was Dr. Akram Aldroubi.

G. T. Zhou and R. W. Schafer are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA.

W. R. Schafer is with the Department of Biology, University of California–San Diego, La Jolla, CA 92093-0116 USA.

Publisher Item Identifier S 1053-587X(98)07071-8.

[1] In genetics, comparisons are often made between different genetically homogeneous populations of animals (or plants). Populations that are composed of "normal" animals (i.e, animals that are like the ones that are found in nature) are called "wild-type," and populations that differ from normal in a particular gene are called "mutant."

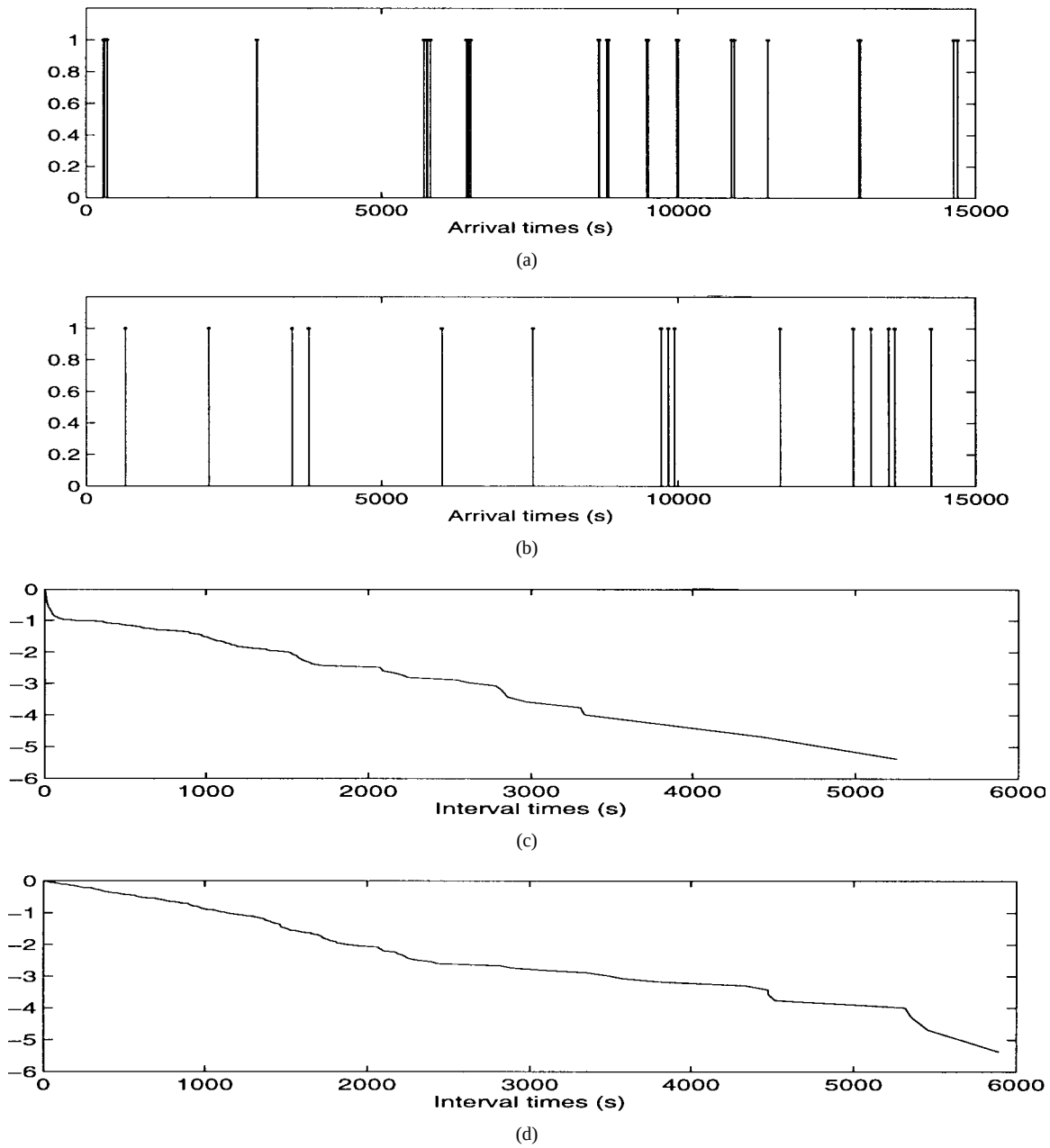

Fig. 2. (a) Temporal pattern of egg-laying behavior in wild type *C. elegans*. Vertical lines indicate the egg arrivals. (b) Simulated arrivals using a homogeneous Poisson model with $\lambda = 8.1 \times 10^{-4}\,\mathrm{s}^{-1}$. (c) Estimated log tail probability of the interarrival times from the actual egg-laying data (estimated slope for large intervals is -8.98×10^{-4}). (d) Estimated log tail probability of the interarrival times from the simulated data (estimated slope for large intervals is -8.63×10^{-4}).

laying might be more accurately modeled as a random point process.

The Poisson process is a popular choice for modeling arrival processes. The interarrival times (denoted by random variable X) of a homogeneous Poisson process follow an exponential distribution with parameter λ (see e.g., p. 57 of [7])

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0 \tag{1}$$

and it is straightforward to show that its mean is $E[X] = 1/\lambda$ and its tail probability is $\Pr(X \geq x) = e^{-\lambda x}$. Since the log tail probability $\ln \Pr(X \geq x) = -\lambda x$ is linear in x with slope $-\lambda$, linearity of the log tail probability curve usually is a good indication of conformity of the data to the homogeneous Poisson model. In Fig. 2(b), we show some simulated arrivals whose intervals were generated according to (1) with $\lambda = 8.1 \times 10^{-4}\,\mathrm{s}^{-1}$. The arrivals are not as clustered as those in Fig. 2(a). Given a finite number of intervals, we can estimate the log tail probability using the rank statistic [8]. In Fig. 2(c) and (d), we show, respectively, the estimated log tail probability of the actual egg-laying data and those of the simulated arrivals. The two exhibit similarity only for long intervals (e.g., $x > 100$ s). In Fig. 2(c), since the slope for the long intervals does not extend to the short intervals, we realize that a simple Poisson model cannot capture the separate mechanisms that account for the short and long intervals in the data, and a more complex model is required.

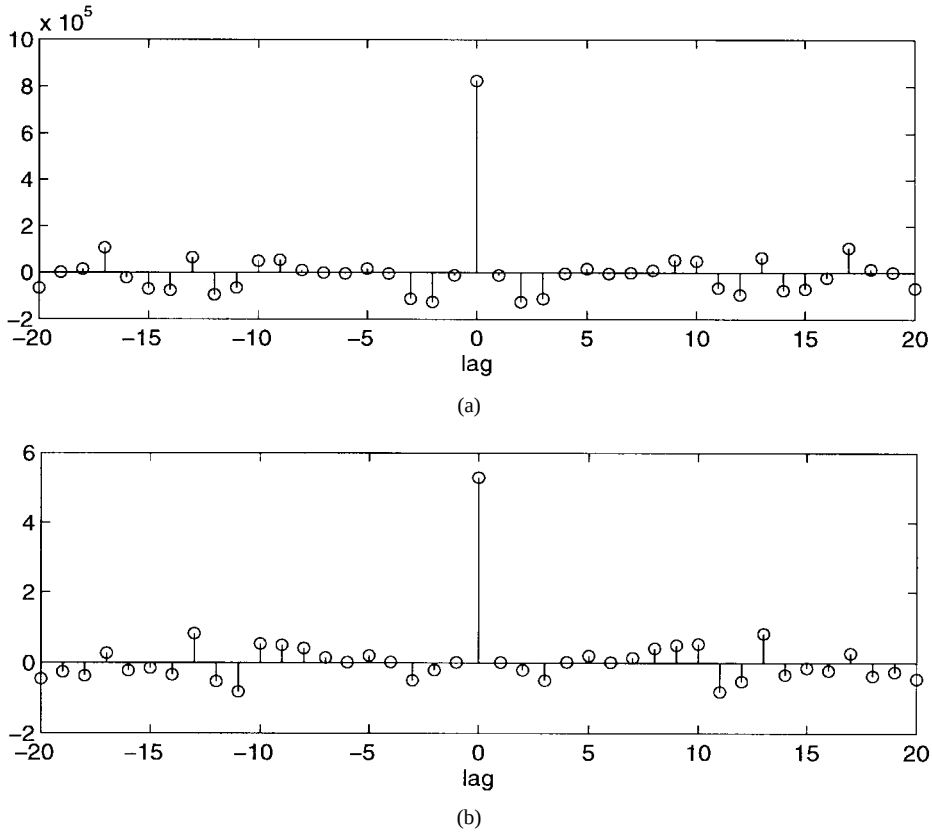

Fig. 3. (a) Covariance between the egg-laying intervals. (b) Covariance between the log intervals.

One possibility, suggested by the differing time scales of the long and short intervals, is that egg-laying behavior could be described by a more complex point process, in which the intracluster and intercluster arrival times are governed by different rate parameters. As noted above, the long intervals (i.e., >100 s) show a linear log tail distribution, indicating that they are exponentially distributed. Similarly, when the short intervals (i.e., <100 s) are analyzed separately, their log tail distribution is also linear. This indicated that taken separately, both the long and short intervals each model as Poisson point processes with different rate constants and that egg-laying behavior as a whole could be modeled by coupling these Poisson processes. It is therefore reasonable to attempt coupling two Poisson processes with different time scales to describe the probabilistic structure of the intervals. We show in Fig. 2(c) and (d), however, only the marginal (log tail) probabilities, and it is important to examine the estimated correlation structure of the intervals as well. In Fig. 3(a) and (b), we show, respectively, the estimated covariance function of the intervals and the log intervals. Since these covariance functions peak only at the zeroth lag, it is safe to assume that the intervals as independent, identically distributed (i.i.d.) and concern ourselves with the derivation of one-dimensional (1-D) probability density functions (pdf's) only.

One of our objectives is to parameterize the data in order to find a parsimonious representation of the intervals. A reasonable "guess" is to model the pdf as a weighted sum of two exponential pdf's

$$f_X(x) = a\mu_1 e^{-\mu_1 x} + (1-a)\mu_2 e^{-\mu_2 x}, \qquad x \geq 0 \qquad (2)$$

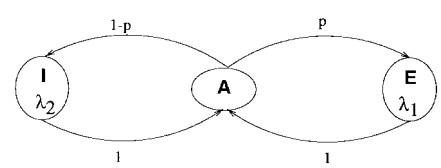

Fig. 4. Three state model. I = inactive state. A = active state. E = egg-laying state. The $A \rightarrow E$ transition occurs with probability p, and its time constant is λ_1. The $A \rightarrow I$ transition occurs with probability $q = 1-p$, and its time constant is λ_2.

where $0 < a < 1$ is the weight factor, and μ_1 and μ_2 are the time constants for the short and long intervals. Although this model by itself does not tell us the biological mechanism that generates the egg-laying pattern, it does provide important clues. For example, the formal states in the model (to be presented next) may correspond to functional states in the nematode's nervous system.

B. Model Description

A three-state diagram representing our model is shown in Fig. 4, and probabilities of state transitions are indicated on the branches of the diagram. Before an egg is released, the material (referred to as the egglet here) travels through three states:

1) inactive state (I);
2) active state (A);
3) egg-laying state (E).

At the A state, the egglet can either enter the E state with probability p and materialize as an egg after some delay

represented by an exponential random variable (r.v.) with parameter λ_1 or enter the I state with probability $q = 1 - p$ and be dormant for a period of time described by a second exponential r.v. with parameter λ_2. We emphasize that the A state acts as a "switch," and there is no time delay associated with that state. Once an egg is laid following the $A \to E$ transition, another egglet enters the A state and faces a chance mechanism of going through the $A \to E$ transition and eventually materializing as an egg or returning to the I state, as described above. Therefore, for any egg that is released, the muscles have existed in the following sequence of states:

$$\underbrace{A \to I \to \cdots \to A \to I}_{k \text{ visits to the } I \text{ state}} \to A \to E \tag{3}$$

where k is an integer that takes on values $0, 1, \cdots, \infty$. The probability of a path resulting in an egg laying event that involves k visits to the I state is pq^k, and the mean duration of that path is $k/\lambda_2 + 1/\lambda_1$.

C. Probabilistic Analysis of the Model

Our goal here is to derive the pdf of the intervals between the egg-laying events. The characteristic function (cf) defined as

$$\phi_X(u) = E[e^{juX}], \qquad j = \sqrt{-1} \tag{4}$$

will turn out to be very useful. For the exponential pdf in (1), it can be shown that (see, e.g., [6, p. 116])

$$\phi_X(u) = \frac{\lambda}{\lambda - ju}. \tag{5}$$

Suppose that the muscles have visited the I state k times before their entry into the E state. Then, the duration of the entire path depicted in (3) is an r.v. $X_k = X_1 + \Sigma_{i=1}^{k} X_{2,i}$, where X_1 and $\{X_{2,i}\}_{i=1}^{k}$ are mutually independent r.v.'s with X_1 having an exponential pdf $f_1(x)$ with parameter λ_1 and each $X_{2,i}$ having the same exponential pdf $f_2(x)$ with parameter λ_2. The pdf of X_k is then the convolution of individual pdf's (see e.g., [6, p. 195])

$$f_1(x) * \underbrace{f_2(x) * f_2(x) \cdots * f_2(x)}_{k \text{ times}}.$$

Since the probability of this particular path (with k visits to the I state) is pq^k, the pdf of the intervals between the eggs laid is

$$f_X(x) = \sum_{k=0}^{\infty} pq^k f_1(x) * \underbrace{f_2(x) * f_2(x) \cdots * f_2(x)}_{k \text{ times}}. \tag{6}$$

The above expression can be much simplified with the help of the cf. First, we realize that the cf for (6) is the product of the individual cf's

$$\begin{aligned}\phi_X(u) &= \sum_{k=0}^{\infty} pq^k \phi_1(u)[\phi_2(u)]^k \\ &= \sum_{k=0}^{\infty} pq^k \frac{\lambda_1}{\lambda_1 - ju}\left[\frac{\lambda_2}{\lambda_2 - ju}\right]^k.\end{aligned} \tag{7}$$

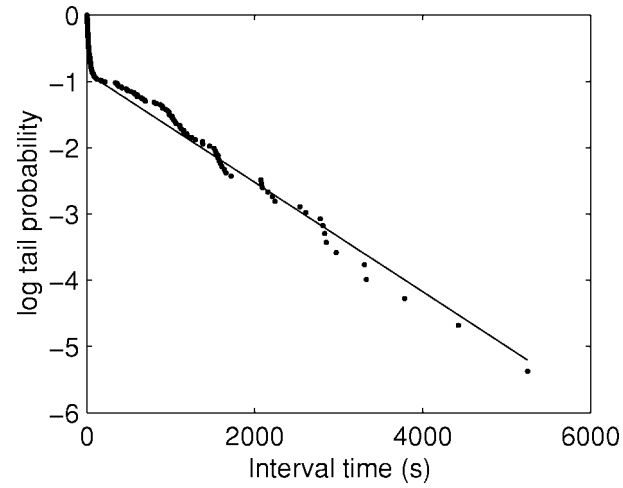

Fig. 5. Estimated log tail probability (dotted line) from the wild type data and the theoretical log tail probability (solid line) that corresponds to $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}, \lambda_2 = 0.0014\,\mathrm{s}^{-1}$.

Since $|\lambda_2/(\lambda_2 - ju)| < 1$ and $0 < q < 1$, we express the above geometric sum as

$$\phi_X(u) = \frac{p\lambda_1}{\lambda_1 - ju} \sum_{k=0}^{\infty} \left[\frac{q\lambda_2}{\lambda_2 - ju}\right]^k \tag{8}$$

$$= \frac{p\lambda_1}{\lambda_1 - ju} \frac{1}{1 - \dfrac{q\lambda_2}{\lambda_2 - ju}} \tag{9}$$

$$= \frac{\lambda_2 - ju}{\lambda_1 - ju} \frac{p\lambda_1}{p\lambda_2 - ju}. \tag{10}$$

Next, we assume $\lambda_1 \neq p\lambda_2$ and perform partial fraction expansion on (10) to obtain

$$\phi_X(u) = \frac{p(\lambda_1 - \lambda_2)}{\lambda_1 - p\lambda_2} \frac{\lambda_1}{\lambda_1 - ju} + \frac{\lambda_1(1-p)}{\lambda_1 - p\lambda_2} \frac{p\lambda_2}{p\lambda_2 - ju}. \tag{11}$$

In this case, the pdf of X is simply a weighted sum of two exponential pdf's[2] with respective parameters λ_1 and $p\lambda_2+$, i.e.,

$$f_X(x) = k_1\lambda_1 e^{-\lambda_1 x} + k_2(p\lambda_2)e^{-(p\lambda_2)x}, \qquad x \geq 0 \tag{12}$$

where

$$k_1 = \frac{p(\lambda_1 - \lambda_2)}{\lambda_1 - p\lambda_2} \tag{13}$$

$$k_2 = \frac{\lambda_1(1-p)}{\lambda_1 - p\lambda_2}. \tag{14}$$

It follows easily that the tail probability of X is

$$\Pr(X \geq x) = k_1 e^{-\lambda_1 x} + k_2 e^{-(p\lambda_2)x}. \tag{15}$$

In Fig. 5, we show the log tail probability estimated from the actual egg-laying data (dotted line) and the logarithm of (15) (solid line, using $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}, \lambda_2 = 0.0014\,\mathrm{s}^{-1}$; these parameters were estimated from the data using the methods described in Section III). Close agreement between the two is observed.

[2]When $\lambda_1 \neq p\lambda_2$, we can interpret (12) as an "exponential mixture." Just as "Gaussian mixtures" are sometimes postulated to handle outliers, the exponential mixture serves to model both short and long interval times here.

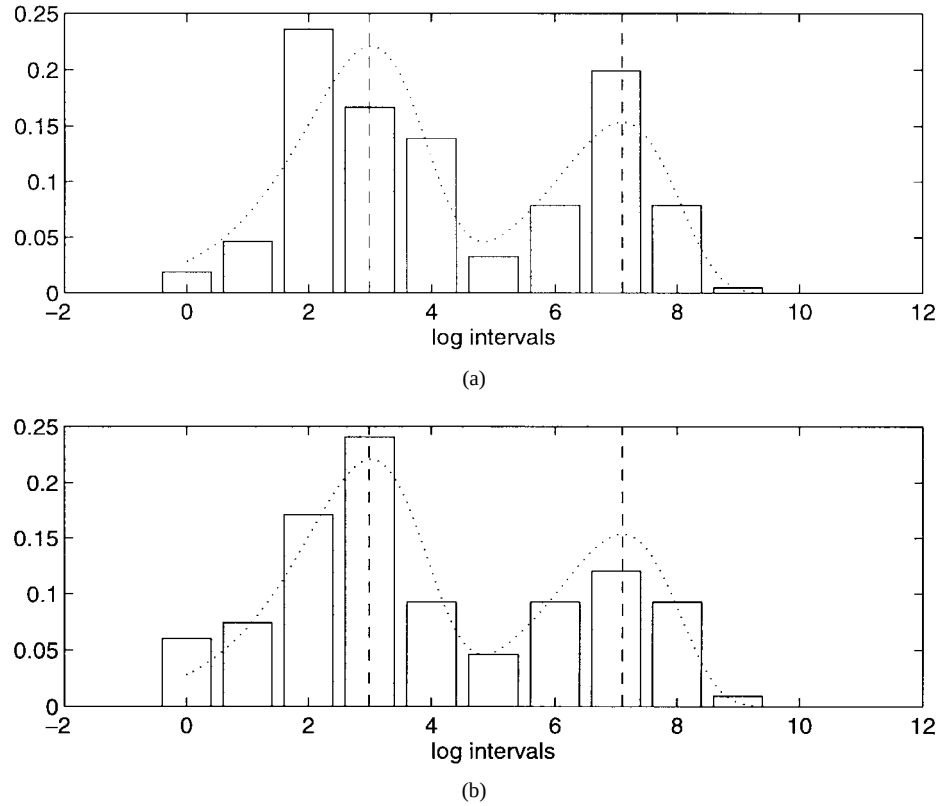

Fig. 6. (a) Histogram (bar graph) of the log intervals from the wild type data. (b) Histogram (bar graph) of the log intervals from the simulated data generated according to the three-state model. Dotted lines correspond to $f_Y(y)$ with $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}, \lambda_2 = 0.0014\,\mathrm{s}^{-1}$. Dashed lines show the values $-\ln(\lambda_1)$ and $-\ln(p\lambda_2)$.

Since the intervals between the eggs laid are clustered at short intervals and sparse at long intervals, their logarithm $Y = \ln(X)$ will have a more uniform spread. In Section III, we will show that the distribution of the log intervals is a useful basis for estimating the parameters of the model.

Recall that the pdf of Y is related to that of X via (see e.g., [6, p. 93])

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|. \tag{16}$$

Substituting $x = e^y$ and $dx/dy = e^y$, we obtain the pdf of the log intervals

$$f_Y(y) = [k_1\lambda_1 e^{-\lambda_1 e^y} + k_2(p\lambda_2)e^{-(p\lambda_2)e^y}]e^y. \tag{17}$$

Let us focus on the first term in (17)

$$g_1(y) = (k_1\lambda_1 e^{-\lambda_1 e^y})e^y. \tag{18}$$

It can be shown that

$$g_1(y)|_{y=-\ln\lambda_1} = \frac{k_1}{e} \tag{19}$$

$$\left.\frac{\partial g_1(y)}{\partial y}\right|_{y=-\ln\lambda_1} = 0 \tag{20}$$

$$\left.\frac{\partial^2 g_1(y)}{\partial y^2}\right|_{y=-\ln\lambda_1} < 0. \tag{21}$$

Therefore, $g_1(y)$ peaks at $y = -\ln\lambda_1$ with peak magnitude k_1/e. Similarly, the second term in (17)

$$g_2(y) = (k_2 p\lambda_2 e^{-p\lambda_2 e^y})e^y \tag{22}$$

peaks at $y = -\ln(p\lambda_2)$ with peak magnitude k_2/e. When λ_1 and $p\lambda_2$ are sufficiently different, the overall pdf $f_Y(y) = g_1(y) + g_2(y)$ exhibits a bimodal pattern with peak locations around $-\ln\lambda_1$ and $-\ln(p\lambda_2)$ and peak heights approximately k_1/e and k_2/e. This property can be explored for parameter estimation purposes, as will be seen in Section III-A.

In Fig. 6(a), we show the histogram of the log intervals for recorded egg-laying data and in Fig. 6(b), that of the simulated data generated according to the three-state model with parameters $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}$, and $\lambda_2 = 0.0014\,\mathrm{s}^{-1}$. In both cases, the total number of intervals is 216. We also show with dotted lines, the pdf of the log intervals (17) that corresponds to the above parameters. Both histograms show a bimodal pattern with peaks located around $-\ln(\lambda_1)$ and $-\ln(p\lambda_2)$ (dashed lines).

A special case worth mentioning is when $\lambda_1 = p\lambda_2$. In that case, we take the limit of (12) with $p\lambda_2 \to \lambda_1$ and obtain

$$f_X(x) = (\lambda_1^2(1-p)x + p\lambda_1)e^{-\lambda_1 x}, \qquad x \geq 0. \tag{23}$$

The pdf of the log intervals is $f_Y(y) = e^y f_X(x)|_{x=e^y}$ and can be shown to exhibit a single peak.

D. Statistical Analysis of the Model

Since the closed-form expression for the pdf of the intervals is available, we can calculate many probabilities of interest. For example, if a long (L) interval is defined as being at least a seconds long, then the probability that any given interval

is long is

$$p_L = k_1 e^{-\lambda_1 a} + k_2 e^{-p\lambda_2 a} \tag{24}$$

and the probability that any given interval is short (S) is

$$p_S = 1 - p_L. \tag{25}$$

Moreover, if we define clustering as M or more consecutive short intervals followed by a long interval, then the probability of clustering is $(p_S)^M$.

We are also interested in finding out the average duration of the short and long intervals. First, we deduce that their respective conditional pdf's are

$$f_X(x|X \le a) = \frac{f_X(x)}{p_S}, \qquad 0 \le x \le a \tag{26}$$

$$f_X(x|X \ge a) = \frac{f_X(x)}{p_L}, \qquad x \ge a. \tag{27}$$

Substituting (12) into (26), we find the average duration of the short intervals

$$\begin{aligned}\overline{X}_S &= E[X|X \le a] \\ &= \int_0^a x f_X(x|X \le a)\,dx \\ &= \frac{1}{p_S}\left[\frac{k_1}{\lambda_1} - k_1\left(\frac{1}{\lambda_1} + a\right)e^{-\lambda_1 a} + \frac{k_2}{p\lambda_2}\right. \\ &\quad \left. - k_2\left(\frac{1}{p\lambda_2} + a\right)e^{-p\lambda_2 a}\right].\end{aligned} \tag{28}$$

Similarly, we find the average duration of the long intervals

$$\begin{aligned}\overline{X}_L &= E[X|X \ge a] \\ &= \int_a^\infty x f_X(x|X \ge a)\,dx \\ &= \frac{1}{p_L}\left[k_1\left(\frac{1}{\lambda_1} + a\right)e^{-\lambda_1 a} + k_2\left(\frac{1}{p\lambda_2} + a\right)e^{-p\lambda_2 a}\right].\end{aligned} \tag{29}$$

From their complete probabilistic description (i.e., the pdf), we can obtain arbitrary order moment or cumulant expressions of the intervals as well. These statistics can yield methods for model parameter estimation and will also be useful for performance analysis of the sample estimates.

It turns out that the cumulants of the intervals X have simpler expressions than the moments. Refer to [2, Sec. 2.3] for definitions and properties of cumulants. The so-called cumulant generating function is the logarithm of the cf, i.e., $\psi_X(u) = \ln \phi_X(u)$. For our three-state model, we infer from (11) that the cumulant generating function is

$$\begin{aligned}\psi_X(u) = &\ln(\lambda_2 - ju) - \ln(\lambda_1 - ju) \\ &+ \ln(p\lambda_1) - \ln(p\lambda_2 - ju).\end{aligned} \tag{30}$$

By induction, we can prove that the kth-order derivative of (30) is

$$\begin{aligned}\frac{\partial^k \psi_X(u)}{\partial u^k} = &(-1)^{k-1}(k-1)!\left[\frac{(-j)^k}{(\lambda_2 - ju)^k} - \frac{(-j)^k}{(\lambda_1 - ju)^k}\right. \\ &\left. - \frac{(-j)^k}{(p\lambda_2 - ju)^k}\right].\end{aligned} \tag{31}$$

Next, we find the kth-order cumulant of X as

$$\begin{aligned}c_{kx} &= \frac{1}{j^k}\frac{\partial^k \psi_X(u)}{\partial u^k}\bigg|_{u=0} \\ &= -(k-1)!\left[\frac{1}{\lambda_2^k} - \frac{1}{\lambda_1^k} - \frac{1}{(p\lambda_2)^k}\right] \\ &= \left(\frac{1}{p^k} - 1\right)\frac{(k-1)!}{\lambda_2^k} + \frac{(k-1)!}{\lambda_1^k}.\end{aligned} \tag{32}$$

Specifically, we find the first- through third-order cumulants of X as

$$c_{1x} = E[X] = (1/p - 1)/\lambda_2 + 1/\lambda_1 \tag{33}$$

$$c_{2x} = E[(X - c_{1x})^2] = (1/p^2 - 1)/\lambda_2^2 + 1/\lambda_1^2 \tag{34}$$

$$c_{3x} = E[(X - c_{1x})^3] = (2/p^3 - 2)/\lambda_2^3 + 2/\lambda_1^3. \tag{35}$$

They are referred to as the mean, variance, and skewness,[3] respectively, [5, p. 16]. These expressions can be used together with their sample statistics to generate estimates for the parameters p, λ_1 and λ_2.

III. Parameter Estimation

Our goal here is to estimate the model parameters[4] $\boldsymbol{\theta} = [p, \lambda_1, \lambda_2]'$ from N observations of the intervals $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]'$. Since closed-form expressions for the pdf $f_X(x)$ of the intervals and $f_Y(y)$ of the log intervals are available, various estimation procedures can be followed. We shall first use a simple, noniterative method to obtain good initial estimates for the parameters and then refine them using a maximum likelihood approach.

A. Estimation Based on $f_Y(y)$: Peak Picking

Since the pdf (17) of the log intervals Y peaks around $-\ln \lambda_1$ and $-\ln(p\lambda_2)$ with approximate peak strengths k_1/e and k_2/e, we first form the histogram of the log intervals at uniformly spaced bins y_i. The histogram values are then normalized by the sample size so that the scaled histogram serves as an estimate of $f_Y(y)$. Next, we estimate $\hat{\lambda}_1$ and $\widehat{(p\lambda_2)}$ from the peak locations of the histogram and $\hat{k}_1/e$ and $\hat{k}_2/e$ from the corresponding peak heights. Afterwards, we find [cf., (13)]

$$\widehat{(p\lambda_1)} = \hat{k}_1[\hat{\lambda}_1 - \widehat{(p\lambda_2)}] + \widehat{(p\lambda_2)} \tag{36}$$

and

$$\hat{p} = \widehat{(p\lambda_1)}/\hat{\lambda}_1, \quad \hat{\lambda}_2 = \widehat{(p\lambda_2)}/\hat{p}. \tag{37}$$

Since histograms are calculated at a set of discrete bins y_i, we may wish to interpolate around the bins to obtain a smooth looking $\hat{f}_Y(y)$ as an estimate of $f_Y(y)$, prior to peak-picking. This ensures good resolution of the y-axis. Provided that the bimodal pattern is evident, estimates obtained from the peaks of $\hat{f}_Y(y)$ are usually reasonably accurate.

[3] Skewness is sometimes defined as c_{3x} divided by the cube of the standard deviation to make it dimensionless. However, the normalization factor is immaterial for modeling purposes.

[4] We may also define the parameter vector as $\boldsymbol{\theta} = [k_1, \lambda_1, \bar{\lambda}_2 = p\lambda_2]'$, estimate it, and then find out p and λ_2.

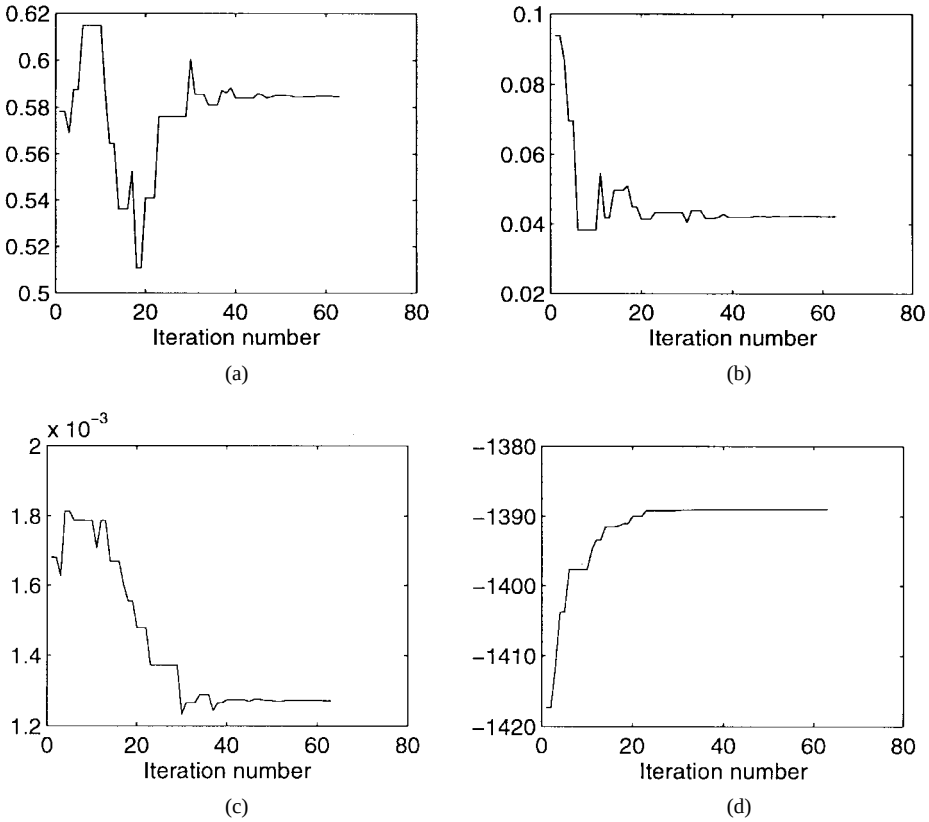

Fig. 7. (a)–(c). Intermediate ML estimates of p, λ_1, and λ_2, respectively. (d) The log likelihood function increases and attains its maximum at iteration 63.

TABLE I
MONTE CARLO SIMULATIONS OF THE PEAK PICKING ALGORITHM TO ESTIMATE THE PARAMETERS OF THE THREE-STATE MODEL. RESULTS ARE OBTAINED USING 100 INDEPENDENT REALIZATIONS OF 216 INTERVALS EACH

Parameters	True	Estimated (mean $\pm$ std)
p	0.5891	$0.6006 \pm 0.6186 \times 10^{-1}$
λ_1	0.5010×10^{-1}	$0.5531 \times 10^{-1} \pm 0.1491 \times 10^{-1}$
λ_2	0.1400×10^{-2}	$0.1617 \times 10^{-2} \pm 0.4794 \times 10^{-3}$

TABLE II
MONTE CARLO SIMULATIONS OF THE ML ALGORITHM TO ESTIMATE THE PARAMETERS OF THE THREE-STATE MODEL. RESULTS ARE OBTAINED USING 100 INDEPENDENT REALIZATIONS OF 216 INTERVALS EACH

Parameters	True	Estimated (mean $\pm$ std)
p	0.5891	$0.5967 \pm 0.3559 \times 10^{-1}$
λ_1	0.5010×10^{-1}	$0.4969 \times 10^{-1} \pm 0.5921 \times 10^{-2}$
λ_2	0.1400×10^{-2}	$0.1396 \times 10^{-2} \pm 0.2137 \times 10^{-3}$

Example 1: We simulated[5] 216 samples of interval data according to our three-state model with parameters $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}, \lambda_2 = 0.0014\,\mathrm{s}^{-1}$. From the peak locations and peak heights of the histogram of the log intervals, we can obtain estimates for p, λ_1, and λ_2 as described above. The mean and standard deviation from 100 independent realizations are shown in Table I. □

To improve the accuracy of the peak-picking approach, we may consider a nonlinear least squares matching criterion and seek a model that best fits the given histogram data. The method to be described next, however, is preferred because it yields maximum likelihood estimates.

[5]There are two ways to generate X that has the pdf (12). One way is to use one Bernoulli random number generator and two exponential random number generators and to "synthesize" their outputs according to the three-state model. An alternative approach is to first generate a uniform (in [0, 1]) r.v. U and then obtain X through transformation $X = F_X^{-1}(U)$, where $F_X(x) = 1 - k_1 e^{-\lambda_1 x} - k_2 e^{-p\lambda_2 x}$ is the cumulative distribution function of X, i.e., the indefinite integral of (12). The X so generated will be i.i.d. and have the pdf given by (12).

B. Estimation Based on $f_X(x)$: Maximum Likelihood Method

Recall that the intervals X are i.i.d. random variables with pdf specified in (12). If a record of N intervals $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]'$ is observed, the likelihood function is then

$$f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} [k_1 \lambda_1 e^{-\lambda_1 x_i} + k_2 (p\lambda_2) e^{-(p\lambda_2) x_i}] \tag{38}$$

and the log likelihood function is

$$\ln f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln [k_1 \lambda_1 e^{-\lambda_1 x_i} + k_2 (p\lambda_2) e^{-(p\lambda_2) x_i}]. \tag{39}$$

The maximum likelihood (ML) estimate $\hat{\boldsymbol{\theta}}$ is defined as the $\boldsymbol{\theta}$ that maximizes the likelihood function $f_X(\boldsymbol{x}|\boldsymbol{\theta})$ or, equivalently, the log likelihood function $\ln f(\boldsymbol{x}|\boldsymbol{\theta})$.

Since the (log) likelihood function is nonlinear in the parameters $\boldsymbol{\theta}$, convergence to local maxima is possible, and good initial guesses are critical. We can use the estimates obtained from the peak picking method as initial guesses. Although the ML algorithm is nonlinear, our experience has

Fig. 8. *Caenorhabditis elegans* laying eggs. Shown is a wild-type *C. elegans* hermaphrodite (strain N2 Bristol) and recently laid eggs under standard experimental conditions (NGM agar plates seeded with *E. coli* OP50 bacteria as a food source).

shown that its convergence is fairly fast. The ML method is adopted in this paper to analyze the real data records.

Example 2: We simulated 216 samples of interval data according to our three-state model with the same parameters used in Example 1. Using the peak picking method, we obtained initial estimates $\hat{p} = 0.5507, \hat{\lambda}_1 = 0.1044\,\mathrm{s}^{-1}$, and $\hat{\lambda}_2 = 0.0016\,\mathrm{s}^{-1}$ for one particular realization. We then used them as initial guesses to start the ML algorithm. After 63 iterations, the ML algorithm converged and yielded final estimates $\hat{p} = 0.5847, \hat{\lambda}_1 = 0.0422\,\mathrm{s}^{-1}$, and $\hat{\lambda}_2 = 0.0013\,\mathrm{s}^{-1}$ for the same realization. Fig. 7 illustrates how the ML algorithm converges to its final solution. To verify the accuracy of our ML procedure, we generated 100 independent realizations of the intervals and obtained the mean $\pm$ standard deviation of the estimates as summarized in Table II. Comparing with Table I, we see that both the mean and std of the parameter estimates have noticeably improved over the peak picking method. $\square$

IV. Results on Real Data

The model we have formulated for egg-laying in *C. elegans* has a number of important implications concerning the biological basis for this behavior and has important applications for future investigations of this process. Perhaps most importantly, the three state model implies the existence of discrete behavioral states for egg laying, including an inactive state, during which the animals are refractory to egg laying, and an active state, during which clusters of eggs are laid. Additionally, the finding that the intercluster and intracluster intervals are determined by coupled, independent Poisson processes suggests that induction of the "active state" and egg laying within the active state are caused by distinct biological processes whose occurrence is stochastic rather than periodic. Finally, application of our model to the analysis of real egg-laying data has allowed us to characterize the roles of specific neurons and neurotransmitters in the nematode nervous system in generating patterns of egg-laying behavior. By extracting model parameters from egg-laying data obtained from wild-type animals, mutant animals, and animals containing precise neuronal ablations, we can identify which feature of the egg-laying pattern is controlled by a particular neuron or neurotransmitter. Although the specific conclusions drawn from such studies are described in detail elsewhere [10], [11], examples of how our parameter estimation methods can be applied to studies of this type are described below.

To gather data on the timing of egg-laying events, we followed individual nematodes using an automated tracking system and recorded their behavior on videotape. By replaying the recordings, we were able to determine the point in time that each egg was laid; this information was then analyzed in the context of our model to extract behavioral parameters.

To illustrate the use of our model in analyzing experimental data, we present two examples below. These and the results of other experiments are reported in more detail in [10] and [11] along with an interpretation of the mathematical model as it relates to the physiology of the nematode.

A. On Wild-Type Data

We first obtained from a common laboratory wild-type strain of *C. elegans*, N2 (Bristol), under constant conditions that were favorable to egg-laying (isotonic nematode growth medium in the presence of abundant food). A picture of the worm is shown in Fig. 8. From 216 data points, the ML algorithm yielded parameter estimates $p = 0.5891, \lambda_1 = 0.0501\,\mathrm{s}^{-1}, \lambda_2 = 0.0014\,\mathrm{s}^{-1}$, and $p\lambda_2 = 8.0352 \times 10^{-4}\,\mathrm{s}^{-1}$. If we regard $x \leq a = 50$ s [this is where the curve turned in Fig. 2(c)] as a short interval, then out of the 216 intervals,

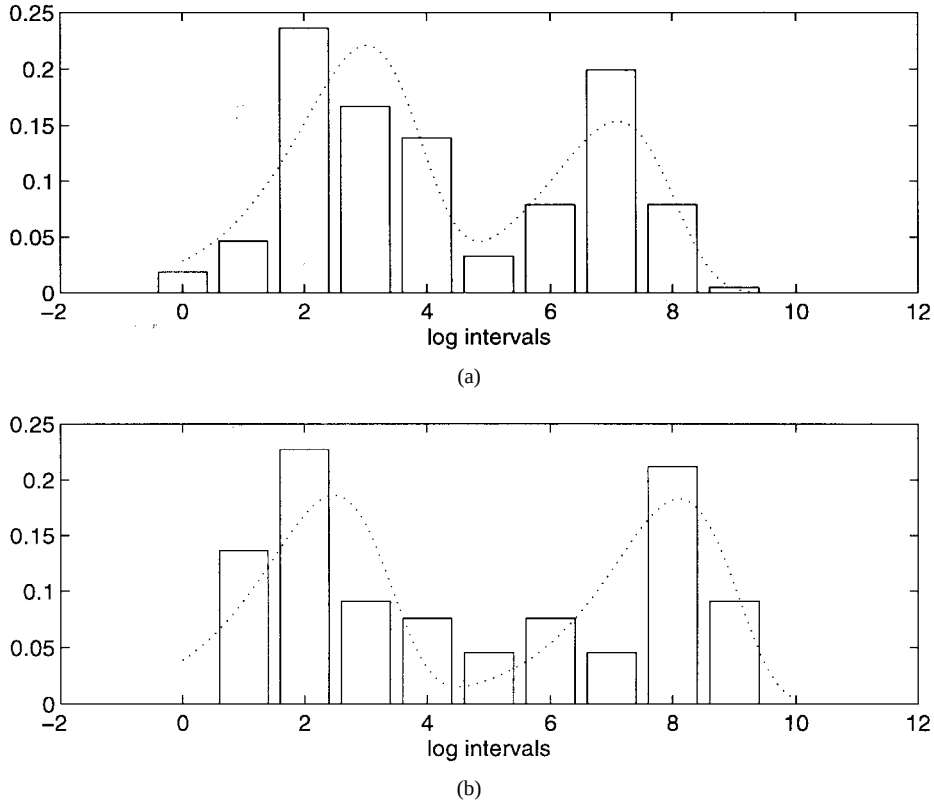

Fig. 9. Histograms of the log intervals (bar graphs) and the $f_Y(y)$ curves generated using the ML parameter estimates. (a) Wild type. (b) HSN ablated. From the shifts in the peak positions, we infer that λ_1 increased slightly while $p\lambda_2$ decreased greatly in the HSN ablated animals.

119 were classified as short yielding $\hat{p}_S = 0.5509$, and their average duration was $\hat{\overline{X}}_S = 15.8319$ s. On the other hand, the average duration of the other 97 long intervals $(\hat{p}_L = 97/216 = 0.4491)$ was $\hat{\overline{X}}_L = 1164$ s. These results are corroborated by our theoretical expressions (24), (25), (28), and (29), where we found $p_L = 0.4485, p_S = 0.5515, \overline{X}_S = 15.7981$ s, and $\overline{X}_L = 1136$ s (or approximately 20 min), based on our ML parameter estimates.

B. Effect of the HSN Neurons

The model described here has also been useful in describing quantitatively the effects of anatomical or genetic lesions on the pattern of egg laying. For example, we were able to derive model parameters from animals in which a specific class of neurons [the hermaphrodite-specific neurons (HSN's)] had been ablated through laser microsurgery [10]. We used the ML method to obtain the parameter estimates. It is difficult to find a theoretical expression for the confidence intervals, so we obtained the standard deviations from numerical simulations (cf., Section III). The parameter estimates with their estimated standard deviations bounds are: $p = 0.5037 \pm 0.0506, \lambda_1 = (0.0838 \pm 0.0268)\,\mathrm{s}^{-1}, \lambda_2 = (0.0006 \pm 0.0002)\,\mathrm{s}^{-1}$. Comparing these estimates with those of nonablated animals, we observe that the λ_1 value is increased, whereas the λ_2 value is decreased. Thus, elimination of the HSN's altered the egg-laying pattern by causing a substantial lengthening of the intercluster times (see Fig. 9 for comparison between the results obtained from the wildtype and HSN ablated worms). Interestingly, the HSN's contain a neurotransmitter molecule called serotonin, which appears to induce egg-laying clusters [10]. Thus, the effects of HSN ablation on λ_2 correlates with the activity of a specific neurotransmitter (serotonin) normally released by the HSN's.

V. Conclusions

We have proposed a new multistate point process model that is based on two coupled point processes. The model is capable of representing a wide range of conditions in neurobiology experiments that are aimed at discovering the relationships between neural and molecular mechanisms and their resulting manifestations in temporal patterns of behavior. Although it is significantly more complex than a single point process, the model is surprisingly amenable to analysis. We have derived closed-form expressions for the probability distribution of intervals between events, their kth-order cumulants, and distributions for the log intervals. Furthermore, we have developed a maximum likelihood method for estimating the parameters of the model from interval data that can be measured by experiment.

The model has succeeded in providing a quantitative framework for a wide range of neuro-biology experiments on egg-laying behavior in nematodes [10], [11], and it also may be applicable to other investigations directed at linking cellular-level phenomena to behavior in simple animals. Furthermore, the model should be adaptable to many other situations in which discrete events occur in clusters.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ananthram Swami for his careful reading of the original manuscript and many helpful suggestions. Thanks also go to L. Waggoner, who obtained the data for Fig. 9

REFERENCES

[1] L. Avery and J. H. Thomas, "Feeding and defecation," in *C. elegans II*, D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess, Eds. Cold Spring Harbor, NY: Cold Spring Harbor Lab., 1997.

[2] D. R. Brillinger, *Time Series: Data Analysis and Theory*. San Francisco, CA: Holden-Day, 1981.

[3] ——, "Maximum likelihood approach to the identification of neuronal firing systems," *Ann. Biomed. Eng.*, vol. 16, no. 1, pp. 3–16, 1988.

[4] D. H. Johnson and A. Swami, "The transmission of signals by auditory-nerve fiber discharge patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 2, pp. 493–501, 1983.

[5] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectral Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[7] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*, 2nd ed. New York: Springer-Verlag, 1991.

[8] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

[9] R. G. Turcott, S. B. Lowen, D. H. Johnson, C. Tsuchitani, and M. C. Teich, "Nonstationary Poisson point process describes the sequence of action potentials over long time scales in lateral-superior-olive auditory neurons," *Biol. Cybern.*, vol. 70, no. 3, pp. 209–217, 1994.

[10] L. E. Waggoner, G. T. Zhou, R. W. Schafer, and W. R. Schafer, "Control of alternative behavioral states by serotonin in *Caenorhabditis elegans*," *Neuron*, vol. 21, no. 1, pp. 203–214, July 1998.

[11] L. E. Waggoner, D. Pode, D. S. Ramirez, and W. R. Schafer, "Function and adaptation of nicotinic acetylcholine receptors in C. elegans egg-laying behavior," in preparation.

G. Tong Zhou (M'95) received the B.Sc. degree in biomedical engineering and instrumentation from the Tianjin University, Tianjin, China, in July 1989. From September 1989 to May 1995, she was with the University of Virginia (UVA), where she received M.Sc. degree in biophysics in May 1992, the M.Sc. degree in electrical engineering in January 1993, and Ph.D. degree in electrical engineering in January 1995.

She has been with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, since September 1995 as an Assistant Professor. Her research interests are in the general areas of statistical signal processing and biomedical engineering. Specific interests include nonlinear system identification, cyclostationary signal processing, higher order statistical analysis, magnetic resonance imaging, and biophysical signal analysis.

Dr. Zhou was awarded the 1995 Allan Talbott Gwathmey Memorial Award for outstanding research in the physical sciences at UVA based on her Ph.D. dissertation. In 1997, she received the National Science Foundation Faculty Early Career Development (CAREER) Award.

William R. Schafer received the A.B. degree in biology from Harvard University, Cambridge, MA, in 1986 and the Ph.D. degree in biochemistry from the University of California, Berkeley (UC Berkeley), in 1991.

From 1991 to 1992, he was a postgraduate researcher in the Division of Genetics at UC Berkeley, and from 1992 to 1995, he was a Life Sciences Research Foundation Postdoctoral Fellow with the Department of Biochemistry and Biophysics at the University of California, San Francisco. His research concerns the cellular and molecular basis of nervous system function and behavior in simple organisms.

Dr. Shafer is a member of the Society for Neuroscience and the Genetics Society of America and has been the recipient of a Beckman Young Investigator Award, an Alfred P. Sloan Fellowship, and a Klingenstein Fellowship in Neuroscience.

Ronald W. Schafer (F'77) received the B.S.E.E. and M.S.E.E. degrees from the University of Nebraska, Lincoln, in 1961 and 1962, respectively, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1968.

From 1968 to 1974 he was a member of the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, where he was engaged in research on speech analysis and synthesis, digital signal processing techniques, and digital waveform coding. Since 1974, he has been on the faculty of the Georgia Institute of Technology (Georgia Tech), Atlanta, where he is now John and Marilu McCarty Professor and Institute Professor of Electrical and Computer Engineering. His current research interests include speech and video processing, nonlinear signal processing systems, applications of signal processing in multimedia communication systems, and applications of signal processing in biology and medicine. He is coauthor of six textbooks, including *Discrete-Time Signal Processing* and the recently published *DSP First: A Multimedia Approach*.

Dr. Schafer is a Fellow of the Acoustical Society of America and he is a member of the National Academy of Engineering. He was awarded the Achievement Award and the Society Award of the IEEE ASSP Society in 1979 and 1983, respectively, the 1983 IEEE Region III Outstanding Engineer Award, and he shared the 1980 Emanuel R. Piore Award with L. R. Rabiner. In 1985, he received the Class of 1934 Distinguished Professor Award at Georgia Tech, and he received the 1992 IEEE Education Medal. He has been active in the affairs of the IEEE Acoustics, Speech, and Signal Processing Society, having served as Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, a member of several committees, Vice-President and President of the Society, and Chairman of the 1981 ICASSP.