Inference in a Partially Observed Queuing Model with Applications in Ecology

Kevin Winner[1] KWINNER@CS.UMASS.EDU
Garrett Bernstein[1] GBERNSTEIN@CS.UMASS.EDU
Daniel Sheldon[1,2] SHELDON@CS.UMASS.EDU
[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA 01002, USA
[2]Department of Computer Science, Mount Holyoke College, South Hadley, MA 01075, USA

Abstract

We consider the problem of inference in a probabilistic model for transient populations where we wish to learn about arrivals, departures, and population size over all time, but the only available data are periodic counts of the population size at specific observation times. The underlying model arises in queueing theory (as an $M_t/G/\infty$ queue) and also in ecological models for short-lived animals such as insects. Our work applies to both systems. Previous work in the ecology literature focused on maximum likelihood estimation and made a simplifying independence assumption that prevents inference over unobserved random variables such as arrivals and departures. The contribution of this paper is to formulate a latent variable model and develop a novel Gibbs sampler based on Markov bases to perform inference using the correct, but intractable, likelihood function. We empirically validate the convergence behavior of our sampler and demonstrate the ability of our model to make much finer-grained inferences than the previous approach.

1. Introduction

We consider the problem of inference in a probabilistic model for transient populations where we wish to learn as much as possible about the complete state of the population over time, including arrivals, departures, and population size, but the only available data are periodic counts of the population size at specific observation times.

Our work applies to a simple probabilistic model that arises in two distinct places. It is most precisely described in queuing theory, where it is known as the $M_t/G/\infty$ queue (Eick et al., 1993) and describes a situation where: (1) customers arrive at a queue over time according to a Poisson process with arbitrary intensity function, (2) they are assigned to a server immediately upon arriving at the queue, and (3) their service time is drawn independently from an arbitrary, shared, service-time distribution. In this terminology, our paper addresses the problem of making inferences about arrivals and departures from the queue when only the total number of customers in service is observable and only at discrete observation times.

We are primarily motivated by a specific application in ecology. Zonneveld (1991) proposed what is in essence an $M_t/G/\infty$ queue for analyzing transient or short-lived animal populations, specifically, insects. Adults enter the population (arrive at the queue) either by advancing from a previous life stage or immigrating from outside the survey area and then remain in the survey area for a specified amount of time (service time) before dying or leaving the area (departing the queue). Counts of abundance (number of customers in service) are made over time at the survey location. From this information, ecologists would like to make inferences about life history events such as migration, birth, and death that correspond to arrivals and departures from the queue.

Information about arrivals and departures in insect populations is important for several reasons. First, the timing of arrivals (e.g., the emergence of adult butterflies from cocoons) is linked to climate. Shifts in timing are important to detect because they may result from climate change and have the potential to disrupt the synchrony of ecosystems. Second, understanding lifespans (departure rates) is key to monitoring population size and trends over time, because lifespans are confounded with abundance when interpreting survey counts. For example, it can be hard to distinguish between a population where many individuals arrive but die quickly from one where few arrive but individuals

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

are long-lived.

The main contributions of this work are the formulation of a latent variable model for this problem and the development of a novel Gibbs sampler for the challenging problem of inference in the model. Our work improves upon the statistical treatment in the ecology literature. Zonneveld and almost all subsequent authors have made a simplification to the likelihood that (wrongly) assumes independence between observations at different times. While this is convenient for estimating parameters, it relies on a false assumption, and, more importantly, because the relevant random variables are replaced by their expected values, it impoverishes the model in a way that prevents inference over hidden aspects of the process.

Our model works by dividing time into intervals based on the observation times and then binning all individuals according to their birth and death intervals. The number of individuals in each bin is unknown and treated as a latent variable. We then seek to infer the values of the latent variables from (potentially noisy) observations of abundance. The problem is particularly challenging when the observations are exact, because this imposes hard constraints on the latent variables. For the task of inference over the hidden variables, we contribute a novel Gibbs sampler that uses a set of update "moves" to resample the latent variables. We prove that these moves form a *Markov basis*—i.e., they lead to an ergodic sampler—even in the presence of hard constraints. We also prove that the univariate distributions encountered by our sampler are log-concave, which allows for highly efficient sampling even in large populations.

We empirically validate our theoretical result to show that, when there are hard constraints, our novel moves are required for ergdocity, and that they *accelerate* convergence even when there are no hard constraints. We also demonstrate the scalability benefits of log-concavity and present a case study to demonstrate the value of our latent variable model for making inferences about the hidden aspects of partially observed transient populations.

2. Related Work

We briefly mention some related work. Eick et al. (1993) give a detailed mathematical analysis of the $M_t/G/\infty$ queue. Their reasoning about the covariance of queue size at two different times (Theorem 2) uses a scheme similar to our latent variable model to partition individuals by their birth and death intervals. Several queueing papers touch on the idea of parameter estimation from partial observations. Ross et al. (2007) estimate the parameters of an $M/M/c$ queue from length data. That model differs from ours in that it has a finite number of servers, and arrival and departure rates are constant over time so there are only two parameters to estimate. Blanghaps et al. (2013) estimate the service-time (lifespan) distribution of an $M/G/\infty$ model from partial data about arrivals and departures. This differs from our work because arrival rates are assumed constant, and the data and inferential goals are different.

In the ecology literature, Gross et al. (2007) share our motivation of addressing the simplifying assumptions made by Zonneveld (1991). Their emphasis is parameter estimation, but to better estimate confidence intervals, they develop an MCMC method to sample from the correct probabilistic model. Their approach differs from ours in several ways. First, they do not fully represent the latent process: in particular, they aggregate all individuals alive on a day regardless of when they entered the population. This is valid only for exponential lifespan distributions (constant death rates), while our method applies more generally. By aggregating, their model also loses the ability to answer inference queries about lifespans. Second, they do not consider the problem of perfect observations and the hard constraints they impose on the sampler. Finally, their sampler operates in discretized time and moves individual emergence times by one unit at a time, which scales poorly with population size. We work in continuous time and exploit log-concavity to scale to very large populations efficiently.

Our sampling approach draws on the concept of Markov bases and is closely related to samplers for contingency tables (Diaconis & Sturmfels, 1998). Our idea to exploit log-concavity for efficient sampling in very large populations is based on ideas from sampling in collective graphical models (Sheldon & Dietterich, 2011).

3. Generative Model

In this section, we first introduce the underlying probabilistic model for lifespans of individuals in a transient population and describe the model of repeated observations of populations size. Then, we describe how to formulate the entire process as a generative model with discrete latent variables over which we will perform inference.

The model assumes that N individuals will be born or enter the study area during a fixed time interval (e.g., for insects, N is the total number from a single generatation) and that individuals are independent and identically distributed. The ith individual is born at time S_i and has lifespan Z_i. Birth times are drawn independently from a distribution with density $f_S(s)$ and lifespans are drawn independently from a distribution with density $f_Z(z)$. In our experiments, we use the normal density for f_S and exponential density for f_Z to mimic Zonneveld's setup, though the method can work with arbitrary distributions. This model differs very slightly from the $M_t/G/\infty$ queue because it assumes a fixed number of individuals. However, it becomes identical if we assume that $N \sim \text{Poisson}(\lambda)$, in which case the birth times follow a Poisson process with intensity function $\lambda f_S(s)$. Our methods apply to that case with very minor modifications, which we describe in Section 4.5.

Our observation model assumes we cannot directly observe the births or lifespans. Instead, we make T measurements of abundance (population size) at times $\{t_1, t_2, \ldots, t_T\}$. Let n_k be the actual abundance at time t_k. We assume that each individual in the population is observed independently with probability α to yield the noisy count $y_k \sim \text{Binomial}(n_k, \alpha)$ for some $0 \leq \alpha \leq 1$.

To formulate the latent variable model, it is useful to notice that the observation times partition the real line into intervals $\{I_0, I_1, \ldots, I_T\}$ (e.g., see Figure 1) and we can use these intervals to aggregate lifespan events. The joint probability of an individual being born at some point in interval I_i and leaving the population at some point in interval I_j is

$$p(i,j) := \int_{t_i}^{t_{i+1}} f_S(s) \int_{t_j - s}^{t_{j+1}-s} f_Z(z)\, dz\, ds.$$

Similarly, let $q(i,j)$ be a random variable denoting the total number of individuals who are born during interval I_i and die during I_j, and let $\mathbf{p}$ and $\mathbf{q}$ be the vector concatenation of the $p(i,j)$ and $q(i,j)$ values, respectively. (Later, we will view $\mathbf{p}$ and $\mathbf{q}$ as matrices when it is convenient to do so.) Since individuals are i.i.d. and each is counted in exactly one cell of $\mathbf{q}$, the marginal distribution of $\mathbf{q}$ is $\text{Multinomial}(N, \mathbf{p})$.

The count variables $\mathbf{q}$ suffice as latent variables to determine the abundance at sampling times. In particular, the abundance n_k at observation time t_k is:

$$n_k = \sum_{i<k} \sum_{j \geq k} q(i,j). \tag{1}$$

This is the number of individuals that were born in an interval prior to t_k and die in an interval after t_k. Note that individuals that are born and die in the same interval, i.e. they are counted in a diagonal entry $q(i,i)$, are not included in any n_k because they were never alive during an observation time. Alternatively, we can write Eq. (1) as $n_k = \mathbf{a}_k^T \mathbf{q}$, where $\mathbf{a}_k$ is the vector with entries

$$a_{k,(ij)} = \begin{cases} 1 & i < k \leq j \\ 0 & \text{otherwise} \end{cases}.$$

Then we can stack the vectors $\mathbf{a}_k^T$ into the rows of the matrix A to write the abundance values compactly as $\mathbf{n} = A\mathbf{q}$.

This provides enough information to succinctly write the full generative model:

$$\begin{aligned} \mathbf{q} &\sim \text{Multinomial}(N, \mathbf{p}), \\ \mathbf{n} &= A\mathbf{q}, \\ y_k &\sim \text{Binomial}(n_k, \alpha). \end{aligned}$$

The full joint probability of latent variables $\mathbf{q}$ and noisy observations $\mathbf{y}$ is then the product of the multinomial prior and the binomal likelihood:

$$p(\mathbf{q}, \mathbf{y}) = N! \prod_{i,j} \frac{p(i,j)^{q(i,j)}}{q(i,j)!} \prod_k \frac{n_k!}{y_k!(n_k - y_k)!} \alpha^{y_k} (1-\alpha)^{n_k - y_k}$$

In this equation, it is understood that n_k is a deterministic function of $\mathbf{q}$. Direct computation of the marginal probability $p(\mathbf{y})$ is intractable because it requires summing over all possible values of $\mathbf{q}$.

In contrast to this model, Zonneveld used the following tractable approximation:[1]

$$\begin{aligned} \boldsymbol{\rho} &= A\mathbf{p} \\ y_k &\sim \text{Binomial}(N, \alpha\rho_k) \end{aligned}$$

This model makes two major simplifications. First, the latent random varables $\mathbf{q}$ are replaced by the deterministic quantities $\boldsymbol{\rho}$, where ρ_k is the probability an individual is alive at time t_k. This model is therefore incapable of performing inference over the latent process. Second, because the observation y_k now only depends on the deterministic quantity ρ_k, the observations at different times become mutually independent. In the true model, observations are correlated due to the lifespans of individuals that span multiple observation times. The primary motivation of our work is to develop an inference procedure for the more difficult, but correct, model in which $\mathbf{q}$ is preserved as a latent variable.

4. Inference

In this section our goal will be to draw samples from the conditional distribution of $\mathbf{q}$ given observations $\mathbf{y}$ and known density functions f_S and f_Z (and hence known cell probabilities $\mathbf{p}$). Exact calculation of the likelihood is intractable because it involves summing over all possible configurations of $\mathbf{q}$, but sampling is a tractable alternative.

4.1. Hard Constraints

Our method is based on Markov Chain Monte Carlo (MCMC) sampling, but a key difficulty arises as $\alpha \to 1$ due to the presence of hard constraints in the probability distribution. To see this, note that a typical approach for

[1] Zonneveld wrote this using a Poisson likelihood $y_k \sim \text{Poisson}(\alpha N \rho_k)$, but we write it as Binomial to make a more direct comparison. This is appropriate when individuals are counted only once during a single survey.

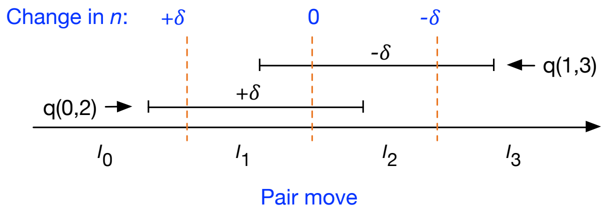


Figure 1. Illustration of a pair move on timeline. Observation times (vertical lines) divide time into intervals I_0, I_1, I_2, I_3. This particular move subtracts δ individuals from $q(1,3)$ and adds them to $q(0,2)$. Note that this move does not preserve the abundances at observation times t_1 and t_3.

MCMC in a multinomial would be to simultaneously resample the counts in two cells of $\mathbf{q}$, with the effect that one increases by δ and the other decreases by δ. Figure 1 shows an example where $q(0,2)$ increases by δ and $q(1,3)$ decreases by δ. It is easy to see in this example that n_1 and n_3 also change from their original values (by $+\delta$ and $-\delta$ respectively) so this modification is not possible under the constraint on abundance at observation times.

Overall, there are four constraints we must consider when proposing a new value for $\mathbf{q}$. The first two come from the multinomial distribution: $\mathbf{q}$ must always remain non-negative and $\mathbf{q}$ must sum to N, which we write compactly as $\mathbf{q} \geq \mathbf{0}$ and $\mathbf{1}^T\mathbf{q} = N$. The third constraint comes from the binomial likelihood: for all k, the observed value y_k may not exceed the true queue length at time t_k, which we write as $A\mathbf{q} \geq \mathbf{y}$. Finally, when $\alpha = 1$, the observations specify the *exact* abundance values, so the previous constraint becomes an equality constraint: $A\mathbf{q} = \mathbf{y}$.

Our approach will then be based on "moves" that carefully modify more than two entries of $\mathbf{q}$ so the constraints are always preserved. A move is a vector $\mathbf{z}$ of the same size as $\mathbf{q}$, with entries in the set $\{-1, 0, +1\}$. A new configuration $\mathbf{q}' = \mathbf{q} + \delta\mathbf{z}$ is obtained by first selecting a move $\mathbf{z}$, and then choosing an integer move amount δ.

We describe below how the moves are designed to always preserve the equality constraints. The inequality contraints place bounds on the possible value of δ. First, to ensure that all entries of $\mathbf{q}'$ remain non-negative, δ must be at least $L_1 = -\min\{q(i,j) : z(i,j) = +1\}$ and at most $U_1 = \min\{q(i,j) : z(i,j) = -1\}$. Second, to ensure that $A\mathbf{q}' \geq \mathbf{y}$, we require $\delta A\mathbf{z} \geq \mathbf{y} - A\mathbf{q}$, which provides:

$$\delta \geq L_2 := \max\{(\mathbf{y} - A\mathbf{q})_k/(A\mathbf{z})_k : (A\mathbf{z})_k > 0\},$$
$$\delta \leq U_2 := \min\{(\mathbf{y} - A\mathbf{q})_k/(A\mathbf{z})_k : (A\mathbf{z})_k < 0\}.$$

The overall constraints are $\delta \geq L := \max\{L_1, L_2\}$ and $\delta \leq U := \min\{U_1, U_2\}$.

The value of δ is chosen by sampling from the induced univariate distribution:

$$p(\delta) \propto p(\mathbf{q} + \delta\mathbf{z} \mid \mathbf{y}), \quad \delta \in \{L, \ldots, U\}. \tag{2}$$

The values of $p(\delta)$ are proportional to the joint probability $p(\mathbf{q} + \delta\mathbf{z}, \mathbf{y})$, which can be computed efficiently.

Designing moves in this way leads to a form of Gibbs sampler (Geman & Geman, 1984): the proposed configuration $\mathbf{q}'$ is drawn from the restricted set $\{\mathbf{q} + \delta\mathbf{z} : L \leq \delta \leq U\}$ with probability proportional to $p(\mathbf{q}')$. Just as in standard Gibbs sampling, the ratio of the proposal density to the true density is equal to one and the move is always accepted.

4.2. Markov Basis

A challenge is to design a set of moves such that the sampler is ergodic. Let $\mathcal{Q}$ be the set of configurations that satisfy all hard constraints. We require that, for any two configurations $\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{Q}$, there is a valid sequence of moves that leads from $\mathbf{q}_1$ to $\mathbf{q}_2$. A move set $\mathcal{M}$ that satisfies this property is called a *Markov basis* with respect to $\mathcal{Q}$ (Diaconis & Sturmfels, 1998). We next describe several patterns of moves that we will use to construct Markov bases. When selecting a move $\mathbf{z}$, we first select one of these patterns uniformly at random, then select the indices for the move uniformly at random.

A **pair move** $\mathbf{z} \in \mathcal{M}_{\text{pair}}$ is of the form illustrated in Figure 1. It is specified by four indices $i \leq j$, $k \leq \ell$ such that $(i,j) \neq (k,\ell)$. It has one positive entry $z(i,j) = +1$ and one negative entry $z(k,\ell) = -1$, with the effect of moving one individual from cell (k,ℓ) to cell (i,j). Pair moves do not in general preserve the abundance values $\mathbf{n} = A\mathbf{q}$.

A **shuffle move** is the special case of pair moves that occurs when $i = j$ and $k = \ell$, so only unobserved individuals are "shuffled". Such a move *does* preserve abundance values $\mathbf{n} = A\mathbf{q}$ at observation times. We denote the set of all shuffle moves by $\mathcal{M}_{\text{shuffle}}$.

A **cycle move** $\mathbf{z} \in \mathcal{M}_{\text{cycle}}$ (Figure 2, top) is specified by four indices $i \leq i' \leq j \leq j'$. It has four non-zero entries $z(i,j) = z(i',j') = +1$ and $z(i,j') = z(i',j) = -1$ with the effect of taking two overlapping lifetimes and swapping their end intervals, e.g.: $(i,j), (i',j') \leftrightarrow (i,j'), (i',j)$. It is straightforward to see that a cycle move preserves the abundance values $\mathbf{n} = A\mathbf{q}$ at observation times. Cycle moves are well-known from the contingency table literature (Diaconis & Sturmfels, 1998). When viewed as a matrix, a cycle move modifies a pair of rows and columns of $\mathbf{q}$ in the pattern illustrated at the right. From this it is clear that, in addition to preserving the abundance values $\mathbf{n} = A\mathbf{q}$, a cycle move also preserves the row and column sums of $\mathbf{q}$, i.e., the numbers of individuals that are born and die in each interval.

$$\begin{array}{c|cc|} & j & j' \\ \hline i & + & - \\ i' & - & + \\ \hline \end{array}$$

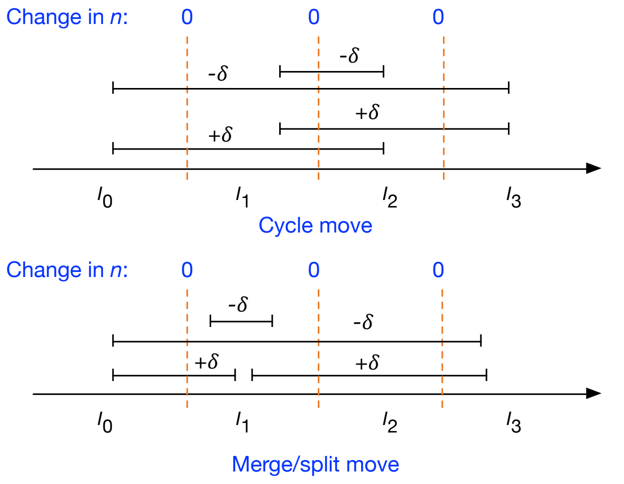


Figure 2. Moves that preserve the constraint $A\mathbf{q} = \mathbf{y}$. Top: cycle. Bottom: merge/split.

A **merge/split move** (Figure 2, bottom) is the special case of cycle moves when $i' = j$. It has the effect of either merging two short lifetimes into a longer one and adding an unobserved individual or splitting one longer lifetime into two shorter ones and eliminating an unobserved individual.

Our main result is that the moves above can be combined to form a Markov basis for both the cases $\alpha < 1$ and $\alpha = 1$.

Theorem 1. *When $\alpha < 1$, the feasible set (i.e, the support of the distribution $p(\mathbf{q} \mid \mathbf{y})$) is $\mathcal{Q} = \{\mathbf{q} : \mathbf{q} \geq \mathbf{0}, \mathbf{1}^T\mathbf{q} = N, A\mathbf{q} \geq \mathbf{y}\}$. The set of all pair moves is a Markov basis with respect to $\mathcal{Q}$. When $\alpha = 1$, the feasible set is $\mathcal{Q}_\mathbf{y} = \{\mathbf{q} : \mathbf{q} \geq \mathbf{0}, \mathbf{1}^T\mathbf{q} = N,\ A\mathbf{q} = \mathbf{y}\}$. The set of all cycle and shuffle moves is a Markov basis with respect to $\mathcal{Q}_\mathbf{y}$.*

Although Theorem 1 implies that cycle moves are only strictly required when $\alpha = 1$, we will show empirically that they can subtantially improve mixing time even when $\alpha < 1$. Before proving Theorem 1, we first state several useful lemmas.

Lemma 1. *Any configuration $\mathbf{q}$ such that $A\mathbf{q} = \mathbf{y}$ satisfies $\sum_{k>j} q(j,k) - \sum_{i<j} q(i,j) = y_{j+1} - y_j$ for all j.*

Proof. This simply states that the change in abundance $y_{j+1} - y_j$ between the start and end of the jth interval is equal to the number of new individuals $\sum_{k>j} q(j,k)$ (those that are born during interval j and make it until the end of the interval) minus the number of lost individuals $\sum_{i<j} q(i,j)$ (those that were alive at the start of the interval but died during interval j). □

Lemma 2. *Suppose $A\mathbf{q} = \mathbf{y}$. The following conditions are equivalent:*

(i) No merge moves can be performed on $\mathbf{q}$,

(ii) For all j, either $\sum_{i<j} q(i,j) = 0$ or $\sum_{k>j} q(j,k) = 0$,

(iii) For all j, we have $\sum_{i<j} q(i,j) = \max\{0, y_j - y_{j+1}\}$ and $\sum_{k>j} q(j,k) = \max\{0, y_{j+1} - y_j\}$.

Proof. It is easy to see that if (ii) is not satisifed then some merge move can be performed, so (i) implies (ii). If (ii) is satisfied, then each interval has either births or deaths, but not both. Thus, the total numbers of births and deaths are determined by Lemma 1 and the sign of $y_{j+1} - y_j$: the number of deaths is equal to $\max\{0, y_j - y_{j+1}\}$ and the number of births is equal to $\max\{0, y_{j+1} - y_j\}$. This shows that condition (ii) implies (iii). Finally, if (iii) is true, then it is clear that no merge moves can be performed, so (iii) implies (i). □

Lemma 3. *Let $\mathbf{q}$ and $\mathbf{q}'$ be any two configurations such that $A\mathbf{q} = A\mathbf{q}' = \mathbf{y}$, the diagonal entries of $\mathbf{q}$ and $\mathbf{q}'$ are the same, and no merge moves can be performed in either configuration. Then $\mathbf{q}$ and $\mathbf{q}'$ have the same row and column sums.*

Proof. The ith row sum of q is $q(i,i) + \sum_{j<i} q(i,j)$. We have assumed that $q(i,i) = q'(i,i)$. By Lemma 2 and the assumption that no merge moves are possible in either configuration, we have $\sum_{j>i} q(i,j) = \sum_{j>i} q'(i,j)$, since both of these are deterministic functions of $\mathbf{y}$ (condition (iii) of the Lemma). Therefore the row sums of $\mathbf{q}$ and $\mathbf{q}'$ are the same. The case of column sums is similar. □

Proof of Theorem 1. Let $\mathbf{q} \xrightarrow{\mathcal{M}} \mathbf{q}'$ indicate that there is a valid sequence of moves from $\mathbf{q}$ to $\mathbf{q}'$ in $\mathcal{M}$. This means there are moves $\mathbf{z}_1, \ldots, \mathbf{z}_M \in \mathcal{M}$ such that $\mathbf{q}' = \mathbf{q} + \mathbf{z}_1 + \ldots + \mathbf{z}_M$ and $\mathbf{q} + \mathbf{z}_1 + \ldots + \mathbf{z}_m \in \mathcal{Q}$ for all $0 \leq m \leq M$. It suffices to consider moves with $\delta = 1$ for the purposes of this proof. We wish to show that $\mathbf{q} \xrightarrow{\mathcal{M}} \mathbf{q}'$ for all $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$.

The case $\alpha < 1$ is easy. Starting from $\mathbf{q}$, execute pair moves to move all individuals to cell $(0, T)$, which guarantees that all individuals are present at each observation and hence $A\mathbf{q} \geq \mathbf{y}$ remains satisfied. Then, for each other cell (i,j) we execute pair moves to move $q'(i,j)$ individuals from cell $(0,T)$ to cell (i,j).

For $\alpha = 1$, it is straightforward to verify that each cycle and shuffle move does not change the abundance at observation times or the total number of individuals, so the equality constraints $A\mathbf{q} = \mathbf{y}$ and $\mathbf{1}^T\mathbf{q} = N$ always remain satisfied. We need to demonstrate that the moves can be constructed to also preserve the non-negativity constraints.

We will proceed by first transforming $\mathbf{q}$ and $\mathbf{q}'$ into configurations $\mathbf{r}$ and $\mathbf{r}'$ that have the same number of births and deaths in each interval, and then transforming $\mathbf{r}$ into

$\mathbf{r}'$ through another sequence of moves. Put together, this will show that $\mathbf{q} \xrightarrow{\mathcal{M}} \mathbf{r}$, $\mathbf{r} \xrightarrow{\mathcal{M}} \mathbf{r}'$, and $\mathbf{q}' \xrightarrow{\mathcal{M}} \mathbf{r}'$. Because moves are reversible ($\mathcal{M}$ is closed under negation), we can also conclude that $\mathbf{r}' \xrightarrow{\mathcal{M}} \mathbf{q}'$ and hence $\mathbf{q} \xrightarrow{\mathcal{M}} \mathbf{q}'$.

To transform $\mathbf{q}$ into $\mathbf{r}$, first apply merge moves until no more are possible, and then perform shuffle moves to move all unobserved individuals to cell $(0, 0)$. Do the same to transform $\mathbf{q}'$ into $\mathbf{r}'$. Now, the conditions of Lemma 3 apply, and we can conclude that $\mathbf{r}$ and $\mathbf{r}'$ have the same row and column sums.

We will now show that there is a sequence of cycle moves leading from $\mathbf{r}$ to $\mathbf{r}'$. The reasoning is very similar to the argument that cycle moves form a Markov basis for contingency tables with fixed row and column sums (Diaconis & Sturmfels, 1998)—however, we have the additional restriction that $\mathbf{q}$ is upper triangular, so our result does not follow directly from that result.

Let $\Delta = \mathbf{r}' - \mathbf{r}$. We wish to create a sequence of moves that add up to Δ. It is enough to find one cycle move $\mathbf{z}$ such that $\|\Delta - \mathbf{z}\|_1 < \|\Delta\|_1$, which means that applying the move $\mathbf{z}$ to $\mathbf{r}$ moves us strictly closer to $\mathbf{r}'$. We can then apply an inductive argument.

Since $\mathbf{r}$ and $\mathbf{r}'$ have the same row and column sums, we know that Δ has row and column sums that are identically zero. Identify a cycle move using Δ as follows: first, let $\Delta(i_1, j_1)$ be a negative entry of Δ, which must exist as long as $\mathbf{r} \neq \mathbf{r}'$. Since the j_1 column-sum of Δ is zero, there must also be a positive entry $\Delta(i_2, j_1)$ in the same column. Now, since the i_2 row sum is zero, there must be negative entry $\Delta(i_2, j_2)$ in the same row. Construct a cycle move using these four indices. This gives the following:

$$\begin{aligned}
\Delta(i_1, j_1) &< 0, \quad & z(i_1, j_1) &= -1 \\
\Delta(i_1, j_2) &> 0, \quad & z(i_1, j_2) &= +1 \\
\Delta(i_2, j_2) &< 0, \quad & z(i_1, j_2) &= -1 \\
\Delta(i_2, j_1) &= ?, \quad & z(i_2, j_1) &= +1
\end{aligned}$$

Since Δ is integer-valued, it is clear that subtracting $\mathbf{z}$ from Δ reduces the sum of absolute values of Δ by three for the first three cells, and increases by at most one for the last cell. We conclude that $\|\Delta - \mathbf{z}\|_1 < \|\Delta\|_1$, as desired.

The final thing to check is that the non-negativity constraint $\mathbf{r}+\mathbf{z} \geq \mathbf{0}$ remains satisfied. For the (i_1, j_1) cell, we observe that $r(i_1, j_1) > r'(i_1, j_1) \geq 0$, so decreasing $r(i_1, j_1)$ by one cannot violate the constraint. The (i_2, j_2) entry is similar. The (i_2, j_1) and (i_1, j_2) entries both increase, which cannot violate non-negativity. □

4.3. Log-Concavity and Efficient Sampling

As discussed in Section 4, the probability $p(\delta)$ for a specific move can be calculated efficiently (Eq. (2)). However, after fixing a move $\mathbf{z}$, the number of possible values for δ can grow very large as the population size N increases. Thus, the running time of a naive sampling method that computes $p(\delta)$ for all possible values and then samples from this discrete distribution scales poorly with N. To alleviate this issue, we prove that $p(\delta)$ is log-concave, which allows us to apply the discrete adaptive random sampling (ARS) algorithm (Gilks & Wild, 1992; Sheldon, 2013) to sample from $p(\delta)$ in time that depends only very mildly on the number of possible values, which nearly eliminates the dependence of running time on population size and allows us to scale to very large populations.

Theorem 2. *The distribution $p(\delta)$ is log-concave, i.e., $p(\delta)^2 \geq p(\delta-1)p(\delta+1)$ for all $\delta \in \mathbb{Z}$.*

A proof is provided in the supplementary material.

4.4. Initialization with Canonical Form

Before running our sampler, we must initialize the latent variables $\mathbf{q}$ in a way that satisfies all of the constraints. We initialize $\mathbf{q}$ to a canonical form that always satisfies the equality constraint $A\mathbf{q} = \mathbf{y}$ (and thus the inequality constraint $A\mathbf{q} \geq \mathbf{y}$) by using the reasoning of Lemma 2. In particular, we iterate over over observation times t_k while maintaining a "supply" of individuals that have lived from previous intervals; in each interval, we either decrement the supply (by ending the lifetime of some individuals) or increase it (by spawning new individuals) to explain the difference between y_k and y_{k-1}. At the end of the process, any remaining individuals are created to be "unobserved" and distributed uniformly along the diagonal of $\mathbf{q}$.

4.5. Modifications for Poisson Model

We briefly return to the discussion of the $M_t/G/\infty$ queue, which is obtained from our model when $N \sim \text{Poisson}(\lambda)$ instead of being a fixed constant. Minor technical differences arise in this case. First, in the generative model, the distribution of $\mathbf{q}$ is no longer multinomial; instead, by standard Poisson thinning arguments, it now has entries that are *independent* Poisson random variables: $q(i,j) \sim \text{Poisson}(\lambda p(i,j))$. In the sampler, the hard constraint $\mathbf{1}^T\mathbf{q} = N$ that arises from the multinomial distribution becomes unnecessary and invalid. As a result, pair moves are no longer necessary, and are replaced in the sampler by moves that change only a single entry of $\mathbf{q}$. All other constraints remain valid, and cycle moves, which are designed to preserve the hard constraints $A\mathbf{q} = \mathbf{y}$ when $\alpha = 1$, remain valid and necessary.

5. Experiments

We now report several experiments to evaluate the performance of our sampler and demonstrate the advantages

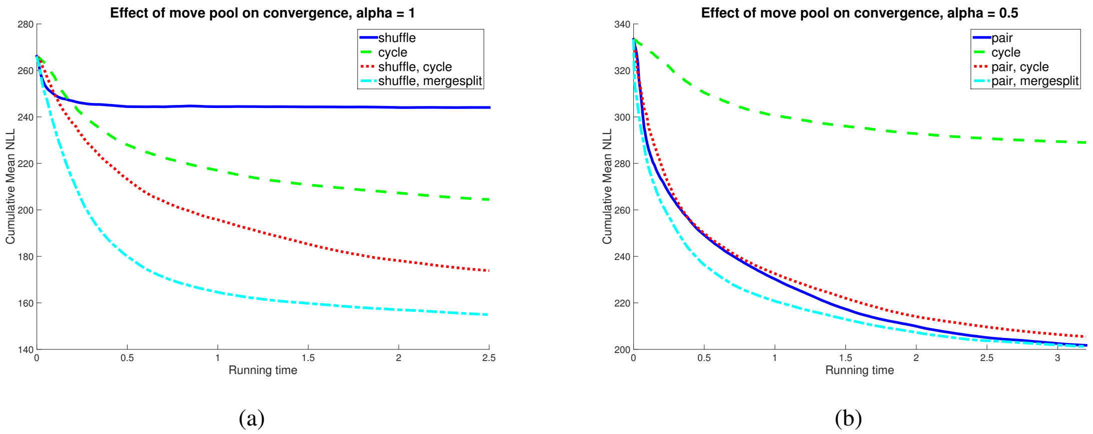


Figure 3. Effect of move types on convergence for two values of the observability probability: (a) $\alpha = 1$, (b) $\alpha = 0.5$. Plots show cumulative mean negative log-likelihood of MCMC iterates vs. number of seconds.

of having a latent variable model. Our first two experiments empirically confirm the ergodicity result of Theorem 1 and demonstrate the improvement in mixing time resulting from adding supplemental moves to the sampler. Our third experiment demonstrates the running-time advantages gained by exploiting the log-concavity of the likelihood function within the sampler. We also provide a case study that compares the inference capabilities of our latent variable method compared to the previous approach of (Zonneveld, 1991).

Effect of move types on convergence when $\alpha = 1$. To evaluate the convergence of our Gibbs sampler under the hard constraints imposed when $\alpha = 1$, we generated data for a population of size $N = 100$ from a model with emergence density $f_S(s) \sim \text{Normal}(\mu = 8, \sigma = 4)$ and lifespan density is $f_Z(z) \sim \text{Exp}(\tau = 3)$ (parameterized by the mean τ) and computed observations at times $t = \{1, 2, 3, \ldots, 20\}$. We then performed MCMC from the intial configuration described in Section 4 using different subsets of the full move pool.

Figure 3(a) shows the convergence of the cumulative mean negative log likelihood (NLL) of the first 1500 MCMC iterates. When $\alpha = 1$, the sampler with only pair moves converges to a mean NLL that is much higher than the other samplers: this is evidence that pair moves are insufficient for the sampler to reach higher probability configurations. Similarly, the sampler with only cycle moves cannot explore the whole space because it cannot adjust the lifespans of unobserved individuals (diagonal entries of $\mathbf{q}$). In contrast, the samplers that use both pair and cycle moves are able to explore the complete space, and converge to a much lower mean NLL. Note that "pair, mergesplit" converges to the same mean NLL as "pair, cycle". It is possible to show that merge/split moves can be used to simluate any cycle move, so "pair, mergesplit" is also an ergodic sampler.

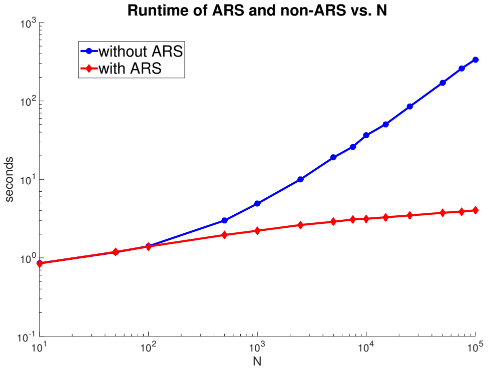
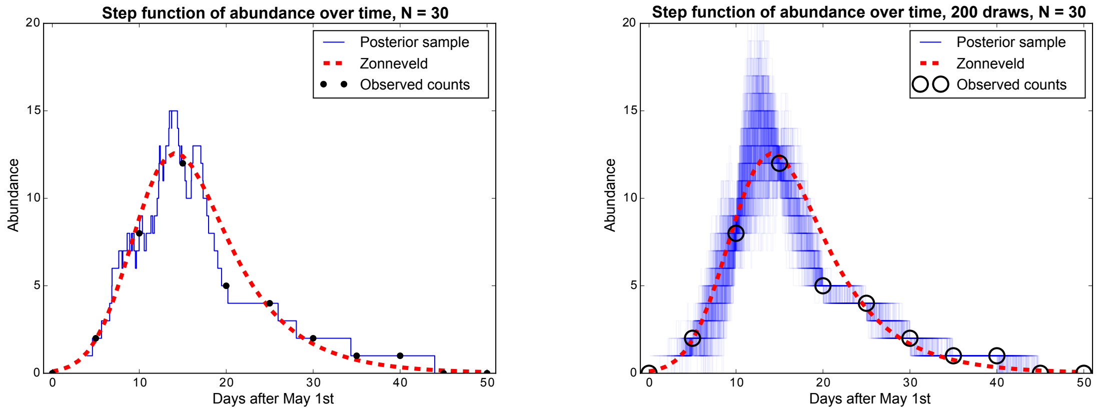


Figure 4. Running time of 1000 MCMC iterations vs. population size for sampler with and without ARS.

Effect of move types on convergence when $\alpha < 1$. Figure 3(b) shows results of the same experiment for $\alpha = 0.5$. In this case, pair moves alone are sufficient for convergence, so all samplers that include pair moves converge to the same mean NLL. In contrast, cycles moves are not sufficient for convergence, because they preserve the initial value of $A\mathbf{q}$ (abundance at sampling times), which is not a valid constraint when $\alpha < 1$. Adding cycle moves alone to the pair moves does not improve the speed of convergence. However, adding merge/split moves, which are a subset of cycle moves, does improve convergence speed. This demonstrates the fact that our more sophisticated moves are valuable even when hard constraints are not present.

Impact of Log-Concavity on Efficiency of Sampler. To evaluate the running-time improvements of ARS over the naive sampling method for $p(\delta)$ we recorded the running time of 1000 MCMC iterations using the entire pool of moves with and without ARS. For this experiment, we fixed the parameters $\mu = 8.0$, $\sigma = 4.0$, $\tau = 3.0$, $\alpha = 0.5$,



(a)

(b)

Figure 5. Plots of abundance over time as samples from the posterior on $\mathbf{q}$. In both cases, $f_S(s) \sim \text{Normal}(\mu = 10.5, \sigma^2 = 7.84)$ and $f_Z(z) \sim \text{Exponential}(\tau = 7)$ with $N = 100$ and $t = \{0, 5, 10, \ldots, 50\}$.

$t \in \{1, 2, \ldots, 20\}$, and varied the population size N to study the scalability of the sampler with respect to population size. Figure 4 shows the results, averaged over 10 trials for each value of N. The naive method scales approximately linearly (note that both axes are log-scale) with N, as expected, while the running time of the ARS-based algorithm grows very slowly with population size. As a result, it can scale to very large populations and outperforms the naive method by orders of magnitude as N increases.

Benefits of Latent Variable Model. Our method provides a number of unique advantages over Zonneveld's approximation to the likelihood. In particular, by retaining the latent variables, we can query the posterior distribution of life history events given observations. Figure 5(a) illustrates this comparison. Given a set of observed counts, Zonneveld's likelihood approximation can be used to estimate model parameters; this then provides a mean abundance curve under those parameters (red line).

Our method allows much finer-grained inference. For example, the blue line shows a sample from the joint posterior over abundance over the entire interval given the observations. This is an integer-valued curve that increases or decreases by one at arbitrary points in time when a new individual is born or an existing individual dies. In this case, $\alpha = 1$ so the curve must exactly match the observations.

To generate this sample, we first use our MCMC sampler to generate $\mathbf{q}$ from the posterior distribution given observations $\mathbf{y}$. This specifies how many individuals are born in I_i and die in I_j for all i, j. We then generate lifespans for each individual as follows. For each (i, j), we generate $q(i, j)$ lifespans from the conditional distribution of S and Z given $S \in I_i$ and $S + Z \in I_j$. This is done by a simple rejection sampler.

Figure 5(b) illustrates the entire posterior distribution by showing 200 semi-transparent samples from the posterior as in Figure 5(a). The samples in Figure 5(b) were obtained by running or MCMC sampler over $\mathbf{q}$ until convergence, and then thinning to obtain approximately independent samples. Notice that each of the sampled abundance curves converges to each of the observations, since $\alpha = 1$. The increased spread of the samples between observation times gives a sense of the increased variability in the model as it interpolates between points.

This case study illustrates the advantages of having a true latent variable model together with an efficient method to draw samples from the posterior distribution.

6. Conclusion

This paper introduces a novel latent variable model for inference in transient populations when only periodic observations of population size are available. The population model arises both in queueing theory as an $M_t/G/\infty$ queue and in ecological models for insect populations. Previous approaches in the ecology literature have made a simplifying assumption to make the likelihood tractable. Instead, we present a Gibbs sampler for the correct, but intractable, likelihood. The Gibbs sampler employs specially-designed moves to preserve the hard constraints present in this problem, and we prove that these lead to an ergodic sampler. We empirically validate the ergodicity result and show that special moves lead to faster mixing even when hard constraints are not present. Finally, we demonsrate the utility of this model over existing work with a comparative case study.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1125228 and by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1052875.

References

Blanghaps, Nafna, Nov, Yuval, Weiss, Gideon, and Others. Sojourn time estimation in an $M/G/\infty$ queue with partial information. *Journal of Applied Probability*, 50(4): 1044–1056, 2013.

Diaconis, Persi and Sturmfels, Bernd. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 1998.

Eick, Stephen G., Massey, William A., and Whitt, Ward. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, 1993.

Geman, Stuart and Geman, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Gilks, W. R. and Wild, P. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society*, 41(2):337–348, 1992.

Gross, Kevin, Kalendra, Eric J., Hudgens, Brian R., and Haddad, Nick M. Robustness and uncertainty in estimates of butterfly abundance from transect counts. *Population Ecology*, 49(3):191–200, 2007.

Ross, J. V., Taimre, T., and Pollett, P. K. Estimation for queues from queue length data. *Queueing Systems*, 55: 131–138, 2007.

Sheldon, Daniel. Discrete adaptive rejection sampling. Technical Report UM-CS-2013-012, School of Computer Science, University of Massachusetts, Amherst, Massachusetts, May 2013.

Sheldon, Daniel and Dietterich, Thomas. Collective graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1161–1169, 2011.

Zonneveld, Cor. Estimating death rates from transect counts. *Ecological Entomology*, 16(1):115–121, 1991.

Supplementary Material:
Inference in a Partially Observed Queuing Model with Applications in Ecology

Kevin Winner[1] KWINNER@CS.UMASS.EDU
Garrett Bernstein[1] GBERNSTEIN@CS.UMASS.EDU
Daniel Sheldon[1,2] SHELDON@CS.UMASS.EDU
[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA 01002, USA
[2]Department of Computer Science, Mount Holyoke College, South Hadley, MA 01075, USA

1. Proof of Log-Concavity

Proof of Theorem 2.

Proof. We can factor the pmf of δ as follows:

$$\begin{aligned}
p(\delta) &= a(\delta)b(\delta)\prod_{n_k\in n^+} c_k(\delta)e(\delta)\prod_{n_l\in n^-} d_l(\delta)e(\delta)^{-1}\\
a(\delta) &= \prod_{p_i\in p^+, q_i\in q^+} \frac{p_i^{q_i+\delta}}{(q_i+\delta)!}\\
b(\delta) &= \prod_{p_j\in p^-, q_j\in q^-} \frac{p_j^{q_j-\delta}}{(q_j-\delta)!}\\
c_k(\delta) &= \frac{(n_k+\delta)!}{(n_k+\delta-y_k)!}\\
d_l(\delta) &= \frac{(n_l-\delta)!}{(n_l-\delta-y_l)!}\\
e(\delta) &= (1-\alpha)^{\delta}
\end{aligned}$$

Where $\{p^+, q^+, n^+\}$ and $\{p^-, q^-, n^-\}$ represent the subsets of $\mathbf{p}, \mathbf{q}$, and $\mathbf{n}$ which change positively and negatively under $\mathbf{z}$ accordingly. Since the product of log concave functions is also log concave, it is thus sufficient to demonstrate that each of the factors of $\mathcal{L}(\delta)$ is log concave. Observe that the inner part of $a(\delta)$ is the form of $e^{\lambda}\text{Poisson}_{\lambda}(k)$ where $\lambda = p_i$ and $k = q_i + \delta$. Since the Poisson is log concave and so is e^{λ}, $a(\delta)$ is also log concave in δ. By an identical argument, so is $b(\delta)$.

For $e(\delta)$, note that $\log(1-\alpha)^{\delta} = \delta\log(1-\alpha)$, which is linear in δ and therefore $e(\delta)$ is log concave in δ, as is $e(\delta)^{-1}$.

The proof of concavity for $c_k(\delta)$ and $d_l(\delta)$ is below:

$$c_k(\delta) = \frac{(n_k+\delta)!}{(n_k+\delta-y_k)!}$$

Let $n' = n_k + \delta$

$$c_k(n') = \frac{n'!}{(n'-y_k)!}$$

by construction, $n' = n_k + \delta \geq y_k$

to show $c_k(n')$ is log concave, we must show:

$$\begin{aligned}
c_k(n')^2 &\geq c_k(n'-1)c_k(n'+1)\\
\frac{c_k(n')}{c_k(n'+1)} &\geq \frac{c_k(n'-1)}{c_k(n')}\\
\frac{n'!}{(n'-y_k)!}\frac{(n'+1-y_k)!}{(n'+1)!} &\geq \frac{(n'-1)!}{(n'-1-y_k)!}\frac{(n'-y_k)!}{n'!}\\
\frac{n'+1-y_k}{n'+1} &\geq \frac{n'-y_k}{n'}\\
1-\frac{y_k}{n'+1} &\geq 1-\frac{y_k}{n'}
\end{aligned}$$

Thus $c_k(n')$ and, by extension, $c_k(\delta)$ are log concave. A similar argument shows that $d_l(\delta)$ is log concave as well. Then we have shown that $p(\delta)$ is a product of log concave functions and therefore $p(\delta)$ is also log concave. □