# In de verdediging

| In defence


**Statistical Inverse Problems for Population Processes**
*Birgit Sollie*

In June 2021 Birgit Sollie from the Vrije Universiteit Amsterdam successfully defended her PhD thesis. Her thesis has title *Statistical Inverse Problems for Population Processes* and she carried out her research under the supervision of Prof. dr. Mathisca de Gunst (VU) and Prof. dr. Michel Mandjes (UvA).

During her PhD Birgit worked on population processes. In particular, she studied population processes where the parameters depend on a background process. She also developed a model, relying on quasi birth-death processes, to model RNA transcription.

**A population evolving**

Population processes describe how a population evolves over time by modelling their growth as a function of random events. Population processes can be used to study various kinds of populations, of people or animals, but also of molecules in a cell, or even the number of visits to a website. An example is a population where individuals die at rate $\mu$ and new individuals are born at rate $\lambda$. In this model, if $\mu > \lambda$ then the population will eventually die out since individuals die at a higher rate than new individuals are born.

Birgit studied population processes where the parameters depend on some unobserved *background* process. In general, populations evolve in an environment, think for example of microorganisms living in the human body. Complete knowledge of the environment in which a population grows is in most cases unavailable, and measurements come mainly from the population. This is why in many models the environment is modelled as a background process which affects the population and is in principle considered to be unobservable.

In practical applications, where the environment affects the birth and death rates, the two-parameter model above may be inadequate. Think for example of a population of animals whose welfare depends on food availability. By adding a background process to the model this dependence can be taken into account. The background process could be considered to have two states, one for the case of high food availability and one for low. When the background process is in state 'low' then the birth parameter $\lambda$ takes some value $\lambda_1$ and the death parameter some value $\mu_1$. When in state 'high' then these parameters are equal to $\lambda_2$ and $\mu_2$, respectively. This background process alternates between the two states at random moments in time. The time between two switches is considered to follow an exponential distribution. For the interested readers, the background process is thus a two-state Markov process. In Figure 1 below we present a realization of such a population process.

*Pas gepromoveerden brengen hun werk onder de aandacht. Heeft u tips voor deze rubriek of bent u zelf pas gepromoveerd? Laat het weten aan onze redacteur.*

*Redacteur: Nicolaos Starreveld*
*FNWI, Universiteit van Amsterdam*
*Postbus 94214*
*1090 GE Amsterdam*
*verdediging@nieuwarchief.nl*

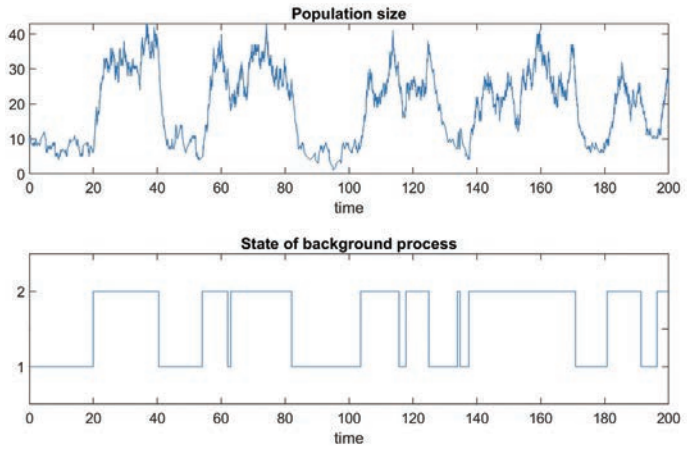


**Figure 1** Upper panel: number of individuals in the population. Lower panel: state of the background process. Depending on the state of the background process the population process grows differently.

You can think of many more examples of external conditions, like temperature affecting the spread of bacteria or weather conditions affecting the mobility of individuals.

**A Markovian world**

Birgit worked with a class of models called Markov-modulated Independent Sojourn Processes (MMIS). These models resemble the model described above but are more general. In an MMIS the population process has two parameters, a birth parameter $\lambda$, which determines the rate of births of new individuals, and a lifetime parameter $\mu$, which determines how long individuals live. Moreover, new individuals are born according to a Poisson Process with parameter $\lambda$ and have a lifetime that follows an exponential distribution with parameter $\mu$.

In an MMIS there is also a background process, which is called the modulating process. In Birgit's work the modulating process is considered to be a continuous time Markov-process, denoted by $\{X(t)\}_{t\geq 0}$, which takes values in a finite state space $\{1,\ldots,d\}$. This process determines the value of the birth rate $\lambda$. As long as $X$ is in state $i$, then new individuals are born according to a Poisson Process with parameter $\lambda_i$. In Birgit's research the lifetime parameter $\mu$ is taken to be constant, thus not varying as the background process changes.

The question that is central in Birgit's research is the following: if we have observations of the population process $\{M(t)\}_{t\geq 0}$, where $M(t)$ denotes the number of individuals in the population at time $t$, is it possible to make accurate estimates of the parameters $\lambda_i$, $\mu$, and the parameters (transition rates and stationary probabilities) of the background process?

Birgit developed an algorithm, based on the Expectation-Maximization algorithm (EM), to determine such estimates. Let's have a look at the basic ideas behind the algorithm. All the parameters that need to be estimated, i.e. the rates $\lambda_i$ and $\mu$, the transition rates and the stationary probabilities of the background Markov process, are collected in one vector-parameter $\theta$. Furthermore, only finitely many observations of the population process are available, denoted by $m_n = (M(t_1),M(t_2),\ldots,M(t_n))$. Two more quantities are also considered, namely $(X,A)$, which describe the behavior of the background process and of all the births that occurred in the time intervals $[t_i,t_{i+1}], i = 1,\ldots,n-1$. These are both random variables since they describe the unobserved behavior in between each pair of observations of the population process. The key in Birgit's analysis is to extract as much information as possible for $(X,A)$ from the finitely many observations in $m_n$.

The EM algorithm works iteratively. Each iteration consists mainly of two steps, first considering an expectation and afterwards solving a maximization problem. The algorithm will work with a likelihood function which describes the joint probabilities of the observations $m_n$ and the random variables $(X,A)$, namely,

$$\mathcal{L}(\theta \mid m_n,A,X) = P_\theta(M = m_n,A,X) = P_\theta(M = m_n \mid A,X)P_\theta(A,X).$$

This likelihood function is a function of the parameter $\theta$, and it is also a random variable since $(X,A)$ are random. As a reminder, the vector-parameter $\theta$ contains all the parameters of the model that need to be estimated from the observations. The addition of the random vector in the likelihood function is necessary. A typical likelihood function with respect only to $m_n$ is not easy to work with in this model due to the unobserved data.

Each iteration of the algorithm will yield a better estimate for the parameter $\theta$. In each iteration we start with an estimate $\theta'$ for $\theta$, which is obtained from the previous iteration of the algorithm. Then an expectation of the log-likelihood function is estimated over the random variables $(X,A)$, but with respect to the new estimate $\theta'$. This expectation is given below,

$$\mathbb{E}_{\theta'}(\log \mathcal{L}(\theta \mid m_n,A,X)). \tag{1}$$

This is a deterministic function of $\theta$. In the maximization step the algorithm updates the estimate $\theta'$ to $\tilde{\theta}$, which will be used in the next iteration. The new estimate $\tilde{\theta}$ is determined as the solution to a maximization problem, namely,

$$\tilde{\theta} = \operatorname{argmax}_\theta(\mathbb{E}_{\theta'}(\log \mathcal{L}(\theta \mid m_n,A,X))). \tag{2}$$

The idea is thus to find the value of the parameter that maximizes the expected likelihood to obtain the observed data $m_n$ under the assumption that the population evolves as an MMIS. For the exact technical steps in order to compute the expectation in (1) and the parameter $\tilde{\theta}$ in (2) we refer the interested reader to Birgit's thesis.

**Populations come in contact**

In the previous section we discussed the case of a single population which does not interact with other populations. In practice, populations interact with each other, as in the case of migration. Think for example of bacteria which can migrate from one cell to other cells. Birgit also studied a network-based model, where each node corresponds to a population process. Individuals from each node can migrate to other populations depending on the underlying structure of the network, namely, individuals can move between two populations if there is an edge connecting the corresponding nodes. The network-based model that Birgit studied is a discrete-time model, and not a continuous-time model as the single population model discussed above. In such a model a vector $\mathbf{M}(k) = (M_1(k),\ldots,M_n(k))$ describes the population process at time $k$, and $M_j(k)$ describes the population process at time $k$ and node $j$.

As before, in this network-based model there is a background process affecting the parameters of the population processes. The question is again to estimate the birth-lifetime parameters, and the transition rates and stationary probabilities of the background pro-

cess, solely from observations of the population process. Unfortunately, the EM algorithm described above can't be implemented because some additional complications arise due to the network structure. The major implication concerns estimating the transition probabilities of the vector-population process, defined as

$$t_i(\mathbf{m}'\,|\,\mathbf{m}) = P(\mathbf{M}(k+1)=\mathbf{m}'\,|\,\mathbf{M}(k)=\mathbf{m}, X=i), \quad i=1,\ldots,d,\ k=1,2,\ldots, \tag{3}$$

where $X$ denotes the background process which can be in one of the states $\{1,\ldots,d\}$. To compute these probabilities for an arbitrary network all possible transitions between populations need to be taken into account. In single population processes this difficulty is absent, and in small-scale networks an exact analysis is possible. But as the size of the network grows, new techniques are needed In order to deal with the increasing complexity.

Birgit relied on the *saddlepoint method* to develop an accurate approximation for the probabilities $t_i(\mathbf{m}'\,|\,\mathbf{m}), i=1,\ldots,d$. Shortly stated, the saddlepoint method approximates a random variable's probability mass function through its moment generating function. When analyzing Markovian populations interacting on a network it is usually possible to determine such moment generating functions, which makes the saddlepoint method suitable to approximate the desired transition probabilities in (3).

Using a suitably defined likelihood function, and the approximation obtained from the saddlepoint method, estimates for the desired parameters can be obtained. The likelihood function used in the network-based model is the typical likelihood function with respect to the observations of the vector-population process. This is a major difference when compared to the likelihood function considered in the single population in continuous time, defined in (1). When using the EM algorithm, a likelihood function is used that is based on both observed and unobserved data. And the likelihood function is maximized iteratively. In the network-based model it is not possible to use the EM algorithm due to the complexity of the model. Hence the likelihood function is based only on the observed data, and maximized numerically to obtain the parameter estimates.

**RNA transcription**

An application that Birgit worked on concerns modelling the population of RNA molecules in single cells. The synthesis (or birth) of a single RNA molecule, which is called transcription, is a stochastic process regulated by an on/off mechanism and follows a sequential process consisting of multiple steps. These steps are the following:

- The molecule RNA polymerase binds to the DNA and slides along the DNA to find a transcription start site, called promoter.
- Once it has found a start site it binds firmly and the transcription begins.
- The RNA polymerase moves along the gene while copying the genetical code step by step.
- Once it reaches the stop site, it releases itself and the new RNA transcript from the DNA.

This process can be repeated to produce more RNA molecules. The RNA transcription can be controlled by a process called gene repression. The promoter can bind to repressors for a period of time in which RNA polymerase cannot reach the start site to initiate transcription. This causes the promoter to switch between an active state, free from repressors, and an inactive state, bound by repressors (which is the background on/off mechanism in the mathematical model). Using a class of models called quasi-random birth-death processes, which are similar to, but more general models than the MMIS described above, Birgit developed and analyzed a model for the dynamics of a population of RNA molecules in single living cells.

**The more personal aspect**

Before we conclude this article we would like to give the word to Birgit.

*Birgit, would you like to share some memories with us?*

"I have many nice memories from my PhD years. I enjoyed all the collaborations, and meeting so many nice people. I was lucky to be part of the NETWORKS program, where I have met a lot of fellow-researchers. I have nice memories of the various NETWORKS events, where I learned about interesting topics, but were also full of fun. We went for nice walks and played many boardgames together. I remember one training week in a conference center with bowling lanes. We played every night, which even caused me a sore wrist for a few months."

*Were you also involved in some other activities you would like to share with the readers?*

"I have been a board member of European Women in Mathematics – The Netherlands (EWM-NL). It was my first experience as a board member and I learned a lot from it. I co-organized for example a workshop on Work-Life balance, and I was coordinating the EWM-NL mentor network. Through this network, female mathematicians can sign up to be matched with a mentor for advice on things related to their career."

**Concluding**

Birgit's research focused on population processes where the parameters depend on a background process. She developed methods, relying on the Expectation-Maximization algorithm and the saddlepoint method, in order to estimate the parameters of both the population and the background process. Finally, Birgit developed and analyzed a model for RNA transcription using quasi birth-death processes.

Since April Birgit is working as a postdoctoral researcher at the Department of Epidemiology and Data Science of the Amsterdam UMC. Her current research focuses on HPV (human papillomavirus), which is a sexually transmitted virus that causes various diseases, most prominently cervical cancer. Birgit works on HPV transmission models and cost-effectiveness assessments of HPV vaccination strategies. We wish Birgit all the best with her further research.