Hierarchical Bayesian Inference of Stochastic Biochemical Processes

by
Mark Jayson V. Cortez

A dissertation submitted to the Department of Mathematics,

College of Natural Sciences and Mathematics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Mathematics

Chair of Committee: Krešimir Josić

Committee Member: William Ott

Committee Member: Robert Azencott

Committee Member: Jae Kyoung Kim

University of Houston
May 2022

Sections 3 and 4 of the present work contain text and images appearing in the following publication: Cortez, M., Hong, H., Choi, B., Kim, J., and Josić, K. Hierarchical Bayesian models of transcriptional and translational regulation processes with delays. Bioinformatics 38(1) (2022), 187–195.

DEDICATION/EPIGRAPH

This work is dedicated to the memory of Cecilia Cortez, who taught me the value of learning and dreaming.

ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the constant supervision of Dr. Krešimir Josić, whose invaluable insights allowed me to advance this work to the end. Special thanks to Dr. Jae Kyoung Kim, Dr. Boseung Choi, and Mr. Hyukpyo Hong whose comments and suggestions led to the betterment of this research. Finally, I would like to express my appreciation to my friends and family who gave both support and inspiration throughout my Ph.D. journey.

ABSTRACT

Data from population measurements of gene network dynamics have shown that cells exhibit variability even in clonal lines. A reliable mathematical reconstruction of a biological process requires the inference of parameters characterizing this process in a single cell while considering the observed heterogeneity of the population from which data was obtained. Parameter inference, however, is complicated by the fact the outcomes of constituent reactions in a gene circuit are only partially observed in time or are detected indirectly in experiments. One approach is to replace unobserved reactions with time delays, a technique that also simplifies inference through the reduction of model dimension. This simplification, however, results in a non-Markovian model that requires the development of new inference methods. Here, we propose a hierarchical Bayesian inference framework for quantifying the variability of cellular processes within and across cells in a population in a non-Markovian setting, such as a reaction system with delays. We demonstrate our framework using a delayed birth-death process with birth delays which are either fixed or distributed, and show that a model with distributed delays is better when dealing with experimental systems since inference assuming fixed delays lead to underestimation when the true delays are variable. Using synthetic and experimental data, we show that the proposed hierarchical framework is robust and leads to improved estimates as compared to its non-hierarchical analog. We apply our method to data obtained using time-lapse microscopy and infer the parameters that describe the dynamics of fluorescent protein production at the individual cell and population level.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

No two biological systems are exactly alike. The same principle applies, in a fundamental level, to cells and cellular processes that are inherently variable, not only in time, but also across individuals in a population. While each cell in a population is unique, we expect typical cells to exhibit a range of predictable behaviors when presented with the same stimulus. Hence, a detailed understanding of cellular behaviors requires a characterization of the observed variability within and across cells, as well as the quantification of how features of cellular processes covary with phenotype and genotype. The estimation of such covariability often involves the analysis of time series of measurements from different cells across a clonal population [33, 47, 87]. These observations must be paired with an inference framework that can extract features from individual cell processes, and can explain the cell-to-cell variability of these features. In this study, we aim to develop such an inference framework: a mathematically principled approach to infer individual- and population-level properties from a collection of observations of a cell population.

Before we detail how we propose to do inference, we first look at appropriate representations of the type of processes that we will analyze in this study. Biochemical processes, such as those occurring in cells like protein production and gene regulation, are stochastic in nature [58]. The stochasticity in these processes can be attributed to extrinsic noise sources brought by the interaction of a cell with its environment [9], or to intrinsic noise of the chemical reactions which are particularly prominent at low reactant concentrations [39]. Some stochastic phenomena, such as noise–induced bistability [55], cannot be captured by deterministic chemical rate equation models, and appropriate representations of these processes must be used for analysis. One such approach is to model by a *biochemical reaction network* (BCRN) that is endowed by a kinetics describing the rate at which reactions occur.

A BCRN describes the evolution of u species Y_1, Y_2, ..., Y_u, in a volume or domain, through

a set of v chemical reactions R_1, R_2, ..., R_v. We represent this system as:

$$\begin{aligned} R_1 &: p_{11}Y_1 + p_{12}Y_2 + \ldots + p_{1u}Y_u \rightarrow q_{11}Y_1 + q_{12}Y_2 + \ldots + q_{1u}Y_u \\ R_2 &: p_{21}Y_1 + p_{22}Y_2 + \ldots + p_{2u}Y_u \rightarrow q_{21}Y_1 + q_{22}Y_2 + \ldots + q_{2u}Y_u \\ &\vdots \\ R_v &: p_{v1}Y_1 + p_{v2}Y_2 + \ldots + p_{vu}Y_u \rightarrow q_{v1}Y_1 + q_{v2}Y_2 + \ldots + q_{vu}Y_u \end{aligned} \tag{1}$$

where p_{kj} is the stoichiometric constant associated with reactant j in reaction k, and q_{kj} is the stoichiometric constant associated with product j in reaction k. Given a vector of molecular counts of all chemical species at time t, denoted $y\left(t\right) = \left(y_1\left(t\right), y_2\left(t\right), \ldots, y_u\left(t\right)\right)$, for each reaction R_k, there is a stochastic rate constant θ_k and function $h_k(y(t), \theta_k)$ that describes the instantaneous hazard of reaction R_k occurring under some kinetic law. Hence, in a sufficiently small time interval $(t, t + \Delta t]$, the probability of reaction k occurring is

$$P(\text{reaction } k \text{ occurs in } (t, t + \Delta t]) = h_k(y(t), \theta_k)\Delta t + o(\Delta t).$$

A common assumption about the system described by a BCRN is that the chemical species are well-mixed and equally likely to be found anywhere in the domain. This assumption leads to *mass-action* kinetics in which the rate of a chemical reaction is proportional to the concentrations of its substrates raised to the power of its stoichiometry [27, 85]. Specifically, the reaction hazard for reaction k takes the form

$$h_k\left(y\left(t\right), \theta_k\right) = \theta_k \prod_{j=1}^{u} \left[y_j\left(t\right)\right]^{p_{kj}}.$$

As an example, let us look at Michaelis-Menten kinetics [63] that describes a catalytic reaction which involves the chemical species E, S, ES, and P, respectively pertaining to the enzyme, substrate, enzyme-substrate complex, and the product. The following reaction system models the reversible binding of the enzyme E to the substrate S to form a complex ES, and the eventual

production of P that consumes S and releases E:

$$\begin{aligned} R_1: &\quad E + S \xrightarrow{\theta_1} ES \\ R_2: &\quad ES \xrightarrow{\theta_2} E + S \\ R_3: &\quad ES \xrightarrow{\theta_3} E + P. \end{aligned}$$

Here, the stochastic rate constants are θ_1, θ_2, and θ_3. Given the vector of molecular counts at time t, $y(t) = (E(t), S(t), ES(t), P(t))$, the resulting reaction hazards are

$$\begin{aligned} h_1\left(y(t), \theta_1\right) &= \theta_1 E(t) S(t) \\ h_2\left(y(t), \theta_2\right) &= \theta_2 ES(t) \\ h_3\left(y(t), \theta_3\right) &= \theta_3 ES(t). \end{aligned}$$

Since the probability of a reaction occurring is determined by the collection of reaction hazards $h_k\left(y(t), \theta_k\right)$, then it can be shown that at time t, the time to the next reaction is exponentially distributed with rate parameter

$$h_0\left(y\left(t\right), \theta\right) = \sum_{k=0}^{v} h_k\left(y\left(t\right), \theta_k\right),$$

and this reaction is of type k with probability $\dfrac{h_k\left(y(t), \theta_k\right)}{h_0\left(y(t), \theta\right)}$ [38, 85]. This observation can be used to develop an exact simulation method called the *stochastic simulation algorithm*[1] [27] that can be used to generate a sample path from the stochastic process defined by the BCRN. Hence, we can associate to a BCRN a Markov jump process in which each reaction occurs at a particular rate that depends only on the current state of the system. Other algorithms to simulate a time-evolution of the process include approximate methods like tau- and K-leaping [12, 28], and the solution of chemical Langevin equations [38].

[1] This is also called Gillespie algorithm.

Many cellular processes are composed of a sequence of substeps, which are often not of interest in themselves or are unobservable experimentally. In models, we can often replace such reaction chains by a single reaction, at the expense of introducing a delay [5, 30, 52]. For instance, consider a chain of mono-molecular reactions [5, 41, 84] composed of the chemical species $Y_1, Y_2, \ldots, Y_N$:

$$\begin{aligned} \emptyset &\xrightarrow{\theta_1} Y_1 \\ Y_1 &\xrightarrow{\theta_2} Y_2 \\ &\vdots \\ Y_{N-1} &\xrightarrow{\theta_{N-1}} Y_N \\ Y_N &\xrightarrow{\theta_N} \emptyset \end{aligned} \tag{2}$$

The entire process for the formation of Y_N can also be described more coarsely as a single reaction, which, once initiated, takes a random time to complete [57]. As a simplification of the reaction chain given in (2), we can consider a birth-death process in which the species $Y := Y_N$ is formed with delay:

$$\emptyset \xrightarrow[\tau]{A} Y \xrightarrow{B} \emptyset. \tag{3}$$

This delay birth-death process has production rate A, degradation rate B, and birth delay time τ that may be random or constant. For the modified reaction described by (3), the first $N-1$ reactions are lumped into a single birth reaction with delayed completion resulting in the removal of the intermediate species $\{Y_i\}_{i=1,\ldots,N-1}$ in the dynamics. This simplification, however, is done so that the dynamics of Y_N production is the same on both the processes (2) and (3). We will be seeing the delay birth-death reaction given in (3) repeatedly in this study as it is our primary model for analysis and is the model we will use to describe an experimental study of non-instantaneous production of a regulator protein that involves a sequence of reactions including transcription, translation, and post-translational steps [29, 46]. While the introduction of delay reduces the number of reaction and hence simplify the analysis of BCRNs, the resulting dynamics is no longer Markovian. The

time-evolution of the chemical species being dictated by the completion of a delayed reaction after a time τ after initiation implies that the system has memory and is non-Markovian.

To answer both practical and theoretical questions, stochastic BRNC models often require that the rates of constituent reactions and other parameters of interest be estimated from experimental data. This is one of the key features of the inference technique that we develop in this study: *an inference framework for BCRNs with delayed reactions using partial observations of a biological process as data*. A straightforward way to perform such estimation is by taking the deterministic analog of the stochastic system and use ordinary least squares or maximum likelihood approaches. Such approaches, however may not always work very well as the deterministic model may not capture some features arising from the stochasticity of the original system [76]. One principled method to perform estimation, which we employ in this study, is Bayesian inference which quantifies knowledge about the true parameter values based both on data and prior belief about the parameters [49, 84]. This is expressed mathematically through Bayes' Theorem,

$$\pi(\theta|Y_{obs}) = \frac{\pi(Y_{obs}|\theta)\ \pi(\theta)}{\pi(Y_{obs})}.$$

With model parameter θ and observations Y_{obs}, the factors in the equation above are described as follows:

- $\pi(\theta|Y_{obs})$ is the *posterior distribution* which quantifies what we know about the parameter, given the data and any prior information;
- $\pi(Y_{obs}|\theta)$ is the *likelihood*, i.e. the probability of observing the data Y_{obs} given the parameter θ; and
- $\pi(\theta)$ is the *prior distribution*, which represents our knowledge about the parameters before taking into account the observed data.

The denominator is the normalizing constant $\pi(Y_{obs}) = \int_{\theta} \pi\left(Y_{obs}\,|\theta\right)\pi\left(\theta\right)\ d\theta$.

The goal, in our case, therefore is to infer the posterior distributions of the parameters of the BCRN. Depending on the likelihood function and the choice of prior distribution, the posterior may not always be a standard density function or may be difficult to obtain analytically. In these cases, there are computational methods to characterize the posterior that allows for the computation of its conditional and marginal distributions, and their moments. Assuming that the likelihood is tractable and can be evaluated, a simple way to estimate the posterior distribution is through the use of sampling algorithms. One such approach is to use one of the Markov Chain Monte Carlo (MCMC) algorithms which are based on simulating an m-step Markov chain of samples, $\theta^{(0)}, \theta^{(1)} \ldots, \theta^{(m)}$, whose stationary distribution is the target posterior. A commonly used method for perform this task is based on the *Metropolis-Hastings algorithm* [34, 62]. At any step n, this iterative method relies on a proposal distribution, q, that is used to generate a new sample θ^* by perturbing the sample from a previous step $\theta^{(n)}$. Specifically, this method proceeds as follows:

1. Initialize $n = 0$ and select starting point $\theta^{(0)}$.
2. Generate a proposal sample, $\theta^* \sim q(\theta|\theta^{(n)})$.
3. Calculate the acceptance probability $\alpha = min\left(1, \frac{\pi(Y_{obs}|\theta^*)\pi(\theta^*)q(\theta^{(n)}|\theta^*)}{\pi(Y_{obs}|\theta^{(n)})\pi(\theta^{(n)})q(\theta^*|\theta^{(n)})}\right)$.
4. with probability α, set $\theta^{(n+1)} = \theta^*$, and with probability $1 - \alpha$, set $\theta^{(n+1)} = \theta^{(n)}$.
5. update step, $n = n + 1$.
6. if $n > m$, terminate simulation, otherwise go to step 2.

Whether the new sample θ^* is accepted depends on the likelihood ratio between the old value $\theta^{(n)}$ and the new proposed value θ^*, as generated by the proposal distribution q. It is hence critical to choose q so that a good amount of proposals are accepted while maintaining sufficiently fast convergence. If the variance of q is too low, then the acceptance rate is high but the parameter space may not be explored effectively thus resulting in poor convergence. On the other hand, if the variance of q is too high, proposal will be frequently rejected, again resulting in poor convergence and wastage of computational resources. A common proposal kernel for a continuous random

variable is the symmetric Gaussian kernel, $\mathcal{N}(\theta^{(n)}, \sigma)$, with mean $\theta^{(n)}$ and a user-defined standard deviation σ. Note that the acceptance probability, α, in step 3 of the Metropolis–Hastings algorithm simplifies to

$$\alpha = min\left(1, \frac{\pi(Y_{obs}|\theta^*)\pi(\theta^*)}{\pi(Y_{obs}|\theta^{(n)})\pi(\theta^{(n)})}\right)$$

when using a Gaussian proposal, since symmetry gives $q(\theta^{(n)}|\theta^*) = q(\theta^*|\theta^{(n)})$.

Aside from the choice of proposal distribution, there are other key issues in parameter estimation: size and resolution of data from experiments, overfitting, among others. Both of these relate to matters of parameter identifiability, which is the unique determination of parameter values from available data [23]. In this study, we address these issues and discuss checks for their detection and some possible solutions.

The Metropolis-Hastings method loses efficiency as the number of parameter increases, and a variant, called *Gibbs sampling* [26, 79], is oftentimes implemented instead when the number of parameters is large. Instead of sampling from the joint posterior of $\theta = \{\theta_1, \theta_2, \ldots, \theta_k, \ldots, \theta_u\}$, Gibbs sampling brings the sampling step to one dimension by picking one of the parameters, θ_k, fixing all other parameters $\theta_{k'}$ for all $k' \neq k$, and then sampling from the one-dimensional conditional posterior

$$\pi\left(\theta_k \left| \{\theta_{k'}\}_{k' \neq k}, Y_{obs}\right.\right).$$

This is implemented by the following algorithm:

1. Initialize $n = 0$ and select starting point $\theta^{(0)}$.
2. For each $k = 1, 2, \ldots, u$, sample θ_k^* from $\pi\left(\theta_k \left| \left\{\theta_{k'}^{(n)}\right\}_{k' \neq k}, Y_{obs}\right.\right)$ using the steps 2–4 of the Metropolis-Hastings algorithm. Every new sample, therefore, has the form $\theta^* = \{\theta_1^{(n)}, \theta_2^{(n)}, \ldots, \theta_k^*, \ldots, \theta_u^{(n)}\}$.
3. Update step, $n = n + 1$.
4. If $n > m$, terminate simulation, otherwise go to step 2.

Other alternative methods for sampling in high dimensions, including, but not limited to, Hamiltonian Monte Carlo [59, 67] and variational approaches [44, 81], are also widely used in practice, but are not discussed in this study.

In this study, we consider realizations of a BCRN with delayed reaction completion, collected from different cells in a population. Hence, the Bayesian inference framework and sampling techniques presented thus far need to be tailored to the inference of parameters of such a BCRN while also considering population variability in the observations. The development of a framework that is consistent with these requirements, and compatible with real experimental data requires us to consider the following:

- likelihood functions must incorporate the correlations brought by the introduction of delays in the reactions;
- appropriate approximations of the likelihood function must be used since observations of the biological system from experiments are incomplete and only available at discrete times; and
- inference must be done so that individual and population estimates are obtained simultaneously to ensure inference robustness.

The introduction of delays results in a non-Markovian process, making inference challenging. Several studies have addressed issues in the inference of such a process. Stochastic delay differential equations are often used to model processes with delays. The exact stationary probability densities of these equations have been used to identify the components of the systems under study [20, 21]. Progress has also been made by using artificial neural networks to approximate delay chemical master equations [42], an approach which works well for the inference of parameters of a birth-death process with a delayed death reaction. Likelihood-based inference using the chemical Langevin equation descriptions of the delayed process [36], and linear noise approximations [13] have also been used. These approaches, however, are effective only when molecule counts are high and stochastic differential equations accurately capture system dynamics [32]. Choi et al. [2020]

developed an alternative Bayesian approach using non-Markovian models to develop inference algorithms for rate and delay parameters in common biochemical reactions. This approach works well with synthetic and experimental data, and is effective even when molecule counts are low because the model is based on the chemical master equation. However, it relies on treating measurements from different cells as independent, identically distributed observations of a single cellular process, thus compounding uncertainty in parameter estimates with variability across the population of cells.

In order to characterize population variability, we take a *hierarchical Bayesian modelling* approach to the inference problem. This framework provides a systematic way to analyze population level data, and characterize both cell-to-cell and within cell variability [37, 78], as well as improves the robustness of the estimates of parameters that describe processes within individual cells, by assuming that these parameters follow an underlying population distribution [17, 25]. For instance, suppose we are observing a population of cells (Fig. 1), each producing a protein Y instantaneously through the birth reaction

$$\emptyset \xrightarrow{A_n} Y,$$

where every reaction is individually indexed by n, and A_n is the rate at which the protein Y is produced by cell n. Since the cells are members of the same clonal population, one expects them to produce Y at similar rates that can be described by some probability distribution $\pi(A\,|\omega_A)$ that is parameterized by ω_A. The individual production rates, A_n, can be viewed as independent samples from the same distribution $\pi(A\,|\omega_A)$. In this study, we will refer to individual-level characteristics, like A_n, as parameters, and population-level properties, like ω_A, as hyperparameters. As we will explain in what follows, the Bayesian inference using such a multi-level model no longer requires a specification of a prior for each A_n, as $\pi(A\,|\omega_A)$ already serves that role. What remains is to specify the *hyperprior* distribution $\pi(\omega_A)$ indicating the belief about the hyperparameter ω_A. We describe subsequently how the likelihood and posterior distribution can be formulated in a similar scenario, but with delayed reactions.

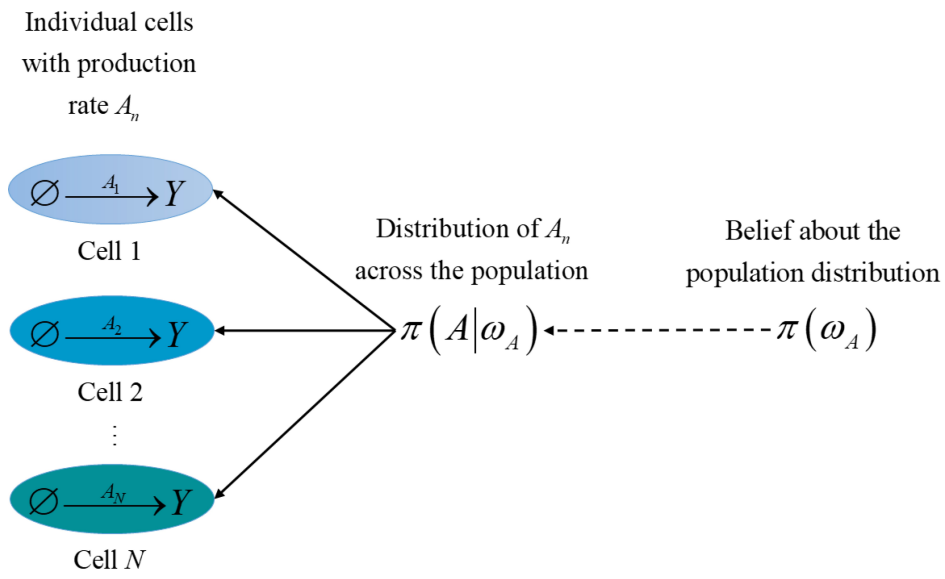

Figure 1: A hierarchical model where the production rate of a protein Y, A_n, is assumed to follow a distribution $\pi(A\,|\omega_A)$. A hyperprior distribution, $\pi(\omega_A)$, describes the belief about the distribution of A_n in the population.

We organize the presentation of our results as follows:

- In **Chapter 2**, we present the basic theoretical framework for the inference of a BCRN with delayed reactions, and derive the corresponding likelihood function for the process defined by this reaction network. We also discuss an approximation to this likelihood function that can accommodate discrete-time observations of the process under study.

- In **Chapter 3**, we develop a general hierarchical modelling algorithm for a BCRN with delayed reactions that can simultaneously estimate the posterior distribution of parameters characterizing processes within individual cells, as well as the distribution of these parameters across the population. We demonstrate the advantages and shortcomings of our approach using a delayed birth-death process, which, although it may not fully describe the underlying biophysical processes, captures the main effects of protein production, and can serve as a building block for more complex systems. Specifically, we consider two types of delay: fixed and distributed.

- In **Chapter 4** we apply our hierarchical model to experimentally-derived data of protein

production in *E. coli*. We also introduce checks and assessments for parameter estimates to rule out model overfitting.

- In **Chapter 5**, we summarize the findings of this study, and present ideas for future research direction.

2 Bayesian inference of a biochemical network

In this chapter we present the general theoretical framework for the inference of parameters defining the dynamics of biochemical reaction networks (BCRN). While our approach, utilizing Bayesian inference, has already been discussed in detail and implemented in many other applications, we find it practical to present here basic concepts like the likelihood function and posterior distributions when put in the context of BCRNs. We also discuss the work of Boys et al. [10] on inference using stochastic models and the extension of this work to models with delays made by Choi et al. [16], which both serve as the foundation for our hierarchical inference model.

2.1 Likelihood function of a stochastic BCRN

While it is almost straightforward to implement the sampling algorithms introduced in Chapter 1, the process is incomplete without the construction of a likelihood function which describes the process which generates the observations, Y_{obs}. We also need to specify the prior distribution for this data. We focus on the construction of the likelihood function for a stochastic BCRN.

Following the discussion of Boys et al. [10], consider the BCRN (1) describing the evolution of u species Y_1, Y_2, ..., Y_u through a set of v chemical reactions R_1, R_2, ..., R_v. Recall that a reaction k takes the form

$$R_k : p_{k1}Y_1 + p_{k2}Y_2 + \ldots + p_{ku}Y_u \rightarrow q_{k1}Y_1 + q_{k2}Y_2 + \ldots + q_{ku}Y_u,$$

with reactant and product stoichiometric constants p_{kj} and q_{kj}. Given a vector of molecular counts of all chemical species at time t, denoted $y(t) = (y_1(t), y_2(t), \ldots, y_u(t))$, for each reaction R_k, there is a stochastic rate constant θ_k and function $h_k(y(t), \theta_k)$ that describes the instantaneous hazard of reaction R_k occurring under some kinetic law.

First assume that the entire process is fully observable on the time interval $[0, T]$. We assume that time is rescaled so that the reactions are observed at unit intervals. We then partition a finite interval of length T into subintervals $(i, i+1]$, for $i = 0, 1, \ldots, T-1$. We assume that all

reactions are observed with a complete recording of the time of every reaction occurrence and the corresponding molecular counts for every chemical species. Denote by r_{ki} be the number of reactions of type k that are completed in the time subinterval $(i, i+1]$, and let $\rho_i = \sum_{k=1}^{v} r_{ki}$. Every reaction, j, occurring within the subinterval $(i, i+1]$ is associated to reaction time and type (t_{ij}, k_{ij}), for $j = 1, 2, \ldots, \rho_i$. The likelihood function for the complete molecule count trajectory $\mathbf{y} = \{y(t)\}_{t\in[0,T]}$ given the parameter $\theta = \{\theta_k\}_{k=1,\ldots,u}$, is then given by

$$L(\mathbf{y}\,|\theta) = \left[\prod_{i=0}^{T-1}\prod_{j=1}^{\rho_i} h_{k_{ij}}\left(y\left(t_{i,j-1}\right), \theta_{k_{ij}}\right)\right] \times \exp\left[-\int_0^T h_0\left(y\left(t\right), \theta\right) dt\right] \tag{4}$$

where

$$h_0\left(y\left(t\right), \theta\right) = \sum_{k=0}^{v} h_k\left(y\left(t\right), \theta_k\right).$$

This formulation follows from the result presented in Chapter 1, that at time t, the time to the next reaction is exponentially distributed with rate parameter $h_0\left(y\left(t\right), \theta\right)$, and that the reaction is of type k with probability $\dfrac{h_k\left(y(t), \theta_k\right)}{h_0\left(y(t), \theta\right)}$. If the reaction law is mass-action, that is,

$$h_k\left(y\left(t\right), \theta_k\right) = \theta_k \prod_{j=1}^{u} \left[y_j\left(t\right)\right]^{p_{kj}},$$

then the likelihood function can factorized so that assuming independent Gamma priors, $\Gamma(a_k, b_k)$ for each parameter θ_k results in a posterior that is a product of Gamma distributions [85]. In particular, the posterior for each parameter is given by

$$\theta_k|\mathbf{y} \sim \Gamma\left(a_k + r_k, b_k + \int_0^T \prod_{j=1}^{u} \left[y_j\left(t\right)\right]^{p_{kj}}\, dt\right),$$

where r_k is the number of completed reactions of type k throughout $[0, T]$. In such a case, we say that the Gamma prior is a *conjugate prior*[2] to the given likelihood.

[2]The term *conjugate prior* is often used to mean that the posterior is in the same probability distribution family as the prior probability distribution. In the succeeding chapters, we will use the term more loosely to mean that the prior leads to a posterior that is a known distribution, not necessarily belonging to the same family as the prior.

2.2 Inference for a delayed BCRN

Although Bayesian inference has been particularly promising in fitting stochastic models to data from single-cell assays, the complexity of a biological process may require that a large number of model parameters be estimated, which can make the inference process difficult: The likelihood function may be too complex to be tractable, or may be computationally demanding to evaluate. One remedy to this problem is the simplification of chemical reaction networks by replacing chains of chemical reactions with a delay distribution [5, 30, 52].

This simplification results in a chemical reaction that does not complete instantaneously. The introduction of delay accounts for the aggregation of reactions, which individually are not necessarily instantaneous and almost always take time to complete. For context, take the production of a certain protein which commences at transcription, but can only be considered as a mature functional protein once translation and post-translational modifications have been completed. To introduce delay in a chemical reaction, we suppose that once it is initiated at time t_{initial}, it takes a fixed or random time to complete. If the completion time is t_{final}, then the delay in reaction completion is given by $t_{\text{initial}} - t_{\text{final}}$, which, in the context of protein synthesis, is the total time between the initiation of transcription and protein maturation. As we have pointed out in Chapter 1, the introduction of delays to a chemical reaction makes the system non-Markovian, thereby warranting a careful treatment of the inference process.

2.2.1 Likelihood function of a reaction with delay

As a delayed system is non-Markovian, the likelihood function presented in Section 2.1 no longer applies in the present case. Here, we develop a likelihood analogous to (4) that takes into account the correlation between observations of a system due to the introduction of delays. We follow the outline of steps as presented by Choi et al. [16]. To each reaction R_k with delayed completion, we associate a delay measure, η_k, with support on $[0, \infty)$. Suppose that the delay distribution is independent of the time or the state of the system, and that it only depends on a vector of parameters $\Delta_k = (\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kl_k})$. Schlicht and Winkler [72] have proven the existence of

reaction completion propensities defined by

$$f_k\left(t, \mathbf{y}, \theta_k, \Delta_k\right) = \int_0^t h_k\left(y\left(t-s\right), \theta_k\right) d\eta_k\left(s\right).$$

These completion propensities define the effective rate of reaction k at a given time t. We can now analogously develop the likelihood function as in the case with no delay. Integrating with time, we have

$$\Lambda_k\left(t, \mathbf{y}, \theta_k, \Delta_k\right) = \int_0^t f_k\left(\tau, \mathbf{y}, \theta_k, \Delta_k\right) d\tau,$$

$$\Lambda_0\left(t, \mathbf{y}, \theta, \Delta\right) = \sum_{k=0}^{v} \Lambda_k\left(t, \mathbf{y}, \theta_k, \Delta_k\right),$$

where $\Delta = \{\Delta_k\}$ is the collection of all delay parameter vectors, and $\theta = \{\theta_k\}$ is the set of all reaction constants. If complete knowledge of a process $\mathbf{y}$ is available for every $t \in [0, T]$, then the likelihood function for the parameters θ is given by

$$L\left(\mathbf{y}\left|\theta\right., \Delta\right) = \left[\prod_{i=0}^{T-1} \prod_{j=1}^{\rho_i} f_{k_{ij}}\left(t_{ij}, \mathbf{y}, \theta_{k_{ij}}, \Delta_{k_{ij}}\right)\right] \times \exp\left[-\Lambda_0\left(T, \mathbf{y}, \theta, \Delta\right)\right]. \tag{5}$$

While the likelihood in Eq. (5) is very similar to the likelihood for delay-free systems in Eq. (4) derived in Section 2.1, the former is obtained by adopting a *backward view* of delayed chemical kinetics, assuming that only the reaction completion times are known, and that the corresponding unobserved reaction initiation times which occurred in the past are random quantities [16]. This contrasts the *forward view* in delay-free systems, where only the current reaction times are known and the time to the next reaction is a random quantity, dictated by the current state of the system.

The evaluation of the likelihood given in Eq. (5) relies on the experimentally infeasible assumption that the entire process $\mathbf{y}$ is known. In practice, experimental data is usually recorded in equally-spaced intervals thereby giving rise to discretely-observed recordings. Suppose that the process $\mathbf{y}$ is observed only at the discrete time points, $t = 0, 1, \ldots, T-1, T$, and denote by $\mathbf{y}_d$ the subset of observations $\{y(0), y(1), \ldots, y(T-1), y(T)\}$. If only the discrete time observations $\mathbf{y}_d$ are

available, an approximation of the completion propensity f_k will have to be derived. In order to do so, we will take an approximation $\hat{f}_k$, which is a propensity that is constant between observations, obtained by averaging f_k over each unit time interval $[i, i+1]$. We thus obtain

$$\begin{aligned}\hat{f}_k(i, \mathbf{y}_d, \theta_k, \Delta_k) &= \int_i^{i+1} \int_0^t h_k\left(t-s\right) d\eta_k\left(s\right) dt \\ &= \sum_{n=0}^{i} \int_i^{i+1} \int_{t-(n+1)}^{t-n} h_k\left(t-s\right) d\eta_k\left(s\right) dt \\ &= \sum_{m=0}^{i} \int_m^{m+1} \int_{\hat{t}-1}^{\hat{t}} h_k\left(\hat{t}+(i-m)-s\right) d\eta_k\left(s\right) d\hat{t}\end{aligned}$$

by substitutions $\hat{t} = t - n$ and $m = i - n$. The reader is directed to the Supplementary Material of [16] for a detailed derivation. The last step specifies how the reaction hazard h_k is to be evaluated between observations. Since h_k is to be evaluated from $1 + (i - m)$ to $i - m$ as s ranges from $\hat{t} - 1$ to $\hat{t}$, linear interpolation yields the final approximate completion propensity

$$\hat{f}_k(i, \mathbf{y}_d, \theta_k, \Delta_k) = \sum_{m=0}^{i} \int_m^{m+1} \int_{\hat{t}-1}^{\hat{t}} \left[(s+1-\hat{t})h_k(i-m) + (\hat{t}-s)h_k(i-m+1)\right] d\eta_k(s)dt. \tag{6}$$

Conditioned on the entire system history up to time i, we see that the number reactions of type k which completed in $(i, i+1]$, r_{ki}, is Poisson-distributed with mean equal to $\hat{f}_k(i, \mathbf{y}_d, \theta_k, \Delta_k)$. As such, using the approximate completion propensity in Eq. (6), as is also seen in [32], the likelihood given by Eq. (5) can finally be approximated as

$$\hat{L}\left(\mathbf{y}_d \,|\theta\,, \Delta\right) = \left[\prod_{i=0}^{T-1} \prod_{k=1}^{v} \frac{\hat{f}_k(i, \mathbf{y}_d, \theta_k, \Delta_k)^{r_{ki}}}{r_{ki}!}\right] \times \exp\left(-\hat{\Lambda}_0\left(T, \mathbf{y}, \theta, \Delta\right)\right), \tag{7}$$

where

$$\hat{\Lambda}_0\left(T, \mathbf{y}, \theta, \Delta\right) = \sum_{k=1}^{v} \sum_{i=0}^{T-1} \hat{f}_k(i, \mathbf{y}_d, \theta_k, \Delta_k),$$

and r_{ki} is the number of reactions of type k that was completed in the time interval $(i, i+1]$.

2.2.2 Likelihood for pooled observations

Molecule count trajectories from experimentally-derived data usually come from the observation of a cell population. When cell-to-cell variability can be ignored, the cells can be assumed to be identical with the reaction rates, and other parameters equal across the population. As a consequence, the trajectories can be treated as independent realizations of the same stochastic process thereby limiting the sources of differences in the observations to measurement and intrinsic noise. We refer to this strategy as *data pooling*[3] [2, 16].

Consider the case of discrete-time observations $\mathbf{y}_d$ introduced in Subsection 2.2.1. As the observations are assumed to be independent, the likelihood function for a delayed system can be obtained by multiplying the likelihood, Eq. (7), for a single trajectory to obtain:

$$\begin{aligned}\hat{L}_p\left(\mathbf{y}_d\,|\theta\,,\Delta\right) &= \prod_{n=1}^{N}\hat{L}\left(\mathbf{y}_{d,n}\,|\theta\,,\Delta\right)\\ &= \prod_{n=1}^{N}\left\{\left[\prod_{i=0}^{T-1}\prod_{k=1}^{v}\frac{\hat{f}_k(i,\mathbf{y}_{d,n},\theta_k,\Delta_k)^{r_{nki}}}{r_{nki}!}\right]\times\exp\left(-\hat{\Lambda}_0\left(T,\mathbf{y}_n,\theta,\Delta\right)\right)\right\}. \qquad (8)\end{aligned}$$

Here, n is the index of an individual trajectory obtained from observing a population of N cells. For a fully observed system, an analogous likelihood function can be formulated by multiplying the likelihoods of the form given in Eq (5).

This approach to dealing with population data is appropriate when the trajectories are indeed observations of the same or nearly identical cells: parameters are more accurately estimated, and posterior distributions are narrow. On the other hand, when we observe a heterogeneous group of cells, population information, such as cell-to-cell variability cannot be captured using this approach. Moreover, it is not always clear how the estimate obtained using data pooling is related to the collection of parameters that define the dynamics of individual cells [2]. Therefore, inferring population-level properties requires a different treatment, one which we develop in this study.

[3]These are also referred to as population-averaged assays in the literature.

2.2.3 The block-updating method

The approximate likelihood given in Eq. (7) remedies the unavailability of complete system information, but requires data that is not readily obtainable from partial molecule count trajectories when there are multiple reactions. This likelihood replaces the unobserved process in the time interval $(i, i+1]$ with the specification of the number of completed reactions that may result in the observations at times i and $i+1$. As the number of reactions of type k that completed in $(i, i+1]$, r_{ki}, in the likelihood given by Eq. (7), are not observed directly, we will infer them by following the block-updating method introduced by Boys et al. [10] in 2008 for a Lotka-Volterra system[4]. Instead of describing the method in general, we will focus on its application to a delayed birth-death process

$$\emptyset \xrightarrow[\pi(\Delta_1)]{A} Y \xrightarrow[\pi(\Delta_2)]{B} \emptyset, \tag{9}$$

which is a reaction of interest for the succeeding illustrations and applications[5]. Here A and B are birth and death reaction constants respectively, while $\pi(\Delta_1)$ and $\pi(\Delta_2)$ are distributions describing the corresponding reaction delays.

A simplification[6] of the block-updating method for the birth-death process (9) uses a random walk proposal on the number of birth reactions. It is implemented using a Metropolis-Hastings algorithm with a random walk chain to sample the number of completions of each type of reaction in a time interval $(i, i+1]$, $i = 0, 1, ...T-1$, given the observed system states $y(i)$ and $y(i+1)$. The birth reaction is indexed with $k = 1$ while the death reaction is indexed with $k = 2$. For the i^{th} interval, the joint conditional posterior of r_{1i} and r_{2i} is given by

$$\pi\left(r_{1i}, r_{2i} \,|\mathbf{y}_d\,, A, B, \Delta\right) \propto \frac{\hat{f}_1(i, \mathbf{y}_d, A, \Delta_1)^{r_{1i}}}{r_{1i}!} \cdot \frac{\hat{f}_2(i, \mathbf{y}_d, B, \Delta_2)^{r_{2i}}}{r_{2i}!} \tag{10}$$

[4]Boys et al., developed this for a three-species stochastic kinetic system.

[5]Refer to Chapter 3.

[6]The simplification used in [16] departs from the original work of Boys et al. by dropping the Radon-Nikodym derivative of the true process in the acceptance probability, as well as the replacement of the simulation of an inhomogeneous Poisson process between observations with just the sampling of the reaction numbers r_{ki}.

which is proportional to the product of the density functions of Poisson random variables. In view of the complete sampling algorithm for the parameters of a BCRN, the updating of reaction numbers, r_{ki}, at an iteration j, is done after the rate and delay parameters sampling step is completed. Hence, the current parameter samples $A := A^{(j)}$, $B := B^{(j)}$, $\Delta_1 := \Delta_1^{(j)}$, and $\Delta_2 := \Delta_2^{(j)}$, alongside $\mathbf{y}_d$, are used as data for the posterior in Eq. (10).

Denote by $r_{1i}^{(j-1)}$ a sample of r_{1i} from the $(j-1)^{\text{th}}$ iteration. The proposal distribution can be chosen to be a discrete random walk in which the current value is augmented by u, whose distribution is the difference of two Poisson random variables with means that are both equal to some λ which is usually a function of $r_{1i}^{(j-1)}$. For instance, Boys et al. [10], reported that the form

$$\lambda = 1 + \frac{r_{1i}^2}{b},$$

for some tuning parameter b, induces good chain mixing. The value u is then used to define the proposed value $r_{1i}^* = r_{1i}^{(j-1)} + u$. In particular, the distribution of the update value u is a Skellam distribution [10, 43] given by

$$p\left(u \left| {r_{1i}}^{(j-1)} \right.\right) = \exp\left(-2{r_{1i}}^{(j-1)}\right) I_u\left(2{r_{1i}}^{(j-1)}\right),$$

where I_u is a regular modified Bessel function of order u. Once r_{1i}^* is chosen, then r_{2i}^* can be uniquely determined using $y(i+1) - y(i) = r_{1i}^* - r_{n2i}^*$. The proposed updates r_{1i}^* and r_{n2i}^* are then accepted with probability α_r, defined by

$$\alpha_r = \min\left\{1, \frac{p\left(u \left| {r_{1i}}^* \right.\right) \frac{\hat{f}_1(i,\mathbf{y}_d,A,\Delta_1)^{{r_{1i}}^*}}{{r_{1i}}^*!} \cdot \frac{\hat{f}_2(i,\mathbf{y}_d,B,\Delta_2)^{{r_{2i}}^*}}{{r_{2i}}^*!}}{p\left(u \left| {r_{1i}}^{(j-1)} \right.\right) \frac{\hat{f}_1(i,\mathbf{y}_d,A,\Delta_1)^{{r_{1i}}^{(j-1)}}}{{r_{1i}}^{(j-1)}!} \cdot \frac{\hat{f}_2(i,\mathbf{y}_d,B,\Delta_2)^{{r_{2i}}^{(j-1)}}}{{r_{2i}}^{(j-1)}!}}\right\},$$

that implicitly assumes that flat priors are given to both the birth and death reaction numbers.

For more general systems, a similar method can be employed by sampling one of the reaction numbers and determining the other number of completed reactions by matching the population sizes at the end of the observation intervals.

2.3 Summary

For a fully observed non-delayed chemical reaction system that is endowed with mass action kinetics, the posterior distribution is given by a gamma distribution when the independent priors for the rate constants are also gamma distributed. This situation, where every reaction is tracked entirely, is idealized as experimental data and has a different structure. The problem of having only discretely-observed data, paired with a complicated reaction network structure forms part of the complexities of inference of BCRNs.

Many cellular process are composed of multiple chemical reactions, and replacing a sequence of chemical reactions with a delay distribution is one approach to simplify such complex reaction networks. The introduction of delay makes the associated stochastic process non-Markovian thereby making the inference process more challenging. Building on the work of Boys et al. [10] and Choi et al. [16] that address the difficulty of parameter inference in discretely-observed systems, approximations of the reaction hazards and the resulting completion propensities can be used to form an approximate likelihood function. This approximate likelihood is dependent on the number of completed reactions between observations, which are not directly observed, but can be inferred using a simplified block-updating method based on a Metropolis-Hastings method with a random walk chain.

3 Hierarchical Models of a Stochastic BCRN

The results described thus far center on inference from observations of a single individual. When put in the context of cell recordings, a molecule count trajectory, as discussed in Sections 2.1 and 2.2, can be thought of as a recording of protein counts that dynamically changes in the process of protein synthesis and degradation. As is common in experimental systems, a collection of recordings from a cell population may exhibit significant differences which may account for both intrinsic and extrinsic noise sources.

Recognizing that a cell population is heterogeneous suggests that we need to estimate the distribution of parameters of interest across an entire recorded population of cells. This problem was avoided by Choi et al. [16] by assuming that all cells in the population are identical and that observations are realizations of the same stochastic process, and thus reaction rates are functions of the same unknown parameters. This is equivalent to the assumption that only intrinsic variability in the protein production accounts for the observed differences in the trajectories. In contrast, in this chapter, we propose a hierarchical approach to infer the distribution characterizing the parameters of a BCRN within individual cells, as well as the distribution of these parameters across the entire population.

As a system with no delay[7] is a special case of a delayed BCRN where the delay distribution is a Dirac delta function centered at 0, we will no longer talk about delay-free systems. The following discussion and derivations are done for systems with delay.

A note on notation: For the parameters A and B and τ, a subscript, n, will refer to the parameter of an individual cell n in the population. The same symbol without a subscript will refer to the collection of parameters across the population.

3.1 General inference process for a heterogeneous cell population

Suppose that we are observing a population of N cells, each one with a distinct sequence of observations $\mathbf{y}_n$ on the time interval $[0, T]$. Indexing an individual cell by n, we have a collection of

[7] In our nomenclature, this falls under the case of a system with fixed delays.

trajectories $\{\mathbf{y}_n\}_{n=1,\ldots,N}$. While the parameter values that characterize every reaction between cells may differ, our observations are assumed to be the product of the same set of reactions in every cell. Hence, every cell is defined by the same number of reactions, with the exact same number and type of parameters which may vary between the individual cells in the population.

All reactions R_k, for $k = 1, \ldots, v$, are present in each cell n, with each reaction R_k endowed with a rate constant θ_{nk}. Hence the reaction rates in a cell n are fully characterized by the rate vector $\theta_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nk}, \ldots, \theta_{nv})$. If a reaction R_k in cell n has delayed completion, then we also couple with that reaction the set of delay parameters $\Delta_{nk} = (\Delta_{nk1}, \Delta_{nk2}, \ldots, \Delta_{nkl_k})$, and denote the set of all delay parameters for cell n by $\Delta_n = \{\Delta_{nk}\}$. We denote by $\boldsymbol{\theta}$ the collection $\{\theta_n\}$ of all rate constants and by $\boldsymbol{\Delta}$ the collection $\{\Delta_n\}$ of parameters that define all delay measures.

For instance, consider the birth-death process with birth delays

$$\emptyset \xrightarrow[\pi(\Delta_{n1})]{A_n} Y \xrightarrow{B_n} \emptyset,$$

where $k = 1$ corresponds to the birth reaction and $k = 2$ to the death reaction. When the delay distribution $\pi(\Delta_{n1})$ is a Gamma distribution, we have $\Delta_{n1} = \{\alpha_{n1}, \beta_{n1}\}$, which are the shape and rate parameters for the distribution. Since the death reaction is instantaneous, $\Delta_{n2} = \emptyset$. Hence, for each cell n, we have the vector of rate and delay parameters as $\theta_n = \{A_n, B_n\}$ and $\Delta_n = \{\alpha_{n1}, \beta_{n1}\}$, respectively. This is much simpler for a fixed delay distribution where $\Delta_{n1} = \{\tau_{n1}\}$, for some non-negative constant τ_{n1}.

Suppose that the process $\mathbf{y}_n$ is fully observed over the interval $[0, T]$. If we assume that all individual parameters are identically distributed and independently sampled from their respective population distributions, then the individual realizations $\mathbf{y}_n$ are also independent and we can define the total likelihood

$$\mathcal{L}\left(\mathbf{y} \,|\boldsymbol{\theta}\,, \boldsymbol{\Delta}\right) = \prod_{n=1}^{N} L\left(\mathbf{y}_n \,|\theta_n\,, \Delta_n\right), \tag{11}$$

which is the product of likelihoods given in Eq. (5)[8] for all N individuals. If only the discrete-time observations $\mathbf{y}_{d,n} = (y_n(0), y_n(1), \ldots, y_n(T-1), y_n(T))$ are available, we can approximate the total likelihood by

$$\hat{\mathcal{L}}(\mathbf{y}_d | \boldsymbol{\theta}, \boldsymbol{\Delta}) = \prod_{n=1}^{N} \hat{L}(\mathbf{y}_{d,n} | \theta_n, \Delta_n), \tag{12}$$

which is a product of likelihoods given in Eq. (7)[9]. The computation of this approximate likelihood, $\hat{\mathcal{L}}$, requires the number of completed reaction of type k on each interval $(i, i+1]$ for each cell n, which we write as r_{nki}. As before, these values are not observed directly, and hence are inferred through the block-updating method discussed in Subsection 2.2.3.

3.1.1 A hierarchical Bayesian model of a cell population

The discrete-time observations $\mathbf{y}_{d,n}$ are a product of the initiation and eventual completion of the chemical reactions $\{R_k\}_{k=1,\ldots,v}$. The observations $\mathbf{y}_{d,n}$ therefore are dependent on the individual-level parameters θ_{nk} and Δ_{nkl}, which we assume follow underlying distributions which are themselves characterized by hyperparameters, ω_{θ_k} and $\omega_{\Delta_{kl}}$, respectively. We also assume that these individual parameters are identically distributed and independent for the given hyperparameters, as we are observing cells that are not closely related.

Bayes' Theorem paired with the assumption of independence among observations allows for the factorization of the joint posterior of parameters and hyperparameters, and hence enables us to take a multilevel approach to inference. Denote the collection $\{\omega_{\theta_k}\}$ and $\{\omega_{\Delta_{kl}}\}$ of rate and delay hyperparameters respectively as ω_θ and ω_Δ. The approximate likelihood expression given in Eq. (12), and Bayes' Theorem allow us to write the posterior over the rate and delay parameters characterizing the biochemical reaction network, to reflect the sequence of observation-parameter

[8] This is the exact likelihood in Subsection 2.2.1

[9] This is the approximate likelihood in Subsection 2.2.1

and parameter-hyperparameter dependencies as

$$\begin{aligned}\pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta}, \omega_\theta, \omega_\Delta \,|\mathbf{y}_d\right) &\propto \pi\left(\mathbf{y}_d \,|\boldsymbol{\theta}, \boldsymbol{\Delta}, \omega_\theta, \omega_\Delta\right) \pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta}, \omega_\theta, \omega_\Delta\right) \\ &= \hat{\mathcal{L}}\left(\mathbf{y}_d \,|\boldsymbol{\theta}, \boldsymbol{\Delta}\right) \pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta} \,|\omega_\theta, \omega_\Delta\right) \pi\left(\omega_\theta, \omega_\Delta\right) \\ &= \hat{\mathcal{L}}\left(\mathbf{y}_d \,|\boldsymbol{\theta}, \boldsymbol{\Delta}\right) \pi\left(\boldsymbol{\theta} \,|\omega_\theta\right) \pi\left(\boldsymbol{\Delta} \,|\omega_\Delta\right) \pi\left(\omega_\theta\right) \pi\left(\omega_\Delta\right), \end{aligned} \tag{13}$$

with

$$\pi\left(\boldsymbol{\theta} \,|\omega_\theta\right) := \prod_{n=1}^{N} \prod_{k=1}^{v} \pi\left(\theta_{nk} \,|\omega_{\theta_k}\right)$$

and

$$\pi\left(\boldsymbol{\Delta} \,|\omega_\Delta\right) := \prod_{n=1}^{N} \prod_{k=1}^{v} \prod_{l=1}^{l_k} \pi\left(\Delta_{nkl} \,|\omega_{\Delta_{kl}}\right)$$

serving as priors for individual rate and delay parameters, and $\pi\left(\omega_\theta\right)$ and $\pi\left(\omega_\Delta\right)$ as the hyperpriors. The factorization of $\pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta} \,|\omega_\theta, \omega_\Delta\right) \pi\left(\omega_\theta, \omega_\Delta\right)$ into $\pi\left(\boldsymbol{\theta} \,|\omega_\theta\right) \pi\left(\boldsymbol{\Delta} \,|\omega_\Delta\right) \pi\left(\omega_\theta\right) \pi\left(\omega_\Delta\right)$ in the last line of (13) comes from the assumption of independence between the priors and the hyperpriors. The factorization of the posterior distribution is the basis of the hierarchical inference algorithm that we present in the next section.

Refer to Fig. 2 for clarity. Let us revisit the birth-death process

$$\emptyset \xrightarrow[\delta(\tau_n)]{A_n} Y \xrightarrow{B_n} \emptyset,$$

now with a single-parameter Dirac delta function as delay distribution, and assume that parameters all are gamma distributed at the population level. An individual n, has rate parameters $\theta_{n,1} = A_n$ and $\theta_{n,2} = B_n$, and delay parameter $\Delta_{n,1,1} = \tau_n$. With a Dirac delta function as delay distribution, the parameter τ_n serves as a fixed time delay between the initiation of the birth reaction and its eventual completion. The n^{th} observation, which is a trajectory of molecule count, is a realization of the above stochastic process that is parameterized by $\theta_n = \{A_n, B_n\}$ and $\Delta_n = \{\tau_n\}$. Each of the parameters (either rate or delay), is assumed to follow a gamma distribution $\Gamma(a_Z, b_Z)$, for

$Z \in \{A, B, \tau\}$. Therefore, the hyperparameters are $\omega_{\theta_1} = \omega_A = \{a_A, b_A\}$, $\omega_{\theta_2} = \omega_B = \{a_B, b_B\}$, and $\omega_{\Delta_{1,1}} = \omega_\tau = \{a_\tau, b_\tau\}$.

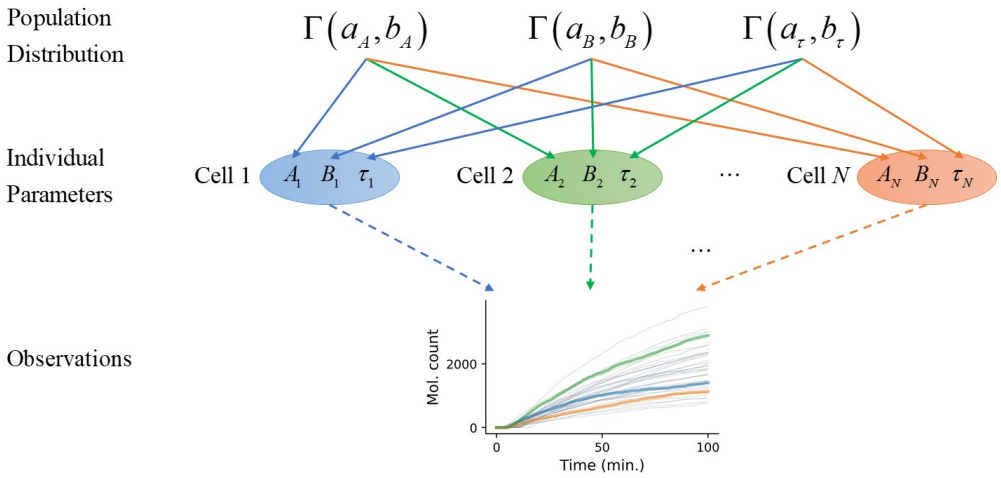

Figure 2: Conceptual model of the observation, parameter and hyperparameter dependencies. Every observation is a realization of an individual birth-death process, that is described by the production rate, A_n, the degradation rate, B_n, and the delay time distribution, $\delta(\tau_n)$. At the population level, we assume that these parameters follow gamma distributions each with corresponding parameter pairs (a_Z, b_Z) for $Z = A, B, \tau$.

3.1.2 General hierarchical model algorithm

We next describe an MCMC algorithm to generate samples from the posterior distribution of the model parameters $(\boldsymbol{\theta}, \boldsymbol{\Delta})$ and corresponding hyperparameters $(\omega_\theta, \omega_\Delta)$. The priors and hyperpriors capture our previous knowledge about the variability of the parameters across the population. As rate parameters are positive, we use gamma distributions as priors, $\pi(\theta_{nk} | \omega_{\theta_k})$. Thus for every reaction k, the set of hyperparameters for the corresponding reaction rate is $\omega_{\theta_k} = (a_{\theta_k}, b_{\theta_k})$, where a_{θ_k} and b_{θ_k} are the shape and rate parameters respectively of a gamma distribution. If the reaction propensity is separable, as in the case of mass-action kinetics where the hazard function can be factored as $h_k(y_n(t), \theta_{nk}) = \theta_{nk} g_k(y_n(t))$, the gamma distribution defines a conjugate prior for the parameters θ_{nk} [85].

As is typical of hierarchical sampling approaches, our algorithm iteratively produces samples

of individual parameters and integrates the result across an ensemble of cells to produce a sample of the hyperparameters that characterize the population distribution. The updated population distribution is then used to generate new samples of individual cell parameters, and the process repeats. To sample from the posterior distribution given by Eq. (13), we use Gibbs sampling: For every individual cell, n, we obtain samples for θ_n and Δ_n from their conditional posterior distributions by using the Metropolis-Hastings algorithm [34, 62]. As described by Choi et al. [16], knowledge about the number of completed reactions of type k in the time interval $(i, i+1]$ for individual n, r_{nki}, is needed in the sampling of these individual parameters. Since the discrete-time measurements do not uniquely determine the number of reactions, we infer r_{nki} in each Gibbs step. To do so, we follow the simplified *block-updating strategy* [10] discussed in Subsection 2.2.3, and infer the number of reactions in each interval $(i, i+1]$ using the posterior distribution generated by the Metropolis-Hastings algorithm. In this scheme, a proposal is generated by augmenting the current value by a random variable from the Skellam distribution [10, 43]. Once samples of both rate and delay parameters for all N cells are obtained, we sample the hyperparameters ω_θ and ω_Δ using the Metropolis-Hastings algorithm and the individual-level parameters as data in the population-level sampling.

The MCMC algorithm to produce samples from the approximate posterior distribution obtained using the hierarchical model given by Eq. (13) can thus be described by the following steps.

1. For each cell $n = 1, 2, \ldots, N$, reaction number $k = 1, 2, \ldots, v$, and time interval $i = 0, 1, \ldots, T-1$, initialize the number of reactions r_{nki}. Initialize the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$, and hyperparameters ω_θ and ω_Δ.

2. For each n,

   (a) Sample, in order, θ_{nk}, $k = 1, 2, \ldots, v$, given all rate hyperparameters ω_θ, other rate constants θ_{nm}, $m \neq k$, delay parameters Δ_n, and reaction numbers. If $y_n(t)$ and θ_{nk} are separable in $h_k\left(y_n\left(t\right), \theta_{nk}\right)$, then sample θ_{nk} from the conjugate gamma distribution. Otherwise, use the Metropolis-Hastings algorithm.

(b) Sample, in order, Δ_{nkl}, $k = 1, 2, \ldots, v$ and $l = 1, 2, \ldots, l_k$, given all delay hyperparameters ω_Δ, other delay constants $\Delta_{nk'l'}$, $(k', l') \neq (k, l)$, rate parameters θ_n, and reaction numbers, using the Metropolis-Hastings algorithm.

(c) Update the number of reactions, r_{nki}, for $k = 1, 2, ...v$ and $i = 0, 1, ..., T-1$, given θ_n, Δ_n, and the observed trajectory y_n using the simplified block-updating method.

3. For every reaction k,

(a) Sample a_{θ_k}, given the rate constants $\{\theta_{nk}\}_n$ from the entire population and the other rate hyperparameter b_{θ_k}.

(b) Sample b_{θ_k}, given the rate constants $\{\theta_{nk}\}_n$ from the entire population and other rate hyperparameter a_{θ_k}.

(c) Sample, in order, $\omega_{\Delta_{kls}}$, $l = 1, 2, \ldots, l_k$, $s = 1, 2, \ldots, |\omega_{\Delta_{kl}}|$ given the delay parameters $\{\Delta_{nkl}\}_n$ from the entire population and other delay hyperparameters $\omega_{\Delta_{kls'}}$, $s' \neq s$.

4. Repeat steps 2-3 until convergence.

When the chemical kinetics specified for the BCRN allows for the separation of $y_n(t)$ and $\theta_{n,k}$ in $h_k\left(y_n\left(t\right), \theta_{nk}\right)$, as in mass-action kinetics, our generative model which assumes that individual parameters are independent samples from a gamma distribution, leads to a conjugate gamma conditional posterior for $\theta_{n,k}$. In this case, step 2(a) of this algorithm simplifies to just sampling from a known probability distribution. Similarly, depending on the choice of hyperpriors, step 3 is either carried out by sampling from conjugate conditional distributions or using the Metropolis-Hastings algorithm. In the succeeding sections we provide all likelihoods and resulting posterior distributions given specific hyperprior distributions and delay measures for a stochastic birth-death process with birth delays.

3.2 Hierarchical model of a stochastic birth-death process with fixed delay

To demonstrate the effectiveness of our approach in the inference of BCRNs, as well as to catch potential issues in accuracy and identifiability, we first demonstrate the inference process using data

generated from a collection of stochastic birth-death processes with fixed birth delays [7, 13, 32, 36],

$$\emptyset \xrightarrow[\tau_n]{A_n} Y \xrightarrow{B_n} \emptyset, \tag{14}$$

where $n = 1, 2, ..., N$ and N is the number of cells (see also Eq. (9)). By fixed delay, we mean that it takes a constant time interval between initiation of a birth reaction and its completion, that is, the delay follows a Dirac delta distribution centered at τ_n.

Although this delayed birth-death process is simple, it can still explain the dynamics of chemical species, such as proteins, that are produced through a sequence of reactions: transcription, translation, protein folding, and maturation. Henceforth, we refer to the product, Y, in Eq. (14) as a protein, consistent with the type of experimental data that we will deal with later in the study. While many models of protein expression assume that production happens in bursts due to rapid translation events [66, 75], we suppose a Poissonian production. This assumption is valid if we are dealing with a system with a high copy gene, as in the experimental systems we will study subsequently.

In every cell, when a birth reaction occurs, it takes a fixed amount of time, that does not change throughout the observation window, before a mature protein Y emerges. For brevity, we say that birth delays, τ_n, are fixed. While constant within a cell, the fixed delays may differ between cells. Later we consider distributed delays (see Section 3.3) that vary between reactions within and between cells.

In the generative model (Fig. 3), a cell is characterized by a triple (A_n, B_n, τ_n), with production rate, A_n, degradation rate, B_n, and fixed birth delay, τ_n. Each of these parameters which describe reaction rate characteristics within individual cells follows a gamma distribution. We assume that the cognate promoters are not leaky. When the population is induced at time $t = 0$, the production rate in each cell n therefore instantaneously changes from 0 to the fully-induced values A_n. Protein count reduction comes from growth-induced dilution or enzymatic degradation, and is described by a death process with rate B_n with an immediately observable effect. We also assume that protein

numbers, Y, are exactly measurable at discrete times.

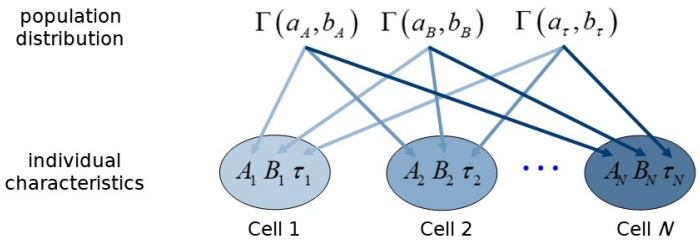

Figure 3: Generative model for the birth-death process with fixed birth delays. Individual birth-death processes, are described by the production rate, A_n, the degradation rate, B_n, and the fixed birth delay time, τ_n. We assumed these parameters follow Gamma distributions each with corresponding parameter pairs (a_Z, b_Z) for $Z = A, B, \tau$.

3.2.1 The posterior distribution

For the birth-death process, each individual n has birth (reaction $k = 1$) parameter A_n and death (reaction $k = 2$) parameter B_n, so that $\boldsymbol{\theta} = \{A_n, B_n\}_{n=1}^N$. Delays are fixed in each experiment and the delay measure $\eta_{n,k}$ is the Dirac point mass measure centered at the fixed delay $\tau_{n,k}$. Since we only consider delays in the birth reaction, henceforth we write η_n for $\eta_{n,1}$ and we write τ_n for $\tau_{n,1}$. Endowing the reaction network with mass-action kinetics, the reaction hazards are given by

$$h_1(y_n(t), A_n) = A_n,$$
$$h_2(y_n(t), B_n) = B_n y_n(t).$$

With only the discrete-time observations, using Eq. (6), the average completion propensity for a birth reaction on the interval $(i, i+1]$ is

$$\begin{aligned}\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n) &= A_n \int_i^{i+1} \int_0^t d\eta_n(s)dt \\ &= A_n \cdot p_{n,i}, \end{aligned} \tag{15}$$

where $\Delta_n = \{\tau_n\}$ and

$$p_{n,i} = \begin{cases} 0 & \text{if } i+1 \leq \tau_n \\ \min(1, i+1-\tau_n) & \text{otherwise} \end{cases}.$$

On the other hand, it is straightforward to compute[10] that the average completion propensity for the death reaction is

$$\begin{aligned} \hat{f}_2(i, \mathbf{y}_{d,n}, B_n) &= \frac{h_2(y_n(i), B_n) + h_2(y_n(i+1), B_n)}{2} \\ &= \frac{B_n y_n(i) + B_n y_n(i+1)}{2}, \end{aligned} \tag{16}$$

which is simply the average of the reaction propensities at time points i and $i+1$.

Using Eq. (15) and (16), we obtain the total likelihood for $\mathbf{y}_d = \{\mathbf{y}_{d,n}\}_n$,

$$\hat{\mathcal{L}}(\mathbf{y}_d \,|\boldsymbol{\theta}\,, \boldsymbol{\Delta}) = \prod_{n=1}^{N} \hat{L}(\mathbf{y}_{d,n}\,|\theta_n\,, \Delta_n), \tag{17}$$

where

$$\begin{aligned} \hat{L}(\mathbf{y}_{d,n}\,|\theta_n\,, \Delta_n) = &\prod_{i=0}^{T-1} \frac{\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n)^{r_{n1i}}}{r_{n1i}!} \exp\left(-\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n)\right) \\ &\times \prod_{i=0}^{T-1} \frac{\hat{f}_2(i, \mathbf{y}_{d,n}, B_n)^{r_{n2i}}}{r_{n2i}!} \exp\left(-\hat{f}_2(i, \mathbf{y}_{d,n}, B_n)\right) \end{aligned}$$

and r_{nki}, for $k = 1, 2$, is the number of reactions which completed in the time interval $(i, i+1]$.

Following the generative model, Fig. 3, we use Gamma priors $\Gamma(A_n|a_A, b_A)$, $\Gamma(B_n|a_B, b_B)$, and $\Gamma(\tau_n|a_\tau, b_\tau)$ for A_n, B_n, and τ_n respectively for $n = 1, ..., N$. We also specify the improper[11] joint hyperpriors $\pi(a_A, b_A) \propto \frac{1}{b_A}$, $\pi(a_B, b_B) \propto \frac{1}{b_B}$, and $\pi(a_\tau, b_\tau) \propto \frac{1}{b_\tau}$. We denote the collection, $\{a_A, a_B, b_A, b_B\}$, of reaction rate hyperparameters as ω_θ, and the collection of delay hyperparameters, $\{a_\tau, b_\tau\}$, as ω_Δ. Putting together all details from Eq. (15), (16), and (17) we arrive at the

[10] Because of the simplicity of the reaction hazard, one need not use the final form of Eq. (6) and can just immediately interpolate the hazards between i and $i+1$ before averaging the completion propensities.

[11] By *improper priors* we mean priors defined by functions whose integral is not equal to 1, and typically diverges. We discuss in Subsection 3.3.2 an interpretation of the improper prior given here.

total posterior distribution over the parameters and hyperparameters

$$
\begin{aligned}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta}, \omega_\theta, \omega_\Delta \left|\mathbf{y}_d\right.\right) \propto & \pi\left(a_A, b_A\right) \pi\left(a_B, b_B\right) \pi\left(a_\tau, b_\tau\right) \hat{\mathcal{L}}\left(\mathbf{y}_d \left|\boldsymbol{\theta}, \boldsymbol{\Delta}\right.\right) \\
& \times \prod_{n=1}^{N} \pi\left(A_n \left|a_A\right., b_A\right) \pi\left(B_n \left|a_B\right., b_B\right) \pi\left(\tau_n \left|a_\tau\right., b_\tau\right) \\
= & \frac{1}{b_A} \frac{1}{b_B} \frac{1}{b_\tau} \prod_{n=1}^{N} \prod_{i=0}^{T-1} \frac{\left(A_n p_{n,i}\right)^{r_{n1i}}}{r_{n1i}!} \exp\left(-A_n p_{n,i}\right) \\
& \times \prod_{n=1}^{N} \prod_{i=0}^{T-1} \frac{\left[\frac{1}{2} B_n\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right]^{r_{n2i}}}{r_{n2i}!} \exp\left(-\frac{1}{2} B_n\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right) \\
& \times \prod_{n=1}^{N} \frac{{b_A}^{a_A}}{\Gamma\left(a_A\right)} {A_n}^{a_A-1} \exp\left(-b_A A_n\right) \frac{{b_B}^{a_B}}{\Gamma\left(a_B\right)} {B_n}^{a_B-1} \exp\left(-b_B B_n\right) \\
& \times \prod_{n=1}^{N} \frac{{b_\tau}^{a_\tau}}{\Gamma\left(a_\tau\right)} {\tau_n}^{a_\tau-1} \exp\left(-b_\tau \tau_n\right). \qquad (18)
\end{aligned}
$$

Using Eq. (18), we derive the conditional posterior distributions of the parameters and hyperparameters. For each A_n and B_n, we obtain the conditional posteriors which belong to the gamma family:

$$
\begin{aligned}
& A_n \left|\mathbf{y}_{d,n}\right., a_A, b_A, \tau_n \sim \Gamma\left(\sum_{i=0}^{T-1} r_{n1i} + a_A, T - \tau_n + b_A\right), \\
& B_n \left|\mathbf{y}_{d,n}, a_B, b_B \sim \Gamma\left(\sum_{i=0}^{T-1} r_{2ni} + a_B\right., \left.\sum_{i=0}^{T-1} \frac{y_n(i+1)+y_n(i)}{2} + b_B\right). \qquad (19)
\end{aligned}
$$

The conditional posterior for a delay parameter, on the other hand, does not follow a known distribution and is proportional to:

$$
\pi\left(\tau_n \left|\mathbf{y}_{d,n}\right., a_\tau, b_\tau, A_n\right) \propto \left(\prod_{i=0}^{T-1} {p_{n,i}}^{r_{n1i}}\right) \exp\left(-A_n\left(T-\tau_n\right)\right) {\tau_n}^{a_\tau-1} \exp\left(-b_\tau \tau_n\right). \qquad (20)
$$

The shape parameters of hyperprior for the reaction constants A and B do not have known distribution as conditional posteriors and are proportional to:

$$\begin{aligned}\pi(a_A|b_A, A) &\propto \frac{b_A^{Na_A}}{\Gamma(a_A)} \prod_{n=1}^{N} {A_n}^{a_A-1}, \\ \pi(a_B|b_B, B) &\propto \frac{b_B^{Na_B}}{\Gamma(a_B)} \prod_{n=1}^{N} {B_n}^{a_B-1}.\end{aligned} \tag{21}$$

The rate parameters of hyperprior for the reaction constants A and B belong to the gamma family:

$$\begin{aligned}b_A|a_A, A &\sim \Gamma(Na_A\ ,\ \sum_{n=1}^{N} A_n), \\ b_B|a_B, B &\sim \Gamma(Na_B\ ,\ \sum_{n=1}^{N} B_n).\end{aligned} \tag{22}$$

In all these conditional posteriors, the symbols A and B are the collections of reaction rate constants for all individuals.

The conditional posteriors for the hyperparameters of delay time follow the same form as the ones for the reaction rate constants:

$$\begin{aligned}\pi(a_\tau|b_\tau, \tau) &\propto \frac{b_\tau^{Na_\tau}}{\Gamma(a_\tau)} \prod_{n=1}^{N} {\tau_n}^{a_\tau-1}, \\ b_\tau|a_\tau, \tau &\sim \Gamma(Na_\tau\ ,\ \sum_{n=1}^{N} \tau_n).\end{aligned} \tag{23}$$

3.2.2 MCMC sampling algorithm for the parameters and hyperparameters of a birth-death process with fixed delays

The MCMC algorithm[12] for the hierarchical model of the stochastic birth-death process with fixed birth delays based on the posterior distribution (18) proceeds as follows.

1. For each n and i, for $n = 1, 2, \ldots, N$ and $i = 0, 1, \ldots, T-1$, initialize the number of reactions by setting $r_{n1i} = y_n(i+1) - y_n(i)$ and $r_{n2i} = 0$ if $y_n(i+1) \geq y_n(i)$, otherwise $r_{n2i} =$

[12] A Python implementation of this algorithm is found at https://github.com/mvcortez/Bayesian-Inference.

$y_n(i+1) - y(i)$ and $r_{n1i} = 0$. Initialize the hyperparameters $a_A, a_B, a_\tau, b_A, b_B, b_\tau$ using appropriate values[13], and initialize A_n and B_n by sampling from their conjugate gamma posterior distributions (Eq. (19)). Set an appropriate value[14] for τ_n.

2. For each n,

   (a) Sample A_n and B_n from their conditional conjugate posterior distribution given by Eq. (19).

   (b) Since the conditional posterior for τ_n does not follow a known distribution (Eq. (20)), use the Metropolis-Hastings algorithm to iteratively draw samples from the conditional posterior $\tau_n \,|\mathbf{y}_{d,n}, a_\tau, b_\tau, A_n\,$. We used the truncated Gaussian distribution with positive support as proposal distribution.

   (c) Conditioned on A_n, B_n, α_n, and β_n, for each time index i, update r_{n1i} and r_{n2i}. As the number of reactions are not observed directly, we will sample over them by following a block-updating method [10] which uses a random walk proposal on the number of birth reactions. Use the Metropolis-Hastings algorithm with a random walk chain to sample the number of completions of each type of reaction in a time interval $(i, i+1]$, $i = 0, 1, ...T-1$, given the observed system states $y_n(i)$ and $y_n(i+1)$. For the i^{th} interval, the joint conditional posterior of r_{n1i} and r_{n2i} is given by

$$\pi\left(r_{n1i}, r_{n2i} \,|\mathbf{y}_{d,n}, A_n, B_n, \Delta_n\right) \propto \frac{\left(\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n\right)^{r_{n1i}}}{r_{n1i}} \frac{\left[B_n\left(y_n\left(i\right) + y_n\left(i+1\right)\right)/2\right]^{r_{n2i}}}{r_{n2i}!},$$

   which is proportional to the product of the density functions of Poisson random variables. Here,

$$\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n) = A_n \cdot p_{n,i}$$

[13]The hyperparameters a_Z, b_Z for $Z = A, B, \tau$, are shape and rate parameters of a gamma distribution, and are hence appropriately initialized with positive constants.

[14]The delay time τ_n is initialized with a non-negative constant.

where $\Delta_n = \{\tau_n\}$ and

$$p_{n,i} = \begin{cases} 0 & \text{if } i+1 \leq \tau_n \\ \min(1, i+1-\tau_n) & \text{otherwise} \end{cases}.$$

Denote by $r_{n1i}^{(j-1)}$ the value of r_{n1i} from the $(j-1)^{\text{th}}$ iteration. The proposal distribution can be chosen to be a discrete random walk in which the current value is augmented by u, that is the difference of two Poisson random variables whose means are both equal to some λ which is usually a function of $r_{n1i}^{(j-1)}$. This value, u, is then used to define the proposed reaction number value $r_{n1i}^* = r_{n1i}^{(j-1)} + u$. In particular, the distribution of the update value, u, is a Skellam distribution [10, 43] given by

$$p\left(u \left| {r_{n1i}}^{(j-1)} \right.\right) = \exp\left(-2{r_{n1i}}^{(j-1)}\right) I_u\left(2{r_{n1i}}^{(j-1)}\right),$$

where I_u is a regular modified Bessel function of order u. Once r_{n1i}^* is chosen, then r_{n2i}^* can be uniquely determined using

$$y_n(i+1) - y_n(i) = r_{n1i}^* - r_{n2i}^*.$$

The proposed updates r_{1i}^* and r_{n2i}^* are then accepted with probability α_r, defined by

$$\alpha_r = \min\left\{1, \frac{p\left(u \left| {r_{1i}}^* \right.\right) \frac{\hat{f}_1(i,\mathbf{y}_d,A,\Delta_1)^{{r_{1i}}^*}}{{r_{1i}}^*!} \cdot \frac{\hat{f}_2(i,\mathbf{y}_d,B,\Delta_2)^{{r_{2i}}^*}}{{r_{2i}}^*!}}{p\left(u \left| {r_{1i}}^{(j-1)} \right.\right) \frac{\hat{f}_1(i,\mathbf{y}_d,A,\Delta_1)^{{r_{1i}}^{(j-1)}}}{{r_{1i}}^{(j-1)}!} \cdot \frac{\hat{f}_2(i,\mathbf{y}_d,B,\Delta_2)^{{r_{2i}}^{(j-1)}}}{{r_{2i}}^{(j-1)}!}}\right\}.$$

3. Sample the hyperparameters which describe the distribution of A_n, B_n, and τ_n across the population.

   (a) As the conditional posteriors (Eq. (21)) of a_A and a_B are not known distributions, draw samples using the Metropolis-Hastings algorithm. We specified as proposal distribution the truncated Gaussian distribution with positive support.

(b) Sample b_A and b_B from their conditional conjugate posterior distributions given by Eq. (22).

(c) With Eq. (23), use the Metropolis-Hasting algorithm with a positively-supported truncated Gaussian proposal distribution to generate a sample of a_τ from its conditional posterior. Sample b_τ from its conjugate gamma conditional posterior.

4. Repeat steps 1-3 until a desired number of samples are generated.

This process is repeated until convergence. However, there is no universal best method to check for convergence, but one current practice involves the generation of multiple chains that are initialized differently, and doing a visual examination of the sample trajectories. The chains overlapping each other is an indication that samples are representative of the posterior distribution.

3.2.3 Inference in a birth-death process with fixed delays

We now test the performance of the fixed delay algorithm using synthetic data that follows Fig. 3 as generative model. To do so, we sampled 40 triplets of the parameters (A_n, B_n, τ_n) from their corresponding gamma distributions and used them to simulate 40 realizations of the birth-death process in a 100-min window using the delayed Gillespie algorithm [6]. We then subsampled the resulting trajectories by recording the molecular counts at unit time intervals (Fig. 4a) in order to mimic experimentally-derived data.

Using our hierarchical inference algorithm and the synthetic data, we estimated the parameters and hyperparameters, and compared the estimates to those used to generate the data. We performed our analysis on two versions of the data: one wherein full discrete trajectories up to the 100^{th} min observation (Fig. 4a) were used to do inference, and another using partial trajectories accounting for the initial 40 min of observation (red box in Fig. 4a). As indicated in Subsection 3.2.1, we used non-informative, rational hyperpriors for all hyperparameters.

When we used full trajectories, the parameters A_n, B_n, and τ_n were all accurately estimated (Fig. 4b and c; blue dots). In contrast, both the birth and death rates (Fig. 4b; orange dots)

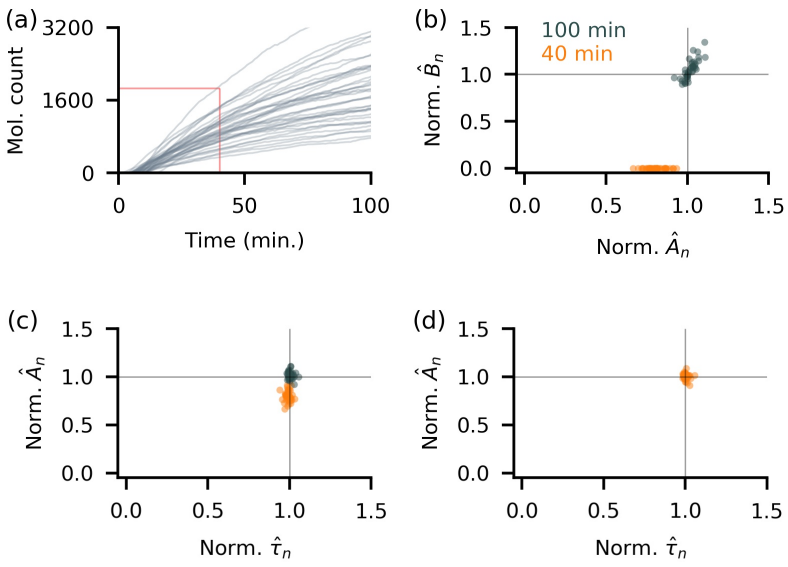

Figure 4: The fixed delay model accurately estimates all individual-level parameters when using molecule count trajectories that reach saturation. (a) Simulated trajectories of birth-death processes with fixed birth delays subsampled every minute. Individual parameters were sampled from the following gamma distributions in order to generate values that are similar to estimates in a previous study [16] for yellow fluorescence protein synthesis: $A_n \sim \Gamma(8, 0.23)$, $B_n \sim \Gamma(9, 625)$, and $\tau_n \sim \Gamma(7, 1)$. To estimate all three parameters per cell, we implemented the algorithm initially using the first 40 min (red box in (a)), and then the entire 100 min of observation. The individual posterior means serve as parameter estimates. In panels b-d we divided each estimate with the true parameter value, so that a perfect match corresponds to 1. (b-c) With 100 min of data, all rates (b) and delays (c) are accurately estimated (blue dots). However, 40 min of data lead to underestimates of the death rates, B_n (b). The birth rates, A_n, were similarly underestimated to compensate for the low death rate estimates (b). Estimates of the fixed delay times were still accurate (c). (d) When the degradation rates, B_n, were assumed known, the estimates for both birth rate and delay for each cell improved and migrated closer to the true parameter values.

were underestimated when we only used the initial segment of the protein count data. This mutual underestimation of both reaction rates stems from the rare occurrence of death reactions in the low protein count regimes: $Y_n(t)$, was initially small thereby resulting to a low value for the death reaction hazard, $B_n \cdot Y_n(t)$. An underestimate of the death rate, B_n, leads to an underestimate of the production rate, A_n, to compensate for the discrepancy. Delay times, on the other hand, were estimated well (Fig. 4c; orange dots) despite the inaccurate reaction rate estimates.

Identifiability problems in estimating the reaction rates may thus arise if we observe only the transient states of the process. This problem is alleviated by using longer observations (Fig. 4b-c;

blue dots), but possibly at the expense of a higher computational cost. To aid in inference when shorter trajectories were used, we assumed that the death rates, B_n, are known. This assumption is not unrealistic as death rates can be estimated from experimental data by measuring dilution rates through the tracking of cell growth [16]. With such an assumption, both A_n and τ_n were accurately estimated even with the shorter trajectories (Fig. 4d). Hence, inferring delays and birth rates from realistic amounts of data (e.g., 40 min in this case) is possible [15] if the death rates, B_n, can be directly measured.

Continuing with the specification of the true death rates in the process and using short 40-min trajectories, we inferred the population distribution of the cell parameters A_n and τ_n. Estimates of the parameter distributions across the population improved with the number of observed trajectories (Fig. 5a). This was characterized by the apparent convergence of posteriors to the true distributions as the number of trajectories increases: Population means (triangular makers) became more accurate and distributions became narrower. A similar improvement in both accuracy and precision of the hyperparameter estimates (Fig. 5b) further strengthened the evidence for the observed convergence: As the number of trajectories increased, hyperparameter medians moved closer to the true value and the posterior distributions become narrower. Our hierarchical inference algorithm, therefore, can be used to simultaneously infer reaction and delay parameters for individual cells, as well as the variability of these parameters across a population[15] from realistic amount and resolution of data.

Although parameter identifiability has been studied thoroughly for deterministic models, it is yet to be adequately established for stochastic systems [11]. One key identifiability issue that comes up in practice, termed *practical identifiability*, is whether a parameter can be estimated from a finite amount of noisy data with better accuracy than what is provided by the prior distribution [40, 70]. The parameters of the delayed birth-death process, at least in the parameter ranges we considered, are practically identifiable in both the data-rich and data-limited situations, with the caveat of having additional information on the death rates on the latter. Our hyperpriors are

[15]See Appendix A for details of the construction of population posterior distributions.

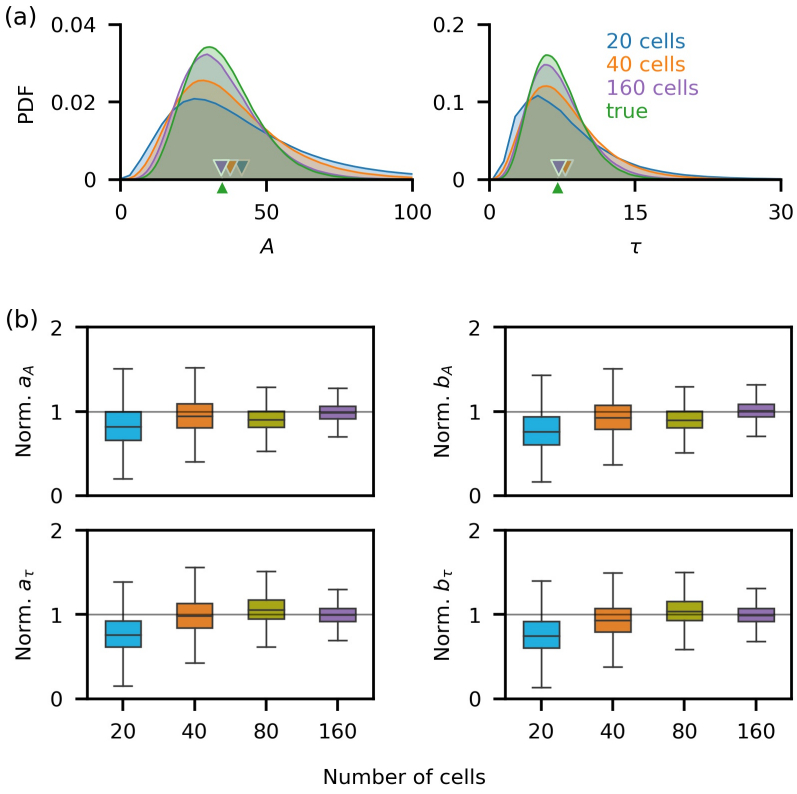

Figure 5: Increasing the number of cells used in hierarchical inference with fixed delays improved hyperparameter estimates and the corresponding population distribution of the parameters. (a) Population-level posterior densities of both the growth rate, A, and delay time, τ, were wider than the true densities, but their means (triangular markers) were close to the true value. The inferred population distributions improved with an increase in the number (from 20 to 160) of observed realizations of the stochastic processes. (b) Samples were normalized by dividing with the true hyperparameter values. Box plots corresponding to hyperparameter posterior distributions obtained using data from an increasing number of cells (from 20 to 160) show the convergence of posteriors to the true hyperparameter values.

uninformative, providing very little to no information on the model hyperparameters. Yet, our population posterior distributions are narrow with accurate measures of central tendencies, that lead to individual parameter estimates that are close to true values. These observations naturally come with questions about the possibility of the posterior distributions eventually converging to point masses given sufficiently large, and at the same time, finely-sampled data. Unfortunately, as in the present case, the finite nature of the time of observations renders such a convergence impossible. This is because the number of observed reactions over a finite interval is finite, regardless of the sampling resolution. Moreover, whether the individual parameter estimates converge to their true values as the observation window diverges, or whether the hyperparameter estimates converge to their true values as the number of cells diverges, as suggested by Fig. 5, are open questions.

3.3 Hierarchical Distributed Delay Model

Models of delayed systems usually incorporate fixed delays to represent the gap between the initiation and completion times of a process. While this is most certainly an oversimplification for most biological systems, the use of fixed delays makes it easier to analyze as well as to infer parameters of a complex model. More specifically, when only mean reaction delays are of interest, the use of fixed delays becomes even more attractive as it balances analytical and computational complexity with a possible loss of accuracy. We therefore asked whether a simple, hierarchical fixed delay model is sufficient to give accurate parameter estimates, even when reaction delays within a cell are not constant. To answer this question, we considered a model in which individual reaction delays followed a gamma distribution, $\tau_n \sim \Gamma(\alpha_n, \beta_n)$ (Fig. 6a), with parameters, α_n and β_n, that could differ between individual cells in the population (Fig. 6b).

3.3.1 The posterior distribution

Similar to the stochastic birth-death process with fixed delays, here the reactions within each individual cell n are characterized by a birth (reaction $k = 1$) parameter A_n and death (reaction $k = 2$) parameter B_n, so that $\boldsymbol{\theta} = \{A_n, B_n\}_{n=1}^{N}$. We assumed that the completion of a birth reaction

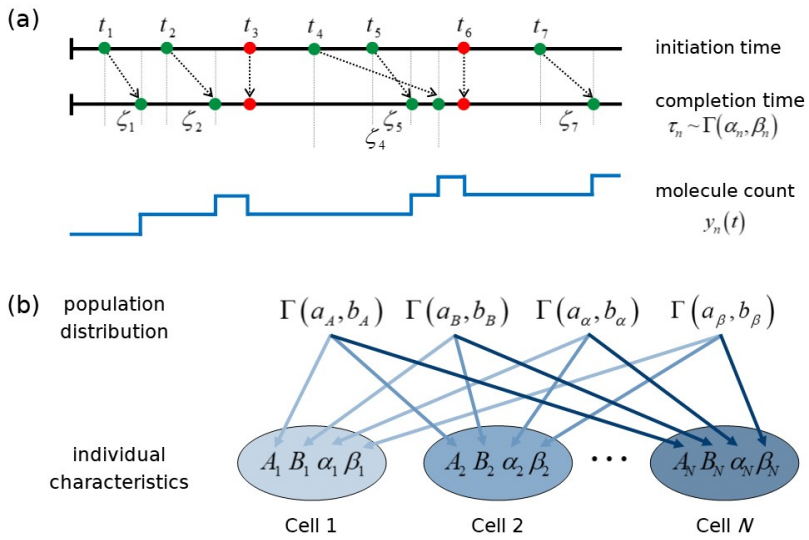

Figure 6: Generative model for the birth-death process with distributed delays. (a) A birth reaction (green) in each cell n is initiated at time t_i and completed after a delay, ζ_i. Each delay is a realization of the random variable τ_n which follows a Gamma distribution with parameters (α_n, β_n). Death reactions (red) are instantaneous. (b) The generative model for a birth-death process with distributed delays. Each individual process, n, is described by four parameters: the production rate, A_n, the degradation rate, B_n, and the two parameters describing the delay distribution, (α_n, β_n). All parameters follow Gamma distributions, with respective hyperparameters.

is delayed by a time τ_n following a gamma distribution $\Gamma(\alpha_n, \beta_n)$, so that $\boldsymbol{\Delta} = \{\alpha_n, \beta_n\}_{n=1}^N$. Since we only consider delays in the birth reaction, we write η_n for the delay distribution $\eta_{n,1} = \Gamma(\tau_n; \alpha_n, \beta_n)$, and we write τ_n for $\tau_{n,1}$. With mass-action kinetics, the reaction hazards are given by

$$h_1(y_n(t), A_n) = A_n,$$

$$h_2(y_n(t), B_n) = B_n y_n(t).$$

In our setup where only discrete-time observations are available, only the birth reaction is delayed so that the corresponding average completion propensity for a birth reaction on the interval

$(i, i+1]$ is

$$\begin{aligned}\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n) &= A_n \int_i^{i+1} \int_0^t d\eta(s)dt \\ &= A_n \int_i^{i+1} \frac{\gamma(\alpha_n, \beta_n t)}{\Gamma(\alpha_n)} dt, \end{aligned} \tag{24}$$

where $\Delta_n = \{\alpha_n, \beta_n\}$ and $\gamma(\alpha_n, \beta_n t)$ is the lower Gamma incomplete function [1]. On the other hand, the death reaction propensity is the same as that in Eq. (16), and is given by

$$\begin{aligned}\hat{f}_2(i, \mathbf{y}_{d,n}, B_n) &= \frac{h_2(y_n(i), B_n) + h_2(y_n(i+1), B_n)}{2} \\ &= \frac{B_n y_n(i) + B_n y_n(i+1)}{2}, \end{aligned}$$

which is the average of the delay-free death reaction hazard between the times i and $i+1$.

We use the approximate propensities given by Eq. (24) and (16) to define the total likelihood which accounts for N individual trajectories, $\mathbf{y}_d = \{\mathbf{y}_{d,n}\}_n$, given by

$$\hat{\mathcal{L}}(\mathbf{y}_d | \boldsymbol{\theta}, \boldsymbol{\Delta}) = \prod_{n=1}^{N} \hat{L}(\mathbf{y}_{d,n} | \theta_n, \Delta_n), \tag{25}$$

where

$$\begin{aligned}\hat{L}(\mathbf{y}_{d,n} | \theta_n, \Delta_n) = &\prod_{i=0}^{T-1} \frac{\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n)^{r_{n1i}}}{r_{n1i}!} \exp\left(-\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n)\right) \\ &\times \prod_{i=0}^{T-1} \frac{\hat{f}_2(i, \mathbf{y}_{d,n}, B_n)^{r_{n2i}}}{r_{n2i}!} \exp\left(-\hat{f}_2(i, \mathbf{y}_{d,n}, B_n)\right)\end{aligned}$$

and r_{nki}, for $k = 1, 2$, is the number of reactions which completed in the time interval $(i, i+1]$.

Following the generative model shown in Fig. 2b, we specify gamma priors $\Gamma(A_n|a_A, b_A)$, $\Gamma(B_n|a_B, b_B)$, $\Gamma(\alpha_n|a_\alpha, b_\alpha)$, and $\Gamma(\beta_n|a_\beta, b_\beta)$ for $n = 1, ..., N$. For the reaction rate hyperparameters, we specified the improper joint hyperpriors $\pi(a_A, b_A) \propto \frac{1}{b_A}$ and $\pi(a_B, b_B) \propto \frac{1}{b_B}$. We leave for later the specification of delay hyperpriors, and write as $\pi(a_\alpha, b_\alpha)$ and $\pi(a_\beta, b_\beta)$ the arbitrary

hyperpriors for α and β respectively. We denote the collection, $\{a_A, a_B, b_A, b_B\}$, of reaction rate hyperparameters as ω_θ, and the collection of delay hyperparameters, $\{a_\alpha, a_\beta, b_\alpha, b_\beta\}$, as ω_Δ. Accounting for Eq. (16), (24), and (25), the joint posterior distribution over the parameters and hyperparameters is given by

$$
\begin{aligned}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\Delta}, \omega_\theta, \omega_\Delta \,|\mathbf{y}_d\right) \propto\; & \pi\left(a_A, b_A\right) \pi\left(a_B, b_B\right) \pi\left(a_\alpha, b_\alpha\right) \pi\left(a_\beta, b_\beta\right) \hat{\mathcal{L}}\left(\mathbf{y}_d \,|\boldsymbol{\theta}, \boldsymbol{\Delta}\right) \\
& \times \prod_{n=1}^{N} \pi\left(A_n \,|a_A\,, b_A\right) \pi\left(B_n \,|a_B\,, b_B\right) \pi\left(\alpha_n \,|a_\alpha\,, b_\alpha\right) \pi\left(\beta_n \,|a_\beta\,, b_\beta\right) \\
= & \frac{1}{b_A} \frac{1}{b_B} \pi\left(a_\alpha, b_\alpha\right) \pi\left(a_\beta, b_\beta\right) \\
& \times \prod_{n=1}^{N} \prod_{i=0}^{T-1} \frac{\left(A_n \int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt\right)^{r_{n1i}}}{r_{n1i}!} \exp\left(-A_n \int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt\right) \\
& \times \prod_{n=1}^{N} \prod_{i=0}^{T-1} \frac{\left[\frac{1}{2} B_n\left(y_n\left(i+1\right) + y_n\left(i\right)\right)\right]^{r_{n2i}}}{r_{n2i}!} \exp\left(-\frac{1}{2} B_n\left(y_n\left(i+1\right) + y_n\left(i\right)\right)\right) \\
& \times \prod_{n=1}^{N} \frac{{b_A}^{a_A}}{\Gamma\left(a_A\right)} {A_n}^{a_A - 1} \exp\left(-b_A A_n\right) \frac{{b_B}^{a_B}}{\Gamma\left(a_B\right)} {B_n}^{a_B - 1} \exp\left(-b_B B_n\right) \\
& \times \prod_{n=1}^{N} \frac{{b_\alpha}^{a_\alpha}}{\Gamma\left(a_\alpha\right)} {\alpha_n}^{a_\alpha - 1} \exp\left(-b_\alpha \alpha_n\right) \frac{{b_\beta}^{a_\beta}}{\Gamma\left(a_\beta\right)} {\beta_n}^{a_\beta - 1} \exp\left(-b_\beta \beta_n\right).
\end{aligned}
\tag{26}
$$

Without specifying hyperpriors for the delay parameters $\boldsymbol{\Delta}$, using Eq. (26), we can derive the conditional posterior of A_n and B_n, which belong to the Gamma family:

$$
\begin{aligned}
A_n \,|\mathbf{y}_{d,n}, a_A, b_A, \Delta_n \;&\sim \Gamma\left(\sum_{i=0}^{T-1} r_{n1i} + a_A,\; \sum_{i=0}^{T-1} \int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)}\, dt + b_A\right), \\
B_n|\mathbf{y}_{d,n}, a_B, b_B \;&\sim \Gamma\left(\sum_{i=0}^{T-1} r_{n2i} + a_B\,,\; \sum_{i=0}^{T-1} \frac{y_n(i+1) + y_n(i)}{2} + b_B\right).
\end{aligned}
\tag{27}
$$

The delay parameters α_n and β_n do not have standard distributions as conditional posteriors which

are proportional to

$$\alpha_n \left| \mathbf{y}_{d,n}, A_n, \beta_n \right. \propto \prod_{i=0}^{T-1} \left[\int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt \right]^{r_{n1i}} \exp\left(-A_n \sum_{i=0}^{T-1} \int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt \right) {\alpha_n}^{a_\alpha - 1} \exp\left(-\alpha_n b_\alpha\right),$$
$$\beta_n \left| \mathbf{y}_{d,n}, A_n, \alpha_n \right. \propto \prod_{i=0}^{T-1} \left[\int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt \right]^{r_{n1i}} \exp\left(-A_n \sum_{i=0}^{T-1} \int_i^{i+1} \frac{\gamma\left(\alpha_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt \right) {\beta_n}^{a_\beta - 1} \exp\left(-\beta_n b_\beta\right). \tag{28}$$

The shape parameters of the hyperpriors for the reaction rate constants A and B do not have conditional posteriors which are known distributions but are proportional to:

$$\begin{aligned} \pi\left(a_A \left| A, b_A \right.\right) &\propto \frac{b_A^{N a_A}}{\Gamma(a_A)^N} \prod_{n=1}^{N} A_n^{a_A - 1}, \\ \pi\left(a_B \left| B, b_B \right.\right) &\propto \frac{b_B^{N a_B}}{\Gamma(a_B)^N} \prod_{n=1}^{N} B_n^{a_B - 1}, \end{aligned} \tag{29}$$

while the rate parameters of the hyperpriors for A and B belong to the gamma family:

$$\begin{aligned} b_A \left| A, a_A \right. &\sim \Gamma\left(N a_A, \sum_{n=1}^{N} A_n \right), \\ b_B \left| B, a_B \right. &\sim \Gamma\left(N a_B, \sum_{n=1}^{N} B_n \right), \end{aligned} \tag{30}$$

where A and B are the collections of reaction rate constants for all individuals.

3.3.2 Hyperpriors for delay hyperparameters

The choice of hyperpriors for the delay hyperparameters dictates the form of the conditional posterior distributions of a_α, a_β, b_α, and b_β. We present derivations using three different choices of delay hyperprior distributions, each representing varying levels of provided information in inference. We first show the cases of the non-informative rational hyperprior and maximal data information prior, and afterwards the informative folded normal distribution.

Rational prior

A typical non-informative joint hyperprior is the rational prior[16] which for the pair (a, b) takes the form

$$\pi(a, b) = \frac{1}{b}.$$

This improper prior is equivalent to flat or uniform priors on the first parameter a and the transformed random variable $\log(b)$. Such a transformation on b and the assumption of independence leads to the joint prior

$$\pi(a, b) \propto 1 \times \frac{1}{b},$$

as indicated. Setting these hyperpriors for the hyperparameters corresponding to both α and β yields conjugate conditional posteriors for b_α and b_β that belong to the gamma family. This choice of hyperprior, however, is not conjugate for both a_α and a_β. The conditional posteriors for the hyperparameters are given by

$$\begin{aligned}
\pi\left(a_\alpha \,|\alpha, b_\alpha\right) &\propto \frac{b_\alpha^{N a_\alpha}}{\Gamma(a_\alpha)^N} \prod_{n=1}^{N} \alpha_n^{a_\alpha - 1}, \\
\pi\left(a_\beta \,|\beta, b_\beta\right) &\propto \frac{b_\beta^{N a_\beta}}{\Gamma(a_\beta)^N} \prod_{n=1}^{N} \beta_n^{a_\beta - 1}, \\
b_\alpha \,|\alpha, a_\alpha &\sim \Gamma\left(N a_\alpha, \sum_{n=1}^{N} \alpha_n\right), \\
b_\beta \,|\beta, a_\beta &\sim \Gamma\left(N a_\beta, \sum_{n=1}^{N} \beta_n\right).
\end{aligned} \tag{31}$$

Maximal data information prior

The maximal data information prior (MDIP) [68, 88] is derived by maximizing the Kullback-Leibler divergence between the data density and the prior distribution. As such, the use of this prior puts emphasis on the information contained in the data density or likelihood function, effectively rendering the information provided by MDIP weaker in comparison. In our generative model,

[16] The use of this uninformative improper prior has roots in the inference of the mean and variance, (μ, σ^2), of the one variable Gaussian distribution. Flat priors on μ and $\log \sigma^2$ lead to the joint prior $\frac{1}{\sigma^2}$.

each of the parameters of an individual delay distribution is sampled from a gamma distribution $\Gamma(x; a, b)$ thereby serving as prior distribution in the hierarchical inference. In this case, the MDIP for the hyperparameters (a, b) becomes

$$\pi\left(a, b\right) = \frac{b}{\Gamma\left(a\right)} \exp\left\{\left(a-1\right)\psi\left(a\right) - a\right\}$$

where $\psi(a) = \dfrac{\Gamma'(a)}{\Gamma(a)}$ is the digamma function [65]. In this form, the corresponding joint posterior density is not proper, so Moala et al. [65] suggested the correction

$$\pi\left(a, b\right) = \frac{b}{\Gamma\left(a\right)} \exp\left\{\left(a-1\right)\frac{\psi\left(a\right)}{\Gamma(a)} - a\right\}, \tag{32}$$

so that a proper posterior density can be obtained.

Using the hyperprior (32), the resulting conditional posterior for a_α and a_β do not follow known distributions, however the MDIP is a conjugate prior for both b_α and b_β whose conditional posteriors belong to the gamma family. The conditional posteriors for the hyperparameters are given by

$$\begin{aligned}
\pi\left(a_\alpha \left| \alpha, b_\alpha \right.\right) &\propto \frac{b_\alpha^{Na_\alpha}}{\Gamma(a_\alpha)^{N+1}} \prod_{n=1}^{N} \alpha_n^{a_\alpha - 1} \exp\left\{\left(a_\alpha - 1\right)\frac{\psi\left(a_\alpha\right)}{\Gamma(a_\alpha)} - a_\alpha\right\}, \\
\pi\left(a_\beta \left| \beta, b_\beta \right.\right) &\propto \frac{b_\beta^{Na_\beta}}{\Gamma(a_\beta)^{N+1}} \prod_{n=1}^{N} \beta_n^{a_\beta - 1} \exp\left\{\left(a_\beta - 1\right)\frac{\psi\left(a_\beta\right)}{\Gamma(a_\beta)} - a_\beta\right\}, \\
b_\alpha \left| \alpha, a_\alpha \right. &\sim \Gamma\left(Na_\alpha + 2, \sum_{n=1}^{N} \alpha_n\right), \\
b_\beta \left| \beta, a_\beta \right. &\sim \Gamma\left(Na_\beta + 2, \sum_{n=1}^{N} \beta_n\right).
\end{aligned} \tag{33}$$

Folded normal distribution

Since the delay hyperparameters are positive, being parameters of a gamma distribution, the joint folded normal distribution [53, 69] is a candidate hyperprior distribution that can effectively define an arbitrarily strong joint hyperprior for these hyperparameters. The bivariate version, which follows naturally from the bivariate Gaussian distribution, describes two non-negative real-valued

random variables X and Y with probability density function given by

$$
\begin{aligned}
g\left(x,y\right) =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
& \times \left\{ \exp\left(-\frac{1}{2\left(1-\rho^2\right)} \left(\frac{\left(x-\mu_1\right)^2}{{\sigma_1}^2} - 2\rho\frac{\left(y-\mu_1\right)\left(x-\mu_2\right)}{\sigma_1\sigma_2} + \frac{\left(y-\mu_2\right)^2}{{\sigma_2}^2} \right) \right) \right. \\
& + \exp\left(-\frac{1}{2\left(1-\rho^2\right)} \left(\frac{\left(x+\mu_1\right)^2}{{\sigma_1}^2} - 2\rho\frac{\left(x+\mu_1\right)\left(x+\mu_2\right)}{\sigma_1\sigma_2} + \frac{\left(y+\mu_2\right)^2}{{\sigma_2}^2} \right) \right) \\
& + \exp\left(-\frac{1}{2\left(1-\rho^2\right)} \left(\frac{\left(x+\mu_1\right)^2}{{\sigma_1}^2} + 2\rho\frac{\left(x+\mu_1\right)\left(y-\mu_2\right)}{\sigma_1\sigma_2} + \frac{\left(y-\mu_2\right)^2}{{\sigma_2}^2} \right) \right) \\
& \left. + \exp\left(-\frac{1}{2\left(1-\rho^2\right)} \left(\frac{\left(x-\mu_1\right)^2}{{\sigma_1}^2} + 2\rho\frac{\left(x-\mu_1\right)\left(y+\mu_2\right)}{\sigma_1\sigma_2} + \frac{\left(y+\mu_2\right)^2}{{\sigma_2}^2} \right) \right) \right\},
\end{aligned}
$$

where $x > 0$, $y > 0$, $\sigma_i > 0$, $\mu_i \in \mathbb{R}$, $i = 1, 2$, and $|\rho| \leq 1$.

This distribution is not conjugate for any of the delay hyperparameters and the conditional posterior distributions resulting from this choice are given by

$$
\begin{aligned}
\pi\left(a_\alpha \,|\{\alpha_n\}_n, b_\alpha\right) &\propto \frac{b_\alpha^{Na_\alpha}}{\Gamma(a_\alpha)^N} \prod_{n=1}^{N} \alpha_n^{a_\alpha - 1} g(a_\alpha, b_\alpha; \mu_{a_\alpha}, \sigma_{a_\alpha}, \mu_{b_\alpha}, \sigma_{b_\alpha}, \rho_\alpha), \\
\pi\left(a_\beta \,|\{\beta_n\}_n, b_\beta\right) &\propto \frac{b_\beta^{Na_\beta}}{\Gamma(a_\beta)^N} \prod_{n=1}^{N} \beta_n^{a_\beta - 1} g(a_\beta, b_\beta; \mu_{a_\beta}, \sigma_{a_\beta}, \mu_{b_\beta}, \sigma_{b_\beta}, \rho_\beta), \\
\pi\left(b_\alpha \,|\{\alpha_n\}_n, a_\alpha\right) &\propto b_\alpha^{Na_\alpha} \exp\left(-b_\alpha \sum_{n=1}^{N} \alpha_n \right) g(a_\alpha, b_\alpha; \mu_{a_\alpha}, \sigma_{a_\alpha}, \mu_{b_\alpha}, \sigma_{b_\alpha}, \rho_\alpha), \\
\pi\left(b_\beta \,|\{\beta_n\}_n, a_\beta\right) &\propto b_\beta^{Na_\beta} \exp\left(-b_\beta \sum_{n=1}^{N} \beta_n \right) g(a_\beta, b_\beta; \mu_{a_\beta}, \sigma_{a_\beta}, \mu_{b_\beta}, \sigma_{b_\beta}, \rho_\beta),
\end{aligned}
\tag{34}
$$

where the corresponding folded normal hyperprior $g(a_Z, b_Z)$ is parameterized by μ_{a_Z}, σ_{a_Z}, μ_{b_Z}, σ_{b_Z}, ρ_Z for $Z \in \{\alpha, \beta\}$. The amount of information about a delay parameter Z is controlled by how close μ_{a_Z} and μ_{b_Z} are to the true values, and the magnitude of σ_{a_Z} and σ_{b_Z}.

3.3.3 MCMC sampling algorithm for the parameters and hyperparameters of a birth-death process with distributed delays

The distributed delay model has a more complex structure than its fixed delay counterpart. While an analogous algorithm specific to a birth-death process with fixed delays was already developed in Subsection 3.2.2, it lacks some components that are tailored for a distributed delay model. Here, we present an MCMC algorithm[17] for the stochastic birth-death process with distributed birth delays.

1. For each n and i, for $n = 1, 2, \ldots, N$ and $i = 0, 1, \ldots, T-1$, initialize the number of reactions by setting $r_{n1i} = y_n(i+1) - y(i)$ and $r_{n2i} = 0$ if $y_n(i+1) \geq y_n(i)$, otherwise $r_{n2i} = y_n(i) - y_n(i+1)$ and $r_{n1i} = 0$. Initialize a_A, a_B, b_A, b_B using appropriate values[18]. Initialize A_n and B_n by sampling from their conjugate gamma posterior distributions (Eq. (27)), and set appropriate values[18] for α_n and β_n.

2. For each n,

   (a) Generate samples A_n and B_n from their conditional conjugate posterior distribution given by Eq. (27).

   (b) Since the conditional posterior for α_n and β_n do not follow known distributions (Eq. (28)), use the Metropolis-Hastings algorithm to draw samples, in order, from the conditional posterior $\alpha_n \,|\mathbf{y}_{d,n}, A_n, \beta_n$ and $\beta_n \,|\mathbf{y}_{d,n}, A_n, \alpha_n$. We used the truncated Gaussian distribution with positive support as proposal distribution for α_n and a gamma proposal for β_n [16].

   (c) The update process for the number of completed reactions r_{nki} is similar to the case of fixed birth delays[19], but change the mean of the Poisson likelihood for the birth reaction to

$$\hat{f}_1(i, \mathbf{y}_{d,n}, A_n, \Delta_n) = \left(A_n \int_i^{i+1} \frac{\gamma\left(a_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt \right).$$

[17] A Python implementation of this algorithm is found at https://github.com/mvcortez/Bayesian-Inference.

[18] The hyperparameters a_θ, b_θ are shape and rate parameters of a gamma distribution, and are hence appropriately initialized with positive constants. The same is true for the delay parameters α and β.

[19] Refer to Subsection 3.2.2.

Hence for the interval $(i, i+1]$ interval, the joint conditional posterior of r_{n1i} and r_{n2i} is given by

$$\pi\left(r_{n1i}, r_{n2i} \left| \mathbf{y}_{d,n}, A_n, B_n, \Delta_n \right.\right) \propto \frac{\left(A_n \int_i^{i+1} \frac{\gamma\left(a_n, \beta_n t\right)}{\Gamma\left(\alpha_n\right)} dt\right)^{r_{n1i}}}{r_{n1i}!} \times \frac{\left[B_n\left(y_n\left(i\right) + y_n\left(i+1\right)\right)/2\right]^{r_{n2i}}}{r_{n2i}!}.$$

3. Generate a sample of the hyperparameters which describe the distribution of A_n, B_n, α_n, and β_n across the population.

   (a) As the conditional posteriors (Eq. (29)) of a_A and a_B are not known distributions, draw samples using the Metropolis-Hastings algorithm. We specified as proposal distribution the truncated Gaussian distribution with positive support.

   (b) Generate samples b_A and b_B from their conditional conjugate posterior distributions given by Eq. (30).

   (c) For rational priors, use Eq. (31) to implement the Metropolis-Hasting algorithm with a positively-supported truncated Gaussian proposal distribution to sample a_α and a_β from their conditional posterior. Sample b_α and b_β from their conjugate gamma conditional posteriors.

   In the case of the MDIP, use Eq. (33) to implement the Metropolis-Hasting algorithm with a positively-supported truncated Gaussian proposal distribution to sample a_α and a_β from their conditional posterior. Generate samples of b_α and b_β from their conjugate gamma conditional posterior.

   If folded normal distributions are used as hyperprior for α_n and β_n, we use Eq. (34), to implement the Metropolis-Hasting algorithm with a positively-supported truncated Gaussian proposal distribution to generate samples of a_α, a_β, b_α, and b_β from their conditional posteriors.

4. Repeat steps 1-3 until a desired number of samples are generated[20].

3.3.4 Inference in a birth-death process with distributed delays

We next test the performance of the algorithm using our hierarchical models on synthetic data produced using generative models with distributed birth delays (Fig. 6b). First, we compare the performance of the fixed and distributed delay models on populations with varying degrees of heterogeneity, and later on examine the advantages that a hierarchical model provides over a non-hierarchical counterpart.

We simulated trajectories using a model in which individual reaction delays followed a gamma distribution, $\tau_n \sim \Gamma(\alpha_n, \beta_n)$, with shape and rate parameters, α_n and β_n respectively (Fig. 6a). These shape and rate parameters, which themselves are samples from their respective gamma distributions, could differ between individual cells in the population (Fig. 6b). We chose three sets of parameters α_n and β_n so that for each set, the mean delay[21] across the population was the same (Table 1), while the variances of the *individual* delay distributions differed between the sets: $\sigma^2_{\tau_n} \approx 3.5$, $\sigma^2_{\tau_n} \approx 7$, $\sigma^2_{\tau_n} \approx 14$ min^2. In each case, we simulated 40 trajectories, each with 40 min of observations at 1 min intervals (Fig. 7).

Table 1: Hyperparameter values used to generate the individual delay parameters (α_n, β_n) that were used to simulate trajectories which served as data for Fig. 7.

σ^2_n	(a_α, b_α)	(a_β, b_β)
3.5	$(84, 6)$	$(10, 5)$
7	$(63, 9)$	$(10, 10)$
14	$(35, 10)$	$(10, 20)$

In all three cases, the same set of reaction rates A_n and B_n were used, with $A_n \sim \Gamma(8, 0.23)$ and $B_n \sim \Gamma(9, 625)$. In all cases, the mean delay, μ_{τ_n}, follows a beta prime distribution, $\beta'\left(a_\alpha, a_\beta, 1, \frac{b_\beta}{b_\alpha}\right)$, with mean 7.78 min.

As it was shown in the fixed delay case that specifying the generative values of the death

[20] As noted above, there is no universal best method to check for convergence, but the same heuristic methods discussed previously can be used here as well.

[21] As both α_n and β_n are gamma distributed, the mean delay is a ratio of two gamma distributed random variables whose distribution is a special case of a beta prime distribution.

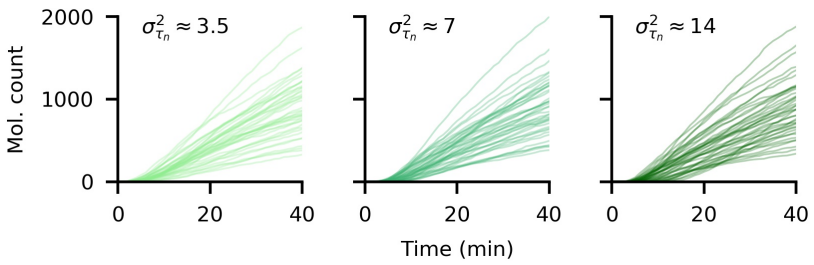

Figure 7: Simulated birth-death trajectories with distributed birth delays that follow a gamma distribution. To generate trajectories, we fixed a set of production and degradation rates, A_n, and B_n, and chose three different sets of delay parameters α_n and β_n. Mean delays were equal for all cases while delay variances within a cell were approximately equal across each population, but differed between the three cases (Table 1). For each parameter, set we simulated 40 trajectories that were subsampled every minute.

rates, B_n, is necessary for identifiability in the data resolution and length at hand, we proceeded by making them available for use in our algorithms during inference. When we applied the hierarchical *fixed delay* algorithm to the current simulated data, we observed that both mean delay times and birth rates were biased towards smaller values, and that the underestimations increased with the variance of the individual delay distributions (Fig. 8a-c; orange dots), that is, the wider the true distribution was, the farther the fixed delay time estimates migrated away from the true delay mean. This is consistent with the findings of Josić et al. [45] who showed that distributed delays can accelerate signaling in genetic networks by reducing the time for a process to reach threshold compared to systems with fixed delay. In the present case, since delay times were distributed, the earliest detectable signal after induction was likely to be observed *before* the mean delay time (Fig. 9). A model with fixed delay interprets the first observation of a molecule of the product species, Y, as the delay time to the completion of the first birth reaction after initiation. In addition, in the usual case when the subsampling interval is less than the first nonzero observation and consequently the mean delay, the discrete-time nature of the observations forces the fixed delay algorithm to reject[22] delay time samples which are larger than the time of the first nonzero observation. As a consequence, this leads to a Markov chain of samples which are all less than the

[22]Refer to the completion propensity (15) and conditional marginal posterior (20) to see this.

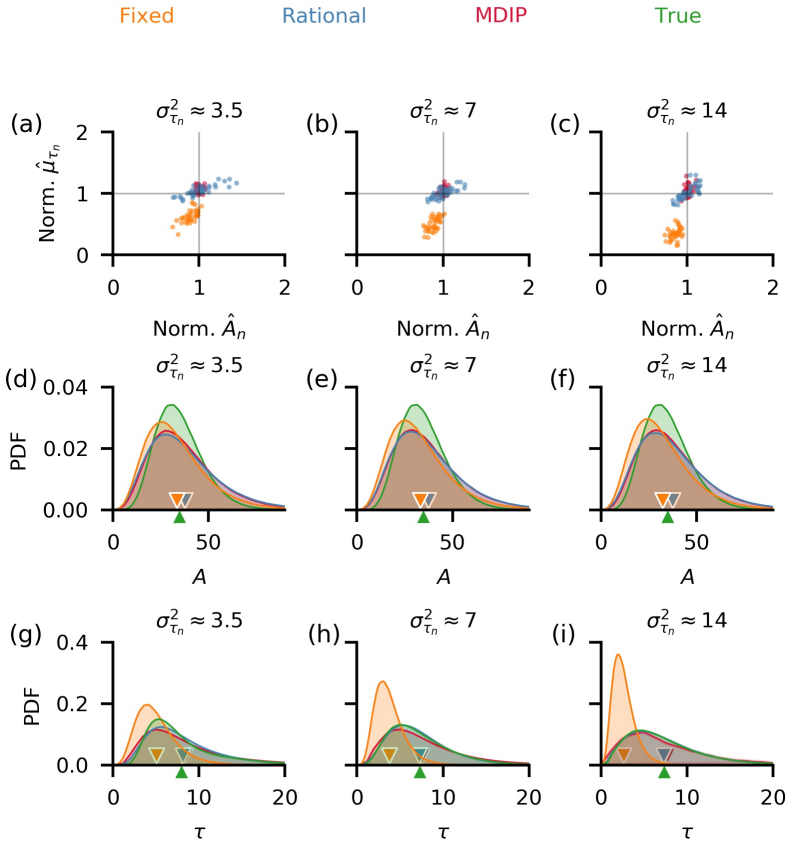

Figure 8: A hierarchical distributed delay model leads to accurate estimates of distributed delays, while a fixed delay model underestimates delays. (a-c) Across all data sets, both the production rates, A_n, and mean delay times, μ_{τ_n}, were underestimated when we used the fixed delay model (orange dots), but were accurately estimated with the distributed delay model with either rational hyperpriors (blue dots) or MDIP (red dots) over the delay hyperparameters. With the fixed delay model, the bias in the mean delay estimate increased with within-cell delay variance, $\sigma^2_{\tau_n}$. For comparison, we normalized the estimated parameters by dividing with the true values. We assumed B_n is known. (d-f) The estimated population distributions of the production rates were similar for the distributed delay models, with means (triangular markers) close to those of the true distributions. The posterior obtained with the fixed delay model gave a slight underestimate of the mean population production rate. (g-i) The pooled posterior delay distributions obtained using the distributed delay model matched the true distribution. The bias in the pooled posterior delay distributions obtained using the fixed delay model increased, and the estimated mean population delay (orange triangular marker) approached zero, as delay variance, $\sigma^2_{\tau_n}$, increased.

time of the first observed increase in protein count, and thus less than the mean of the true delay distribution. These observations suggest a careful treatment and interpretation of inference results for delayed systems (especially those with large delay variances) that are modeled with fixed delay [31, 61, 82], as even the inference of mean delay times generally requires an algorithm based on a distributed delay model.

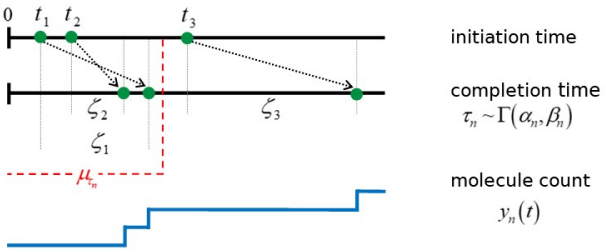

Figure 9: Initial samples, ζ_i from the distribution of delay times, $\Gamma(\alpha_n, \beta_n)$, may be less than the mean delay time μ_{τ_n}. Hence, the molecule count increases earlier than μ_{τ_n} making the fixed delay model biased toward a smaller delay time estimate.

We next asked whether parameters of a population of birth-death processes with distributed delays can be accurately estimated using a matching model (Fig. 6b) that specifies individual gamma distributed delays. With the non-informative rational hyperpriors as in the fixed delay model, inference resulted in accurate estimates of the individual production rates, A_n, and mean delay times, μ_{τ_n} (Fig. 8a-c; blue dots). The delay estimates were further improved by specifying a non-informative maximal data information prior (MDIP) [68, 88] over the delay hyperparameters (Fig. 8a-c; red dots). This advantage of the MDIP may be attributed to it being able to incorporate the dependence structure of the parameters (α_n, β_n) of the Gamma delay distribution [65], which are assumed to be independent when using rational priors. At the population level, the distributions of both the production rate, A, (Fig. 8d-f) and delay time, τ, (Fig. 8g-i) were similar for both non-informative hyperprior choices, and closely matched the true distributions. Henceforth, we used the MDIP as the default non-informative delay hyperprior.

The use of hyperpriors that capture pre-existing knowledge about the delay distribution, that is, specifying informative folded normal priors over the delay hyperparameters, ultimately improved

estimates of reaction rates and delay parameters (Fig. 10). For the folded normal distribution we used a mean that was close to values that were used to generate the trajectories, and a variance such that the generative values are within a standard deviation of the mean (Table 2). While individual mean delay time estimates, $\hat{\mu}_{\tau_n}$, were similar in both the folded normal and MDIP hyperprior cases, the non-informative MDIP resulted to more accurate production rate, A_n, estimates (Fig. 10a-c). This may be due to the fact that the strong folded normal hyperpriors were parameterized with values that are smaller than the true generative values. Individual delay variances, however, were considerably better estimated when folded normal delay hyperpriors were specified, and this effect becomes more prominent when individual delay variances are narrow (Fig. 10d-f). Across the three data sets considered, population posteriors obtained in both cases resemble the true population densities for both the production rate, A (Fig. 10g-i), and delay time, τ (Fig. 10j-l), with posterior means which are close true values (Fig. 10g-l triangular markers).

Table 2: Parameters of the folded normal distribution used to define the informative hyperpriors for the implementation seen in Fig. 10.

Hyperparameter	$\sigma_n^2 \approx 3.5$		$\sigma_n^2 \approx 7$		$\sigma_n^2 \approx 14$	
ω	$(\mu_\omega, \sigma_\omega)$	True value	$(\mu_\omega, \sigma_\omega)$	True value	$(\mu_\omega, \sigma_\omega)$	True value
a_α	$(81, 3)$	84	$(60, 3)$	63	$(32, 3)$	35
b_α	$(6, 3)$	6	$(6, 3)$	9	$(7, 3)$	10
a_β	$(7, 3)$	10	$(7, 3)$	10	$(7, 3)$	10
b_β	$(2, 3)$	5	$(7, 3)$	10	$(17, 3)$	20

The parameters were chosen so that an interval around the mean with radius equal to the variance covers the generative values. In all cases, $\rho = 0$.

Using a fixed delay model to infer mean delay times introduces bias in inference when delay itself varies. We conclude that a distributed delay model should be used for inference whether only the average delay, or more detailed information about the delay distribution, like higher-order statistics [8, 51], are of interest. An algorithm based on the distributed delay model provided accurate estimates of mean delays whether individual delay distributions were wide ($\sigma_n^2 \approx 14$ in Fig. 8a) or point-masses (see Fig. 11a-d). In the latter case, the mean of the distributed delay serving as estimates to the true fixed delay time were overestimated by around 10% (Fig. 11a). Reasoning

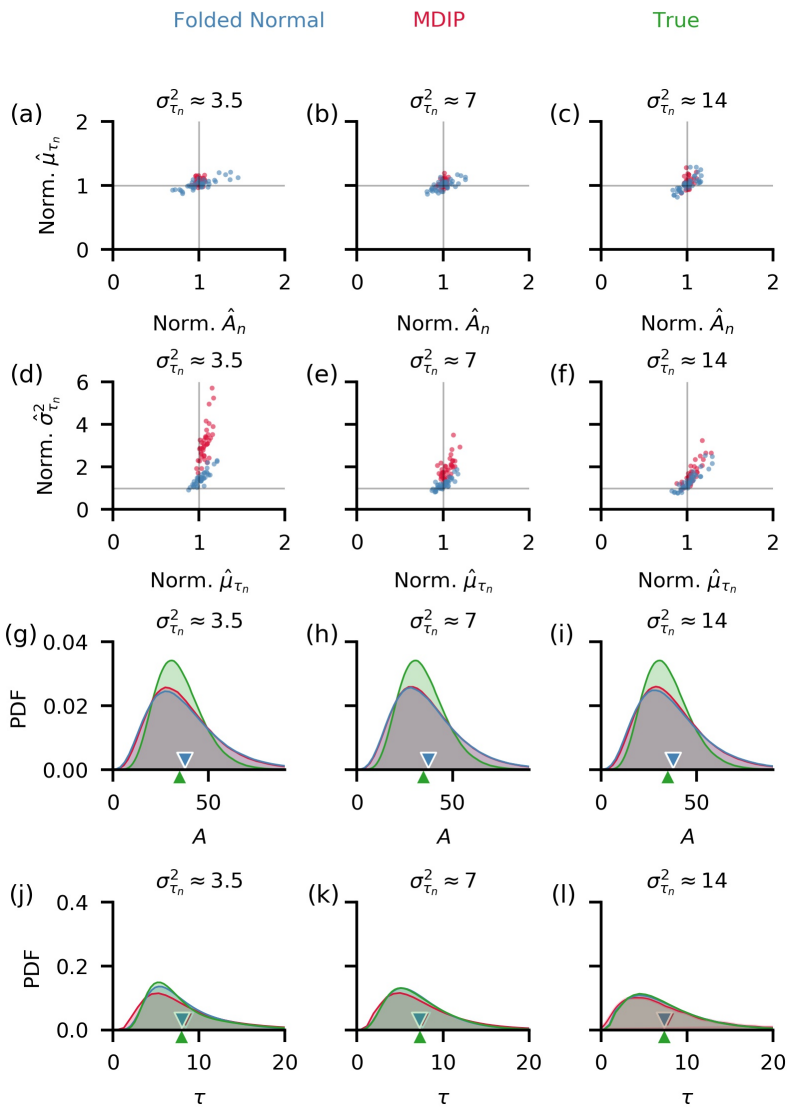

Figure 10: Informative folded normal delay hyperpriors yielded better estimates of delay parameters which consequently led to better estimates of individual delay variances. We considered three data sets with different levels of individual delay variability (See Fig. 8a). We divided the estimates by their true parameter values to facilitate a comparison between different model versions. (a-c). Individual mean delay time estimates, $\hat{\mu}_{\tau_n}$, are similar in both the folded normal and MDIP hyperprior cases, but production rate, A_n, estimates with MDIP are more accurate. (d-f) Individual delay variances were better estimated when we used folded normal delay hyperpriors as compared to when we used MDIP. (g-l) Across the three data sets and hyperpriors considered, population posteriors resemble the true population densities for both the production rate, A (g-i), and delay time, τ (j-l). The mean of the posteriors (triangular markers) are close to true population means.

in the same vein as in the converse situation, when delay time is assumed to be distributed when it actually is fixed, a process with distributed delay should be able to capture the behavior of one with fixed time delay, if the distribution has the right level of variability (Fig. 11b) even when the delay time mean is larger than the true fixed value. Along with all these results at the individual level, population level posteriors (Fig. 8h-j) also had the accurate mean delay times, but were distributed more widely than the parameters used to generate the observations. Thus, our hierarchical model is robust to changes in hyperparameters, and is applicable to populations with varied characteristics.

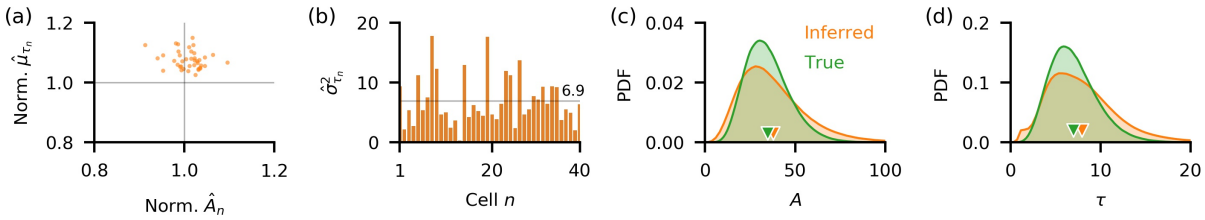

Figure 11: A distributed delay model with non-informative delay hyperpriors overestimates both the production rate and mean delay times when fit to data with fixed birth delays. (a) Even with a misspecified generative model, the distributed delay model is able to accurately infer individual parameters of a process with fixed birth delays with a slight overestimation of both the production rates, A_n, and mean delay times, μ_{τ_n}. (b) Since the delay hyperpriors are wide and uninformative, delay variances are largely overestimated with average variance of approximately 6.9 throughout the population, as compared to the true variance which equaled 0. (c-d) The slight overestimation of A_n and μ_{τ_n} extends to the population distribution whose means (triangular markers) are around 10% larger than the true values.

3.3.5 A comparison with a non-hierarchical analog

Hierarchical methods lead to robust estimates of population-level parameters by shrinking individual parameter estimates towards the average outcome in the population [17]. Despite this advantage, the integration of individual-level information done at the higher level increases model complexity by introducing additional hyperparameters that capture population-level information. An alternative approach is to infer model parameters individually from each observed trajectory and then perform population-level estimation based on the collected individual estimates. While

this approach requires less computational resources, it is typically less robust than hierarchical inference especially when dealing with a few number of individual observations. We therefore asked what advantages hierarchical inference offers over a non-hierarchical approach, and under what circumstances these advantages become apparent.

To compare the two inference approaches, we again considered measurements from a collection of birth-death processes with distributed birth delays (1-min subsampled trajectories with $\sigma_n^2 \approx 7$ in Fig. 7), and used non-informative priors[23] for both cases. While the estimates of the individual birth parameters, A_n, were similar for both models, the hierarchical model provided better estimates of the mean delay times, μ_{τ_n} (Fig. 12a). Although individual delay variances ($\sigma_{\tau_n}^2$) were overestimated in both cases, the hierarchical model still provided a better estimate (Fig. 12b). The hierarchical model also gave better estimates of the production rates and delay times (Fig. 12c and d) at the population level. The inferiority of the non-hierarchical model in estimating individual delay variances became even more pronounced when data resolution was low (Fig. 12e and f).

To confirm these observations, we simulated two additional sets of trajectories generated with parameters that differed from the hyperparameters used to generate the set of trajectories described above: one with a smaller production rate population mean and narrower individual delay distributions (Fig. 13a), and another with a larger production rate population mean and wider individual delay distributions (Fig. 13f). These two data sets resulted in trajectories with different characteristics, but the hierarchical model still resulted in better estimates of individual-level characteristics (Fig. 13b,c,d,h) as well as population variability (Fig. 13d,e,i,j).

We next sought to explain the observed difficulty in estimating the individual-level delay variances using both hierarchical and non-hierarchical approaches. Inference with non-informative delay hyperpriors on the three sets of trajectories found in Fig. 7 encountered the same issue of delay variance overestimation (Fig. 14g-i), as a result of the mutual overestimation of the individual delay parameters, (α_n, β_n) (Fig. 14a-c). An examination of the sample chain at the individual-level revealed a strong correlation between the inferred parameters α_n and β_n, (Fig. 14d-f) for all data

[23] See Appendix B for details.

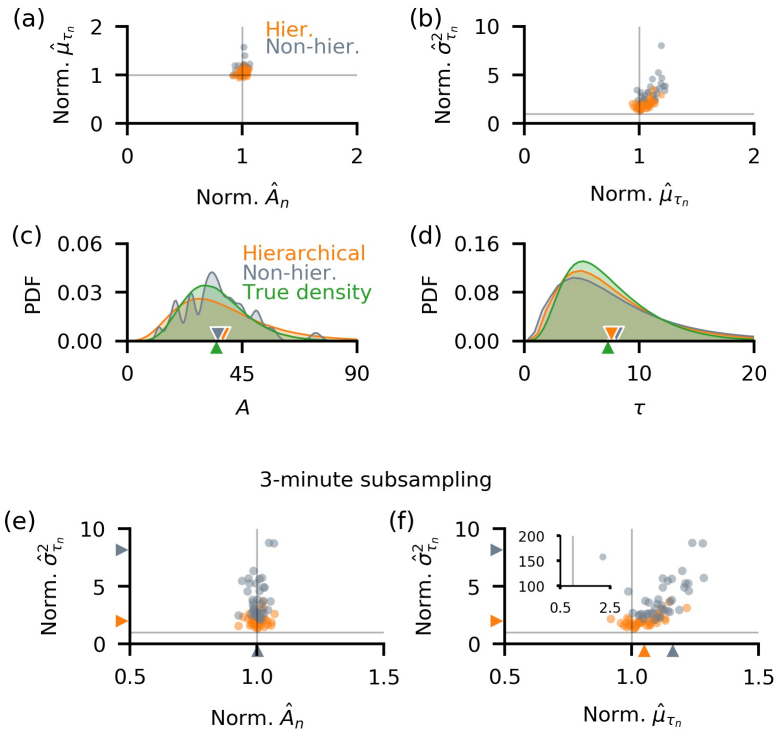

Figure 12: Hierarchical inference outperforms non-hierarchical inference and leads to better estimates of delay variances. Panels (a-b) show the individual parameter estimates normalized by the true values. (a) Although individual production rate estimates were similar for both approaches, the hierarchical model produced better estimates of mean delays with fewer outliers. (b) Delay variances are similarly better estimated by the hierarchical model. (c-d) Comparison of inferred population-level distributions of production rates, A, and delay times, τ, exhibit the same advantages of the hierarchical model. (e-f) In model implementation using 3-min subsampled trajectories, the accuracy of inferred production rates, $\hat{A}_n$ (e), and mean delay times, $\hat{\mu}_{\tau_n}$ (f), is similar, but the hierarchical model has a smaller bias, and produces fewer outlying estimates. With non-hierarchical inference, there was an extreme outlier which corresponded to overestimates of μ_{τ_n} and $\sigma^2_{\tau_n}$ (f inset). The hierarchical model provided better estimates of the individual delay variances. We used non-informative priors over parameters in all cases.

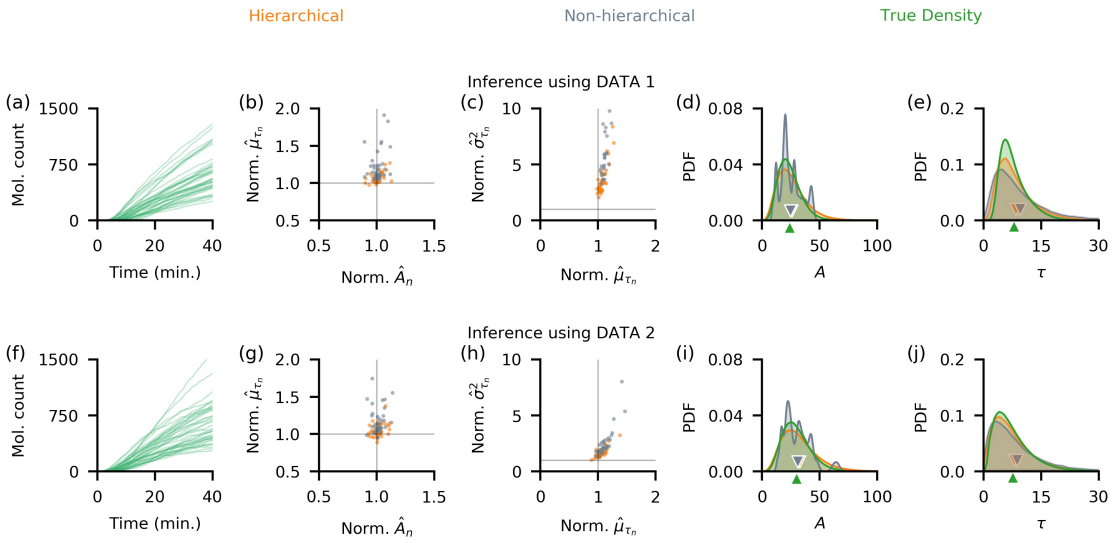

Figure 13: The hierarchical model consistently outperforms its non-hierarchical counterpart on different parameter and hyperparameter sets. (a and f) We generated two additional sets of 40 trajectories, each with 40 min of observation that were subsampled at 1-min intervals. The following population distributions were used to generate individual data: $A_n \sim \Gamma(6, 0.25)$, $B_n \sim \Gamma(9, 625)$, $\alpha_n \sim \Gamma(84, 6)$, and $\beta_n \sim \Gamma(10, 5)$ for data set 1; and $A_n \sim \Gamma(6, 0.2)$, $B_n \sim \Gamma(9, 300)$, $\alpha_n \sim \Gamma(35, 10)$, and $\beta_n \sim \Gamma(10, 20)$ for data set 2. Data set 1 was obtained using a smaller production rate population mean and narrower individual delay distributions compared to the data set used Fig. 12a-d. Data set 2, was obtained using a larger production rate population mean and wider individual delay distributions. (b and g) While individual production rate estimates, $\hat{A}_n$, were similar in both models, the mean delay times were better estimated with a hierarchical model. (c and h) The same advantage of the hierarchical approach also applies to the estimates of delay variances. (d and i) Although population mean of production rate (triangular markers), A, is captured in both approaches, the posterior from the hierarchical model better represent the true density. (e and j) The non-hierarchical model overestimates the population mean of delay times (triangular markers) while the hierarchical model gives a more accurate estimate.

sets. This strong linear relationship between the delay parameters in every individual made an accurate estimation of both individual delay mean and variance difficult. A similar observation was made by Choi et al. [16] for estimates from single trajectories. They proposed pooling multiple recordings to increase estimate accuracy, implicitly assuming that all cells in the population are identical and that the observations are different realizations of exactly the same birth-death process. While this strategy may result in estimates that capture the mean of the parameter across the population, this approach may can also lead to biased parameter estimates when cell populations are heterogeneous with large variability, as we show in the Subsection 3.3.6.

When generating synthetic data we used reaction rates, and delay values within an order of magnitude of those measured in biological systems [16, 86], and sampling rates that were consistent with those obtainable experimentally using time-lapse fluorescence microscopy [14, 24, 29]. In this range, the frequency of measurements has a strong impact on the accuracy of individual parameter estimates. We thus expected that the non-hierarchical model performs worse than its hierarchical counterpart when sampling frequency is low, but both produce similar estimates of parameters within individual cells at high sampling frequencies, as shown in Fig. 12. We further tested this prediction by considering the trajectories in Fig. 7 with $\sigma_n^2 \approx 7$ but changed the sample size and sampling frequency: We decreased the number of individuals to 20 (from the former 40), extended the observations to 60 min (from the former 40), and considered different sampling frequencies ranging from 4 per min up to once every 3 min. The hierarchical model yielded consistent results over the entire range of sampling frequencies we tested: individual parameter estimates were accurate, delay variance estimates did not diverge even with low-resolution data (Fig. 15a-e orange), and KL-divergence between the the delay population posterior and the true density remained small, and depended weakly on sampling frequency (Fig. 15f orange). The non-hierarchical model, on the other hand, produced estimates which decreased in accuracy with the increase in sampling interval (see Fig. 15a-e grey), produced delay population posteriors that were far from the true distributions (see Fig. 15f grey), as well as outlying individual parameter estimates (Fig. 15a-e last column inset). Thus the hierarchical model is more flexible and consistent across different data resolutions

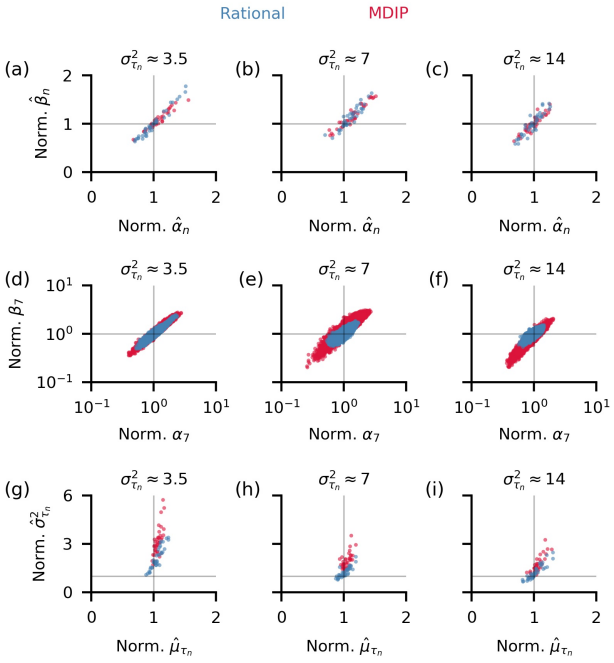

Figure 14: Variance of individual delay distributions are better captured using rational delay hyperpriors but this advantage disappears as true delay distributions become wider. In the distributed delay model, we used two different non-informative delay hyperparameter distributions in three different implementations: rational priors and the MDIP. (a-c) Estimates of individual delay parameters (α_n, β_n) were similar for both choices of non-informative priors. (d-f) Samples of individual estimates of the delay parameters α and β, for cell 7. The posterior distribution of the parameters shows a strong correlation between the two. We observed similar correlations in all cells, both when using the hierarchical, and non-hierarchical model. (g-i) Errors in the estimates of (α_n, β_n) lead to the overestimation of delay variances in model implementations using the rational and MDIP delay hyperpriors. While the errors in the estimates remained small in the case of the rational hyperpriors in all three data sets considered, the estimates improved for the MDIP case, as true individual delay variances become larger.

than its non-hierarchical counterpart. We concluded that the use of a hierarchical model is better for inferring population level characteristics, especially when computational constraints are not a major concern.

3.3.6 A comparison with results from pooled data

Another strategy to deal with heterogeneous cell populations is to pool the data[24] treating each trajectory as a realization of the same stochastic process. While this approach to multi-cell data may be effective for populations with little cell-to-cell variation, it can introduce a large bias when cells exhibit strong heterogeneity, even in estimates of the population central tendencies. We tested this hypothesis by performing hierarchical and pooled inference in three populations with increasing cell-to-cell variability in both the production rates and mean delay times, and keeping the death rates the same across three different cases.

The approximate likelihood function for pooled data sets is very similar to the likelihood (25) for hierarchical inference,

$$\hat{L}_p\left(\mathbf{y}_d\,|A,B,\alpha,\beta\,\right)=\prod_{n=1}^{N}\prod_{i=0}^{T-1}\frac{\left(A\int_i^{i+1}\frac{\gamma\left(\alpha,\beta\hat{t}\right)}{\Gamma(\alpha)}d\hat{t}\right)^{r_{n1i}}}{r_{n1i}!}\exp\left(-A\int_i^{i+1}\frac{\gamma\left(\alpha,\beta\hat{t}\right)}{\Gamma\left(\alpha\right)}d\hat{t}\right)$$
$$\times\prod_{n=1}^{N}\prod_{i=0}^{T-1}\frac{\left[\frac{1}{2}B\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right]^{r_{n2i}}}{r_{n2i}!}\exp\left(-\frac{1}{2}B\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right). \tag{35}$$

The crucial difference in the posterior derivation is in the specification of the the prior distribution, as for each parameter type, all individuals share a common prior distribution unlike in the hierarchical approach where a prior is set individually[25]. With gamma priors for every rate and delay

[24] Refer to Subsection 2.2.2 for likelihood formulation.
[25] Refer to Eq. (26).

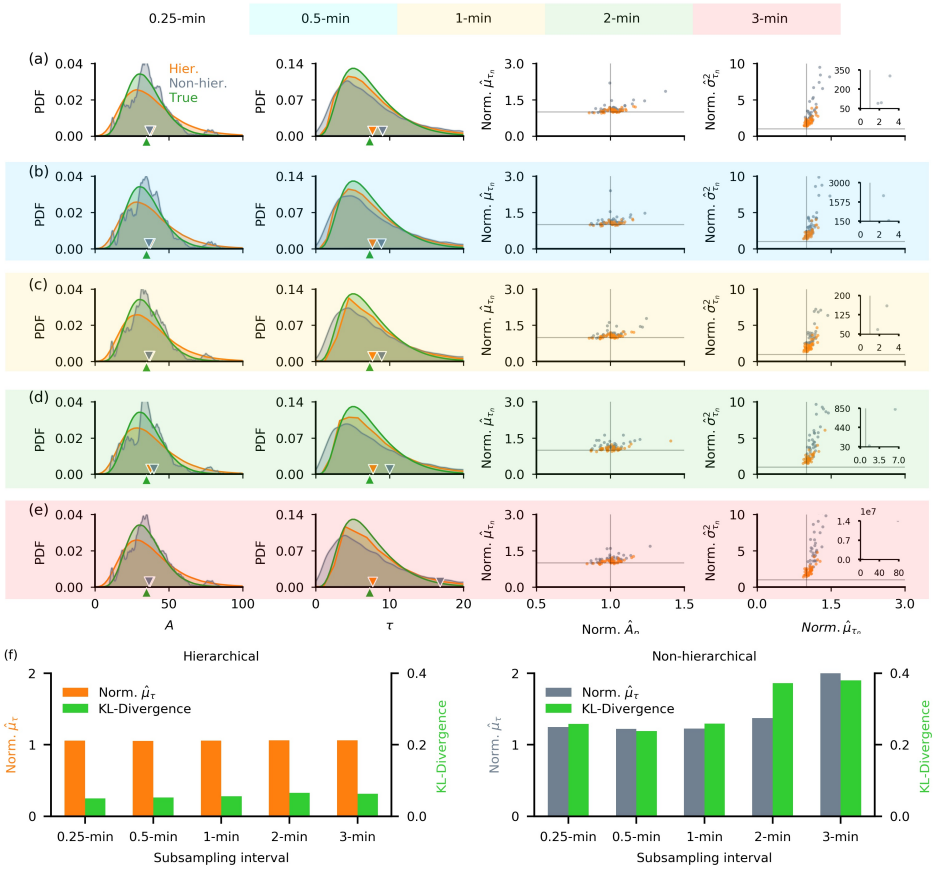

Figure 15: A non-hierarchical model is more sensitive to changes in sampling frequency than a hierarchical model. We implemented the hierarchical model and its non-hierarchical counterpart using 20 min of subsampled data (see $\sigma_n^2 \approx 7$ trajectories Fig. 8a) with decreasing sampling frequency (from 4 per min, i.e. 0.25-min subsampled, to 1/3 per min, i.e. 3-min subsampled). Although population means of the production rate, A, were very similar for both models across all subsampling schemes (triangular markers in a-e 1st column), the accuracy of the estimate of the delay distribution mean from the non-hierarchical model (grey) decreased with sampling frequency while those from the hierarchical model (orange) exhibited a similar accuracy (triangular markers in a-e 2nd column). Across all data subsets we considered, the hierarchical model individual parameter estimates for A_n, μ_{τ_n} (a-e 3rd column), and $\sigma^2_{\tau_n}$ (a-e 4th column) exhibited small deviations. Estimates from the non-hierarchical model, on the other hand, had reduced accuracy especially in terms of $\sigma^2_{\tau_n}$ (a-e 4th column) when we decreased the sampling frequency, with extreme outlying estimates produced at low sampling frequencies (a-e 4th column inset). (f) A comparison of population delay distributions showed that the hierarchical model produced a mean delay estimate (left - orange bars) that was consistently accurate, together with a population posterior with low KL-divergence between the posterior to the true density (left - green bars) that remained approximately constant for the different data subsets we considered. Decrease in sampling frequency resulted in reduced accuracy of the population mean delay estimate from the non-hierarchical model (right - grey bars). The non-hierarchical delay posterior also exhibited KL-divergence that increased with the subsampling interval (right - green bars).

parameter, the resulting posterior[26] over all the parameters is

$$\begin{aligned}\pi\left(A,B,\alpha,\beta\left|\mathbf{y}_d\right.\right) \propto & \prod_{n=1}^{N}\prod_{i=0}^{T-1}\frac{\left(A\int_i^{i+1}\frac{\gamma\left(\alpha,\beta\hat{t}\right)}{\Gamma(\alpha)}d\hat{t}\right)^{r_{n1i}}}{r_{n1i}!}\exp\left(-A\int_i^{i+1}\frac{\gamma\left(\alpha,\beta\hat{t}\right)}{\Gamma\left(\alpha\right)}d\hat{t}\right)\\ & \times\prod_{n=1}^{N}\prod_{i=0}^{T-1}\frac{\left[\frac{1}{2}B\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right]^{r_{n2i}}}{r_{n2i}!}\exp\left(-\frac{1}{2}B\left(y_n\left(i+1\right)+y_n\left(i\right)\right)\right)\\ & \times A^{a_A-1}\exp\left(-Ab_A\right)B^{a_B-1}\exp\left(-Bb_B\right)\\ & \times\alpha^{a_\alpha-1}\exp\left(-\alpha b_\alpha\right)\beta^{a_\beta-1}\exp\left(-\beta b_\beta\right).\end{aligned}$$

Indeed with pooled trajectories, for the population with small cell-to-cell variability (Fig. 16a red dot), both the mean delay time and production rate estimates closely matched the true population means. As the variability increased (Fig. 16b-c red dot), both mean estimates moved farther away from the true populations means with a particular underestimation in the cases considered. Whether this underestimation is a universal property remains to be explored in detail. The population central tendencies obtained using the hierarchical approach (Fig. 16 dark blue dots) were consistently accurate across all degrees of data variability.

3.3.7 The case of data-model incompatibility

The gamma distribution, as specified in our hierarchical model, is frequently used in models that include time delays in the production of mature functional proteins [13, 48, 50, 77]. To show that the assumption that reaction delays are gamma distributed does not strongly bias the estimates of different parameters, we used our model to infer delay parameters when the delay distribution was misspecified: We considered the beta and inverse gamma distributions for delays to generate synthetic data but performed inference assuming that delays are gamma distributed. The gamma distribution generally has infinite support and decays exponentially, while the beta distribution has compact support and the inverse-gamma distribution is heavy-tailed.

[26] See Appendix C for the derivation of corresponding conditional marginal posterior distributions.

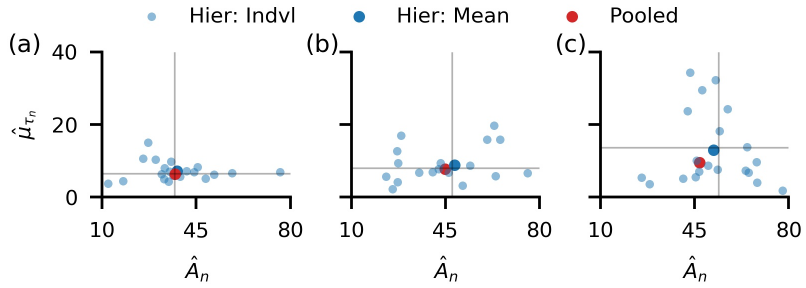

Figure 16: While pooling of data produces good estimates of mean parameter values for data with little variation across the population, errors may increase as cells become more different. Twenty trajectories accounting for 20-min observations of a delayed stochastic birth-death process served a data in this comparison. In order of increasing variability, both in terms of mean delays and production rates, across the population, data 1 (a) has the least variability, next is data 2 (b), while data 3 (c) has the largest. The following population distributions were used to generate individual data: $A_n \sim \Gamma(8, 0.23)$, $B_n \sim \Gamma(9, 625)$, $\alpha_n \sim \Gamma(63, 9)$, and $\beta_n \sim \Gamma(10, 10)$ for data 1; $A_n \sim \Gamma(8, 0.16)$, $B_n \sim \Gamma(9, 625)$, $\alpha_n \sim \Gamma(7, 1)$, and $\beta_n \sim \Gamma(5, 5)$ for data 2; and $A_n \sim \Gamma(8, 0.16)$, $B_n \sim \Gamma(9, 625)$, $\alpha_n \sim \Gamma(3.3, 0.6)$, and $\beta_n \sim \Gamma(2, 2.5)$ for data 3. As the variability increases, the estimates from model with data pooling migrate farther away from the true population means (vertical and horizontal lines in each plot), while the means of the hierarchical model estimates remain accurate.

With trajectories generated using beta and inverse-gamma delays, our algorithm produced both population and individual estimates which were overall accurate, but tended to slightly overestimate individual delay variances (see Fig. 17), as was observed previously when the model used for inference matched the model used to generate the data (Fig. 8). The overestimation of variances is larger in the case of the beta delays, which was expected since the beta distribution is finitely supported while the gamma distribution is supported on the semi–infinite interval $[0, +\infty)$.

3.4 Summary

The observed differences between cell phenotype in a population may be attributed to intrinsic and extrinsic noise. In the presence of such noise sources inference models must account for cell-to-cell variability not only to characterize population characteristics *per se*, but to also improve estimation at the level of individual cells by using population information. To this end, we developed a general,

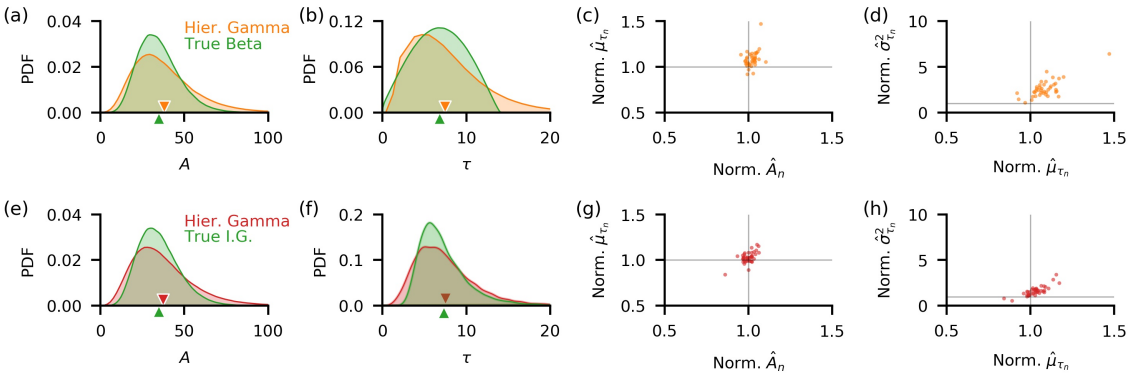

Figure 17: The hierarchical model provides accurate estimates even when the delay distribution is mismatched. We fit the hierarchical model with gamma distributed individual cell birth delays to data generated using beta (a-d) and inverse-gamma (e-h) distributions for the same. Even when the delay distributions in the model and data are not matched, population posteriors obtained in both cases closely resemble the true population densities for both the production rate, A (a and e), and delay time τ (b and f). The mean of the posteriors (triangular markers) are close to true population means. Individual estimates of the mean delay, μ_{τ_n} (c and g), are accurate, while delay variances, $\sigma^2_{\tau_n}$ (d and h), are slightly overestimated, as in when the distributions in the model and data are matched (Fig. 8).

multi-level approach to inference, a hierarchical model that can be used to infer within cell and cell-to-cell variability.

We applied our algorithm to a collection of observations of a stochastic birth-death process with birth delays. First, we developed a model for processes with fixed delays. We showed that our model is able to accurately infer all parameters, both reaction rate and delay, when data is sufficient in amount and resolution. When only transient states are known, the fixed delay model underestimated the reaction rates, a problem that was remedied by specifying the death rates during inference. The fixed delay model resulted to population posteriors which tended to the true distributions as the number of observed trajectories were increased. When the delays are in fact distributed, a fixed delay model underestimated both the production rates and delay times.

Next, we developed an algorithm that is applicable when delay times vary. We assumed that the death (or dilution) rates are known. We used this model to infer individual reaction rates and mean delay times accurately. However, using this method of inference produced overestimates in the delay variance due to the strong correlation between the samples of the delay parameters.

Regardless of hyperpriors used (be it rational, MDIP, or folded normal), these observations were verified across a range of generative hyperparameter values that determined population variability. We showed that our approach performs better than a non-hierarchical analog, especially when data resolution is low. We obtained better performance with our approach in exchange for additional complexity and higher computational cost, compared to inference with a non-hierarchical model. We also demonstrated that our hierarchical approach outperforms data pooling strategies even when only the population mean of delays and production rates are of interest. Even when the data does not match the model, as when the delay distribution is misspecified, our model was still able to accurately infer individual parameters and population distributions, with still a particular overestimation of delay variances.

4 Hierarchical inference of transcriptional and translational regulation

In the previous chapters, we developed a hierarchical algorithm for inferring within cell and cell-to-cell variability from discrete-time observations of a stochastic process. We have established the strengths and weaknesses of the hierarchical model using synthetically generated observations of a birth-death process with delays, and explored how the model performs under different degrees of population heterogeneity. In the present chapter, we use the hierarchical model with distributed delays to characterize the variability of delays and production rates in an experimentally observed clonal population of *E. coli*. We also examine the quality of parameter estimates by cross-validation and overfitting analysis. As previously, we assume that protein expression is delayed but Poissonian.

4.1 The YFP circuit

We use a birth-death process with distributed birth delays as a model of yellow fluorescent protein (YFP) production within individual cells. As with all proteins, the production of YFP is not instantaneous, as there can be a gap between the time of gene activation until the emergence of a mature functional protein.

In our analysis, we use fluorescence microscopy data obtained by Cheng et al. [15] in two independent experiments that measures YFP fluorescence intensity. These experiments used a P_{BAD} reporter-only circuit in *E. coli*, constructed by placing the YFP gene under the control of the P_{BAD} promoter. In this circuit (Fig. 18[27]), when Arabinose (ARA) is added to the media, ARA binds with AraC, which in turn promotes the constitutive transcription of YFP [22, 60]. Induction is followed by a sequence of steps including transcription, translation, protein folding, and maturation, which result in the production a mature YFP after a certain delay.

Although cell growth does not have a strong impact on YFP production [4, 22, 60], it drives the decrease in YFP through dilution. YFP is a relatively stable protein and was not tagged

[27]This image is based on a figure found in [16]

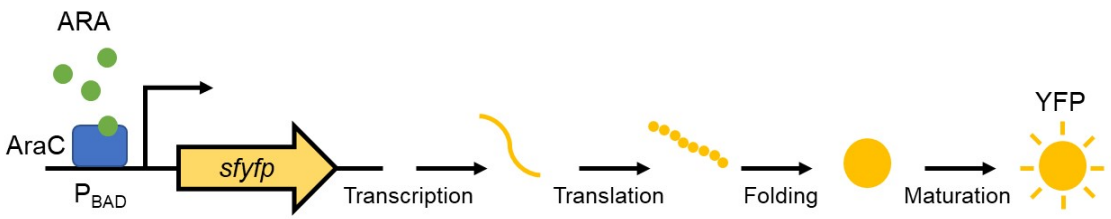

Figure 18: Formation of mature YFP. In the presence of Arabinose (ARA), AraC is activated and promotes the constitutive transcription of YFP. The process of synthesis involves transcription, translation, protein folding and maturation, which accounts for the delay in the emergence of a mature, fluorescing YFP.

for enzymatic degradation in these experiments [3]. Therefore, dilution was the main driver of the decrease in protein number within a cell. Considering the mechanism of constitutive YFP synthesis that involved delays, and the process of protein degradation through dilution, Eq. (9) provides a good representation of YFP dynamics.

4.2 Estimation using YFP trajectories

We worked with data obtained from time-lapse fluorescent microscopy of a population of *E. coli* that expresses a YFP upon induction [15] by Arabinose. In two independent experiments, the population was observed in a microfluidic device, allowing for the recording of flourescence intensity at 1-min intervals from 39 cells and 27 cells, respectively (Fig. 19). As noted above, the addition of Arabinose to the media at time $t = 0$ induced the transcription of YFP within all cells. Following previous work [15, 16], the recorded fluorescent signal was assumed proportional to the number of mature YFP molecules, thus allowing us to estimate the delay in the formation of the mature, fluorescing proteins after induction. In an earlier study, Choi et al. [16] performed a similar analysis assuming that cells in the population are identical. This assumption allowed for increased inference accuracy through the pooling of data across cells observed in an experiment. However, the assumption that all cells are identical may lead to biases in population estimates, and did not allow for the estimation of the variability in reaction rates and delays across the population.

YFP did not saturate in either of the experiments (Fig. 19). As noted earlier in Subsection 3.2.3, we generally need to observe the saturation in protein number to be able to identify both the

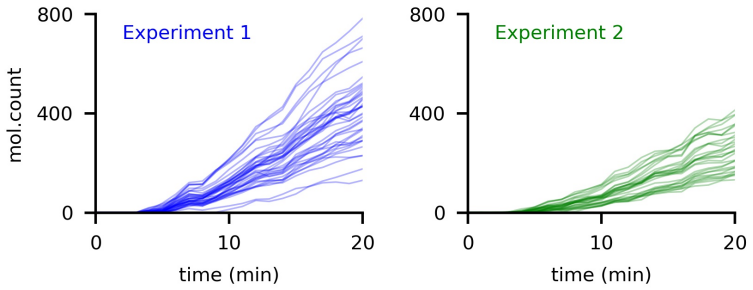

Figure 19: Data from time-lapse images of YFP expression from two independent experiments performed previously by Cheng et al [15]. Trajectory of estimated YFP molecule number were obtained by dividing the total fluorescence level of each cell by a conversion constant.

production and dilution rate. Hence, we specified individual dilution rates B_n, that were separately measured, and used the hierarchical model to estimate the individual production rates, A_n, and birth delays, τ_n. We used the individual dilution rate estimates obtained by Choi et al. [16] for the same data set obtained by tracking the rate of cell growth and division.

We performed inference using our hierarchical model with distributed birth delays. Measurement of fluorescence intensity, and thus of mature YFP protein count, in the two experiments were considerably different, with the second (Fig. 19 green) lower than the first (Fig. 19 blue). As explained by Choi et al. [16], the cause of this discrepancy in the measurements may be due to factors affecting the experimental setup. This difference in the measured fluorescence levels in the two experiments, eventually led to production rate estimates across the population that were higher in the first experiment (Fig. 20a). This difference in the estimated production rates in the two experiments, however did not extend to the mean delay times, as when averaged across the population, the estimates were close: 9.43 and 9.80 min (Fig. 20b).

Our production rate and delay time estimates were higher than estimates previously obtained using a non-hierarchical model with pooling strategies [16]. In particular, our delay time estimates were about 3 min longer (Fig. 21a) than previously reported values. We refer to Subsection 3.3.6, where we applied the two approaches to synthetic data from populations with varying heterogeneity, to explain this observed discrepancy in estimates. As we have shown in Fig. 16, a hierarchical model produced good estimates of individual cell parameters, while a model with pooling strategies led

to an underestimate of both the population birth rate and average delay in the case of a highly heterogeneous cell population. The underestimates produced in the latter case may explain the discrepancy between estimates produced with pooled and unpooled data. This thus provided evidence for the difference in inference method used as a primary contributing factor for the estimate discrepancy.

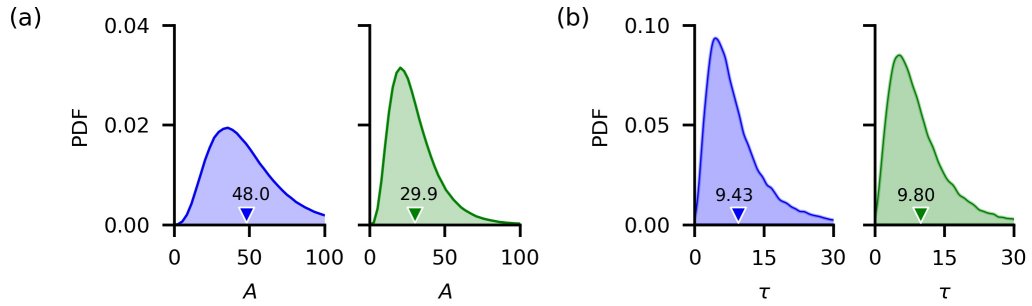

Figure 20: Consistent estimates of the time delay distribution of YFP synthesis after induction. (a-b) We estimated the production rates, A_n, and mean delay times, μ_{τ_n}, for each cell as the mean of the individual posterior distributions, obtained by fixing the dilution rate B_n estimated previously [16]. Because the molecular counts in the first were higher than in the second experiment, the population posterior mean for A was higher for the first. The population mean of the delay distributions are similar in the two experiments (9.43 and 9.80 min, respectively).

As the YFP fluorescence data were gathered from separate but identical experiments, we expected some similarity in the inferred parameters and their variability across the population. The population distribution of mean delay times in the two experiments were very similar, but the posterior distribution the production rates were different (Fig. 20a and b). To quantify this observation, we measured the distance between the population posterior densities of both the production rate, A, and delay time, τ, from the two experiments, by estimating the Kullback-Liebler (KL) divergence from the posterior samples[28] [83]. The KL divergence between the production rate posterior distributions (from the first experiment to the second) was large at 0.43, mainly due to the considerable difference in YFP levels (Fig. 20a) in the two experiments. In contrast, the posterior delay times (Fig. 20b) were almost identical with low KL divergence of 0.005, thereby showing a consistency in the estimation of mean delay times despite the difference in reaction rates. We also found that

[28]The estimator of divergence developed by Wang et al. [83] is based on k-nearest neighbor distances.

at the population level, the birth rate, A, has coefficients of variation (CVs) of 0.52 and 0.55 in the first and second experiment, respectively, while the collection of estimated mean delays, $\hat{\mu}_{\tau_n}$ (Fig. 20d), has CVs of 0.31 and 0.21, respectively. The latter is similar to the CV of 0.20 reported for mean maturation times of YFP measured directly using fluorescent microscopy [15] .

To test for dependence between different parameter pairings we computed the Pearson correlation coefficients (Fig. 21). We did not find any consistent relationships between the parameters, except for a moderate positive correlation between CVs and mean delay times, $\hat{\mu}_{\tau_n}$, of the individual delay distribution (Fig. 20b).

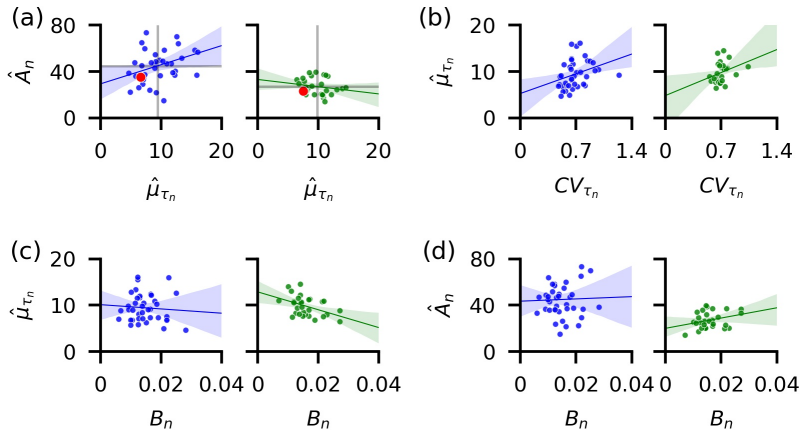

Figure 21: Pearson correlation coefficients reveal no consistent linear relationships between individual parameters in both experiments. (a) Both the average of the production rates and mean delay times (gray lines) are higher than previously reported by Choi et al. (red dots). We found no consistent correlation between $\hat{A}_n$ and $\hat{\mu}_{\tau_n}$ ($\rho = 0.33$ and $\rho = -0.17$) in the two experiments. (b) Individual CVs and $\hat{\mu}_{\tau_n}$ are moderately positively correlated ($\rho = 0.31$ and $\rho = 0.30$ in the first and second experiment, respectively). (c) The dilution rate, B_n, and $\hat{\mu}_{\tau_n}$ have ρ equal to -0.08 and -0.43 in the two experiments, respectively. (d) The reaction rates B_n and A_n, show no clear evidence of correlation with $\rho = -0.03$ and $\rho = 0.30$ in the first and second experiment, respectively. Shaded regions show the 95% confidence interval for the regression estimate.

To further verify the correctness of our estimates, we also generated an ensemble of trajectories using the delay Gillespie algorithm and parameters obtained from samples from the posterior distributions for each individual cell (Fig. 22 and 23). Simulated trajectories matched the experimental data well, with mean trajectories (solid lines in Fig. 22 and 23) that overlap with experimental data.

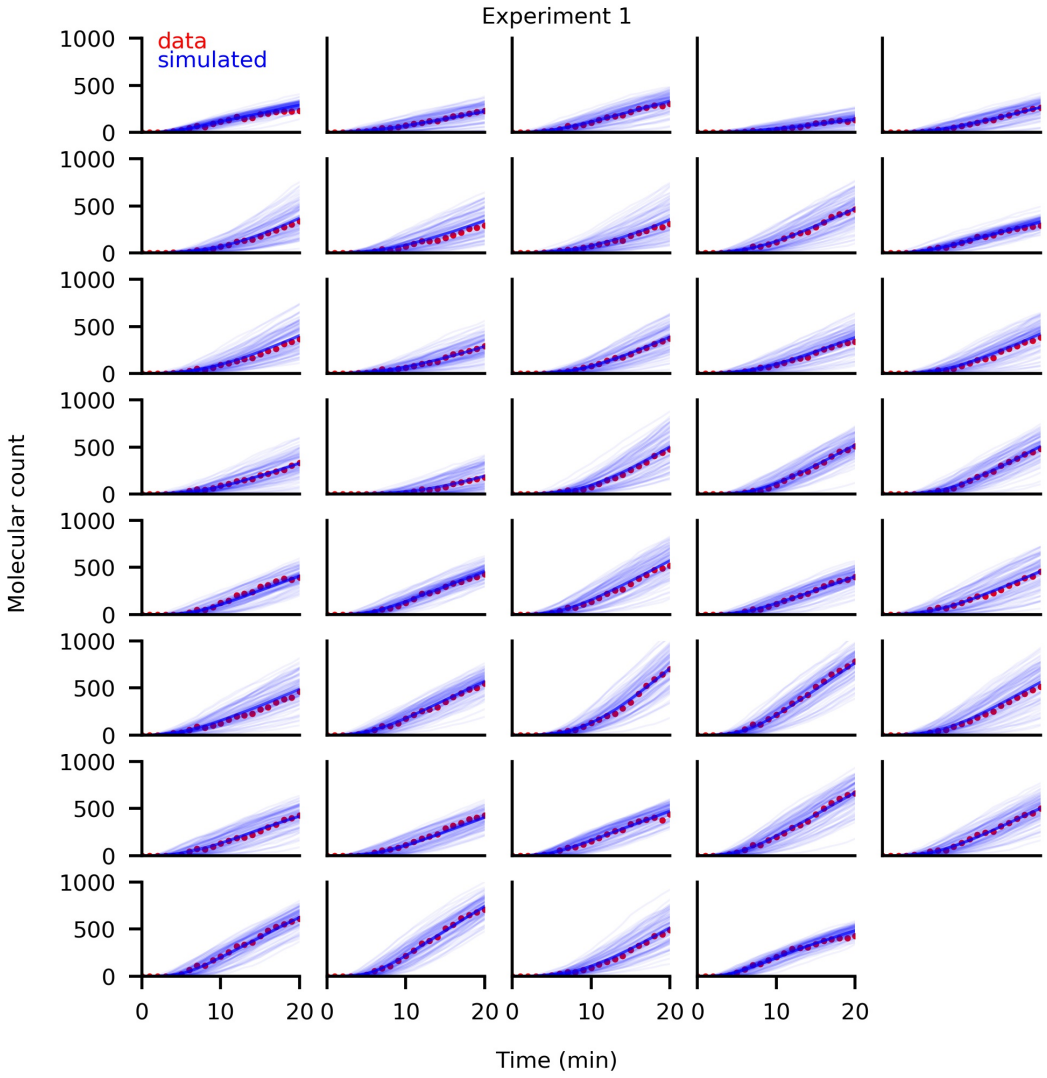

Figure 22: Simulated realizations with estimated parameters fit individual YFP trajectories from the first experiment. We simulated 100 trajectories for each cell by sampling the parameters from the 95% high density interval (HDI) of the posterior distributions, using the delayed Gillespie algorithm [6]. The mean of the realizations (solid lines), per cell, fit the experimental data very well.

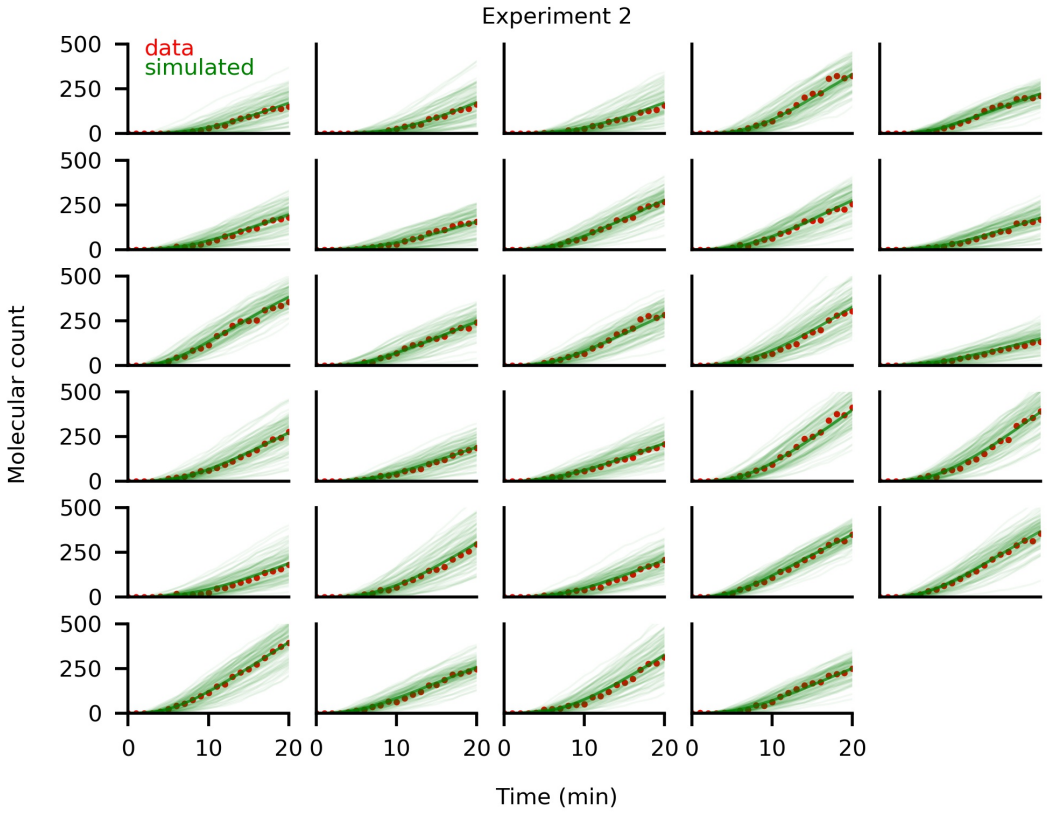

Figure 23: Simulated realizations with estimated parameters fit individual YFP trajectories from the second experiment. We simulated 100 trajectories for each cell by sampling the parameters from the 95% high density interval (HDI) of the posterior distributions, using the delayed Gillespie algorithm [6]. The mean of the realizations (solid lines), per cell, fit the experimental data very well.

4.3 Overfitting analysis

Although trajectories simulated using our parameter estimates showed good agreement with experimental data, this evaluation of the quality of our estimates does not rule out the possibility of overfitting. When dealing with *in silico* data, estimates can be easily compared to generative values to perform assessment, but in the present case, the true parameter values characterizing the experimental system are unknown and thus additional cross-validation and overfitting detection have to be performed.

An important consideration in the subsequent analyses is the complexity of our inference model: In the case of synthetically generated data, the complexity of the model we are fitting is similar to that of the generative model in many cases we considered. As the model we proposed is usually an exact description of the generative process, we do not expect our algorithm to overfit. We typically expect overfitting when the model we fit to data is more flexible or complicated than the the process it is supposed to describe [35]. This is, for instance, the case when there is no delay in the generative model but there is a delay in the model we fit to the data, or when the delay is fixed but the model we are fitting specifies one that is distributed. We have seen the latter scenario previously in Fig. 11, where the inference algorithm overestimated the production rates to accommodate a delay with non-zero variance. On the other hand, if the model is insufficiently flexible or complex, we can expect underfitting, as characterized by the model being unable to explain certain features of the data.

We subscribe to the principle of parsimony: The models we considered provide a minimal description of the underlying biological processes with delays. We therefore did not expect these models to overfit the YFP data, and our tests indicate this is true. We looked at three versions of the hierarchical model:

1. Model HF: the hierarchical model with *fixed* delays inferring the individual delay times, τ_n, and production rates, A_n, and their population distributions;

2. Model HD1: the hierarchical model with *distributed* delays inferring τ_n and A_n, and their

population distributions; and

3. Model HD2: the hierarchical model with *distributed* delays inferring τ_n, A_n, and individual dilution rates, B_n, and their population distributions.

To test performance, we looked at how the predictions obtained using these models generalize to previously unseen data and measured the interpolation and extrapolation errors. We considered three subsets of the experimental data:

1. full 20 min trajectories that are observed at 1-min intervals;

2. full 20 min trajectories that are observed at 2-min intervals; and

3. shorter 15 min trajectories that are observed at 1-min intervals.

The first data subset served as control to show how models perform in a data-rich scenario, and the second and third served to show model interpolation and extrapolation capabilities, respectively.

The fixed delay model, Model HF, failed to capture system behavior in significant portions of the data set, indicating high bias in inference (Fig. 24a and b) and underfitting. Although individual parameter estimates remained realistic and consistent in all data subsets considered (Fig. 24c), simulated trajectories with these parameters did not exhibit the correct initial development of the YFP counts. When we fitted the distributed delay model with the death rates, B_n, specified and fixed, Model HD1 extrapolated well to the unobserved data points (Fig. 25a and b), and was minimally sensitive to small changes in input data (Fig. 25c). With Model HD2, inference of the full parameter set, which now included the death rates, B_n, resulted in good recovery of the trajectories (Fig. 26a and b), but unrealistically large parameter estimates[29] (Fig. 26c), which suggested overfitting.

To measure interpolation and extrapolation errors, we computed the root mean square error of the mean simulated trajectories against the experimental data per individual cell and averaged over

[29]Previous estimates of production rate pegged μ_A at around 35 min^{-1} [16] while experimental estimates for YFP variant VENUS maturation is 7 ± 2.5 min [86]. Estimate from HD2 for μ_A is set at around 180 min^{-1} and μ_τ at around 20 min, which is more than triple the previous estimate for the production rate, and more than double the maturation time from experimental observations.

all cells. Model HD1 produced estimates that were consistently best with the lowest interpolation and extrapolation errors, among the three model versions considered (Fig. 27a orange). Errors for the fixed delay model, Model HF, on the other hand, were consistently high in all data subsets (Fig. 27a blue). Model HD2 showed accuracy that is comparable to Model HD1 in terms of interpolation errors, but was considerably inferior in extrapolation capabilities (Fig. 27a grey). Next, we compared how individual parameter estimates vary across the data subsets we considered. We computed the coefficient of variation of the parameter estimates (Fig. 25c, 24c, and 26c) across the three data subsets per individual, then averaged over all individuals. Model HD1 was largely insensitive to small changes in input data (Fig. 27b orange) with moderate variation in estimates, while the fixed delay model, Model HF, showed small changes (Fig. 27b blue) especially in second experiment. Using the distributed delay model, Model HD2, to infer the full parameter set, produced results that were sensitive to input data resulting in dramatically different parameter estimates for the different data sets (Fig. 27b grey).

These observations pointed to the efficacy of the hierarchical distributed delay model when the death rates, B_n, are specified in inferring parameters of the system under study. Model HD1 extrapolated well to unobserved data, and was largely insensitive to small changes in input data. This confirmed that the hierarchical distributed delay model generalized well and did not overfit, when provided with enough information and data. The fixed delay model, Model HF, showed clear signs of underfitting: extrapolation and interpolation errors were high, and parameter estimates no longer improved even when presented with better data amount and resolution. The hierarchical model that infers the full parameter set, Model HD2, resulted in unrealistically high parameter estimates that exhibited large variations when the data resolution and amount were changed, low interpolation errors, and high extrapolation errors, which when taken together served as evidence for overfitting. This indicated that parameters may not be recovered, and that the model can overfit the data when all parameters need to be inferred [54].

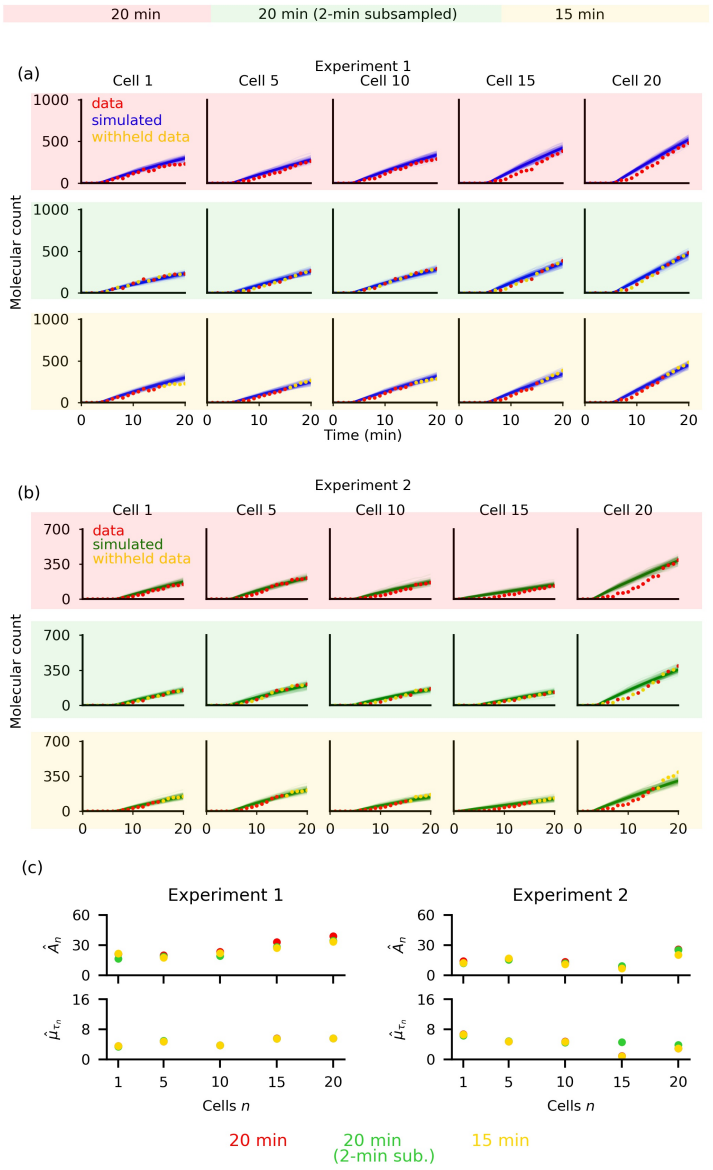

Figure 24: Simulated realizations of the delayed birth–death process with estimated parameters from the hierarchical *fixed* delay model do not exhibit the sigmoidal trajectories that characterize the YFP data. Setting the death rates, B_n, to their true values during inference, we fit the fixed delay model to subsets of the experimental data: full 20 min (red background), 20 min with data subsampled at 2-min intervals (green background), and the first 15 min (yellow background) data. In both experiments 1 (a) and 2 (b), simulated trajectories closely matched the initial and final data points but deviated from the data in the middle of the trajectory. (c) Inference results of five randomly selected cells are shown. Individual estimates of production rates, A_n, and mean delay time, μ_{τ_n}, showed small deviations with the change in data amount and resolution. See Fig. 27 for the analysis of the inference using all cells.

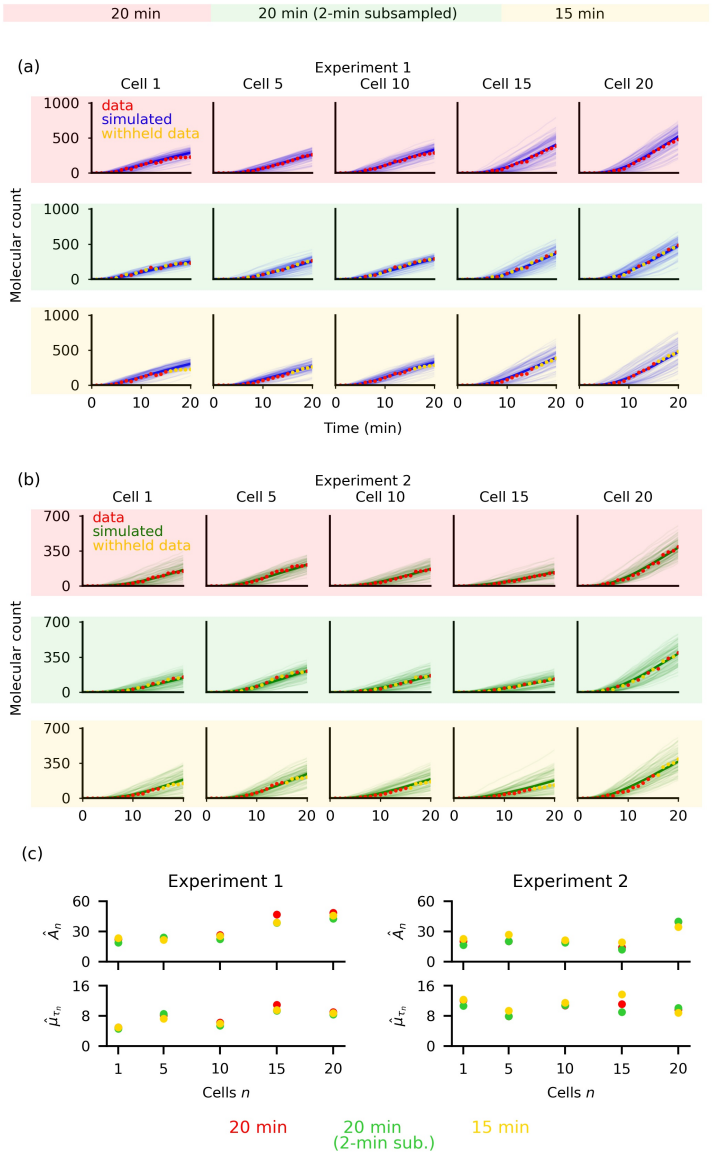

Figure 25: Simulated realizations with estimated parameters from the hierarchical distributed delay model fit individual YFP trajectories even when some data points were withheld during inference. Setting the death rates, B_n, to their true values during inference, we fit the model to subsets of the experimental data: full 20 min (red background), 20 min with data subsampled at 2-min intervals (green background), and the first 15 min (yellow background) data. Simulated trajectories for experiments 1 (a) and 2 (b) using the inferred parameters in all the settings we considered fit data well. (c) Inference results of five randomly selected cells are shown. Individual estimates of production rates, A_n, and mean delay time, μ_{τ_n}, showed small deviations with changes in the data set indicating that inference is robust. See Fig. 27 for the analysis of the inference using all cells.

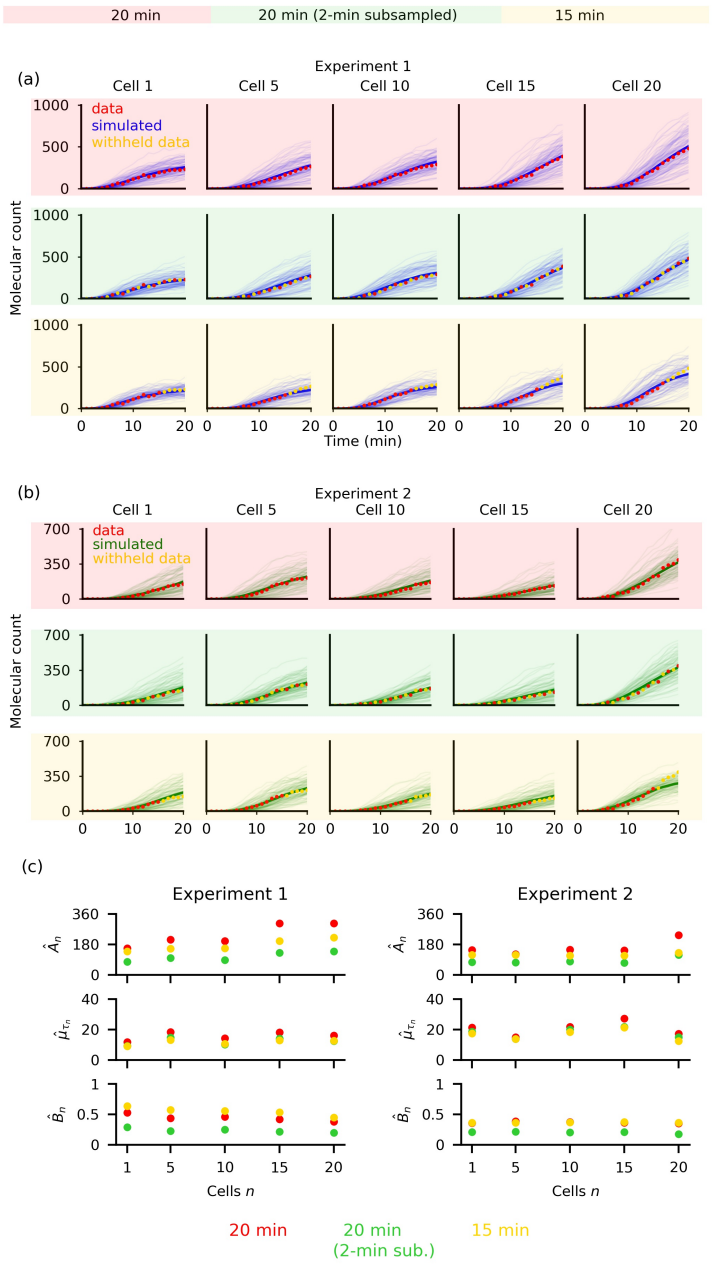

Figure 26: Full parameter set estimation using the hierarchical distributed delay model resulted in unrealistically large estimates that produced simulated realizations which fit individual YFP trajectories well. We fit the model to subsets of the experimental data: full 20 min (red background), 20 min with data subsampled at 2-min intervals (green background), and the first 15 min (yellow background) data. Simulated trajectories for experiments 1 (a) and 2 (b) using the inferred parameters across all settings we considered fit data well. (c) Inference results of five randomly selected cells are shown. Individual estimates of production rates, A_n, mean delay time, μ_{τ_n}, and death rate, B_n, all are unrealistically large. See Fig. 27 for the analysis of the inference using all cells.

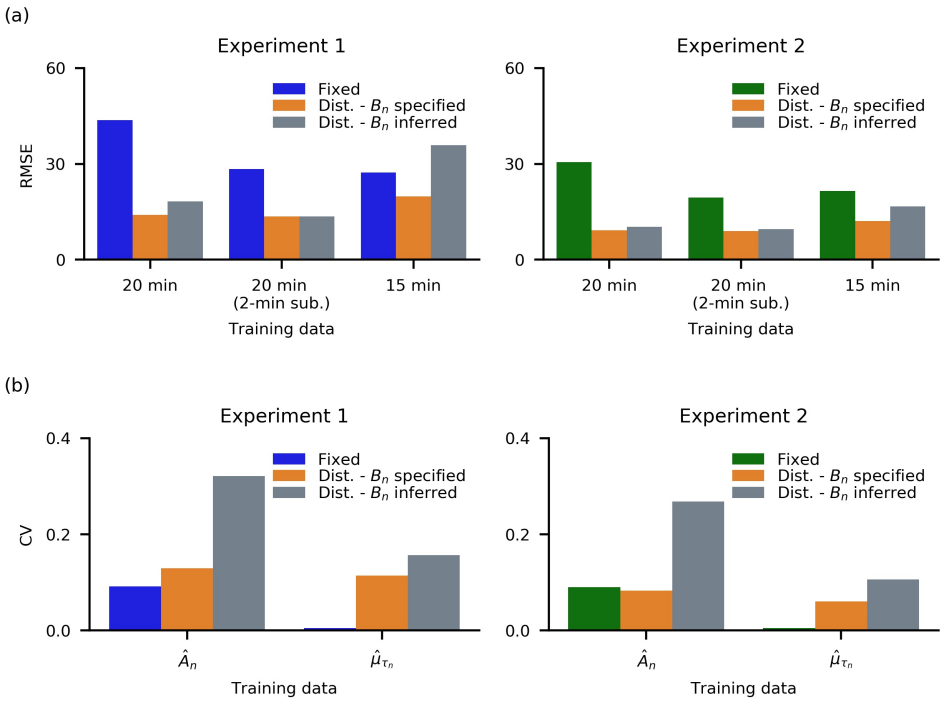

Figure 27: Fixed delay and unspecified death rate lead to underfitting and overfitting respectively. (a) We computed the root mean square error (RMSE) of the mean simulated trajectories from the experimental data per individual cell, and averaged over all cells. In both experiments 1 (left) and 2 (right), the RMSE remained low with small changes in the case of the distributed delay model where B_n was specified. In the case of the fixed delay model the error unexpectedly increased with the amount of data used to infer the parameters and hyperparameters, indicating a larger bias. Inference of the full parameter set (including B_n) using the distributed delay hierarchical model resulted in larger RMSEs compared to when B_n was specified. (b) We computed the coefficient of variation (CV) of the parameter estimates (Fig. 24c, 25c, and 26c) across the different data subsets per individual, then averaged over all individuals. The fixed delay model showed the least variation among the models then followed by the distributed model with B_n specified. The distributed model where all parameters were inferred exhibited the largest variation in all parameters.

4.4 Summary

We tested the performance of our hierarchical model on fluorescence data from an experimentally observed clonal population of *E. coli*. In this experiment [15], when Arabinose is added to the media at time $t = 0$, the transcription of YFP is induced, followed by translation and post-translational modifications. The time for the completion of this sequence of reactions after induction comprises the delay time in the formation of mature fluorescing proteins. The number of mature YFP is not directly counted, and is instead assumed to be proportional to the observed fluorescence intensity.

The dynamics of YFP count can mainly be described by two processes: protein production with delayed completion, and decrease in count due to dilution. As such, a delayed birth-death process is a good model of this dynamics. We dealt with data from two independent experiments. During the observation window, the trajectories did not saturate and so we needed to specify the dilution rates instead of inferring them simultaneously with the other parameters. We used the dilution rate estimates obtained by Choi et al. [16] in a previous study. Individual parameters we obtained for the production rates and delay times were higher than previously reported values derived from an implementation of a non-hierarchical model using pooling strategies [16]. We argued that this difference in estimates is caused by the difference in the inference methods used, as the same level of discrepancy can be observed when dealing with *in silico* data with high population heterogeneity. A pairwise test for linear dependence between the individual parameters showed no strong relationships between the parameters.

We characterized the population distribution of both production rates and delay times and compared the results in the two experiments by estimating the KL-divergence of the population posterior distributions. Because of the difference in fluorescence intensities in the two setups, the distance between the distributions of the production rates was high. Despite this difference, the distribution of delay times were almost identical with very low KL-divergence.

As a form of evaluation of estimate quality, we generated an ensemble of realizations of the birth-death process for every individual using samples from the obtained posterior distributions of the parameters. The trajectories showed good agreement with the experimental data, but additional

checks for overfitting still had to be performed. We found that when the dilution rates are specified, a hierarchical model with distributed delay does not overfit as it is highly insensitive to changes in input data, and produces estimates that result to low interpolation and extrapolation errors. A fixed delay model, on the other hand, underfitted the data as it cannot explain a considerable segment of the YFP trajectories. When we tried to estimate the full parameter set, the hierarchical model with distributed delay displayed signs of overfitting, as characterized by low interpolation but high extrapolation errors.

5 Conclusions

In this work, we have developed a hierarchical Bayesian model for the inference of parameters of a biochemical chemical reaction network with delays, and their distribution in a cell population. We have shown that our inference framework produces accurate and robust estimates of reaction rates, reaction completion time delays, and their population variability even when the stochastic process of interest is discretely-observed, as in experimental measurements of gene regulatory networks obtained via fluorescent microscopy.

We considered a stochastic birth-death process with birth delays to demonstrate the performance and limitations of our method. This process, although simple, provides a minimal model for the experimental system we studied, and also is a building-block for more complex biochemical reaction networks. In this simple setting, we have shown that our method can be used for the simultaneous inference of the individual-level parameters characterizing the cells, and in the quantification of the cell-to-cell variability of these characteristics. Our method for deriving the approximate likelihood and posterior distribution is general, so that the techniques we developed for the inference of a birth-death process readily extends to more complex systems. In systems with more complicated structures, however, some other challenges in computation and identifiability may arise, ones which we may not have encountered in our test scenarios.

In our analyses, we looked at models with fixed and distributed delays. The fixed delay model has fewer parameters, is easier to implement, faster to run, and performs better when measurements are generated using fixed delay processes. However, using the fixed delay model for inference leads to underestimates of birth reaction rates and delay times at the individual and population levels when reaction delays are in fact variable in time. The hierarchical distributed delay model, on the other hand, is applicable more widely. It works well with high-resolution data, but at the expense of high computational cost. We also showed when data is sparsely sampled, the ensemble estimation allows our hierarchical model to outperform its non-hierarchical counterpart.

Our inference model is applicable to experimental data that can be obtained using fluorescent

microscopy. In such experiments, cells can vary in growth rates, size, plasmid copy number, and other factors [2, 64, 71, 73, 74]. Hierarchical models explicitly account for such heterogeneity and thus provide a suitable framework for inferring the variability and covariability of biochemical rates and delays across the population [37, 78, 87]. Our estimates of the reaction rates and delays were higher than previous estimates [16]. This discrepancy may be due to biases introduced by data pooling without considering cell heterogeneity, as in the case of model implementation with synthetic data. Despite this difference in the parameter estimates, we were able to closely recreate the observed YFP molecular counts for all the cells using the resulting values from our hierarchical approach. Thus, we expect that our robust hierarchical approach can be extended to quantify the variability and covariability of complex gene network dynamics from experimental data.

Even in the simple cases that we considered, computational cost already comes up as an issue. A distributed delay model, for instance, has four parameters per individual that needed to be inferred for dozens of cells, alongside population distribution hyperparameters that characterize each parameter type. Model dimension increases with the complexity of the biological system, and with it also grow the required computational resources to perform inference. Related to the issue of model dimension increase is the selection of a sampling algorithm that can efficiently explore the parameter space. As one may immediately conclude, an inefficient algorithm may encounter problems in the convergence of the sample chain to the posterior distribution. Alternative algorithms such as Hamiltonian Monte Carlo [59, 67], variational approaches [44, 81], or machine learning methods [18] can be considered by future studies to address these issues.

One key assumption in our model is that cells are independent: We assumed individual cell parameters are independent samples from some population distribution. While this assumption worked well in the experimental system that we studied, where cells are not closely related, our method does not apply to cell lineages where daughter cells have been shown to exhibit correlated gene expression for a considerable length of time after division [80]. Such a setting necessitates a significant change in the paradigm that we followed in this study: For instance, the collection of individual rate parameters for the same reaction can be viewed as a realization of a random

vector that is described by a multi-dimensional probability distribution, with some covariance function that accounts for spatio-temporal relationships. This approach, however, requires strong assumptions about the spatio-temporal structure of the parameters [19]. Research focus now shifts from determining how different cell parameters are across the population, to how long parameter correlations are maintained in the lineage.

Cell populations may exhibit different levels of heterogeneity. The population may be homogeneous where all cells are identical, or cell-to-cell variable as we assumed in this study. While we have established the effectiveness of our approach in the later case, we have not explored how the model performs in cases when subpopulations [56] are present. As descriptions for subpopulations are commonly given in terms of mixture distributions, developing an inference framework for such a scenario may require particular steps that are not present in this work.

In sum, we have developed a general hierarchical Bayesian inference framework for delayed chemical reaction systems that resulted to accurate and robust estimates for the cases we considered. While we limited our model analysis to a stochastic birth-death process, this simple case is a fundamental component of more complex biochemical networks and our methods easily extend to these cases. Future work can look into applications in different processes, address issues related to model dimensionality, and consider exploring inference for cell lineages and different levels of heterogeneity.

Bibliography

[1] Abramowitz, M., and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover Books on Advanced Mathematics, New York, 1965.

[2] Altschuler, S., and Wu, L. Cellular heterogeneity: do differences make a difference? *Cell 141(4)* (2010), 559–563.

[3] Andersen, J., Sternberg, C., Poulsen, L., Bjorn, S., Givskov, M., and Molin, S. New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol. 64(6)* (1998), 2240–2246.

[4] Austin, D., Allen, M., McCollum, J., Dar, R., Wilgus, J., Sayler, G., Samatova, N., Cox, C., and Simpson, M. Gene network shaping of inherent noise spectra. *Nature 439* (2006), 608–611.

[5] Barrio, M. Reduction of chemical reaction networks through delay distributions. *J. Chem. Phys. 138* (2013), 104114.

[6] Barrio, M., Burrage, K., Leier, A., and Tian, T. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Comput. Biol. 2(9)* (2006), e117.

[7] Bel, G., Munsky, B., and Nemenman, I. The simplicity of completion time distributions for common complex biochemical processes. *Phys. Biol. 7(1)* (2010), 016003.

[8] Blyuss, K., and Kyrychko, Y. Stability and bifurcations in an epidemic model with varying immunity period. *Bull. Math. Biol. 72(2)* (2010), 490–505.

[9] Bowsher, C., and Swain, P. Identifying sources of variation and the flow of information in biochemical networks. *PNAS 109(20)* (2012), E1320–E1328.

[10] Boys, R., Wilkinson, D., and Kirkwood, T. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing 18(2)* (2008), 125–135.

[11] Browning, A., Warne, D., Burrage, K., Baker, R., and Simpson, M. Identifiability analysis for stochastic differential equation models in systems biology. *J. R. Soc. Interface 17* (2020), 20200652.

[12] Caia, X., and Xu, Z. K-leap method for accelerating stochastic simulation of coupled chemical reactions. *The Journal of Chemical Physics 126* (2007), 074102.

[13] Calderazzo, S., Brancaccio, M., and Finkenst adt, B. Filtering and inference for stochastic oscillators with distributed delays. *Bioinformatics 35(8)* (2019), 1380–1387.

[14] Chen, Y., Kim, J., Hirning, A., Josić, K., and Bennett, M. Emergent genetic oscillations in a synthetic microbial consortium. *Science 349(6251)* (2015), 986–989.

[15] Cheng, Y., an K. Josić, A. H., and Bennett, M. The timing of transciptional regulation in synthetic gene circuits. *ACS Synth. Biol. 6(11)* (2017), 1996–2002.

[16] Choi, B., Cheng, Y., Cinar, S., Ott, W., Bennett, M., Josić, K., and Kim, J. Bayesian inference of distributed time delay in transcriptional and translational regulation. *Bioinformatics 36(2)* (2020), 586–593.

[17] Congdon, P. *Bayesian hierarchical models with applications using R, 2nd ed.* CRC Press, New York, 2020.

[18] Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *PNAS 117(48)* (2020), 30055–30062.

[19] Cressie, N. A. C. *Statistics for spatial data, revised edition.* John Wiley & Sons, Inc, New Jersey, 1993.

[20] Frank, T., and Beek, P. Stationary solutions of linear stochastic delay differential equations: Applications to biological systems. *Phys. Rev. E 64* (2001), 021917.

[21] Frank, T., Beek, P., and Friedrich, R. Fokker-Planck perspective on stochastic delay systems: Exact solutions and data analysis of biological systems. *Phys. Rev. E 68* (2003), 021912.

[22] Fritz, G., Megerle, J., Westermayer, S., Brick, D., Heermann, R., Jung, K., Rädler, J., and Gerland, U. Single cell kinetics of phenotypic switching in the arabinose utilization system of *E. Coli*. *PLoS One 9(2)* (2014), 2140–2145.

[23] Gábor, A., Villaverde, A., and Banga, J. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Syst. Biol. 11* (2017), 54.

[24] Garcia, H., Tikhonov, M., Lin, A., and Gregor, T. Quantitative imaging of transcription in living drosophila embryos links polymerase activity to patterning. *Curr. Biol. 23(21)* (2013), 2140–2145.

[25] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis, 3rd ed.* Chapman and Hall/CRC, 2013.

[26] Geman, S., and Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6* (1984), 721–741.

[27] Gillespie, D. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry 81* (1977), 2340–2361.

[28] Gillespie, D. Approximate accelerated simulation of chemically reacting systems. *The Journal of Chemical Physics 115* (2001), 1716–1733.

[29] Golding, I., Paulsson, J., Zawilski, S., and Cox, E. Real-time kinetics of gene activity in individual bacteria. *Cell 123(6)* (2005), 1025–1036.

[30] Gomez, M. The effects of time-varying temperature on delays in genetic networks. *SIAM J. Appl. Dyn. Syst. 15* (2016), 1734–1752.

[31] Gopalakrishnan, A., Kaisare, N., and Narasimhan, S. Incorporating delayed and infrequent measurements in extended kalman filter based nonlinear state estimation. *Journal of Process Control 21(1)* (2011), 119–129.

[32] Gupta, C. Modeling delay in genetic networks: From delay birth-death processes to delay stochastic differential equations. *J. Chem. Phys. 140* (2014), 204108.

[33] Hasenauer, J., Waldherr, S., Radde, N., Doszczak, M., Scheurich, P., and Allg ower, F. A maximum likelihood estimator for parameter distributions in heterogeneous cell populations. *Procedia Computer Science 1* (2012), 1655–1663.

[34] Hastings, W. Monte carlo sampling methods using markov chains and their applications. *Biometrika 57* (1970), 97–109.

[35] Hawkins, D. The problem of overfitting. *J. Chem. Inf. Comput. Sci. 44(1)* (2004), 1–12.

[36] Heron, E., Finkenst adt, B., and Rand, D. Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics 23(19)* (2007), 2596–2603.

[37] Heydari, J., Lawless, C., Lydall, D., and Wilkinson, D. Bayesian hierarchical modelling for inferring genetic interactions in yeast. *J. R. Stat. Soc. Ser. C Appl. Stat. 65(3)* (2016), 367–393.

[38] Higham, D. Modeling and simulating chemical reactions. *SIAM Review 50(2)* (2008), 347–368.

[39] Hilfinger, A., and Paulsson, J. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *PNAS 108(29)* (2011), 12167–12172.

[40] Hines, K., Middendorf, T., and Aldrich, R. Determination of parameter identifiability in nonlinear biophysical models: a Bayesian approach. *J. Gen. Physiol. 143(3)* (2014), 401–416.

[41] Jahnke, T., and Huisinga, W. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology 54* (2007), 1–26.

[42] Jiang, Q., Fu, X., Yan, S., Li, R., Du, W., Cao, Z., Qian, F., and Grima, R. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun. 12* (2021), 2618.

[43] Johnson, N. L., and Kotz, S. *Discrete distributions. In: Distributions in Statistics.* Wiley, New York, 1969.

[44] Jordan, M., Z. Ghahramani, T. J., and Saul, L. Introduction to variational methods for graphical models. *Machine Learning 37* (1999), 183–233.

[45] Josić, K., López, J., Ott, W., Shiau, L., and Bennett, M. Stochastic delay accelerates signaling in gene networks. *PLoS Comput. Biol. 7(11)* (2011), e1002264.

[46] Kaern, M., Elston, T., Blake, W., and Collins, J. Introduction to variational methods for graphical models. *Nat. Rev. Genet. 6* (2005), 451–464.

[47] Koeppl, H., Zechner, C., Ganguly, A., and Pelet, S. Accounting for extrinsic variability in the estimation of stochastic rate constants. *Int. J. Robust Nonlinear Control 22(10)* (2012), 1103–1119.

[48] Korsbo, N., and Jönsson, H. It's about time: Analysing simplifying assumptions for modelling multi-step pathways in systems biology. *PLoS Comput. Biol. 16(6)* (2020), e1007982.

[49] Kruschke, J. K. *Doing Bayesian data analysis : a tutorial with R, JAGS, and Stan, 2nd ed.* Elsevier Inc., 2015.

[50] Krzyzanski, W. Ordinary differential equation approximation of gamma distributed delay model. *Pharmacokinet. Pharmacodyn. 46(1)* (2020), 53–63.

[51] Kyrychko, Y., Blyuss, K., and Schöll, E. Amplitude and phase dynamics in oscillators with distributed-delay coupling. *Phil. Trans. R. Soc. A 371* (2013), 20120466.

[52] Leier, A. The effects of time-varying temperature on delays in genetic networks. *J. R. Soc. Interface 11* (2014), 20140108.

[53] Leone, F., Nelson, L., and Nottingham, R. The folded normal distribution. *Technometrics 3(4)* (1961), 543–550.

[54] Lever, J., Krzywinski, M., and Altman, N. Model selection and overfitting. *Nat. Methods 13* (2016), 703–704.

[55] Loinger, A., Lipshtat, A., Balaban, N., and Biham, O. Stochastic simulations of genetic switch systems. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 75(2)* (2007), 021904.

[56] Loos, C., Moeller, K., Fr ohlich, F., Hucho, T., and Hasenauer, J. A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Systems 6(5)* (2018), 593–603.

[57] MacAdams, H., and Shapiro, L. Circuit simulation of genetic networks. *Science 269(5224)* (1995), 650–656.

[58] McAdams, H., and Arkin, A. Stochastic mechanisms in gene expression. *PNAS 94(3)* (1997), 814–819.

[59] McKay, D. J. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, New York, 2003.

[60] Megerle, J., Fritz, G., Gerland, U., Jung, K., and Rädler, J. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys. J. 95(4)* (2008), 2103–2115.

[61] Mehrkanoon, S., Mehrkanoon, S., and Suykens, J. Parameter estimation of delay differential equations: An integration-free ls-svm approach. *Commun. Nonlinear Sci. Numer. Simul. 19(4)* (2014), 830–841.

[62] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21* (1953), 1087–1092.

[63] Michaelis, L., and Menten, M. Die kinetik der invertinwirkung. *Biochem. Z. 49* (1913), 333–369.

[64] Mitchell, S., Roy, K., Zangle, T., and Hoffmann, A. Nongenetic origins of cell-to-cell variability in b lymphocyte proliferation. *PNAS 115(12)* (2018), E2888–E2897.

[65] Moala, F., Ramos, P., and Achcar, J. Bayesian inference for two-parameter gamma distribution assuming different noninformative priors. *Rev. Colomb. Estad. 36(2)* (2013), 321–338.

[66] Munsky, B., Neuert, G., and van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science 336(6078)* (2012), 183–188.

[67] Neal, R. M. *MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo.* Chapman & Hall/CRC, New York, 2011.

[68] Pradhan, B., and Kundu, D. Bayes estimation and prediction of the two-parameter gamma distribution. *J. Stat. Comput. Simul. 81(9)* (2011), 1187–1198.

[69] Psarakis, S., and Panaretos, J. On some bivariate extensions of the folded normal and folded t distributions. *Journal of Applied Statistical Sciences 10(2)* (2001), 119–136.

[70] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingm uller, U., and Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics 25(15)* (2009), 1923–1929.

[71] Rinott, R., Jaimovich, A., and Friedman, N. Exploring transcription regulation through cell-to-cell variability. *PNAS 108(15)* (2011), 6329–6334.

[72] Schlicht, R., and Winkler, G. A delay stochastic process with applications in molecular biology. *J. Math. Biol. 57* (2008), 613–648.

[73] Sherman, M., Lorenz, K., Hunter, M., Barak, L., and Cohen, A. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell Syst. 1(5)* (2016), 315–325.

[74] Snijder, B., and Pelkmans, L. Origins of regulated cell-to-cell variability. *Cell Biol. 12* (2011), 119–125.

[75] Thattai, M., and van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *PNAS 98(15)* (2001), 8614–8619.

[76] Tian, T., Xu, S., Gao, J., and Burrage, K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics 23(1)* (2007), 84–91.

[77] Tokuda, I., Akman, O., and Locke, J. Reducing the complexity of mathematical models for the plant circadian clock by distributed delays. *J. Theor. Biol. 463* (2019), 155–166.

[78] Tonner, P., Darnell, C., Bushell, F., Lund, P., Schmid, A., and Schmidler, S. A Bayesian non-parametric mixed-effects model of microbial growth curves. *PLoS Comput. Biol. 16(10)* (2020), e1008366.

[79] Turner, B., and Zandt, T. V. Hierarchical approximate Bayesian computation. *Psychometrika 79(2)* (2014), 185–209.

[80] Veliz-Cuba, A., Hirning, A., Atanas, A., Hussain, F., Vancia, F., and Josić, K. Sources of variability in a synthetic gene oscillator. *PLoS Comput. Biol. 11(12)* (2015), e1004674.

[81] Wainwright, M., and Jordan, M. Graphical models, exponential families, and variational inference. *JABES 1(1-2)* (2008), 1–305.

[82] Wang, L., and Cao, J. Estimating parameters in delay differential equation models. *JABES 17* (2012), 68–83.

[83] Wang, Q., Kulkarni, S., and Verdú, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distance. *IEEE Trans. Inf. Theory 55(5)* (2009), 2392–2405.

[84] Warne, D., Baker, R., and Simpson, M. Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *arXiv:1812.05759v1* (2018).

[85] Wilkinson, D. J. *Stochastic modelling for systems biology.* Addison-Wesley Professional, 2011.

[86] Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, S. Probing gene expression in live cells, one protein molecule at a time. *Science 311(5767)* (2006), 1600–1603.

[87] Zechner, C., Unger, M., Pelet, S., Peter, M., and Koeppl, H. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat. Methods 11(2)* (2014), 197–202.

[88] Zellner, A. *Bayesian methods and entropy in economics and econometrics. In, Grandy W.T. and Schick,L.H. (eds), Maximum entropy and Bayesian methods*, vol. 43. Springer, Dordrecht, 1991.

A Sampling from population distributions and individual delay distributions

Here we present the algorithm for generating the posterior population distributions and the individual delay distributions found throughout the main text. Population distributions are not directly sampled in the algorithm, and we instead sample from the posterior distribution of the hyperparameters. Here, we present the algorithm we used to sample from the population marginal posteriors of the parameters A, α, β, and the delay time τ. We also apply the same strategy to sample values from the individual delay distributions.

Through the MCMC algorithm we obtain the hyperparameter posterior distributions $\pi(a_Z)$ and $\pi(b_Z)$, for $Z \in \{A, B, \tau\}$ in the fixed delay, and $Z \in \{A, B, \alpha, \beta\}$ in the distributed delay case. We employed the algorithm below to sample from the respective population distributions.

1. Take m samples a_Z^s from $\pi(a_Z)$.
2. Take m samples b_Z^s from $\pi(b_Z)$.
3. For each pair (a_Z^s, b_Z^s), take p samples from $\Gamma(a_Z^s, b_Z^s)$.
4. Combine all the mp samples taken from step 3. This pooled samples are realizations of the population distribution of the reaction rate or delay parameter Z.

In the distributed delay case, a similar algorithm was also applied to sample from the individual delay posterior distributions. For an individual n, the algorithm infers the posteriors $\pi(\alpha_n)$ and $\pi(\beta_n)$. We sample from $\pi(\tau_n)$ as follows.

1. Take m samples α_n^s from $\pi(\alpha_n)$.
2. Take m samples β_n^s from $\pi(\beta_n)$.
3. For each pair (α_n^s, β_n^s), take p samples from $\Gamma(\alpha_n^s, \beta_n^s)$.
4. Combine all the mp samples taken from step 3. The pooled samples are realizations of the marginal posterior of the individual delay τ_n.

We typically used $m = 1,000,000$ and $p = 1$ to generate the figures.

B Non-informative hyperpriors for the hierarchical distributed delay model and non-informative priors for its non-hierarchical counterpart

The hierarchical model requires the specification of hyperpriors for all the hyperparameters that describe population variation. In the generative model for the distributed delay case (Fig. 6b), we have four pairs of hyperparameters (a, b), one for each of A, B, α, and β. As information about these hyperparameters may be scarce, especially in real biological systems, specifying non-informative hyperpriors will sometimes be appropriate.

In model implementation, we assumed that the death rates B_n are known and so no longer inferred their population distributions. For the rate parameter, A_n, we specified a rational prior with form $\pi(a_A, b_A) = \dfrac{1}{b_A}$. For the delay parameters α and β, we first considered rational priors of the form $\pi(a_\alpha, b_\alpha) = \dfrac{1}{b_\alpha}$ and $\pi(a_\beta, b_\beta) = \dfrac{1}{b_\beta}$, and afterwards tested changes in estimate accuracy when these are replaced by the MDIP (32).

In the comparison done between the hierarchical and non-hierarchical models (See Fig. 12), we implemented both cases all with non-informative hyperpriors and priors, respectively. For the hierarchical model, we specified a rational joint hyperprior for the pair (a_A, b_A), and MDIP for both the pairs (a_α, b_α) and (a_β, b_β). Similar to Choi et al. [16], we specified non-informative gamma priors, $\Gamma(0.001, 0.001)$, for all parameters A_n, α_n, and β_n, in the implementation of the non-hierarchical model.

C Derivation of marginal posterior distributions for a non-hierarchical model with data pooling

Consider a collection, $\mathbf{y} = \{\mathbf{y}_n\}_{n=1,\ldots,N}$, of N independent realizations of a birth-death process with delays,

$$\emptyset \xrightarrow[\tau]{A} Y \xrightarrow{B} \emptyset,$$

where the birth delay τ follows a gamma distribution $\Gamma(\alpha, \beta)$. For each realization n, we denote its subset of discrete-time observations as $\mathbf{y}_{d,n} = (y_n(0), y_n(1), \ldots, y_n(T-1), y_n(T))$. We assign the following gamma priors for each of the parameters:

$$A \sim \Gamma(a_A, b_A)$$
$$B \sim \Gamma(a_B, b_B)$$
$$\alpha \sim \Gamma(a_\alpha, b_\alpha)$$
$$\beta \sim \Gamma(a_\beta, b_\beta).$$

We denote by r_{n1i} the number of birth reaction which completed at interval $(i, i+1]$ in the n^{th} realization, and by r_{n2i} the number of death reactions. Following the likelihood (35), the resulting posterior distribution with the above-specified priors is

$$\begin{aligned}
\pi(A, B, \alpha, \beta \,|\mathbf{y}_d) \propto & \prod_{n=1}^{N}\prod_{i=0}^{T-1} \frac{\left(A\int_i^{i+1} \frac{\gamma(\alpha, \beta\hat{t})}{\Gamma(\alpha)} d\hat{t}\right)^{r_{n1i}}}{r_{n1i}!} \exp\left(-A\int_i^{i+1} \frac{\gamma(\alpha, \beta\hat{t})}{\Gamma(\alpha)} d\hat{t}\right) \\
& \times \prod_{n=1}^{N}\prod_{i=0}^{T-1} \frac{\left[\frac{1}{2}B(y_n(i+1) + y_n(i))\right]^{r_{n2i}}}{r_{n2i}!} \exp\left(-\frac{1}{2}B(y_n(i+1) + y_n(i))\right) \\
& \times A^{a_A - 1}\exp(-Ab_A) B^{a_B - 1}\exp(-Bb_B) \\
& \times \alpha^{a_\alpha - 1}\exp(-\alpha b_\alpha)\beta^{a_\beta - 1}\exp(-\beta b_\beta).
\end{aligned}$$

With this, we derive the conditional marginal posteriors for the rate parameters which both follow a gamma distribution:

$$A \sim \Gamma\left(\sum_{n=1}^{N}\sum_{i=0}^{T-1} r_{n1i} + a_A, N\sum_{i=0}^{T-1}\int_i^{i+1}\frac{\gamma(\alpha,\beta\hat{t})}{\Gamma(\alpha)}d\hat{t} + b_A\right)$$
$$B \sim \Gamma\left(\sum_{n=1}^{N}\sum_{i=0}^{T-1} r_{n2i} + a_B, \sum_{n=1}^{N}\sum_{i=0}^{T-1}\frac{1}{2}\left(y_n(i) + y_n(i+1)\right) + b_B\right).$$

The posteriors for the delay parameters, on the other hand, do not follow known distributions but are proportional to

$$\alpha\,|\mathbf{y}_d, A, B, \beta \propto \prod_{n=1}^{N}\prod_{i=0}^{T-1}\left[\int_i^{i+1}\frac{\gamma(\alpha,\beta\hat{t})}{\Gamma(\alpha)}d\hat{t}\right]^{r_{n1i}}\exp\left(-NA\sum_{i=0}^{T-1}\int_i^{i+1}\frac{\gamma(\alpha,\beta\hat{t})}{\Gamma(\alpha)}d\hat{t}\right)\alpha^{a_\alpha-1}\exp(-\alpha b_\alpha)$$
$$\beta\,|\mathbf{y}_d, A, B, \alpha \propto \prod_{n=1}^{N}\prod_{i=0}^{T-1}\left[\int_i^{i+1}\frac{\gamma(\alpha,\beta\hat{t})}{\Gamma(\alpha)}d\hat{t}\right]^{r_{n1i}}\exp\left(-NA\sum_{i=0}^{T-1}\int_i^{i+1}\frac{\gamma(\alpha,\beta\hat{t})}{\Gamma(\alpha)}d\hat{t}\right)\beta^{a_\beta-1}\exp(-\beta b_\beta).$$