Computational Statistics and Data Analysis 213 (2026) 108251

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Inferring the dynamics of quasi-reaction systems via nonlinear local mean-field approximations

Matteo Framba [a], Veronica Vinciotti [a,*], Ernst C. Wit [b]

[a] Department of Mathematics, University of Trento, Trento, Italy
[b] Institute of Computing, Università della Svizzera italiana, Lugano, Switzerland

ARTICLE INFO

*Keywords:*
Quasi-reaction systems
Ordinary differential equations
Mean-field approximation
Taylor approximation

ABSTRACT

Parameter estimation of kinetic rates in stochastic quasi-reaction systems can be challenging, particularly when the time gap between consecutive measurements is large. Local linear approximation approaches account for the stochasticity in the system but fail to capture the intrinsically nonlinear nature of the mean dynamics of the process. Moreover, the mean dynamics of a quasi-reaction system can be described by a system of ODEs, which have an explicit solution only for simple unitary systems. An approximate analytical solution is derived for generic quasi-reaction systems via a first-order Taylor approximation of the hazard rate. This allows a nonlinear forward prediction of the future dynamics given the current state of the system. Predictions and corresponding observations are embedded in a nonlinear least-squares approach for parameter estimation. The performance of the algorithm is compared to existing methods via a simulation study. Besides the generality of the approach in the specification of the quasi-reaction system and the gains in computational efficiency, the results show an improvement in the kinetic rate estimation, particularly for data observed at large time intervals. Additionally, the availability of an explicit solution makes the method robust to stiffness, which is often present in biological systems. Application to Rhesus Macaque data illustrates the use of the method in the study of cell differentiation.

## 1. Introduction

Reaction networks are an efficient modeling framework used to describe the population evolution in many biological and biochemical phenomena. These systems are typically modeled using stochastic differential equations, which effectively capture the inherent uncertainty and randomness of the underlying biological structure (Golightly and Wilkinson, 2005). Understanding the dynamics of a process requires a thorough knowledge of the evolution of its moments, typically obtained through the chemical master equation (Schnakenberg, 1976). The primary objective of many studies is to infer the parameters governing these moments. In the context of SDE-based methods, this is often achieved using Local Linear Approximation (LLA) due to its efficiency and ease of implementation. However, LLA methods have been found to be inaccurate when data are obtained within very small time intervals, due to collinearity, or across very large time intervals, due to nonlinearity. The former problem was recently addressed by means of a state space formulation involving modelling latent reactions (Framba et al., 2024). With respect to the latter problem, existing methods exhibit significant estimation bias because of poor approximations (Shoji, 2013). This is a serious problem, as large observation intervals are

* Corresponding author.
*E-mail addresses:* veronica.vinciotti@unitn.it (V. Vinciotti), wite@usi.ch (E.C. Wit).

https://doi.org/10.1016/j.csda.2025.108251
Received 17 October 2024; Received in revised form 12 July 2025; Accepted 14 July 2025

Available online 22 July 2025
0167-9473/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

typical in many practical experimental settings, such as gene therapy clonal studies, where blood sampling occurs monthly so as to align with the months-long lifespan of blood cells and their production cycle (Pellin et al., 2023).

To date, only few studies have explored the challenges of parameter inference in quasi-reaction models for widely-spaced-data. Pellin et al. (2023) and Milner et al. (2013) proposed moment-closure methods that numerically solve the differential equations of the first and second moments of the process. As a result, these methods are limited to systems of up to second order reactions. Xu et al. (2019) proposed a method-of-moments algorithm that matches second order moments. However, also this approach is limited to simple systems where the theoretical second moments can be calculated analytically. Moreover, due to the statistical instability of the empirical moments, this approach fails in highly stochastic or nonlinear systems. The Bayesian inference approach in Boys et al. (2008) is generic when it comes to the system under consideration and it works well in data-poor scenarios, but it is computationally expensive due to the use of sampling techniques that require stochastic convergence.

Mean-field approximation techniques (Baccelli et al., 1992) offer a viable alternative by providing explicit solutions for the first moments of state distributions while maintaining the nonlinearity of the process. However, in the case of quasi-reaction systems, these approaches are limited to unitary systems, where each reaction involves the transformation of a single element into one or more products. Indeed, only in this case, the hazard rates of the reactions are linear in the state concentrations, resulting in a first-order system of ODEs for the mean of the process which admits an explicit analytical solution. Unitary systems are, however, rare, as real-world models are better characterized by nonlinear dynamics, which capture complex behaviors like logistic growth (Tsoularis and Wallace, 2002), bifurcations (Hale and Koçak, 2012), and limit cycles (Ye and Cai, 1986).

Various methods have employed Taylor approximations to solve kinetic rate equations. Kennealy and Moore (1977) utilized the Taylor series expansion for numerical integration of chemical kinetics, leveraging the ease of obtaining higher-order derivatives from the specific form and symmetries of differential equations in chemical systems. However, the need for frequent adjustments to the step size to ensure convergence remains a challenge. Córdoba-Torres et al. (1998) introduced a method for optimizing the initial parameters of the Taylor integrator by controlling local errors. This approach allows for larger step sizes without increasing computational complexity significantly, yet it demands an accurate analytical expression for the local errors, which can be difficult to obtain in complex systems. Lente et al. (2022) developed an algorithm based on Taylor's theorem for solving kinetic differential equations, using polynomial expansions of concentration-time functions. However, its applicability is limited by the requirement for suitable time transformations to maintain the polynomial nature of the rate equations, which may not always be feasible.

In contrast to these approaches, an efficient method is proposed that extends the mean-field approach using a Taylor expansion not in time, as in the above methods, but in concentration. This is in line with commonly used approaches of linearization of ODEs (Bequette, 1998). Starting from the chemical master equation and using the system of nonlinear equations describing the dynamics of the process mean, a linear approximation of the rate function is derived. This leads to an approximation of the system of ordinary differential equations (ODEs) with an explicit solution. By embedding this solution within a nonlinear least-squares method, it is possible to perform inference of the parameters governing the rate equations.

The paper is structured as follows. Section 2 formalizes the statistical modeling of quasi-reaction systems and introduces the generic mean-field approach. The proposed local mean-field approximation method is defined and illustrated via an example. Its resistance to stiffness is then studied, by comparing the performance of the proposed method against several numerical approaches. The nonlinear least-squares procedure for parameter estimation is described in section 3, both from a methodological and computational point of view. Section 4 is reserved for simulation studies. A comparison of the proposed algorithm with other state-of-the-art approaches shows how the proposed approach strikes a good balance between accuracy in parameter estimation, computational efficiency and generality of implementation for any quasi-reaction system. Gains in estimation accuracy are particularly pronounced as the time interval between consecutive observations increases. Section 5 presents an illustration of the method on the analysis of a cell differentiation study by Wu et al. (2014). Using BIC-based model selection, the pattern of cell differentiation of 5 major blood cells is studied, and the parameters regulating the underlying reaction process are estimated.

## 2. Local Mean-field Approximation (LMA)

In this section, the proposed method for estimation of parameters in a quasi-reaction system is described. To this end, section 2.1 first describes concisely a general quasi-reaction system, with the aim of deriving the general ODE formulation of the conditional mean of this process. Section 2.2 shows that this system of ODEs can be solved explicitly for unitary systems. For generic systems, the solution does not exist. However, by using the solution from a unitary system, a generic approximation for any quasi-reaction system is derived in section 2.3, and is applied to an example in section 2.4. Section 2.5 compares the approach to other prominent ODE solvers in terms of computational time and robustness to stiffness. Finally, section 2.6 shows how the approach can be extended to other types of stochastic differential equations.

### 2.1. Quasi-reaction models

Consider a closed system with $p$ interacting species $\{Q_1, \dots, Q_p\}$. Every interaction between the substrates is caused by the occurrence of a quasi-reaction $R_j$, described as

$$k_{1j}Q_1 + \dots + k_{pj}Q_p \xrightarrow{\theta_j} s_{1j}Q_1 + \dots + s_{pj}Q_p \qquad j \in 1, \dots, r. \tag{1}$$

The occurrence of the $j$-th reaction leads to a change of $v_{lj} = s_{lj} - k_{lj}$ substrates for particle type $Q_l$. Let $V$ denote the net effect matrix having $v_{lj}$ as $lj$-th element, and $K = \{k_{lj}\}$ the reactant matrix. Let $\boldsymbol{Y}(t) = (Y_1(t), \dots, Y_p(t))^T \in \mathbb{N}_0^p$ denote the continuous time

counting process, with $Y_l(t)$ the number of $Q_l$ particles present in the system at time $t \in [0, T]$. Under standard homogeneous mixing assumptions, the conditional rate for the $j$-th reaction in such system is given by

$$\lambda_j(\boldsymbol{Y}(t); \boldsymbol{\theta}) = \theta_j \prod_{l=1}^{p} \binom{Y_l(t)}{k_{lj}}, \tag{2}$$

with $\binom{Y_l(t)}{k_{lj}} := 0$, for all $Y_l(t) < k_{lj}$. The rate parameters $\boldsymbol{\theta} \in \mathbb{R}_+^r$ govern the dynamics of the process. Let $\Theta$ be a diagonal matrix with $\boldsymbol{\theta}$ on the diagonal, and $\boldsymbol{\kappa}(t)$ the vector with $j$-th element $\kappa_j(t) = \prod_{l=1}^{p} \binom{Y_l(t)}{k_{lj}}$. Thus, the hazard function can be compactly written as $\lambda(\boldsymbol{Y}(t); \boldsymbol{\theta}) = \Theta \boldsymbol{\kappa}(t)$.

The stochastic process $\boldsymbol{Y}(t)$ obeys the Markov property. Given an initial condition $\boldsymbol{y}(t_0)$, it is possible to determine the probability density $p_t(\boldsymbol{Y})$ of the system being in the state $\boldsymbol{Y}$ at time $t$. The temporal evolution of the Markov process transition kernel is governed by the Kolmogorov's forward equations (Wilkinson, 2018). In the context of a stochastic kinetic process, these are commonly referred to as the chemical master equation, which is given by

$$\frac{dp_t(\boldsymbol{y})}{dt} = \sum_{j=1}^{r} p_t(\boldsymbol{y} - \boldsymbol{v}_{\cdot j}) \lambda_j(\boldsymbol{y} - \boldsymbol{v}_{\cdot j}; \theta_j) - p_t(\boldsymbol{y}) \lambda_j(\boldsymbol{y}; \theta_j), \quad \forall\ \boldsymbol{y} \in \mathbb{N}_0^p. \tag{3}$$

Solving the above equation involves assessing the evolution of $p_t(\boldsymbol{y})$ over the entire range of possible configurations for the process. This approach clearly does not offer a feasible solution for systems of realistic size and complexity. Nevertheless, valuable insights regarding the dynamics of characteristic statistical features of the system can be derived from (3).

In particular, one can obtain from the chemical master equation a set of ODEs describing the temporal evolution of population concentration averages. To this end, let $\boldsymbol{m}(t+s|t)$ describe the evolution of $\mathbb{E}[\boldsymbol{Y}(t+s)|\boldsymbol{Y}(t) = \boldsymbol{y}(t)] = \sum_{\boldsymbol{y}} \boldsymbol{y} p_{t+s|t}(\boldsymbol{y})$. As shown in detail in Appendix A, the following ODEs system describing the dynamics of the mean of $\boldsymbol{Y}(t+s)$ conditional on $\boldsymbol{Y}(t)$ can be derived (Pellin et al., 2023):

$$\frac{dm_l(t+s|t)}{ds} = \sum_{j=1}^{r} v_{lj} \mathbb{E}[\lambda_j(\boldsymbol{Y}(t+s); \boldsymbol{\theta}) \mid \boldsymbol{Y}(t)], \quad l = 1, \ldots, p. \tag{4}$$

There have been several proposals for solving (4) using approximation methods. These, however, come with significant limitations. The system size expansion method by Van Kampen (1992) tends to lose accuracy in systems with small populations or complex dynamics. The moment closure approximation (Grima, 2012) often loses information due to the arbitrary truncation of higher moments, while the diffusion approximation method by Golightly and Wilkinson (2005) can be unreliable in cases with low molecule counts and highly nonlinear behavior. The next section shows how an explicit mean-field solution can be found for unitary systems, and how this can then be used as the basis of a new approximation method.

### 2.2. Explicit mean-field solution for unitary systems

Unitary systems are quasi-reaction systems where each reaction needs at most one particle from each reactant in order to occur. In such systems, the hazard rate (2) is linear in $\boldsymbol{y}$, so

$$\mathbb{E}[\lambda_j(\boldsymbol{Y}(t+s;\boldsymbol{\theta})) \mid \boldsymbol{Y}(t) = \boldsymbol{y}(t)] = \lambda_j(\mathbb{E}[\boldsymbol{Y}(t+s) \mid \boldsymbol{Y}(t) = \boldsymbol{y}(t)]; \boldsymbol{\theta}) = \lambda_j(\boldsymbol{m}(t+s|t); \boldsymbol{\theta}).$$

Plugging this expression into equation (4) and using the hazard rates from equation (2), it follows that

$$\begin{aligned}
\frac{dm_l(t+s|t)}{ds} &= \sum_{j=1}^{r} v_{lj} \lambda_j(\boldsymbol{m}(t+s|t); \boldsymbol{\theta}) \\
&= \sum_{j=1}^{r} v_{lj} \left[ \Theta_{j\cdot} K^T \boldsymbol{m}(t+s|t) + \theta_j \mathbb{1}_{\{K_{\cdot j} = 0\}} \right] \\
&= \sum_{j=1}^{r} v_{lj} \Theta_{j\cdot} K^T \boldsymbol{m}(t+s|t) + \sum_{j=1}^{r} v_{lj} \theta_j \mathbb{1}_{\{K_{\cdot j} = 0\}}.
\end{aligned}$$

Let $P_\theta = V \Theta K^T$ be the coefficient matrix and $b_\theta = (\sum_j v_{lj} \theta_j \mathbb{1}_{\{K_{\cdot j} = 0\}})_l$ the inhomogeneous term. These are functions of the vector $\boldsymbol{\theta}$. The former relates to the reactions that involve exactly one reactant, whereas the latter refers to spontaneous reactions that do not involve any reactants. With this notation and by adding the initial condition $\boldsymbol{m}(t|t) = \boldsymbol{y}(t)$, the above equation results in the following first order Cauchy differential equation,

$$\begin{cases} \dfrac{d\boldsymbol{m}(t+s|t)}{ds} = P_\theta \boldsymbol{m}(t+s|t) + \boldsymbol{b}_\theta \\ \boldsymbol{m}(t|t) = \boldsymbol{y}(t). \end{cases} \tag{5}$$

If the $P_\theta$ matrix is invertible, the system (5) has the explicit solution

$$\boldsymbol{m}(t+s|t) = \exp\left(sP_{\boldsymbol{\theta}}\right)\boldsymbol{y}(t) + P_{\boldsymbol{\theta}}{}^{-1}\left(\exp\left(sP_{\boldsymbol{\theta}}\right) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}}. \tag{6}$$

Although this approach only works for unitary systems, it inspires the proposal for generic quasi-reaction systems explained in the next section.

### 2.3. LMA: local mean-field approximation for generic systems

Most biological quasi-reaction systems involve complex, higher order interactions that make finding an analytical solution to (4) typically impossible. The idea is to linearize the hazard function with respect to the abundance vector $\boldsymbol{Y}$ so that any quasi-reaction system can be approximated as a unitary system, thus obtaining an explicit solution of the ODEs system of the form (6). Omitting the dependence of the hazard function on the parameters $\boldsymbol{\theta}$ for the sake of readability, a first order Taylor expansion of $\lambda(\boldsymbol{Y}(t+s))$ around $\boldsymbol{Y}(t)$ is given by

$$\lambda(\boldsymbol{Y}(t+s)) = \lambda(\boldsymbol{Y}(t)) + \Lambda(\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)) + \boldsymbol{\eta}(t). \tag{7}$$

Here, $\eta_j$ is the approximation error for the $j$-th component, which, from Taylor's theorem, is of the form

$$\eta_j = \frac{1}{2}\frac{\partial^2 \lambda_j(\tilde{\boldsymbol{Y}})}{\partial Y_l(t)\partial Y_k(t)}(Y_l(t+s) - Y_l(t))(Y_k(t+s) - Y_k(t)), \quad l,k = 1,\dots,p \tag{8}$$

with $\tilde{\boldsymbol{Y}} \in (\boldsymbol{Y}(t), \boldsymbol{Y}(t+s))$. The Jacobian matrix $\Lambda = \dfrac{\partial \lambda}{\partial \boldsymbol{y}}$ is explicitly defined by the proposition below, whose proof can be found in Appendix B.

**Proposition.** *Given the intensity function* $\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\theta}) = \theta_j \prod_{l=1}^{p} \binom{Y_l(t)}{k_{lj}}$, *the* $jl$*-th element of the Jacobian matrix* $\Lambda(\mathbf{Y}(t);\boldsymbol{\theta}) \in \mathbb{R}^{r\times p}$ *is given by*

$$\Lambda_{jl} = \underbrace{\theta_j \prod_{i=1}^{p} \binom{Y_i(t)}{k_{ij}}(1-\delta_{il})\binom{Y_l(t)}{k_{ij}}\left(\psi(Y_l(t)+1) - \psi(Y_l(t)-k_{lj}+1)\right)}_{H_{jl}}$$

*where* $\psi(x) = \dfrac{d}{dx}\log(\Gamma(x))$ *is the digamma function, i.e., the logarithmic derivative of the gamma function.*

Using now approximation (7) in (4), the conditional mean ODEs system is given by

$$\begin{aligned}
\frac{d}{ds}\boldsymbol{m}(t+s|t) &= V\mathbb{E}\left[\lambda(\boldsymbol{Y}(t)) + \Lambda\big(\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)\big) \mid \boldsymbol{Y}(t) = \boldsymbol{y}(t)\right] \\
&= V\Lambda\mathbb{E}[\lambda(\boldsymbol{Y}(t+s))|\boldsymbol{Y}(t) = \boldsymbol{y}(t)] + V\lambda(\boldsymbol{y}(t)) - V\Lambda\boldsymbol{y}(t) \\
&= \underbrace{V\,\Theta\,H}_{P_{\boldsymbol{\theta}}}\boldsymbol{m}(t+s|t) + \underbrace{V\,\Theta\,(\boldsymbol{\kappa}(t) - H\,\boldsymbol{y}(t))}_{\boldsymbol{b}_{\boldsymbol{\theta}}} \\
&= P_{\boldsymbol{\theta}}\boldsymbol{m}(t+s|t) + \boldsymbol{b}_{\boldsymbol{\theta}}.
\end{aligned} \tag{9}$$

If $|P_{\boldsymbol{\theta}}| \neq 0$, the explicit solution of the ODEs system above is defined as in (6).

In terms of convergence to the solution of the original ODEs system (4), the approximation does not depend on the time step but on the difference between concentrations as indicated by the relation (8). Since the expansion is up to first order, and if at most second-order reactions are considered, then the approximation error is determined by the norm of the Hessian matrix of the hazard function.

### 2.4. Example: cyclic chemical reaction network

This section considers the example of a cyclic chemical reaction network involving three particle types $(A, B, C)$ with abundances $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$. The reactions are given by

$$\begin{aligned}
2\mathrm{A} &\xrightarrow{\theta_1} 2\mathrm{B} \\
\mathrm{A} + \mathrm{B} &\xrightarrow{\theta_2} 3\mathrm{C} \\
2\mathrm{C} &\xrightarrow{\theta_3} 2\mathrm{A}.
\end{aligned}$$

In particular, this network is a closed loop where the products of one reaction act as the reactants for another, creating a continuous cycle of chemical transformations. Such cyclic networks are central in understanding metabolic cycles and oscillatory behavior in

biological systems (Gillespie, 2007). The first reaction describes the conversion of two molecules of species A into two molecules of species B and could be seen as a simple process where A is transformed into B in the presence of a catalyst. The second reaction, where A and B combine to form three molecules of species C, can be viewed as a synthesis reaction often seen in polymerization processes. The final reaction regenerates species A from C, completing the cycle and ensuring the continuity of the process.

The reactant and net effect matrix associated to this system are given, respectively, by

$$K = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad V = \begin{bmatrix} -2 & -1 & 2 \\ 2 & -1 & 0 \\ 0 & 3 & -2 \end{bmatrix}.$$

The first is used in the definition of the hazard rates in (2), while the second, together with the hazard function, defines the dynamics of the first moments of the concentrations in (4).

The quasi-reaction system is clearly non-unitary, as more than one particle type is used in each reaction. This leads to a hazard that is not linear in $\boldsymbol{Y}$ and to no analytical solution to the ODEs system in (4). The local mean-field approximation method described in section 2.3 provides an explicit approximate solution to the system. In order to show this, let $\boldsymbol{y} = (y_1, y_2, y_3)$ represent the observation of the continuous process at time $t$. The reaction events are associated with the hazard function

$$\lambda(\boldsymbol{y}) := \lambda(\boldsymbol{Y}(t); \boldsymbol{\theta})\Big|_{\boldsymbol{Y}(t)=\boldsymbol{y}} = \begin{bmatrix} \theta_1 y_1 (y_1 - 1)/2 \\ \theta_2 y_1 y_2 \\ \theta_3 y_3 (y_3 - 1)/2 \end{bmatrix},$$

for which a Taylor expansion in $\boldsymbol{y}$ is considered and then its first order approximation. In particular, the Jacobian matrix, evaluated at $\boldsymbol{y}$, is given by

$$\Lambda = \frac{\partial \lambda(\boldsymbol{Y}(t); \boldsymbol{\theta})}{\partial \boldsymbol{Y}}\Big|_{\boldsymbol{Y}(t)=\boldsymbol{y}} = \begin{bmatrix} \theta_1 (y_1 - 0.5) & 0 & 0 \\ \theta_2 y_2 & \theta_2 y_1 & 0 \\ 0 & 0 & \theta_3 (y_3 - 0.5) \end{bmatrix}.$$

These quantities are now plugged into (7), leading to the ODEs system approximation (9) with

$$\begin{aligned} P_{\boldsymbol{\theta}} &= V\Lambda = \begin{bmatrix} -2 & -1 & 2 \\ 2 & -1 & 0 \\ 0 & 3 & -2 \end{bmatrix} \begin{bmatrix} \theta_1 (y_1 - 0.5) & 0 & 0 \\ \theta_2 y_2 & \theta_2 y_1 & 0 \\ 0 & 0 & \theta_3 (y_3 - 0.5) \end{bmatrix} \\ &= \begin{bmatrix} -2\theta_1 (y_1 - 0.5) - \theta_2 y_2 & -\theta_2 y_1 & 2\theta_3 (y_3 - 0.5) \\ 2\theta_1 (y_1 - 0.5) - \theta_2 y_2 & -\theta_2 y_1 & 0 \\ 3\theta_2 y_2 & 3\theta_2 y_1 & -2\theta_3 (y_3 - 0.5) \end{bmatrix}, \\ \boldsymbol{b}_{\boldsymbol{\theta}} &= V(\lambda(\boldsymbol{y}) - \Lambda \boldsymbol{y}) \\ &= \begin{bmatrix} -2 & -1 & 2 \\ 2 & -1 & 0 \\ 0 & 3 & -2 \end{bmatrix} \left( \begin{bmatrix} \theta_1 y_1 (y_1 - 1)/2 \\ \theta_2 y_1 y_2 \\ \theta_3 y_3 (y_3 - 1)/2 \end{bmatrix} - \begin{bmatrix} \theta_1 (y_1 - 0.5) & 0 & 0 \\ \theta_2 y_2 & \theta_2 y_1 & 0 \\ 0 & 0 & \theta_3 (y_3 - 0.5) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \right) \\ &= \begin{bmatrix} \theta_1 (y_1)^2 + \theta_2 y_1 y_2 - \theta_3 (y_3)^2 \\ -\theta_1 (y_1)^2 + \theta_2 y_1 y_2 \\ -3\theta_2 y_1 y_2 + \theta_3 (y_3)^2 \end{bmatrix}. \end{aligned}$$

Substituting the quantities above into equation (6) returns the explicit form of the mean process values after a time interval $s$. These constitute a nonlinear forward prediction of the system at time $t + s$. It should be noted that $\det(P_{\boldsymbol{\theta}}) = 12\theta_1\theta_2\theta_3(y_1 - 0.5)y_2(y_3 - 0.5)$ is non-zero if and only if $\boldsymbol{\theta} \neq 0$ and $\boldsymbol{y} \neq (0.5, 0, 0.5)$.

### 2.5. Computational time for calculating ODE solution and robustness to stiffness

The main advantage of the proposed method is that the ODEs system in (9), used to approximate the original system in (4), has an explicit solution (6). As this section will show, classical numerical methods for solving ODEs systems may be computationally more efficient than calculating this analytical solution, but they are less stable in many scenarios.

A traditional numerical algorithm for solving ODEs systems, well known for its simplicity of implementation, is the explicit Euler method. This has a slow linear convergence with respect to the width of the time steps $\Delta t$ but a linear computational cost with respect to the total number of subintervals $T_c = T/\Delta t$, with $t \in [0, T]$. A good alternative is the explicit fourth-order Runge-Kutta method. This offers a fourth-order convergence, by evaluating the function at multiple points within each time step, but consequently leading to a four times higher computational cost (Krijnen and Wit, 2022). The balance between accuracy and computational cost makes Runge-Kutta methods preferable for many practical applications where accuracy is a priority (Butcher, 2016). As for the proposed LMA, in order to solve the ordinary differential equations (9) analytically, it is necessary to compute the exponential of a $p \times p$ matrix. For this, the Pade approximation with scaling and squaring method is utilized due to its efficiency with dense matrices. This approach has a computational complexity of $\mathcal{O}(p^3)$. Additionally, the inversion of the same matrix needs to be computed, which incurs a similar

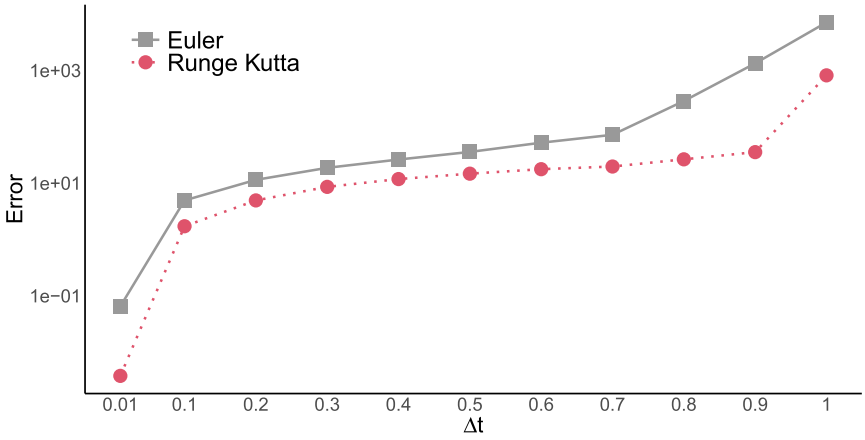



**Fig. 1.** Mean absolute error between the numerical solution from each method and the analytical LMA solution in (6), as $\Delta t$ increases: both Euler and Runge-Kutta methods experience a significant degradation in accuracy as $\Delta t$ increases.

**Table 1**
Comparison of computational costs, error convergence, and performance in stiff problems for the explicit Euler, fourth-order Runge-Kutta, and the proposed LMA method, in terms of the number of particles $p$, the width of time intervals $\Delta t$ and the number of intervals $T_c$.

| Method | Computational Cost | Convergence | Stiffness |
|---|---|---|---|
| Explicit Euler | $\mathcal{O}(pT_c)$ | $\mathcal{O}(\Delta t)$ | Unsuitable |
| Runge-Kutta | $\mathcal{O}(4pT_c)$ | $\mathcal{O}(\Delta t^4)$ | Unsuitable |
| LMA | $\mathcal{O}(p^3T_c)$ | $\mathcal{O}(1)$ | Robust |

computational cost. Consequently, the overall cost for calculating the analytical solution in (9) using the LMA approach amounts to $\mathcal{O}(p^3)$, which is higher than both the Euler and Runge-Kutta methods.

Besides computational time, a further comparison between the methods can be made in terms of performance in the presence of stiffness. It is well known how the explicit Euler and Runge-Kutta methods struggle with problems that are classified as stiff. On the other hand, the availability of an explicit solution makes the method robust also in the presence of stiffness. A system is considered stiff in a given interval if, when applying a numerical method with a finite region of absolute stability, the step length required is excessively small relative to the smoothness of the exact solution (Lambert, 1974). This definition emphasizes the difficulty of maintaining stability with explicit methods, which may necessitate impractically small step sizes to accurately obtain the solution.

In quasi-reaction chemical models, the phenomenon of stiffness arises in situations where very slow and very fast reactions coexist. A well known example of a stiff problem is the kinetics of an autocatalytic system (Robertson, 1967). However, as its associated net effect matrix is not full rank, $P_\theta = V\Lambda$ is not invertible. Instead, the cyclic chemical reaction system described in section 2.4 is considered, with initial value $\boldsymbol{y} = (10, 20, 10)$ and reaction rate $\boldsymbol{\theta} = (2 \cdot 10^{-6}, 10^{-7}, 2 \cdot 10^{-1})$. After a first order Taylor approximation, the ODEs system (9) associated to this network is considered. The matrix associated to this system is invertible. Moreover, the eigenvalues of $P_\theta$, given by $7.6 \cdot 10^{-7}$, $-4.2 \cdot 10^{-5}$, and $-3.8 \cdot 10^{0}$, are different from each other in orders of magnitude, which is taken as an indication of stiffness (Butcher, 2016). Intuitively, it is clear how the third reaction is much faster than the first two.

Fig. 1 shows the robustness of the Euler and Runge-Kutta methods when used to solve numerically the ODEs system (9) of this problem. The error is defined as the average of the absolute difference of the solution obtained by the proposed method and the analytical LMA solution in (6), when both solutions are evaluated at 5 time points. For very small time intervals, the Euler and Runge Kutta methods are accurate. However, they become unstable as $\Delta t$ increases. Table 1 summarizes the comparison of the various methods discussed in this section, in terms of computational costs, convergence and robustness to stiffness. The proposed LMA approach provides a good compromise between computational efficiency and robustness to stiffness.

### *2.6. Beyond quasi-reaction systems*

Although the focus of this paper is on quasi-reaction systems, the approach can be extended to other types of dynamic systems that can be written as an SDE, in particular diffusions. Diffusion models describe the stochastic evolution of a state variable over time. Many important diffusion processes have a drift term that depends on the state of the system. In principle, some of the inference strategies described in the previous sections can be applied to these settings.

The geometric Brownian motion is commonly used in financial mathematics to model stock prices. The stochastic differential equation governing this process is given by:

$$dY(t) = \mu Y(t)dt + \sigma Y(t)dW(t),$$

which, by applying Itô's lemma, has an explicit solution given by

$$Y(t) = Y(0)\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right).$$

In particular, as the drift is linear in the state vector, its mean is given by

$$m(t+s|t) = E[Y(t+s)|Y(t) = y(t)] = y(t)e^{\mu s},$$

which corresponds to the usual first order ODE solution, which has been derived for unitary systems.

There are also non-linear diffusion processes. The logistic diffusion model is used in biological and ecological modeling, particularly for population dynamics where growth is limited by a carrying capacity $K$. The corresponding SDE is given by

$$dY(t) = rY(t)\left(1 - \frac{Y(t)}{K}\right)dt + \sigma dW(t).$$

This model describes a population that grows logistically with stochastic fluctuations. More generally, a multivariate diffusion process with a nonlinear drift term can be considered, given by

$$d\boldsymbol{Y}(t) = \boldsymbol{F}(\boldsymbol{Y}(t))dt + \boldsymbol{B}(t)d\boldsymbol{W}(t)$$

As with the approximation of the hazard function in quasi-reaction systems, one can apply a first-order Taylor approximation to the drift, that is

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{Y}(t+s)) &\approx \boldsymbol{F}(\boldsymbol{Y}(t)) + \Lambda(\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)) \\ &= \Lambda\boldsymbol{Y}(t+s) + \boldsymbol{b}\end{aligned}$$

with the Jacobian matrix given by $\Lambda = \dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}}(t)$ and constant vector $\boldsymbol{b} = \boldsymbol{F}(\boldsymbol{Y}(t)) + \Lambda\boldsymbol{Y}(t)$, where $t$ is assumed a fixed time-point and $s$ describes the subsequent development of the process. If the $\Lambda$ matrix is invertible, the associated ODE system for the conditional mean has the explicit solution

$$\boldsymbol{m}(t+s|t) = \exp\left(s\Lambda\right)\boldsymbol{y}(t) + \Lambda^{-1}\left(\exp\left(s\Lambda\right) - \mathbb{I}_p\right)\boldsymbol{b}.$$

This analytical expression for the conditional mean can be used conveniently in further inference procedures, whether they involve likelihood or method-of-moment approaches.

## 3. Inference

Dynamic processes are often observed at discrete time points, and possibly across several replicates. For example, in gene therapy studies (Del Core et al., 2023), and in hematopoietic clonal dynamics (Pellin et al., 2023), clones, i.e., genetically identical cells, are probed at times that are days or even months apart.

### 3.1. Estimation of reaction rates

A set of $n$ replicates is considered, whereby each replicate $c$ is observed across $T_c$ time points. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_{ci} = \boldsymbol{Y}_c(t_{ci})\}_{c,i}^{n,T_c}$ be the set of $p$-dimensional observations of realizations of particle counts subject to the quasi-reaction system. The time intervals are not necessary equal, which means that the observation times $t_{ci}$ are also indexed by the replicate information. Let $\boldsymbol{m}(\boldsymbol{\theta}) = [\boldsymbol{m}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{m}_n(\boldsymbol{\theta})]$ such that $\boldsymbol{m}_{ci}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{Y}_{ci} \mid \boldsymbol{Y}_c(t_{c,i-1}) = \boldsymbol{y}_c(t_{c,i-1})]$ is the solution of the ODEs system (9) with initial condition $\boldsymbol{Y}_{c,i-1} = \boldsymbol{y}_{c,i-1}$. In other words, the solution of the system of ODEs projects each observation $\boldsymbol{y}_{c,i-1}$ to the expected value at the next time point. The inference procedure for the rates $\boldsymbol{\theta}$ can then be reformulated as the following nonlinear regression problem,

$$\boldsymbol{Y}_{ci} = \boldsymbol{m}_{ci} + \boldsymbol{\epsilon}_{ci}, \tag{10}$$

where $\boldsymbol{\epsilon}_{ci}$ is a $p$-dimensional vector of residuals such that $\mathbb{E}[\boldsymbol{\epsilon}_{ci}] = \boldsymbol{0}$. As likelihood-based approaches are unfeasible due to the computational effort involved in integrating over all possible states between the observation times, a least-squares algorithm is proposed for the estimation of the $\boldsymbol{\theta}$ parameters. The constrained solution is then given by

$$\hat{\boldsymbol{\theta}}_{LMA} = \arg\min_{\boldsymbol{\theta}\geq 0}\left\{ f(\boldsymbol{\theta}) = \sum_{c=1}^{n}\sum_{i=1}^{T_c}\left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}(\boldsymbol{\theta})\right]^T\left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}(\boldsymbol{\theta})\right]\right\}. \tag{11}$$

The optimum is determined using an iterative approach which consists of two steps. First, the approximation of the ODE solution (9) is solved analytically. Then, the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box-constraints is applied to find the optimum. This is an optimization scheme used to solve large-scale optimization problems with simple bounds on the variables (Byrd et al., 1995). A pseudo-code of the algorithm can be found in Appendix C. Taking account of the computational cost for the calculation of $\boldsymbol{m}_{ci}(\boldsymbol{\theta})$ shown in Table 1, the overall computational cost for the estimation of the $r$-dimensional vector $\boldsymbol{\theta}$ is of order

$\mathcal{O}(k_1 nTp^3r^3)$ with $k_1$ the number of iterations needed for the minimization of (11). Given the potentially large number of parameters in the model, it is important to start the minimization of (11) with accurate initial values. A practical starting value is provided by the local linear approximation (LLA) approach, which will also be used in the comparative study. A detailed description of the method can be found in Appendix D.

Compared to other methods, LLA is computationally the most efficient, with a complexity of order $\mathcal{O}(nTpr^3)$. Similarly to the proposed method, the moment-closure method of Pellin et al. (2023) requires both the calculation of the solution of the ODE and the optimization over the parameters, resulting in an overall complexity of $\mathcal{O}(k_2 nTp^3r^3)$, where the constant $k_2$ is determined by the efficiency of the Cplex optimization algorithm used in this method. Also the Bayesian procedure in Boys et al. (2008) has a cost of $\mathcal{O}(k_3 nTp^3r^3)$, but the constant $k_3$ depends on the stochastic convergence of the MCMC sampler. In the case of data augmentation techniques, the constant includes also the number of state assignments between the observed time points. Empirically, in simulation studies, it seems that $k_1 \leq k_2 \leq k_3$, but this may depend on the computational implementations of the various algorithms.

### 3.2. Standard error approximation

In the context of statistical modeling of quasi-reaction systems, it is important to be able to evaluate the uncertainty associated with the estimated rates $\hat{\boldsymbol{\theta}}$. Only few studies provide explicit approximate formulations for this (Framba et al., 2024; Tsugé, 2001). This is done in this section for the proposed method. Under certain regularity conditions the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is approximately the inverse of the observed Fisher information matrix evaluated at $\hat{\boldsymbol{\theta}}$. The latter is the negative of the Hessian matrix $\mathbb{H}(\hat{\boldsymbol{\theta}}) := \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and can be approximated by

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} f(\hat{\boldsymbol{\theta}})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} f(\hat{\boldsymbol{\theta}})\right)^T. \tag{12}$$

The derivative of the objective function $f(\boldsymbol{\theta})$ with respect to the rates involves taking the derivative of the solution of the ODE system (9), i.e., the derivatives of

$$\boldsymbol{m}(t+s|t) = \exp\left(sP_{\boldsymbol{\theta}}\right)\boldsymbol{y}(t) + P_{\boldsymbol{\theta}}{}^{-1}\left(\exp\left(sP_{\boldsymbol{\theta}}\right) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}} \tag{13}$$

with respect to $\boldsymbol{\theta}$. For ease of notation, a single-replicate scenario is considered. The derivatives of $P_{\boldsymbol{\theta}}$ and $\boldsymbol{b}_{\boldsymbol{\theta}}$ defined in (9) are first calculated. This gives

$$\frac{\partial P_{\boldsymbol{\theta}}}{\partial \theta_j} = V\boldsymbol{e}_j H, \qquad \frac{\partial \boldsymbol{b}_{\boldsymbol{\theta}}}{\partial \theta_j} = V\boldsymbol{e}_j \boldsymbol{\kappa}(t) - V\boldsymbol{e}_j H\boldsymbol{y}(t),$$

where $\boldsymbol{e}_j$ is a $r$-dimensional vector with a 1 in the $j$-th position and 0 elsewhere. Using the matrix exponential derivative property and the chain rule,

$$\frac{\partial \exp(sP_{\boldsymbol{\theta}})}{\partial \theta_j} = \int_0^1 \exp((1-u)sP_{\boldsymbol{\theta}})\frac{\partial (sP_{\boldsymbol{\theta}})}{\partial \theta_j}\exp(usP_{\boldsymbol{\theta}})\,du.$$

Using the chain rule, the derivative of the second term in (13) is also obtained,

$$\begin{aligned}\frac{\partial}{\partial \theta_j}\left(P_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}}\right) = {} & \frac{\partial P_{\boldsymbol{\theta}}^{-1}}{\partial \theta_j}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}} + P_{\boldsymbol{\theta}}^{-1}\frac{\partial}{\partial \theta_j}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}} \\ & + P_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\frac{\partial \boldsymbol{b}_{\boldsymbol{\theta}}}{\partial \theta_j}.\end{aligned}$$

Combining everything together and using $\frac{\partial P_{\boldsymbol{\theta}}^{-1}}{\partial \theta_j} = -P_{\boldsymbol{\theta}}^{-1}\frac{\partial P_{\boldsymbol{\theta}}}{\partial \theta_j}P_{\boldsymbol{\theta}}^{-1}$, the partial derivative of the conditional predicted values is

$$\begin{aligned}\frac{\partial \boldsymbol{m}(t+s\mid t)}{\partial \theta_j} = {} & \left(s\int_0^1 \exp((1-u)sP_{\boldsymbol{\theta}})V\boldsymbol{e}_j H\exp(usP_{\boldsymbol{\theta}})\,du\right)\boldsymbol{y}(t) - P_{\boldsymbol{\theta}}^{-1}Ve_j HP_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)b_{\boldsymbol{\theta}} \\ & + P_{\boldsymbol{\theta}}^{-1}s\left(s\int_0^1 \exp((1-u)sP_{\boldsymbol{\theta}})V\boldsymbol{e}_j H\exp(usP_{\boldsymbol{\theta}})\,du\right)\boldsymbol{b}_{\boldsymbol{\theta}} \\ & + P_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\left(Ve_j\boldsymbol{\kappa}(t) - Ve_j H\boldsymbol{y}(t)\right).\end{aligned} \tag{14}$$

Taking into account that the minimization problem relies on $T_c$ observations for each of $n$ clone-type scenario, the $p\times r$-dimensional matrix $\boldsymbol{\xi}_{ci}$ is defined, such that the $j$-th column is the evaluation of (14) at $(t_{ci}, \boldsymbol{Y}_{ci})$. The gradient of the objective function with respect to $\boldsymbol{\theta}_j$ is the $r$-dimensional vector

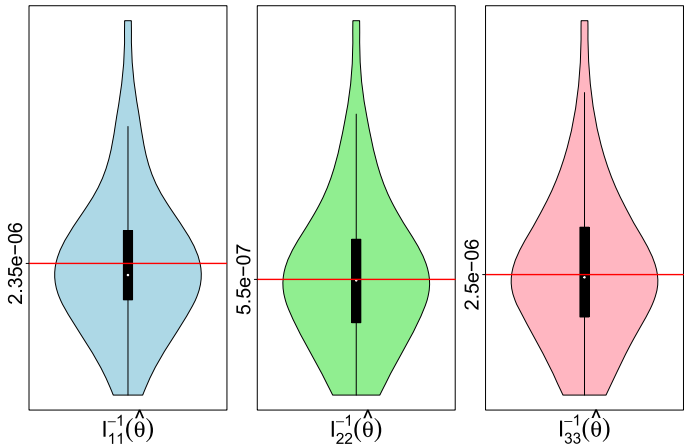



**Fig. 2.** Violin plots showing the distribution across 50 simulations of the diagonal elements of the inverse of the observed Fisher information matrix for the 3 reaction rates of the system in section 2.4. The horizontal line represents the empirical variance of the estimates. The figure shows a good agreement between the theoretically and empirically derived standard errors of the estimated parameters.

$$\frac{\partial f_{ci}(\boldsymbol{\theta})}{\partial \theta_j} = -\left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}\right]^T \boldsymbol{\xi}_{cij}.$$

Using this expression, the variances of the estimated rates $\hat{\boldsymbol{\theta}}$, which are approximated by the diagonal of $\mathcal{I}(\hat{\boldsymbol{\theta}})$ in (12), are given by

$$\mathbb{V}[\hat{\boldsymbol{\theta}}] = \operatorname{diag}\left( \sum_{c=1}^{n} \sum_{i=1}^{T_c} (\boldsymbol{\xi}_{ci})^T \left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}\right] \left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}\right]^T \boldsymbol{\xi}_{ci} \right)^{-1} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The validity of these derivations is illustrated with a numerical example. In particular, the Gillespie algorithm is used to simulate trajectories from the cyclic reaction system defined in section 2.4 with rates $\boldsymbol{\theta} = (0.2, 0.1, 0.2)$ and initial concentration $\mathbf{y}_0 = (100, 100, 100)$. Then, 50 datasets are generated with $c = 100$ replicates for each dataset. Fig. 2 presents violin plots corresponding to the diagonal elements of the inverse of the observed Fisher information matrix. The horizontal lines denote the observed variance of the estimates across the 50 simulations. The figure shows a good agreement between the theoretically and empirically derived standard errors of the estimated parameters.

## 4. Simulation study

This section presents several simulation studies to assess the performance of the proposed method and compare it with existing methods. The experimental setup, based on the system described in section 2.4, reflects conditions typical of numerous biological applications. Section 4.1 evaluates the proposed method by varying the width between the observations $\Delta t$ and the number of time points $T$. Accuracy in parameter estimation is compared with the local linear approximation method, which accounts for the stochasticity of the process. For short time steps, the local linearization is expected to be a serious competitor, whereas for large time steps the nonlinearity of the system will make the mean-field inferential scheme preferable. The same section also compares the proposed approach with the method of Boys et al. (2008), which does not rely on the local linearization approximation. However, the method is computationally expensive due to the use of Bayesian inferential procedures. Section 4.2 further studies how accuracy and computational time of the proposed approach varies by increasing the number of nodes and reactions. Finally, section 4.3 performs a comparison with the method-of-moments formulation introduced by Xu et al. (2019), which relies on matching model-derived and empirical correlations in specific cell type dynamics systems.

### 4.1. Performance as a function of $\Delta t$ and $T$

This section evaluates the performance of the LMA inferential procedure across different time steps $\Delta t$ and number of time points $T$. Assuming homogenous temporal dynamics, time step sizes are taken as a proxy for the extent of missingness of individual reactions that take place between two observation instances. Since the generating process is known, the amount of missingness between time points can be controlled more precisely.

In order to do so, the Gillespie algorithm is first used to simulate trajectories from the cyclic reaction system defined in section 2.4 with rates $\boldsymbol{\theta} = (0.2, 0.1, 0.2)$ and initial concentration $\mathbf{y}_0 = (100, 100, 100)$. Fig. 3 shows an example of trajectories for one of the simulations. Measurements are then extracted from the simulated trajectories, by skipping a certain number of measurements between time points. In particular, 5 different cases are considered where every 10, 30 50, 70, 100 values, respectively, are retained from each trajectory, for a total of $T = 20$ measurements. See Fig. 3 for an example of the last case. As it is clear from this figure, the actual time step will depend on the underlying dynamics. The average time steps $(0.003, 0.007, 0.009, .010, 0.012)$ are taken as a

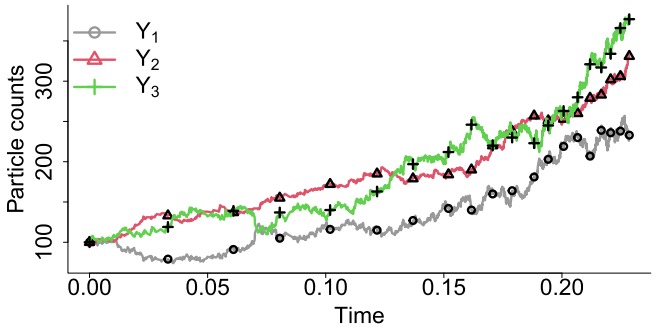



**Fig. 3.** A trajectory generated using the Gillespie algorithm for the model described in section 2.4. The symbols indicate the $T = 20$ observations between every 100 simulated states (average $\Delta t \approx 0.012$).

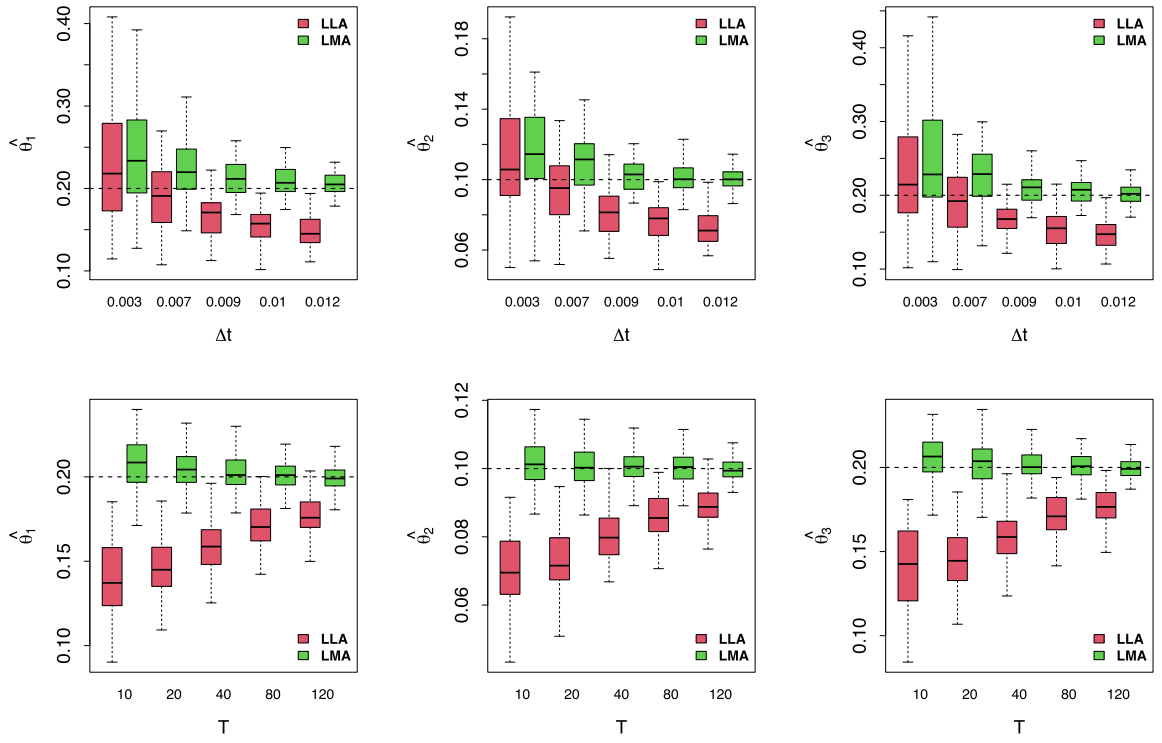



**Fig. 4.** Distributions of the estimated rates from the local linear approximation (LLA, red) and local mean-field approximation (LMA, green) methods across 100 simulations. Top: Keeping $T = 20$ fixed, as the time step $\Delta t$ increases, the LLA estimates, unlike LMA, show increased bias. Bottom: Keeping $\Delta t = 0.012$ fixed, increasing the number of time points $T$ shrinks the standard error of the LMA method in a square-root fashion. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

proxy to distinguish the 5 different scenarios. A second simulation focuses on the last case (skipping 100 values, average $\Delta t = 0.012$) and considers 5 different values for the overall number of time points $T$. The simulations are repeated 100 times. Fig. 4 shows the estimated parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_2)$ for both the proposed LMA method (green) and the alternative LLA method (red). The results show that the performance of the two methods is comparable for small time intervals. However, as $\Delta t$ increases, the LLA approach shows an increasing bias, which is not observed for the LMA method. Furthermore, the standard error of the LMA method shrinks in a roughly $1/\sqrt{T}$ fashion, as the number of time points increases.

The bias of LLA is due to the local linear approximation, which is inaccurate for large $\Delta t$ in the presence of nonlinearity. Fig. 5 compares the proposed LMA approach with a Bayesian inferential procedure, such as the method of Boys et al. (2008). The implementation uses the R package `smfsb` (Wilkinson, 2018) with 5000 MCMC simulations, retaining the last 4000 for estimation. The LMA and Bayesian method perform similarly in terms of estimation (Fig. 5a for $\theta_1$, and a similar behavior for the other two parameters). Although the two methods have the same order of computational cost (see section 3.1), the Bayesian method has a slower stochastic convergence and therefore the overall computational time is significantly higher than for the LMA method (Fig. 5b). This becomes more pronounced the larger $\Delta t$ is, due to an increase in the number of state assignments between the observed time points.

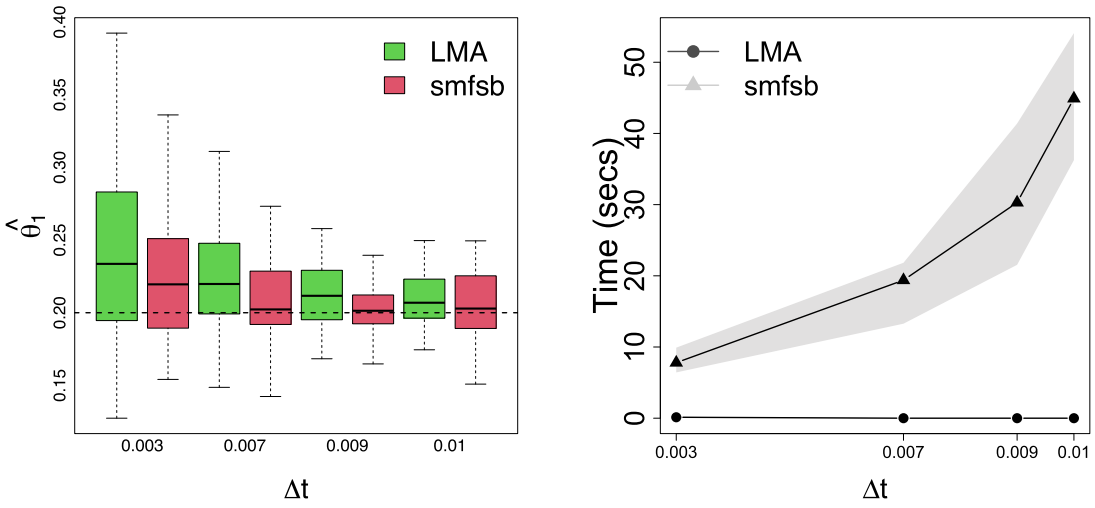



**Fig. 5.** Comparison between LMA and the Bayesian inference algorithm implemented in `smfsb` across 100 simulations, as a function of the time interval length ($\Delta t$), in terms of (a) parameter estimation and (b) computational time. The methods perform similarly in terms of estimation, but the Bayesian method has a higher computational cost.

### *4.2. Accuracy and computational time varying r and p*

This section studies the computational cost of the proposed approach further. In particular, the impact of varying the number of reactions $r$ and the number of particle types $p$ on the estimation accuracy and computational time required to estimate the parameters $\hat{\theta}$ is examined. This analysis is conducted using the same experimental setup of section 4.1, fixing $T = 20$. The simulations are repeated 200 times.

The time complexity is shown with the median time across all simulations, along with the 25th and 75th percentiles. In terms of accuracy of parameter estimation, the Mean Absolute Error (MAE) between the estimated parameter and the corresponding true parameter is considered. This is given by

$$MAE(\theta, \hat{\theta}) = \frac{1}{r} \sum_{j=1}^{r} \left| \theta_j - \hat{\theta}_j \right|.$$

A first simulation varies the number of particle types $p \in \{3, 6, 9\}$. For $p = 3$, the cyclic reaction system in section 2.4 is considered. In order to still have a $P_\theta$ invertible, the settings with a higher $p$ are generated by simply repeating the same system a number of times, as shown in Fig. E.10 of Appendix E. The results of the simulations are shown in Fig. 6. The top-left panel shows a super-linear dependence of computational time on $p$. The bottom-left panel shows that the accuracy of the estimates decreases linearly with respect to the number of $p$ states.

In a second simulation, the number of particles is fixed at $p = 3$ while progressively varying the number of reactions $r \in \{3, 6, 9, 12, 15\}$. The different systems with varying $r$ are generated by following the scheme in Fig. E.11 of Appendix E. The top-right panel of Fig. 6 shows a quasi-cubic dependence of computational time on $r$, which aligns with the theoretical predictions discussed in section 3.1. The bottom right plot shows how the accuracy of individual estimates degrades approximately linearly with the increase in the number of reactions.

### *4.3. Comparison with other moment-based methods*

Finally, a simulation study is conducted by comparing the proposed LMA method with existing moment-based methods. These include the recently developed approaches of Pellin et al. (2023), based on moment-closure, and Xu et al. (2019), involving the generalized method of moments. Differently to LMA, the two methods can be used only on specific systems: Pellin et al. (2023) require systems of up to second order reactions, while Xu et al. (2019) require systems where the theoretical second moment can be calculated analytically. Since the implementation of Pellin et al. (2023) is currently not available, the proposed method is compared with that of Xu et al. (2019).

For the comparison, the human cell differentiation system described in Xu et al. (2019) is considered. The system is shown in Fig. 7a and involves hematopoietic stem cells (HSCs), two progenitors, and five mature cell types. To obtain an analytical solution for the evolution of the moments, Xu et al. (2019) assume linear propensity functions with respect to the state concentrations. This concept is strongly restrictive in the description of cell dynamics as it implies only exponential growth and extinction of the cells, as also noted by Pellin et al. (2023). However, in order to properly compare the two methods, the same assumption has been kept.

As HSCs and progenitor cells are latent in the reference method, the dataset with mature cells is identified only after generating the complete data, in order to maintain the same original setting. In terms of reaction rates, the original values are multiplied by

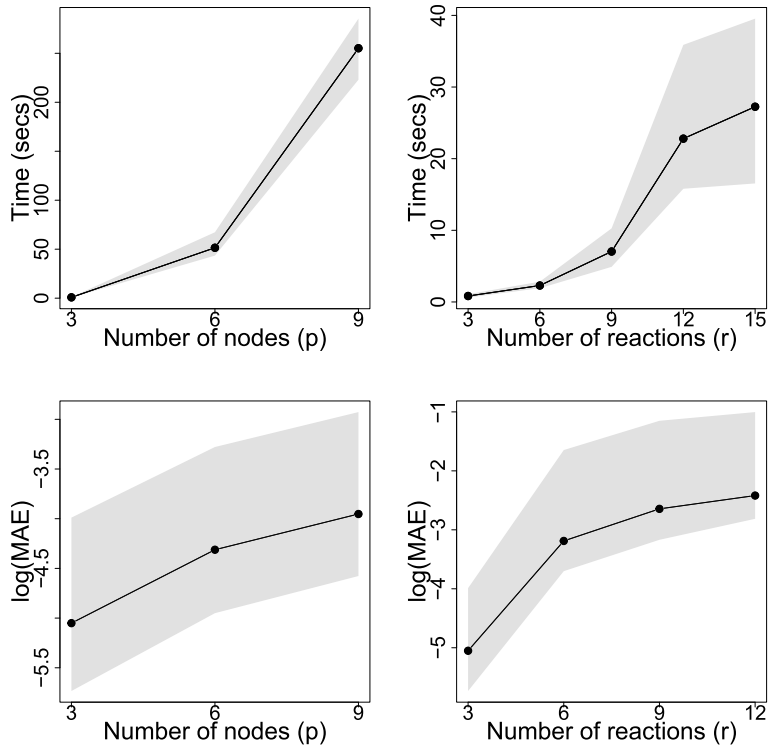



**Fig. 6.** Median computational time (top) and MAE in the log scale (bottom) of the LMA algorithm, as a function of the number of reactions ($r$, right) and states ($p$, left). The shaded area represents the interquartile range across 200 simulations.

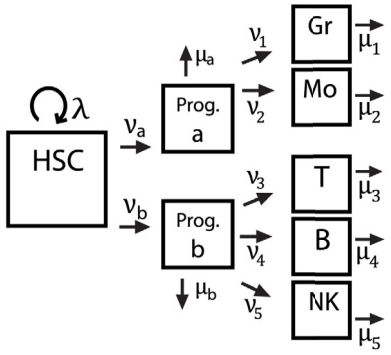
1000 — this merely constitutes a time unit change. The HSC duplication parameter is set to $\lambda = 2850$, and the differentiation in progenitors a-type to $\nu_a = 1400$, and b-type to $\nu_b = 700$, respectively. Such cells have death rates of $\mu_a = 50$ and $\mu_b = 40$, respectively. The type-a progenitor can differentiate into Granulocytes with rate $\nu_1 = 3600$ and into Monocytes with rate $\nu_2 = 1800$. The type-b progenitor differentiates into T-cells with rate $\nu_3 = 1000$, into B-cells with rate $\nu_4 = 2000$ and finally into Natural-Killer (NK) cells with rate $\nu_5 = 1200$. The death rates of the 5 mature cells are given by $\mu_1 = 26, \mu_2 = 13, \mu_3 = 11, \mu_4 = 16, \mu_5 = 9$, respectively.

Each simulation starts from one HSC and no other cell types. From this, $n = 100$ replicates or clones, are simulated in $T = 5$ time steps each. The observation times are stochastically defined, but in order to maintain the biological significance of blood sampling at defined times, the indexed mean time is defined for all clones, resulting in a data span $t = (0.00, 0.08, 0.11, 0.14, 0.17)$. The proposed LMA approach and the generalized method of moments (Xu et al., 2019) are applied to these simulated data. As in the original publication, the death rates are kept fixed, as these values are taken from the biology and immunology literature, while all other parameters are estimated from data. Given that the generalized method of moments relies on an initial value, a sensitivity checking is performed by restarting the inference procedure 100 times for each simulation. The solution corresponding to the minimum cost function is retained as the final result.

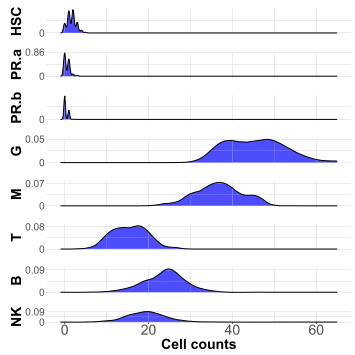
The results of the simulation study, the steady-state distribution process and the parameter distributions with 300 simulations are shown in Fig. 7. The proposed LMA approach outperforms Xu et al. (2019) in several aspects. Firstly, most of the LMA parameter estimates, with the possible exception of $\hat{\nu}_1$, are unbiased, in contrast to the estimates obtained with the generalized method of moments. Secondly, the precision of the LMA estimates is significantly higher than that of the Xu et al. (2019) estimates. The reason for these improvements is that the generalized method of moments in Xu et al. (2019) is based on matching second order moments, which are inherently less stable than first order methods. The moment-closure method of Pellin et al. (2023) considers both first and second moments and has therefore a better performance for these quasi-reaction systems where it can be implemented. This is shown by the simulation study included in Pellin et al. (2023) where parameter estimates appear to have a higher precision than those obtained by LMA.

## 5. Cell lineage barcoding in rhesus macaques

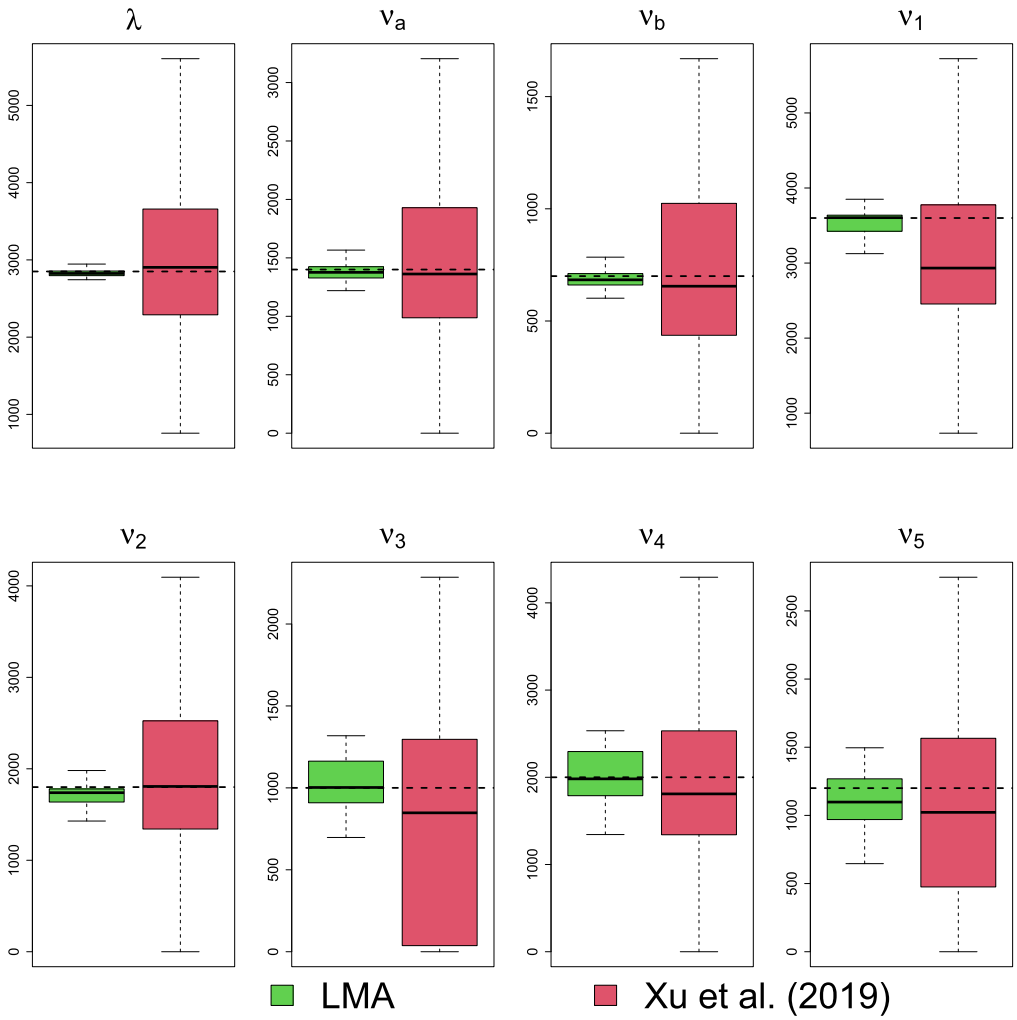
Clonal tracing modeling is commonly used in genetic studies to enhance the understanding of blood cell formation, also known as hematopoiesis. The hematopoietic process is represented as a tree structure with a self-renewing hematopoietic stem cell at its origin. This cell evolves from a pluripotent state into mature blood cells through several intermediate progenitor stages. The regulation of the number of circulating cells in the blood is maintained through the rates of birth, death, and differentiation. An important challenge in studying this system in living organisms is that only mature cells are observable and can be accessed through blood sampling. To gain further insights, recent research has analyzed cell production in non-human primates, which closely mirrors human physiology due




(a)




(b)




(c)

**Fig. 7.** Comparison with Xu et al. (2019) correlation-based M-estimator: a) Blood cell differentiation scheme. b) Cell types multi-modal steady-state distribution calculated using 100 clones. c) Boxplots of the estimated parameter distributions from 300 simulations. The 8 unknown rates are shown, the true values are indicated by a horizontal dashed red line. The proposed method is unbiased and more accurate than the M-estimator by Xu et al. (2019).

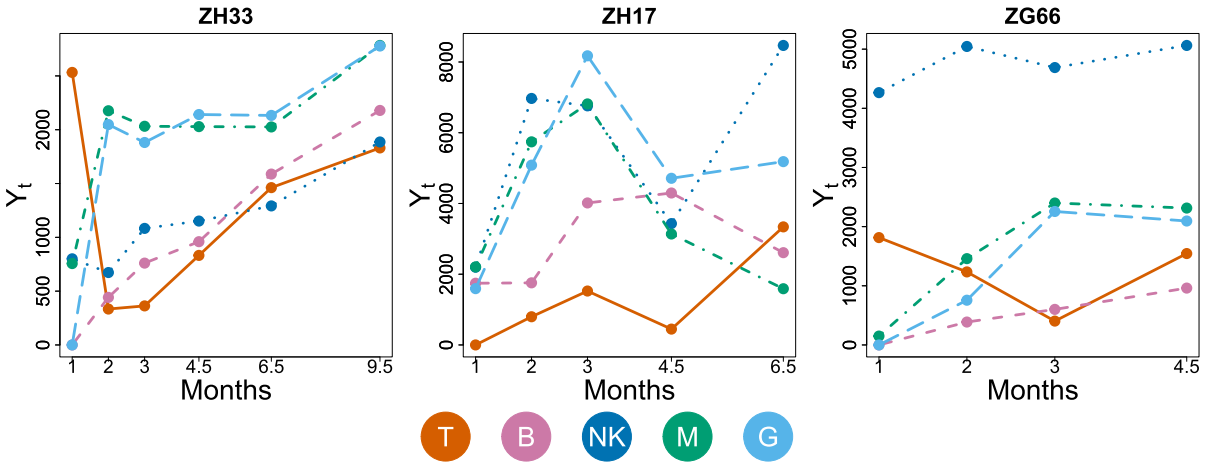



**Fig. 8.** Average concentration over time of each Rhesus Macaque specimen following transplantation. The $p = 5$ cell types are reported with different colors and line styles.

to the similar lifespan and frequencies of hematopoietic stem progenitor cells (HSPCs) (Kim et al., 2000). An in-vivo clonal tracking dataset on Rhesus Macaques (Wu et al., 2014) is considered, with the aim of recovering the rates of birth, death, and differentiation from these data.

*Hematopoietic stem cell gene therapy.* In the (Wu et al., 2014) gene therapy clonal study, unique DNA barcodes IDs were first introduced into autologous CD34+ HSPCs utilizing a high-diversity lentiviral barcode library. Then, such cells were reinfused into the three animals (Shepherd et al., 2007). Once reinfused, the genetically modified HSPCs home to the bone marrow, where they engraft and begin to repopulate the haematopoietic system. These barcoded HSPCs proliferate and differentiate into various blood cell lineages, allowing the tracking of individual clones over time. The integrated barcodes remain stable within the genome of the progeny cells, enabling the detailed monitoring of clonal dynamics, lineage contribution, and the longevity of specific clones across different haematopoietic compartments. In the Wu et al. (2014) study, samples from peripheral blood, bone marrow, and lymph nodes for the three monkeys were collected monthly, enabling the identification of unique clonal patterns and contributions to different cell types, namely Granulocytes (G), Monocytes (M), T, B and Natural-Killer (NK) cells. The entire observation time varies from subject to subject and corresponds to 4.5 months for monkey ZG66, 6.5 months for monkey ZH17, and 9.5 months for monkey ZH33. Each barcoded lineage $\boldsymbol{Y}(t)$ is assumed to be an independent realization of the hematopoietic stochastic model.

The data were imported from the `Karen` library (Del Core et al., 2022). The dataset contains many missing sampling events. As a pre-processing step, the time points when no barcodes were detected are excluded as well as all clones with less than 3 temporal observations. This leads to a total of 555 unique barcodes IDs, which are split between 434, 50, and 19 different clone-types in specimen ZG66, ZH17 and ZH33, respectively.

*Hematopoietic reaction network selection.* Fig. 8 shows the average differentiation trajectories for each specimen. An assumption is made that the dynamics of cell differentiation are universal and thus independent of individual subjects. Consequently, the differentiation process in each subject is assumed to be governed by the same rate parameters. Following the same approach as in Pellin et al. (2023), the death reaction hazard rates are defined as a quadratic function of the underlying abundances, representing a natural saturation effect.

Using the proposed LMA approach, competing models of hematopoiesis are compared from the available experimental data. To this end, the net matrix of the full model is defined, consisting of 30 reactions, of which 10 are birth and death reactions and 20 are all possible differentiations from one cell to another. The reaction rates are initialized with the local linear approximation estimates. Although previous similar studies fix the death parameters according to established chemical regimes (Hellerstein et al., 1999; Xu et al., 2019), such rates are kept variable. A search through the space of possible models is then conducted, going from an initial model $m_1$ consisting of only one reaction out of the 30 possible ones to the full saturated model. For each model, the parameters are estimated using the LMA approach, i.e., by solving the optimization problem (11). At each step of a stepwise procedure, the reaction that most reduces the Bayesian Information Criterion (BIC) is iteratively added and subtracted. This methodology leads to a sequence of models with increased complexity. The results are shown in Fig. 9 (right). For each degree of complexity, the lowest BIC of the optimal model is shown in red, while the BIC values of the unselected models are shown in gray.

The optimal model and complexity level, i.e., the model corresponding to the lowest overall BIC, contains 10 reactions. Table 2 reports the reactions that define this model, the corresponding reaction rates estimated by the proposed LMA method, together with the standard errors calculated as described in section 3.2. Due to the quadratic nature of the death rates and to particle counts in the order of $10^3$ (Fig. 8), reactions involving $M$ and $B$ tend to be the slowest (with rates in the order of 10 days), reactions involving $G$ and $T$ occur at a rate of 1 day, while death rates occur at the fastest rate of $10^{-1}$ days. Standard errors are small with respect to the size of the parameters, suggesting small uncertainty on the estimates.

**Table 2**
Estimated rates, and corresponding standard errors, from Rhesus Macaques gene therapy trial data, based on the optimal model of cell differentiation.

| reaction | reaction rate estimate | Standard errors |
|---|---|---|
| $G \to \emptyset$ | $3.054 \cdot 10^{-7}$ | $2.405 \times 10^{-9}$ |
| $NK \to \emptyset$ | $1.759 \cdot 10^{-7}$ | $3.236 \times 10^{-10}$ |
| $M \to G$ | $1.340 \cdot 10^{-2}$ | $3.593 \times 10^{-4}$ |
| $B \to M$ | $1.183 \cdot 10^{-2}$ | $4.916 \times 10^{-4}$ |
| $M \to B$ | $1.180 \cdot 10^{-2}$ | $3.69 \times 10^{-4}$ |
| $G \to NK$ | $2.339 \cdot 10^{-3}$ | $1.935 \times 10^{-4}$ |
| $G \to M$ | $5.876 \cdot 10^{-3}$ | $4.123 \times 10^{-4}$ |
| $B \to G$ | $7.048 \cdot 10^{-3}$ | $4.690 \times 10^{-4}$ |
| $G \to T$ | $6.413 \cdot 10^{-3}$ | $4.846 \times 10^{-4}$ |
| $T \to B$ | $9.469 \cdot 10^{-3}$ | $1.950 \times 10^{-4}$ |

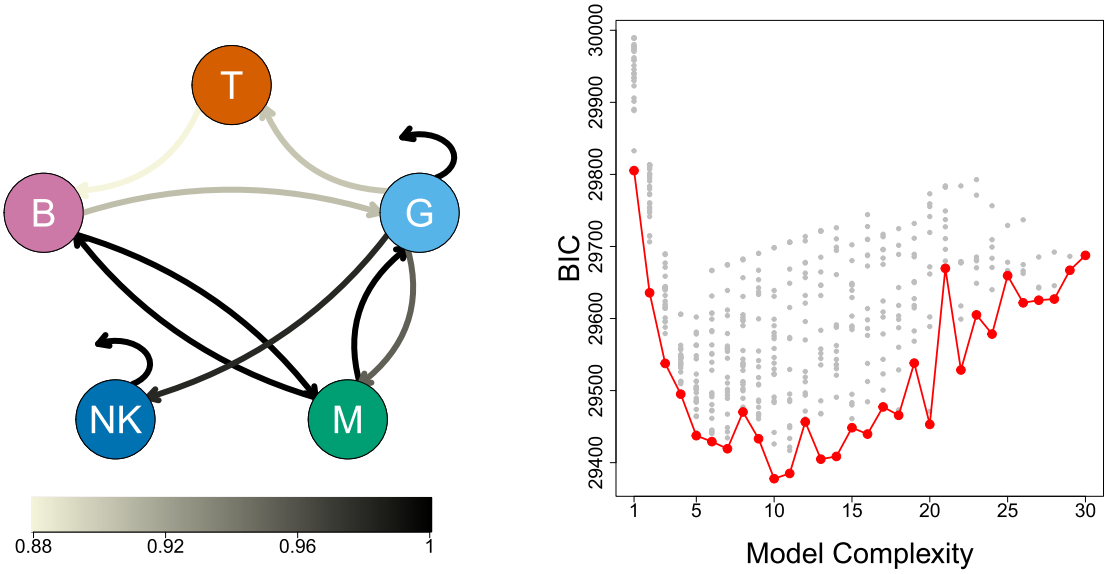



**Fig. 9.** BIC model selection on Rhesus Macaques data. Right: Reaction systems ordered by complexity, from the simplest (single-reaction model) to the most complex (including all possible birth, death, and single-reactant differentiation events). For each complexity level, the BIC of the best model is shown in red, while the others are shown in gray. The optimal overall model contains 10 reactions. Left: HSC differentiation process, where the thickness of each arrow refers to the rescaled BIC-based weights described in (15).

In the search of all models, each reaction may be included in a number of models of varying complexity. As a way of obtaining a measure $p_j$ of overall relevance of each reaction, the evidence that a specific reaction is included is determined by summing the weights of all models that include this reaction. Formally, $p_j = \sum_{i \in \mathcal{M}_j} w_i$, where $\mathcal{M}_j$ is the set of all models that contain the parameter $\theta_j$, and $w_j$ is the rescaled BIC-based weight. This is defined by

$$w_j = \frac{\exp\{-\frac{1}{2}(BIC_j - BIC_{\min})\}}{\sum_{h=1}^{r} \exp\{-\frac{1}{2}(BIC_h - BIC_{\min})\}}, \tag{15}$$

where $BIC_{\min}$ is the minimum BIC value among all the models considered. The cell differentiation network in Fig. 9 (left) has edges whose thickness is proportional to the $p_j$ value. The figure shows, how, although each element appears both as a reactant and a product in the optimal model (Table 2), NK cells do not tend to differentiate into other cell types. Moreover, the less frequent edges form a loop connecting nodes B, T, and G cells, while the reaction events that appear most relevant for explaining the count trajectories involve monocytes.

## 6. Conclusion

A novel methodology for parameter inference in quasi-reaction systems is developed and assessed, with particular focus on situations where observations are made at large time intervals. Traditional local linear approximation methods, while computationally efficient, often fail to capture the complex nonlinear dynamics inherent in biological systems, especially when data are sparse or irreg-

ularly spaced. The proposed approach, which extends mean-field approximation techniques, addresses these limitations by providing an explicit solution for the first moments of the state distributions under a generic quasi-reaction system.

The performance of the proposed local mean-field approximation method is evaluated through an extensive simulation study and compared against alternative methods. The results demonstrate that the method significantly outperforms local linear approximation, particularly as the time interval between observations increases, while retaining computational efficiency in comparison to Bayesian inferential procedures (Boys et al., 2008). Unlike moment-closure methods, the approach can be applied to any quasi-reaction system. Moreover, due to the availability of an explicit solution, the method is robust to stiffness, a common occurrence in biological systems where processes operate at vastly different timescales. The approach is illustrated on a study of cell differentiation from gene therapy clonal tracking data, providing estimates of cell population dynamics and meaningful insights into the underlying biological processes.

The methodology advances the inference of quasi-reaction systems by offering a versatile and reliable tool suited to any generic quasi-reaction system, particularly valuable for applications in compartmental studies and multi-type branching models, where traditional methods may be inadequate or computationally prohibitive.

## Appendix A. Derivation of the first moment equation

The derivation of the first moment equation in (4) is provided for completeness. Following from the definition of $\boldsymbol{m}(t+s|t)$, substituting the master equation (3) and exchanging the order of summation, leads to

$$\begin{aligned}
\frac{dm_l(t+s|t)}{ds} &= \frac{d}{ds}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l p_{t+s}(\boldsymbol{y}) = \sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l \frac{dp_{t+s}(\boldsymbol{y})}{ds} \\
&= \sum_{j=1}^{r}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l\left[p_{t+s}(\boldsymbol{y}-\boldsymbol{v}_{\cdot j})\lambda_j(\boldsymbol{y}-\boldsymbol{v}_{\cdot j};\theta_j) - p_{t+s}(\boldsymbol{y})\lambda_j(\boldsymbol{y};\theta_j)\right] \\
&= \sum_{j=1}^{r}\left[\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}(y_l+v_{lj})\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y}) - \sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y})\right] \\
&= \sum_{j=1}^{r}\left[\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y}) + v_{lj}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y}) - \sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y})\right] \\
&= \sum_{j=1}^{r}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} v_{lj}\lambda_j(\boldsymbol{y};\boldsymbol{\theta})p_{t+s}(\boldsymbol{y}) \\
&= \sum_{j=1}^{r}\mathbb{E}\left[v_{lj}\lambda_j(\boldsymbol{Y}(t+s);\boldsymbol{\theta})|\boldsymbol{Y}(t)\right].
\end{aligned}$$

## Appendix B. Jacobian of the hazard function

**Proposition.** *Given the intensity function* $\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\theta}) = \theta_j\prod_{l=1}^{p}\binom{Y_l(t)}{k_{lj}}$, *the* $jl$*-th element of the Jacobian matrix* $\Lambda(\boldsymbol{Y}(t);\boldsymbol{\theta})\in\mathbb{R}^{r\times p}$ *is given by*

$$\Lambda_{jl} = \theta_j\prod_{i=1}^{p}\binom{Y_i(t)}{k_{ij}}(1-\delta_{il})\binom{Y_l(t)}{k_{lj}}\left(\psi(Y_l(t)+1)-\psi(Y_l(t)-k_{lj}+1)\right).$$

**Proof.** The $jl$-th element of the Jacobian matrix $\Lambda(\boldsymbol{Y}(t);\boldsymbol{\theta})$ is given by

$$\frac{\partial}{\partial Y_l}\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\theta}) = \frac{\partial}{\partial Y_l}\theta_j\prod_{i=1}^{p}\binom{Y_i(t)}{k_{ij}}.$$

Application of the chain rule leads to

$$\frac{\partial}{\partial Y_l}\prod_{i=1}^{p}\binom{Y_i(t)}{k_{ij}} = \left(\prod_{i=1}^{p}\binom{Y_i(t)}{k_{ij}}(1-\delta_{il})\right)\frac{\partial}{\partial Y_l}\binom{Y_l(t)}{k_{lj}},$$

with $\delta_{il}$ an indicator function. Next, recalling the digamma function $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$, the logarithm of the binomial coefficient is differentiated, leading to:

$$\frac{\partial}{\partial Y_l}\log\binom{Y_l(t)}{k_{lj}} = \frac{\partial}{\partial Y_l}\left[\log\Gamma(Y_l(t)+1) - \log\Gamma(k_{lj}+1) - \log\Gamma(Y_l(t)-k_{lj}+1)\right]$$

$$= \psi(Y_l(t)+1) - \psi(Y_l(t)-k_{lj}+1).$$

Applying logarithmic differentiation,

$$\frac{\partial}{\partial Y_l}\binom{Y_l(t)}{k_{lj}} = \binom{Y_l(t)}{k_{lj}}\left(\psi(Y_l(t)+1) - \psi(Y_l(t)-k_{lj}+1)\right).$$

Therefore, the $jl$-th element of the Jacobian matrix $\Lambda(\mathbf{Y}(t);\boldsymbol{\theta})$ is given by

$$\Lambda_{jl} = \theta_j \prod_{i=1}^{p}\binom{Y_i(t)}{k_{ij}}(1-\delta_{il})\binom{Y_l(t)}{k_{lj}}\left(\psi(Y_l(t)+1) - \psi(Y_l(t)-k_{lj}+1)\right). \quad \square$$

## Appendix C. LMA algorithm

Algorithm 1 gives the pseudo-code of the proposed nonlinear local mean-field (LMA) algorithm for parameter estimation in a quasi-reaction model.

**Algorithm 1** Nonlinear local mean-field (LMA) algorithm.

**Data:** $\boldsymbol{Y}$, $K$, $V$
**Initialization:** $\hat{\boldsymbol{\theta}}_0$, $toll$, $maxIter$, $k=0$, $H_0 = \mathbb{I}_r$
Define $f(\boldsymbol{\theta}) = \sum_i^{T_c}\sum_c^{C}(\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci})^T(\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci})$
**while** $\|\nabla f(\boldsymbol{\theta}_k)\| > \epsilon$ **and** $k < maxIter$ **do**
  $\boldsymbol{d}_k = -H_k \nabla f(\boldsymbol{\theta}_k)$
  $\alpha_k = \arg\min_\alpha f(\boldsymbol{\theta}_k + \alpha \boldsymbol{d}_k)$
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \boldsymbol{d}_k$ ▷ update $\boldsymbol{m}$ according to (6)
  $\boldsymbol{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$
  $\boldsymbol{g}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$
  $H_{k+1} = \left(\mathbb{I}_r - \frac{\boldsymbol{s}_k \boldsymbol{g}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}\right) H_k \left(\mathbb{I}_r - \frac{\boldsymbol{g}_k \boldsymbol{s}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}\right) + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}$
  $k = k+1$
$\hat{\boldsymbol{\theta}}_{LMA} = \boldsymbol{\theta}_k$

## Appendix D. Local Linear Approximation

This section describes the Local Linear Approximation (LLA) approach for estimation of the rates $\boldsymbol{\theta}$. The LLA estimates are used in the comparative study in section 4.1 and as initial values for Algorithm 1 that iteratively solves the optimization problem (11).

The local linear approximation approach applies Euler's method to approximate the moments of the process at time $t$ conditional on the history of the process up to time $t$. Using the expression of the conditional moments derived from the chemical master equation (3), and assuming constant hazard rates between consecutive time points, the conditional moments are approximated as follows:

$$\mathbb{E}[\boldsymbol{Y}_{ci} \mid \boldsymbol{Y}_{c,i-1}] \simeq \boldsymbol{Y}_{c,i-1} + \sum_{j=1}^{r} v_{lj}\lambda_j(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta})\Delta t_i \tag{D.1}$$

$$\begin{aligned}\mathbb{E}[Y_{c,i,l}Y_{c,i,k} \mid \boldsymbol{Y}_{c,i-1}] \simeq\ & Y_{c,i-1,l}Y_{c,i-1,k} + \sum_{j=1}^{r} v_{lj}Y_{c,i-1,k}\lambda_j(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta})\Delta t_i \\ & + \sum_{j=1}^{r} v_{kj}Y_{c,i-1,l}\lambda_j(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta})\Delta t_i + \sum_{j=1}^{r} v_{lj}v_{kj}\lambda_j(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta})\Delta t_i.\end{aligned} \tag{D.2}$$

Since the hazard function $\lambda_j(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, the regression model (10) can be rewritten as:

$$\Delta\boldsymbol{Y}_{ci} = M_{ci}\boldsymbol{\theta} + \boldsymbol{\varepsilon}_{ci}, \quad \boldsymbol{\varepsilon}_{ci} \sim \mathcal{N}(\mathbf{0}, \Omega_{ci}),$$

where $\Delta\boldsymbol{Y}_{ci} = \boldsymbol{Y}_{ci} - \boldsymbol{Y}_{c,i-1}$ is the vector of concentration changes between consecutive time points. The matrix $M_{ci}\boldsymbol{\theta} = V\,\mathrm{diag}(\boldsymbol{\lambda}_{ci}(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta}))\Delta t_i$ and $\Omega_{ci} = V\,\mathrm{diag}(\lambda_{ci}(\boldsymbol{Y}_{c,i-1};\boldsymbol{\theta}))V^T\Delta t_i$ represent the discretized drift function and the dispersion matrix of the concentration differentiation process, respectively.

The local linear estimate $\hat{\boldsymbol{\theta}}_{\mathrm{LLA}}$ is then obtained by solving the following constrained generalized least-squares problem:

$$\hat{\boldsymbol{\theta}}_{\mathrm{LLA}} = \arg\min_{\boldsymbol{\theta}\geq\mathbf{0}_r} \sum_{i=1}^{T_c}\sum_{c=1}^{n}\left(\Delta\boldsymbol{Y}_{ci} - M_{ci}\boldsymbol{\theta}\right)^T \Omega_{ci}^{-1}\left(\Delta\boldsymbol{Y}_{ci} - M_{ci}\boldsymbol{\theta}\right).$$

$$p=3 \qquad p=6 \qquad p=9$$

$$\mathcal{R}_3: \begin{cases} 2A_1 \xrightarrow{0.2} 2B_1 \\ A_1+B_1 \xrightarrow{0.1} 3C_1 \\ 2C_1 \xrightarrow{0.2} 2A_1 \end{cases} \quad \mathcal{R}_6: \mathcal{R}_3 \cup \begin{cases} 2A_2 \xrightarrow{0.2} 2B_2 \\ A_2+B_2 \xrightarrow{0.1} 3C_2 \\ 2C_2 \xrightarrow{0.2} 2A_2 \end{cases} \quad \mathcal{R}_9: \mathcal{R}_6 \cup \begin{cases} 2A_3 \xrightarrow{0.2} 2B_3 \\ A_3+B_3 \xrightarrow{0.1} 3C_3 \\ 2C_3 \xrightarrow{0.2} 2A_3 \end{cases}$$

**Fig. E.10.** Dynamic systems with an increasing number of particles ($p=3,6,9$). Rates are reported above each reaction.

$$\begin{array}{llllllll} R_1: & 2A \xrightarrow{0.2} 2B & R_4: & 2A \xrightarrow{0.01} \emptyset & R_7: & 2A \xrightarrow{0.1} 3B & R_{10}: & A \xrightarrow{0.1} C \\ R_2: & A+B \xrightarrow{0.1} 3C & R_5: & C \xrightarrow{0.02} \emptyset & R_8: & C \xrightarrow{0.06} B & R_{11}: & B+C \xrightarrow{0.09} A \\ R_3: & 2C \xrightarrow{0.2} 2A & R_6: & 2C \xrightarrow{0.03} \emptyset & R_9: & C \xrightarrow{0.05} 2A & R_{12}: & B \xrightarrow{0.08} 2A+C \end{array}$$

**Fig. E.11.** Four dynamic systems are illustrated, with the first scenario including $\{R_1, R_2, R_3\}$ reactions. Each subsequent scenario adds 3 additional reactions to the previous one, culminating in a final system with 12 reactions. Reaction rates are reported above each reaction.

## Appendix E. Reaction systems used in the simulation study in section 4.2

Figure E.11 outlines the reaction systems used in section 4.2 to evaluate the computational complexity of the algorithm. Concerning the number of particles ($p$), for ease of notation, the possible states are indicated only by the first three letters. When multiple of the three states are present, these are denoted with a subscript.

## References


Baccelli, F., Karpelevich, F., Kelbert, M., Puhalskii, A., Rybko, A., Suhov, Y., 1992. A mean-field limit for a class of queueing networks. J. Stat. Phys. 66, 803–825.

Bequette, B., 1998. Process dynamics: modeling, analysis, and simulation. In: Prentice-Hall International Series in the Physical and Chemical Engineering Sciences. Prentice Hall PTR.

Boys, R.J., Wilkinson, D.J., Kirkwood, T.B., 2008. Bayesian inference for a discretely observed stochastic kinetic model. Stat. Comput. 18, 125–135.

Butcher, J.C., 2016. Numerical Methods for Ordinary Differential Equations. John Wiley & Sons.

Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16 (5), 1190–1208.

Córdoba-Torres, P., Enriquez, F., Fairén, V., 1998. Optimal start of a Taylor integrator by control of local error. Comput. Phys. 12 (2), 200–207.

Del Core, L., Pellin, D., Grzegorczyk, M., Wit, E., 2022. Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data. bioRxiv. https://doi.org/10.1101/2022.07.08.499353.

Del Core, L., Pellin, D., Wit, E., Grzegorczyk, M., 2023. Scalable inference of cell differentiation networks in gene therapy clonal tracking studies of haematopoiesis. Bioinformatics 39 (10), btad605.

Framba, M., Vinciotti, V., Wit, E.C., 2024. Latent event history models for quasi-reaction systems. Comput. Stat. Data Anal. 198, 107996.

Gillespie, D.T., 2007. Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. 58, 35–55.

Golightly, A., Wilkinson, D., 2005. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics 61 (3), 781–788.

Grima, R., 2012. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. J. Chem. Phys. 136 (15), 154105.

Hale, J.K., Koçak, H., 2012. Dynamics and Bifurcations. Springer-Verlag.

Hellerstein, M., Hanley, M., Cesar, D., Siler, S., Papageorgopoulos, C., Wieder, E., Schmidt, D., Hoh, R., Neese, R., Macallan, D., et al., 1999. Directly measured kinetics of circulating t lymphocytes in normal and hiv-1-infected humans. Nat. Med. 5 (1), 83–89.

Kennealy, J.P., Moore, W.M., 1977. A numerical method for chemical kinetics modeling based on the Taylor series expansion. J. Phys. Chem. 81 (25), 2413–2419.

Kim, H.J., Tisdale, J.F., Wu, T., Takatoku, M., Sellers, S.E., Zickler, P., Metzger, M.E., Agricola, B.A., Malley, J.D., Kato, I., et al., 2000. Many multipotential gene-marked progenitor or stem cell clones contribute to hematopoiesis in nonhuman primates. Blood J. Am. Soc. Hematol. 96 (1), 1–8.

Krijnen, W., Wit, E., 2022. Computational and Statistical Methods for Chemical Engineering. Chapman and Hall/CRC.

Lambert, J., 1974. Two Unconventional Classes of Methods for Stiff Systems. Springer US, Boston, MA, pp. 171–186.

Lente, G., Fursenko, A., Szabo, R., 2022. Use of the Taylor theorem to predict kinetic curves in an arbitrary mechanism. Chem. Eng. J. 445, 136676.

Milner, P., Gillespie, C.S., Wilkinson, D.J., 2013. Moment closure based parameter inference of stochastic kinetic models. Stat. Comput. 23, 287–295.

Pellin, D., Biasco, L., Scala, S., Di Serio, C., Wit, E., 2023. Tracking hematopoietic stem cell evolution in a Wiskott-Aldrich clinical trial. Ann. Appl. Stat. 17 (3), 1841–1860.

Robertson, H., 1967. The Solution of a Set of Reaction Rate Equations. Academic Press, Cambridge, MA, pp. 178–182.

Schnakenberg, J., 1976. Network theory of microscopic and macroscopic behavior of master equation systems. Rev. Mod. Phys. 48 (4), 571–585.

Shepherd, B.E., Kiem, H.-P., Lansdorp, P.M., Dunbar, C.E., Aubert, G., LaRochelle, A., Seggewiss, R., Guttorp, P., Abkowitz, J.L., 2007. Hematopoietic stem-cell behavior in nonhuman primates. Blood J. Am. Soc. Hematol. 110 (6), 1806–1813.

Shoji, I., 2013. Nonparametric estimation of nonlinear dynamics by metric-based local linear approximation. Stat. Methods Appl. 22, 341–353.

Tsoularis, A., Wallace, J., 2002. Analysis of logistic growth models. Math. Biosci. 179 (1), 21–55.

Tsugé, S., 2001. Rate increase in chemical reaction and its variance under turbulent equilibrium. Combust. Sci. Technol. 162 (1), 303–330.

Van Kampen, N., 1992. Stochastic Processes in Physics and Chemistry. Elsevier.

Wilkinson, D., 2018. Stochastic Modelling for Systems Biology, third edition. Chapman and Hall/CRC.

Wu, C., Li, B., Lu, R., Koelle, S.J., Yang, Y., Jares, A., Krouse, A.E., Metzger, M., Liang, F., Loré, K., et al., 2014. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. Cell Stem Cell 14 (4), 486–499.

Xu, J., Koelle, S., Guttorp, P., Wu, C., Dunbar, C., Abkowitz, J., Minin, V., 2019. Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. Ann. Appl. Stat. 13 (4), 2091–2119.

Ye, Y., Cai, S., 1986. Theory of Limit Cycles. Translations of Mathematical Monographs, vol. 66. American Mathematical Society.