University of Trento

Department of Mathematics

Doctor of Philosophy degree in Mathematics

~ · ~

XXXVII cycle

Advanced Statistical Inference for Stochastic Quasi-Reaction Systems

Supervisors
Prof. Veronica VINCIOTTI
Prof. Ernst C. WIT

Candidate
Dr. Matteo FRAMBA

Committee members
Prof. Danilo PELLIN
Prof. Umberto PICCHINI

Final examination date: January 16, 2025

Contents

Summary of Notation

The following are the fixed notations in the text. Capital letters are used for random variables and matrices; lowercase letters are used for random variable realisations. Vectors are written in bold type. Additionally, we use various letters for certain quantities, described below:

Symbol	**Description**
p	Number of particle types in the system
r	Number of reactions
q	Number of covariates, including the intercept
N	Time intervals (Chap. 3-4)
T	Time points (Chap. 5)
C	Replicates (Chap. 5)
k_{lj}	Number of particles of type l consumed in reaction j
s_{lj}	Number of particles of type l produced by reaction j
v_{lj}	Net change in particle count of type l due to reaction j
$\boldsymbol{\theta}$	Reaction rates
$\boldsymbol{Y}(t)$	Particle concentration process at time t
$\boldsymbol{m}(t)$	Mean of the process $\boldsymbol{Y}(t)$ conditional on filtration $\mathcal{F}^-$
$\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta})$	Hazard rate of reaction j when system is in state $\boldsymbol{y}$

Terminology

CME	Chemical Master Equation
LLA	Local Linear Approximation
LMA	Local Mean-Field Approximation

A process cannot be understood by stopping it. Understanding must move with the flow of the process, must join it and flow with it.

— Frank Herbert, *Dune*

Chapter 1

Introduction

1.1 Research Motivation

The growing need to comprehensively interpret the complex dynamics of various natural phenomena such as stem cell differentiation and the spread of infectious diseases, has driven the development and widespread adoption of advanced statistical modeling frameworks. Specifically, quasi-reaction systems effectively describe the temporal dynamics of various biological and biochemical processes, commonly governed by stochastic differential equations (Wilkinson, 2018; Britton et al., 2019; Craigmile et al., 2023). Accurate inference of the unknown parameters which control the system dynamics is important for predicting and characterizing their evolution over time.

Traditional methods for parameter estimation, such as the Local Linear Approximation (LLA), give an explicit approximation of the likelihood function under specific assumptions (Shoji and Ozaki, 1998). However, LLA methods are susceptible to bias when observations are too closely spaced in time (Komorowski et al., 2011; Framba et al., 2024). Several approaches have been suggested to reduce the variance of the estimates in such multi-response non-linear models. To name a few, Fedorov (2013) proposed an optimal design technique that relies on knowledge of the variance-covariance matrix. However, such approach is often computationally inefficient and numerical unstable in certain scenarios. Tikhonov regularization techniques (Engl et al., 1996) offer an alternative way, though when measurements are spaced too closely in time, concentrations may remain constant, resulting in zero standard deviations and making regularization impractical.

A further limitation of the current quasi-reaction framework is the absence of methods incorporating time-varying covariates to model reaction rates. In many applications, such as epidemic models or genetic studies, including extrinsic covariates like environmental conditions, population demographics, and spatial variability can significantly improve the description of the model's dynamics (Britton et al., 2019). Additionally, it is reasonable to expect that reaction rates may fluctuate over time due to changes in these external factors. Existing approaches typically assume constant rates, which limit their ability to fully capture the complexity of the dynamics.

Another significant challenge in inferring reaction rates in stochastic quasi-reaction systems arises when observations are sampled at widely spaced intervals. The system can evolve considerably between observations, making it difficult to accurately infer the reactions that occurred during these gaps (Milner et al., 2011). Traditional methods, such as the Local Linear Approximation mentioned above, are inadequate in these scenarios because their accuracy diminishes as the time intervals increase. Mean-field approximation techniques provide a more effective solution by focusing on the system's average dynamics rather than attempting to track every individual event (Baccelli et al., 1992). These methods are computationally efficient and work particularly well in unitary systems, where each reaction transforms a single element into one or more products. In such cases, it is possible to derive explicit solutions for the ordinary differential equations (ODEs) governing the process's mean. For higher-order reactions, explicit solutions are not feasible, and alternative approximation algorithms have been developed, such as the moment-closure numerical method of Pellin et al. (2023), which approximates the first moments of the system by truncating higher-order terms, and the time scale separation approach of Lente et al. (2022), which applies Taylor series expansions to exploit differences in reaction timescales. However, these approaches involve either solving computationally expensive non-linear systems or working with second-moment equations, which can lead to inaccurate parameter estimates and poor predictions of the system's states. Using series expansions may not fully capture interactions across different timescales. Additionally, these methods do not adequately address the stiffness phenomenon—frequently encountered in biological systems—where reaction rates have very different magnitudes, which leads to numerical instability and further complicates the inference process.

1.2 Outline of the thesis

The thesis is structured as follows:

In the **second chapter**, we provide a thorough description of the quasi-reaction framework, reviewing the theoretical foundations, and including the role of the stochastic analysis in modelling chemical reactions at the particle level. We discuss state-of-the-art inference methodologies, including likelihood-based methods and approximation techniques, alongside their limitations. We present the mathematical results that will be utilized in the subsequent chapters. Our objective is twofold: to provide these findings in advance to facilitate the derivation of later results, while also avoiding overburdening the reader with additional complexities down the line. Furthermore, we introduce the datasets employed in the subsequent applications. We detail the manipulations and preprocessing steps required to prepare these datasets for analysis, explaining how the raw data were structured to align with the computational models. This includes data transformation, and the specific adjustments made to ensure compatibility with the inference methodologies used in the study.

In the **third chapter**, we tackle the challenge of parameter inference for quasi-reaction systems where observations are closely spaced in time. To address the limitations of traditional approaches, we propose a framework that incorporates latent event history models within the quasi-reaction system. Event history models, initially developed for fields such as sociology and medicine, focus on the hidden events driving observable system dynamics, rather than directly modelling the system states themselves. In the context of quasi-reaction systems, such a framework reconstructs the unobserved reactions that take place between consecutive time points. The methodology is based on defining the system as a series of reactions, each governed by a hazard function depending on the system's current state. These reaction rates are assumed to remain constant at sufficiently small time intervals, and the primary challenge is to infer the unobserved reaction events from the observable system state.

To account for the fact that the latent events are not directly observable, we employ an Expectation-Maximization (EM) algorithm for parameter estimation. At the E-step we apply an extended Kalman filter to predict the latent events based on the dynamics observed throughout the time interval. Since the occurred reaction counts follow a Poisson distribution, we first approximate the Poisson-distributed events using a continuous Gamma distribution, which is then transformed into a Gaussian form via marginal transformation. This step enables the use of an extended Kalman filter designed for non-linear systems, facilitating efficient parameter estimation even in the presence of non-linearities in the system dynamics. At the M-step, we optimize the log-likelihood function with respect to both the reaction parameters and the variance-covariance observation matrix, iterating the process until convergence. The model accurately captures the underlying dynamics with minimal bias, particularly in simulated scenarios where the temporal correlation between observations is high due to closely spaced time points.

An illustration of the use of this inferential procedure on epidemiological data, specifically early COVID-19 transmission data, shows how the approach is able to return sensible estimates of the epidemiological parameters, including the basic reproduction number (R0). The approach is applied to three distinct phases of the first year of the COVID-19 epidemic in Italy: the initial acute phase of widespread infection, a summer period with relaxed restrictions, and a third phase during the second lockdown. Our results show that during Phase 2, the spread of infection was significantly limited, primarily as a result of the containment measures implemented during Phase 1. In Phase 3, there was a significant increase in infection rates, especially in the southern regions of Italy, where the disease spread more quickly. Standard errors of R0, calculated via the Delta method, reveal significant differences in the basic reproduction number between consecutive phases. Interestingly, although not explicitly accounted for in the model, the R0 estimates displayed a degree of geographical clustering, which can be attributed to the movement of individuals between neighbouring regions. This clustering aligns with expectations, as human mobility plays an important role in the regional spread of infectious diseases.

In a **fourth chapter**, we extend the latent event history model to include covariates, thereby addressing the limitation of constant reaction rates, which were assumed in the third chapter. To account for time-varying factors, we modify the model to allow the reaction rates to depend on external covariates. The inclusion of covariates is achieved by modelling the log-reaction rates linearly on a vector of covariates. The EM algorithm developed in Chapter 3, is then modified to account for this change, with the Kalman filter in the E-step now taking into account the effect of the covariates on the latent state predictions.

In a simulation study, the enhanced model is applied to a classic SIR epidemic framework, with a binary covariate representing the introduction of a lockdown. The inferential procedure successfully captures the reduction in transmission rates following the implementation of the lockdown, with the estimated parameters closely reflecting the actual changes in the system. The inclusion of the covariate allows the model to reflect the impact of intervention measures, improving both the accuracy of parameter estimates and the overall understanding of the system's behaviour.

When applied to epidemiological data, specifically COVID-19 transmission data from Lombardy, Italy, in 2021, the model provided deeper insights into the effects of covariates such as temperature and government-imposed restrictions on both transmission and recovery rates. The parameter estimates are closely aligned with those reported by the Italian government, demonstrating the practical utility of the proposed approach in epidemic monitoring.

In a **fifth chapter**, we focus on the issue of parameter inference when observations are widely spaced in time, a significant limitation to the existing inferential methods based on stochastic models. In systems with higher-order reactions, mean-field approximation methods often require numerical solutions, as the ordinary differential equations describing the process's mean typically do not have an explicit solution. This also leads to potential numerical instabilities in stiff systems, where reaction rates differ significantly. Systems with at most one reactant per reaction —known as unitary systems— allow for explicit solutions without the need for numerical integration. In this chapter, we present a new approach that generalizes the case of unitary systems to any generic system. In particular, by linearizing the rate function with a first-order Taylor expansion, we approximate any generic system with an ODEs system that has an explicit solution, as for the case of unitary systems. By doing this, we achieve computational efficiency while avoiding the numerical instabilities inherent to traditional numerical methods for solving ODEs, such as explicit Euler and Runge-Kutta. This approach is applied locally for all observations, setting each value in the dataset as an initial condition of the ODEs system and considering the explicit solution of this problem as a projection of the state to the next time point. In this way, we also take into account the non-linearity of the process, leading to high accuracy of parameter estimates.

We further compare our approach with a correlation-based M-estimator, which infers parameters in branching processes by matching the theoretical second moment of cell-type dynamics with the empirical one from the observed data. The proposed local-mean field approximation method outperforms the competitor across multiple metrics, providing more stable and accurate estimates.

When applied to clonal tracking data from Rhesus Macaques, the proposed inferential approach leads to valuable insights into the hematopoietic process. The algorithm effectively captures the differentiation dynamics of key blood cell types, with results that closely align with biological expectations.

Chapter 2

Theoretical and Computational Framework

2.1 Stochastic Chemical Kinetics

In this section, we provide a review of the stochastic chemical kinetics framework. Consider p molecular species P_l in a closed system with volume Ω. We denote by $Y_l = [P_l] \in \mathbb{R}^+$ the concentration of the l-th particle in the system, i.e. the ratio of the number of molecules to the volume. The state of the system is indicated with $\boldsymbol{Y} = ([P_1], \ldots, [P_p])$. The following definition formalizes the notation used for a quasi-reaction system.

Definition 1 (Quasi-reaction equation)**.** A *chemical reaction* is qualitatively described by the following expression

$$k_{1j}P_1 + ... + k_{pj}P_p \xrightarrow{\theta_j} s_{1j}P_1 + ... + s_{pj}P_p, \tag{2.1}$$

where the LHS (*reactants*) correspond to the molecules needed for the j-th reaction, the RHS (*products*) are the results of the chemical transformation. The *stoichiometric coefficients* $k_{lj}, s_{lj} \in \mathbb{Z}_0$ are the number of molecules of the l-th reactant necessary to produce the l-th product in a single reaction step. The concentration change for the occurrence of the j-th reaction is captured by the vector $V_{\cdot j} = k_{\cdot j} - s_{\cdot j}$, which constitutes the j-th column of the net effect matrix $V \in \mathbb{Z}^{p \times r}$. $\theta_j \in \mathbb{R}$ is the j-th *reaction rate*. For most systems, the law of conservation of mass holds, ensuring that the total mass remains constant at all times in a closed system. However, this principle may not apply in cases involving phenomena such as the spontaneous creation of particles, where the concept is generalized using the term ”quasi.”.

Beyond qualitative description of the process, a kinetic mass-action methodology enables a more formal quantitative understanding of the chemical process (Wilkinson, 2018). There are two contrasting views on causal entailment in nature: determinism and randomness. Determinism asserts that natural events follow causal laws that uniquely link the changes in concentrations. That is, the next state of a system is determined without uncertainty by the current and past states of the system. In a large class of quasi-reaction systems, a frequently adopted relation is the *rate law with definite orders* (Mortimer, 2000), in which the concentration of each species decreases in proportion to the number of particles involved in the reaction and increases based on the amount produced, with both changes scaled by the reaction rate. Consequently, the reaction rate is directly proportional to the product of the concentrations of the reactants, each raised to the power of their respective stoichiometric coefficients. A set of ordinary differential equations (ODEs) for the system (2.1) can be written as follows,

$$\frac{d[P_l]}{dt} = \sum_{j=1}^{r} V_{lj}\theta_j [P_1]^{k_{1j}} [P_2]^{k_{2j}} \ldots [P_p]^{k_{pj}} \quad l = 1, \ldots, p. \tag{2.2}$$

Deterministic modeling provides a practical approach by directly translating biochemical reactions into mathematical equations. However, observable phenomena are causally related to numerous interacting factors, many of which may remain unobserved or unidentified. Consequently, different realisations of the same natural process show certain characteristics only to those particular observations. Probabilistic methods might be more suitable to mitigate this loss of information. Furthermore, while the particle dynamics at the macroscale is largely shaped by spatiotemporal changes in the abundance of particle components, at the microscale, particle events are driven by discrete and random interactions among molecules (Ullah and Wolkenhauer, 2011). Given these considerations, the thesis will primarily focus on the stochastic perspective.

Example 1 (Lotka-Volterra). *In the early 20th century, Vito Volterra formulated equations to describe the interaction between two species: a prey and its predator. He aimed to explain population dynamics in the Adriatic Sea. Alfred Lotka, independently, developed a similar model for chemical oscillations (Lotka, 1925). This system became known as the Lotka-Volterra model (Boyce et al., 2017) and is the following*

$$\frac{d}{dt}[X_1] = \theta_1[X_1] - \theta_2[X_1][X_2]$$
$$\frac{d}{dt}[X_2] = \theta_2[X_1][X_2] - \theta_3[X_2]$$

X_1 and X_2 represents respectively the prey and predator populations. The parameter θ_1 denotes the prey's growth rate in the absence of predators, θ_2 the rate of predation, and θ_3 the predator's natural death rate. The corresponding quasi-reaction equations are:

$$\begin{aligned} X_1 &\xrightarrow{\theta_1} 2X_1 \\ X_1 + X_2 &\xrightarrow{\theta_2} 2X_2 \\ X_2 &\xrightarrow{\theta_3} \emptyset \end{aligned} \tag{2.3}$$

Chemical reactions identification. The establishment of the timescale at which the reactions take place is fundamental for providing a comprehensive description of the involved reactions in the process and for achieving a probabilistic understanding of the problem (Golightly and Wilkinson, 2005). Many chemical reactions obtain the final product through several steps involving intermediate reactants that nevertheless do not appear in the reaction coefficients. The detection of such hidden reactants is still an element of interest in numerous studies (Davis and Davis, 2012). When a reaction is described using the same elements present at the molecular level, the time factor corresponds to the *elementary step*. Conversely, if intermediate steps are disregarded, it is referred to as a *stoichiometric reaction*. As elementary steps can be complex and difficult to isolate, while the stoichiometric approach provides a clear balance of reactants and products, we will focus exclusively on the second case.

A second identifiability factor is the reversibility of reactions: a reaction is *reversible* if the conversion of reactants to products and the opposite occurs simultaneously (Davis and Davis, 2012). On the other hand, some reactions occur in both directions but with a predominance of one over the other. We will only consider irreversible reactions in this work to avoid the simultaneity of events and the consequent loss of identifiability. For the same reason, predominance will be achieved by separating the reaction into two but with very different rate magnitudes, thus favouring the occurrence of one over the other.

2.1.1 Mass-action stochastic kinetics

At low concentrations, the deterministic model (2.2) fails to capture the discrete dynamics of molecular interactions, so it is necessary for a more refined approach that accounts for the inherent stochasticity of processes.

There are two sources of randomness in a stochastic chemical system: *extrinsic noise* encompasses environmental factors such as pressure, pH, or temperature, that can cause unpredictability of the model. This often results in the inclusion of a random fluctuation in reaction rates (Wilkinson, 2018). In order to account for these factors in Chapter 4 we consider the effect of covariates on the reaction rates. A second measure of uncertainty is *intrinsic noise* and is based on the discreteness of kinetic systems (Swain et al., 2002). Under reasonable assumptions—namely, that the system is well-stirred, and in thermal equilibrium—statistical mechanics demonstrates that the collision frequency between molecules remains constant, provided that both volume and temperature are fixed. For molecules that are close enough for a reaction to occur, the conditional probability of that reaction occurring becomes independent of the system's volume. This is because, while the chance of molecules being close enough to react depends on the volume, the probability of a reaction firing given their proximity is unaffected by it. As a result, the overall likelihood of a reaction remains unchanged as long as molecules are sufficiently close to interact. A more formal discussion can be found in Gillespie (1992). Throughout this thesis, we will consider the model under the assumption that these conditions apply.

Hazard function We assume that if there are exactly k_{lj} molecules of type l involved, the j-th reaction occurs after a waiting time $T_j \sim \text{Exp}(\theta_j)$, for $j = 1, \dots, r$. Given that collisions occur randomly, it is useful to quantify the instantaneous rate at which a reaction takes place at a given time. This leads to the following definition:

Definition 2 (Hazard function)**.** The hazard function is the probability of the event happening in an infinitesimal interval $[t, t + dt)$, given that it has not occurred until time t, i.e.

$$\lambda_j(t) = \lim_{dt \to 0} \frac{\mathbb{P}(T_j < t + dt \mid T_j > t)}{dt}. \tag{2.4}$$

We will use the above definition with the following result

Proposition 1. *The hazard function of an exponential random variable T_j with parameter θ_j is equal to θ_j.*

Proof. The hazard function in Equation (2.4) can be rewritten as

$$\lambda_j(t) = \lim_{dt \to 0} \frac{\mathbb{P}(T_j < t + dt \mid T_j > t)}{dt} = \lim_{dt \to 0} \frac{\mathbb{P}(t < T_j < t + dt)}{dt \cdot \mathbb{P}(T_j > t)}.$$

Since $\mathbb{P}(T_j > t) = e^{-\theta_j t}$, we have

$$\mathbb{P}(t < T_j < t + dt) = e^{-\theta_j t} - e^{-\theta_j (t+dt)} = e^{-\theta_j t} \left(1 - e^{-\theta_j dt}\right).$$

We substitute these expressions into the definition of the hazard function:

$$\lambda_j(t) = \lim_{dt \to 0} \frac{e^{-\theta_j t} \left(1 - e^{-\theta_j dt}\right)}{dt \cdot e^{-\theta_j t}} = \lim_{dt \to 0} \frac{1 - e^{-\theta_j dt}}{dt}.$$

To evaluate this limit, we use the fact that $\lim_{x \to 0} \frac{1-e^{-x}}{x} = 1$. Setting $x = \theta_j dt$, we have

$$\lim_{dt \to 0} \frac{1 - e^{-\theta_j dt}}{dt} = \lim_{x \to 0} \frac{\theta_j \left(1 - e^{-x}\right)}{x} = \theta_j.$$

□

Substrate concentrations evolve based only on their previous values. Thus, we model it as a continuous-time counting Markov process $\boldsymbol{Y}(t) \in \mathbb{N}_0^p$; the realization $\boldsymbol{y}_t$ is the p-dimensional vector where each component corresponds to the count of molecules of the l-th species at time $t \in [0, T]$. In most of the cases, there may exist a time t such that $y_{lt} > k_{lj}$, meaning there are more molecules of type l in the system than required for the j-th reaction to occur, allowing the reaction to take place in multiple combinatorial ways (Wilkinson, 2018). This leads to the following formulation:

$$T_j \sim \operatorname{Exp}\left(\lambda_j(\boldsymbol{y}_t; \boldsymbol{\theta})\right),$$

where

$$\lambda_j(\boldsymbol{y}_t; \boldsymbol{\theta}) = \theta_j \prod_{l=1}^{p} \binom{y_{lt}}{k_{lj}} \mathbb{1}_{y_{lt} \geq k_{lj}}. \tag{2.5}$$

We assume that the reaction times T_j are independent, so the first time of occurrence follows an exponential distribution and the rate is the sum of the individual hazard functions. While such an assumption of independence is formally ill-defined since the occurrence of one reaction constrains or even prevents others, this remains valid in terms of distribution up to the minimum, after which the clock is reset thanks to the memoryless property of the exponential distribution.

The hazard function depends on the specific reaction mechanism and its order, with different forms of reactions arising based on the number of molecules involved and their interactions. These reactions can be categorized as follows:

1. A **zeroth-order reaction** is described by the quasi-reaction equation

$$R_\emptyset : \quad \emptyset \xrightarrow{\theta_j} P_l$$

While it may seem unlikely for substances to be created from nothing, this definition permits the notion of a constant reaction rate. The hazard function for the j-th reaction is given by $\lambda_j(\boldsymbol{y}_t; \boldsymbol{\theta}) = \theta_j$.

2. A **first-order reaction** is represented as

$$R_{\mathrm{I}}: \quad P_l \xrightarrow{\theta_j} ?$$

The corresponding j-th hazard function is $\lambda_j(\boldsymbol{y}_t, \boldsymbol{\theta}) = \theta_j y_{lt}$. The linear dependence on the concentration vector offers a clear interpretation of how the likelihood of the reaction occurring increases proportionally with the concentration of the reactant.

3. A **higher-order reaction** generalizes the previous case. In this scenario, the number of reactants or their stoichiometric coefficients is strictly greater than one, and the quasi-reaction equation is the following

$$R_{>\mathrm{I}}: \quad \sum_{l=1}^{p} k_{lj} P_l \xrightarrow{\theta_j} ?$$

Higher-order reactions serve as a fundamental tool for examining complex systems with multiple interacting reactants, revealing the interplay and combined effects within multi-component systems. The hazard function is given by

$$\lambda_j(\boldsymbol{y}_t, \boldsymbol{\theta}) = \theta_j \prod_{l=1}^{p} \binom{y_{lt}}{k_{lj}},$$

and is no longer linear with respect to concentrations.

In biochemical terms, these definitions correspond to intuitive classifications: a *duplication reaction* occurs as either a zeroth-order or first-order reaction when the product is identical to the reactant. If the right-hand side (RHS) is empty, the reaction is classified as a *death reaction.* In all other cases, the process is termed a *differentiation reaction.*

2.1.2 The Master Equation

The reaction rates $\boldsymbol{\theta}$ in a quasi-reaction system are typically unknown, and their estimation provides valuable insight into the dynamics of the model, based on an observed dataset. In order to infer these parameters from data, it is necessary to define the underlying probabilistic model. Under the hypothesis of constant volume and thermal equilibrium, every Markov process admits an explicit formulation of the evolution of the probability distribution over time $P(\boldsymbol{y};t)$, given by the Chemical Master Equation (CME)(McQuarrie, 1967; Schnakenberg, 1976; Gillespie, 1992). The CME is the following differential equation system for the process transition probabilities

$$\frac{dP(\boldsymbol{y};t)}{dt} = \sum_{j=1}^{r} \left\{ \lambda_j(\boldsymbol{y} - V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y} - V_{\cdot j};t) - \lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t) \right\}. \tag{2.6}$$

The Chemical Master Equation states that the flow of the probability of being at a state $\boldsymbol{y}$ at time t is equal to the probability of arriving at $\boldsymbol{y}$ due to the occurrence of the j-th reaction, given by $\lambda_j(\boldsymbol{y} - V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y} - V_{\cdot j};t)$, minus the probability of leaving $\boldsymbol{y}$ due to the occurrence of the j-th reaction, given by $\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)$. Considering this over all reactions gives (2.6). A solution of the CME is unfeasible due to the large dimension of possible state configurations and for this reason, various approximation methods have been proposed (Sjöberg et al., 2009; Basile et al., 2013; Gupta et al., 2021).

The first important result derived from the Chemical Master Equation (2.6) is the ability to establish a framework for transitioning from a discrete Markov process to its continuous counterpart. This approach is especially valuable when estimating the parameters that govern the observation process $\boldsymbol{Y}(t)$. One can achieve this by demonstrating the equivalence between the CME and the Fokker-Planck equations. Let introduce the following proposition

Proposition 2. *As $\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)$ are analytic functions in $\boldsymbol{y}$, the Chemical Master Equation can be interpreted as a Kolmogorov Forward Equation, or Fokker-Planck equation, with drift $V\lambda(\boldsymbol{y};\boldsymbol{\theta})$ and diffusion matrix $V\,Diag(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta}))V^T$.*

Proof. Under the regularity assumptions, we can perform a second-order Taylor expansion of the product $\lambda_j(\boldsymbol{y} - V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y} - V_{\cdot j};t)$ around $\boldsymbol{y}$. This gives:

$$\begin{aligned}\lambda_j(\boldsymbol{y}-V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y}-V_{\cdot j};t) = \; & \lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)+\\ & + \nabla_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}((\boldsymbol{y}-V_{\cdot j})-\boldsymbol{y})+\\ & + \frac{1}{2}((\boldsymbol{y}-V_{\cdot j})-\boldsymbol{y})^T\nabla^2_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}((\boldsymbol{y}-V_{\cdot j})-\boldsymbol{y}),\end{aligned}$$

where $\nabla_{\boldsymbol{y}}$ and $\nabla^2_{\boldsymbol{y}}$ represents respectively the Jacobian and Hessian matrices with respect to $\boldsymbol{y}$. Simplifying this, we obtain:

$$\begin{aligned}\lambda_j(\boldsymbol{y}-V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y}-V_{\cdot j};t) = & \lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t) - \nabla_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j}\\ & + \frac{1}{2}V_{\cdot j}^T\nabla^2_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j}.\end{aligned}$$

Thus, we have:

$$\begin{aligned}\lambda_j(\boldsymbol{y}-V_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y}-V_{\cdot j};t) - \lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t) = & - \nabla_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j}\\ & + \frac{1}{2}V_{\cdot j}^T\nabla^2_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j}.\end{aligned}$$

Substituting this expression into the Master equation, we get

$$\frac{dP(\boldsymbol{y},t)}{dt} = \sum_{j=1}^{r}\left\{-\nabla_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j} + \frac{1}{2}V_{\cdot j}^T\nabla^2_{\boldsymbol{y}}\left\{\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}V_{\cdot j}\right\}.$$

This simplifies to:

$$\begin{aligned}\frac{dP(\boldsymbol{y},t)}{dt} = & - \nabla_{\boldsymbol{y}}\left\{V\lambda(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}\\ & + \frac{1}{2}\nabla^2_{\boldsymbol{y}}\left\{V\begin{bmatrix}\lambda_1(\boldsymbol{y};\boldsymbol{\theta}) & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \lambda_r(\boldsymbol{y};\boldsymbol{\theta})\end{bmatrix}V^TP(\boldsymbol{y};t)\right\},\end{aligned}$$

which can be identified as the Kolmogorov forward (Fokker-Planck) equation with drift term $V\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta})$ and diffusion matrix $V\mathrm{Diag}(\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta}))V^T$. □

A Kolmogorov Forward equation for the transition density $P(\boldsymbol{y};t)$ corresponds to a multivariate diffusion process, known as the *Ito process*, with the same drift and diffusion terms. Thus the infinitesimal variation of the process of observations $\boldsymbol{y}_t$ can be rewritten as

$$d\boldsymbol{y}_t = V\boldsymbol{\lambda}(\boldsymbol{y}_t;\boldsymbol{\theta})dt + \left(V\mathrm{Diag}(\boldsymbol{\lambda}(\boldsymbol{y}_t;\boldsymbol{\theta}))V^T\right)^{1/2}d\boldsymbol{W}_t. \tag{2.7}$$

A second result from the Chemical Master Equation is the connection between the continuous deterministic formulation and the expected value of the stochastic kinetic model. In some specific instances, these two are identical. This connection can be demonstrated by deriving the system of differential equations for the expected value of the stochastic kinetic model $\boldsymbol{m}(t) = \mathbb{E}[\boldsymbol{Y}(t)|\boldsymbol{Y}(t_0) = \boldsymbol{y}_0]$. To this end,

$$\frac{dm_l(t)}{dt} = \frac{d}{dt}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l P(\boldsymbol{y};t) = \sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l \frac{dP(\boldsymbol{y};t)}{dt}.$$

Using the Chemical Master Equation (2.6),

$$\frac{dm_l(t)}{dt} = \sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l \sum_{j=1}^{r} \left[\lambda_j(\boldsymbol{y}-\boldsymbol{V}_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y}-\boldsymbol{V}_{\cdot j};t) - \boldsymbol{\lambda}_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right].$$

We swap the summation operators on the RHS as they span all possible state configurations, leading to

$$\frac{dm_l(t)}{dt} = \sum_{j=1}^{r}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N} y_l \left[\lambda_j(\boldsymbol{y}-\boldsymbol{V}_{\cdot j};\boldsymbol{\theta})P(\boldsymbol{y}-\boldsymbol{V}_{\cdot j};t) - \lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right].$$

We make a substitution in the first addend of the RHS term and using the definition of the expectation of a function, and the linearity property,

$$\begin{aligned}\frac{dm_l(t)}{dt} &= \sum_{j=1}^{r}\left\{\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}(y_l+V_{lj})\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)-\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right\}\\ &= \sum_{j=1}^{r}\left\{\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}\left[y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)+V_{lj}\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\right]\right\}\\ &\qquad -\sum_{j=1}^{r}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}y_l\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\\ &= \sum_{j=1}^{r}\sum_{\boldsymbol{y}\in\mathbb{N}_0^N}V_{lj}\lambda_j(\boldsymbol{y};\boldsymbol{\theta})P(\boldsymbol{y};t)\\ &= \sum_{j=1}^{r}V_{lj}\mathbb{E}[\lambda_j(\boldsymbol{Y};\boldsymbol{\theta})].\end{aligned}$$

In general, such ODEs system is not solvable directly, except in *unitary systems*, i.e. in case all reactions have zero or first-order mass action rate law (Wilkinson, 2018). In this case, the hazard function $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta})$ is linear in $\boldsymbol{y}$, thus it is possible to apply the linearity of the expectation, obtaining

$$\frac{dm_l(t)}{dt} = \sum_{j=1}^{r}V_{lj}\lambda_j(\mathbb{E}[\boldsymbol{Y}(t)];\boldsymbol{\theta}) = \sum_{j=1}^{r}V_{lj}\lambda_j(\boldsymbol{m}(t);\boldsymbol{\theta}). \tag{2.8}$$

By considering an initial condition $\boldsymbol{m}(0) = \boldsymbol{m}_0$, the above equation can represented by the following first-order Cauchy problem,

$$\begin{cases}\dfrac{d\boldsymbol{m}(t)}{dt} = P_{\boldsymbol{\theta}}\boldsymbol{m}(t)+\boldsymbol{b}_{\boldsymbol{\theta}},\\ \boldsymbol{m}(0)=\boldsymbol{m}_0.\end{cases} \tag{2.9}$$

The coefficient matrix $P_{\boldsymbol{\theta}}$ and the inhomogeneous term $\boldsymbol{b}_{\boldsymbol{\theta}}$ are both defined in terms of the parameter vector $\boldsymbol{\theta}$. The former corresponds to reactions involving a single reactant, while the latter represents spontaneous reactions, independent of reactants. Utilizing some basic algebraic transformations and defining the *reactant matrix* $K = \{k_{lj}\}$, $P_{\boldsymbol{\theta}}$ and $\boldsymbol{b}_{\boldsymbol{\theta}}$ can be explicitly written as

$P_{\boldsymbol{\theta}} = V\text{diag}(\boldsymbol{\theta})K^T$ and $b_{\boldsymbol{\theta},l} = \sum_j V_{lj}\theta_j \mathbb{1}_{\{K_{\cdot j}=0\}}$. If the matrix $P_{\boldsymbol{\theta}}$ is invertible, then the system (2.9) admits the explicit solution

$$\boldsymbol{m}(t) = \exp(tP_{\boldsymbol{\theta}})\boldsymbol{m}_0 + P_{\boldsymbol{\theta}}^{-1}\left(\exp(tP_{\boldsymbol{\theta}}) - \mathbb{I}_p\right)\boldsymbol{b}_{\boldsymbol{\theta}}. \tag{2.10}$$

In Chapter 5, we will present a new approach that transform any generic quasi-reaction system as unitary, by linearizing with a first-order Talyor expansion the rate function $\boldsymbol{\lambda}(\boldsymbol{y};\boldsymbol{\theta})$ for $\boldsymbol{y}$. Through this approximation, we can always derive an explicit expression for the conditional mean of the process, as the corresponding ODEs system (2.9) admits a closed-form solution. In the following example, we illustrate how an exact solution of the conditional mean of the concentrations process can be defined if the system is unitary.

Example 2 (Reduced Lotka-Volterra)**.** *Consider the previously defined system (2.3), but without interaction between species, to deal only with at most first-order reactions. Such a scheme correspond to the following quasi-reaction equations:*

$$X_1 \xrightarrow{\theta_1} 2X_1$$
$$X_2 \xrightarrow{\theta_3} \emptyset.$$

Note that

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

and the hazard function vector is

$$\boldsymbol{\lambda}(\boldsymbol{Y};\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 Y_1 \\ \theta_3 Y_2 \end{bmatrix}.$$

Thus, we get the following differential equation,

$$\begin{aligned}\frac{d\boldsymbol{m}(t)}{dt} &= V\boldsymbol{\lambda}(\mathbb{E}[\boldsymbol{Y}(t)];\boldsymbol{\theta}) \\ &= V\boldsymbol{\lambda}(\boldsymbol{m}(t);\boldsymbol{\theta}) \\ &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 m_1(t) \\ \theta_3 m_2(t) \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \theta_1 & 0 \\ 0 & -\theta_3 \end{bmatrix}}_{P_{\boldsymbol{\theta}}} \begin{bmatrix} m_1(t) \\ m_2(t) \end{bmatrix}.\end{aligned}$$

The above equation is a first-order ODEs system without the inhomogeneous term $\boldsymbol{b}_{\boldsymbol{\theta}}$. Following (2.10) *and given the initial condition $\boldsymbol{m}_0$, the explicit solution is*

$$\boldsymbol{m}(t) = \exp\left(t\begin{bmatrix} \theta_1 & 0 \\ 0 & -\theta_3 \end{bmatrix}\right)\cdot \boldsymbol{m}_0.$$

2.1.3 The Gillespie Algorithm

The chemical master equation has an exact correspondence with the stochastic simulation algorithm proposed by Gillespie (1977). The Gillespie algorithm provides an exact procedure to simulate such stochastic systems by determining the time of the next reaction and which reaction will occur. This means that the stochastic simulation algorithm produces realisations of the Markov jump process $\boldsymbol{Y}(t)$ whose initial conditional densities are determined by the CME (Gillespie, 1992). Each reaction $j = 1, \dots, r$ occurs at a rate determined by its hazard function $\lambda_j(\boldsymbol{y};\boldsymbol{\theta})$, where θ_j represents the reaction rate constant and $\boldsymbol{y}$ the current state of the system. The algorithm is outlined as follows:

1. Set the initial time $t = 0$, with initial concentrations $\boldsymbol{y}_0$ and rate $\boldsymbol{\theta}$.

2. For each reaction $j = 1, \dots, r$, calculate $\lambda_j(\boldsymbol{y};\boldsymbol{\theta})$ based on the current state $\boldsymbol{y}$.

3. Sample the time to the next reaction τ,

$$\tau \sim \text{Exp}\left(\sum_{j=1}^{r} \lambda_j(\boldsymbol{y}; \boldsymbol{\theta}) \right).$$

4. Select the next reaction j with the following probability

$$\mathbb{P}(\text{"Reaction j occurring"} | \boldsymbol{y}, \boldsymbol{\theta}) = \frac{\lambda_j(\boldsymbol{y}; \boldsymbol{\theta})}{\sum_{i=1}^{r} \lambda_i(\boldsymbol{y}; \boldsymbol{\theta})}.$$

5. Update the state of the system by applying the net effect vector of the selected reaction j:

$$\boldsymbol{y} \leftarrow \boldsymbol{y} + V_{\cdot j}$$

6. Set $t = t + \tau$.

7. Repeat the process from step 2 until a desired stopping time $T_{\max}$ is reached (using the memoryless property of the exponential distribution).

The Gillespie algorithm could be used in other frameworks based on Markov jump processes, such as genetic regulatory systems or queueing networks (Teugels, 2008), where transitions occur without explicitly modeling interactions between components, extending its applicability beyond reaction-based dynamics. However, incorporating reactions explicitly, as in the quasi-reaction systems framework, provides greater versatility for studying the parameters governing the system dynamics. Chapters 3 and 4 investigate the performance of the proposed inference method for several sampling time intervals. It is common practice to set Δt and evaluate the estimates using the inference algorithm. However, as our simulations are performed with the standard Gillespie algorithm, observations are not equidistant: for high concentrations, a reaction is more likely to occur and the intervals between observations are shorter. The data selection strategy used for the simulation studies in Chapter 3 and 4 starts with generating T observations using the Gillespie algorithm, given the initial values $\boldsymbol{y}_0$, and the true parameters $\boldsymbol{\theta}_{true}$. Then, for a fixed k, measurements are retained every k time points from the

Jump	T=10 Occurred Reactions										Selected data	N Intervals
2	●	○	●	○	●	○	●	○	●	○	5	4
3	●	○	○	●	○	○	●	○	○	●	4	3
5	●	○	○	○	○	●	○	○	○	○	2	1

Example of data selection. $T = 10$ time points are generated; the total number of selected observations $(N + 1)$ depends on the *jump* value. The analysis is performed on the dataset indicated by filled dots; empty dots represent latent reactions.

previously sampled trajectories. The value of k is referred to as *jumps*. The greater the jump, the wider the interval between consecutive time points. For ease of notation, we denote by N the number of intervals between the selected points. Table 2.1 provides an example with $T = 10$ occurred events.

In Chapter 5, with a slight abuse of notation, we indicate the number of selected observations (i.e. the size of the dataset) directly with T.

2.2 Inference

The observation process $\boldsymbol{Y}(t)$ is collected at discrete time points. In Chapter 5, multiple replicates of each trajectory are considered, whereas in Chapters 3 and 4, data from a single subject are analyzed. In this section, we will present different inference methods for parameter estimation of the reaction rates θ in stochastic quasi-reaction systems.

In Section 2.2.1 we describe the state-of-the-art Local Linear Approximation (LLA), which discretizes the moments of the continuous Ito process and then estimates parameters using a least-squares approach. In Section 2.2.2, we present an alternative inference approach proposed by Golightly and Wilkinson (2006), which utilizes a Bayesian framework combined with a diffusion approximation. This method relies on the same moment approximations of the Ito process (2.7), but introduces latent unobserved data points between observed time intervals. By augmenting the dataset with these latent observations, the authors were able to enhance the accuracy of parameter estimation, filling gaps between sparse data points. This approach has similarities to the latent variable method discussed in Chapter 3, as both frameworks make use of unobserved states to capture the underlying dynamics of the system. In Section 2.2.3, we illustrate the Xu et al. (2019)'s correlation-based method that minimizes the differences between theoretical, computed from the CME, and empirical, obtained from the observed data, second-order moments. This method serves as a competitor to the approach we present in Chapter 5.

2.2.1 Local Linear Approximation

For each time interval, the Ito process (2.7) can be simplified by applying an Euler-Maruyama approximation of the continuous increments process $d\boldsymbol{y}_t$, and then by conducting a second approximation, that of considering finite time intervals. This leads to the definition of a new process of the variation of concentrations with conditional mean and variance as follows:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}_{i+1} - \boldsymbol{y}_i | \boldsymbol{y}_i] &= \underbrace{V \begin{bmatrix} \lambda_{i1}(\boldsymbol{y}_i;\boldsymbol{\theta}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{ir}(\boldsymbol{y}_i;\boldsymbol{\theta}) \end{bmatrix} (t_{i+1} - t_i)}_{M_i} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_r \end{bmatrix}}_{\boldsymbol{\theta}} \\
\mathbb{V}[\boldsymbol{y}_{i+1} - \boldsymbol{y}_i | \boldsymbol{y}_i] &= \underbrace{V \begin{bmatrix} \lambda_{i1}(\boldsymbol{y}_i;\boldsymbol{\theta}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{ir}(\boldsymbol{y}_i;\boldsymbol{\theta}) \end{bmatrix} V^T (t_{i+1} - t_i)}_{W_i(\boldsymbol{\theta})}
\end{aligned} \tag{2.11}$$

We can write the following regression equation

$$\Delta \boldsymbol{y}_i = \boldsymbol{M}_i \boldsymbol{\theta} + \boldsymbol{\epsilon}_i, \tag{2.12}$$

where $\boldsymbol{\epsilon}_t$ is a p-dimensional random vector with zero mean and variance-covariance matrix $W_i(\boldsymbol{\theta})$. Each block W_i is a $(p \times p)$ matrix that describes the covariance structure related to variations in particle counts at the i-th observation time. A constraint least squares estimator $\hat{\boldsymbol{\theta}}_{LLA}$ is the following:

$$\hat{\boldsymbol{\theta}}_{LLA} = \arg\min_{\theta} (\Delta \boldsymbol{y} - M\boldsymbol{\theta})^T W^{-1} (\Delta \boldsymbol{y} - M\boldsymbol{\theta}) \quad \text{such that } \boldsymbol{\theta} \geq \boldsymbol{0}_r.$$

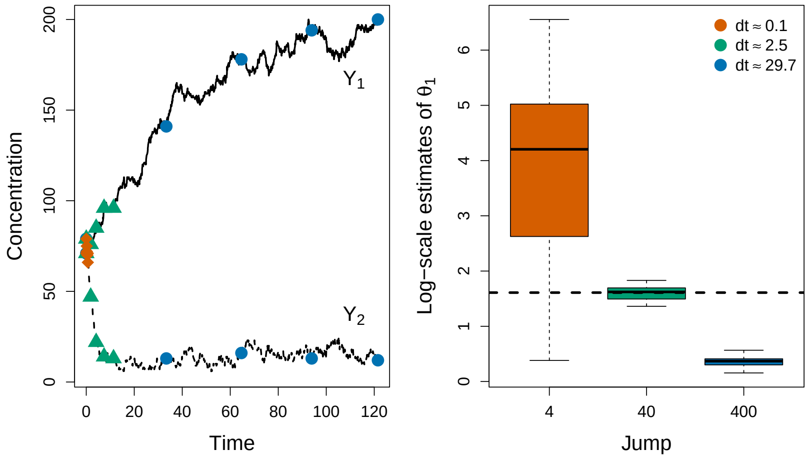


FIG. 2.1 ***Performance of LLA Inference on the Lotka-Volterra model Across Time Intervals.*** *The left panel illustrates the concentration dynamics Y_1 and Y_2, sampled at various time intervals which are represented in different colours. The right panel displays the log-scale estimates of θ_1 corresponding to the jumps values. For very close time intervals (orange triangular points), LLA estimates are biased and lack precision due to numerical issues (boxplot on the left). Conversely, for time intervals that are too distant (blue circular points), the LLA poorly approximates the moments, as its accuracy is heavily dependent on the choice of dt, leading to bad performance (boxplot on the right).*

This method is referred to as *Local Linear Approximation* (LLA) because, after estimating $\hat{\boldsymbol{\theta}}_{LLA}$, conditional on the state of the system at time t_i, it performs a linear forward prediction to estimate the state at the next time point. The resulting trajectory consists of multiple locally linear segments, which collectively approximate the system's overall non-linear behavior. The pseudo-code is presented in the Algorithm 1. The LLA method has two main functions in the course of the thesis: firstly, it provides the values from which we start the optimisation with our proposed methods. Secondly, it is the benchmark against which we compare the proposed methods.

Method limitations and proposed solutions. A necessary condition for a correct modelling of the continuous Ito process is the complete identification of the occurred events, which is given by assuming small time intervals. For timescales that are too wide, such approximation reduces the

Algorithm 1 Local Linear Approximation Method

```
Data: Observations vector $\Delta\boldsymbol{y}$ and predictor matrix $M$
Result: Parameter estimates $\boldsymbol{\theta}_{LLA}$
Initialization: tol $= \epsilon$, $k = 0$
$\boldsymbol{\theta}^{(0)} = \arg\min_{\boldsymbol{\theta}}(\Delta\boldsymbol{y} - M\boldsymbol{\theta})^T(\Delta\boldsymbol{y} - M\boldsymbol{\theta}) \quad \text{s.t. } \boldsymbol{\theta} \geq \mathbf{0}_r$
while $\sum_{j=1}^{r} |\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k-1)}| \geq \epsilon$ do
    $\boldsymbol{W}^{(k)} = \begin{bmatrix} \boldsymbol{W}_0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{W}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{W}_{N-1} \end{bmatrix}$
    where $\boldsymbol{W}^{(k)} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda}(\boldsymbol{y}; \boldsymbol{\theta}^{(k)}))\boldsymbol{V}^T\Delta t$
    $\boldsymbol{\theta}^{(k+1)} = \arg\min_{\boldsymbol{\theta}}(\Delta\boldsymbol{y} - M\boldsymbol{\theta})^T\boldsymbol{W}^{(k)}(\Delta\boldsymbol{y} - M\boldsymbol{\theta}) \quad \text{s.t. } \boldsymbol{\theta} \geq \mathbf{0}_r$
    $k = k + 1$
end while
$\boldsymbol{\theta}_{LLA} = \boldsymbol{\theta}^{(k)}$
```

accuracy of LLA methods, and we propose the algorithm in Chapter 5 as a possible solution. However, from the definition (2.11), we deduce that as the time interval decreases, the change between consecutive concentrations also becomes smaller. Such non-variability of the particle counts may lead to numerical issues, as the covariance matrix might be extremely sparse. The method proposed in Chapter 3 aims to solve this issue by focusing on the events that generate the observations rather than on the observations themselves. All these scenarios are illustrated in Figure 2.1, where we considered the Lotka Model (2.3) and simulate data using the selection strategy described before to create three datasets, each containing five observations but with different time intervals (*jumps*). For large time intervals (blue circular points), the estimates from LLA are highly biased. For small time intervals (orange trapezoidal points), the estimates remain biased and are less precise. Only for specific time intervals (green triangular points), the LLA algorithm demonstrates good performance, as indicated by the boxplot in the center. The method proposed in Chapter 3 will perform better for the case of small time intervals, while the method proposed in Chapter 5 will do better for the case of large time intervals.

2.2.2 The Bayesian inference approach of Golightly and Wilkinson (2006)

Although not the focus of this thesis, reaction rates can be estimated also via a Bayesian approach. In particular, in this section we present the method of Golightly and Wilkinson (2006), who develop a Bayesian framework utilizing a diffusion approximation of the continuous Ito process. They started defining a d-dimensional dataset $\boldsymbol{Y}(t) = (\boldsymbol{X}(t), \boldsymbol{Z}(t))^T$, where $\boldsymbol{X}(t)$ represents the observed data and $\boldsymbol{Z}(t)$ refers to the unobserved or missing observations. Given T measurements at evenly spaced times $t_0, \ldots, t_T$, the observations time interval $[t_0, t_T]$ is divided into $mT + 1$ elements evenly spaced points $t_0 = \tau_0 = \tau_1 < \tau_2 < \cdots < \tau_n = t_T$ for some positive integer m. The resulting $d(nm + 1)$ missing values are filled with $\boldsymbol{y}_i$ simulated values, leading to the augmented dataset $\hat{\boldsymbol{y}}$. The joint posterior distribution for the parameters $\boldsymbol{\theta}$ and the latent process $\boldsymbol{z}$, given the observed data $\boldsymbol{x}$, is:

$$\pi(\hat{\boldsymbol{y}}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{z}_0) \prod_{i=1}^{T} f(\boldsymbol{y}_i | \boldsymbol{y}_{i-1}, \boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{z}_0)$ are the priors on the parameters and the initial latent state respectively, and

$$f(\boldsymbol{y}_i | \boldsymbol{y}_{i-1}, \boldsymbol{\theta}) = |W_{i-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\Delta \boldsymbol{y}_i - M_{i-1}\boldsymbol{\theta})^T W_{i-1}^{-1} (\Delta \boldsymbol{y}_i - M_{i-1}\boldsymbol{\theta})\right)$$

$$\times (\Delta \boldsymbol{y}_i - M_{i-1}\boldsymbol{\theta}).$$

For example, a uniform prior can be chosen to reflect initial non-informativeness about the parameters. Note how the formulation of the equation above is a consequence of an Euler-Maruyama approximation as already seen for the LLA approach. To obtain the distribution of the reaction parameters $\boldsymbol{\theta}$ then Golightly and Wilkinson (2006) employed a *data augmentation procedure*, alternating between simulating the parameters conditioned on the augmented data (including the missing values), and simulating the missing data given the observed data and the current model parameters.

Method limitations and proposed solutions. One significant concern is the computational burden associated with MCMC sampling, particularly as the dimensionality of the parameter space increases with the number of latent variables introduced. This can lead to issues with convergence and mixing, especially in high-dimensional settings. Another limitation of this approach is its reliance on the diffusion approximation, which, while often satisfactory for inference, may not accurately capture the underlying dynamics of the system being modelled. The assumption that the stochastic process can be adequately represented by a continuous approximation may overlook important discrete behaviours inherent in the data, particularly in biological systems where reactions may involve small populations of molecules and significant stochastic fluctuations. Moreover, the efficacy of the MCMC sampling method is highly dependent on the choice of priors for the parameters. Inappropriate priors can lead to biased estimates or poor convergence properties. In Chapter 3 we present a different approach that, rather than estimating the observations between data points, predicts the events that occurred in such time spans. The method uses a state-space model with a Kalman Filter algorithm to efficiently reconstruct the latent occurred reactions.

2.2.3 The correlation-based M-estimator by Xu et al. (2019)

The second problem of the local linear approximation occurs when the measurements are far apart in time. Besides LLA, in Chapter 5, we will compare our approach also with the one of Xu et al. (2019). The methodology was rooted in a correlation-based M-estimator designed to align empirical correlations derived from observed data with those predicted by the model. In particular, the method involves the minimization of a loss function

$$f(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{t_i}\sum_{m}\sum_{n'\neq m}\left[\psi_{mn,i}(\boldsymbol{\theta};\boldsymbol{y}) - \hat{\psi}_{mn,i}(\boldsymbol{y})\right]^2,$$

where $\psi_{mn,i}(\boldsymbol{\theta};\boldsymbol{y})$ denotes the model-based correlation between the counts of particle types m and n at time t_i, and $\hat{\psi}_{mn,j}(\boldsymbol{y})$ signifies the empirical correlation computed from the observed read counts. The model-based correlation is given by

$$\psi_{mn,j}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{\mathrm{Cov}[\boldsymbol{y}_m(t_j),\boldsymbol{y}_n(t_j);\boldsymbol{\theta}]}{\sigma(\boldsymbol{y}_m(t_j);\boldsymbol{\theta})\sigma(\boldsymbol{y}_n(t_j);\boldsymbol{\theta})},$$

where the numerator indicates the covariance between the counts of the specified particle types, and σ represents the standard deviation of these counts, parameterized by the model. The optimization problem that Xu et al. (2019) addressed can be shortly expressed as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta};\boldsymbol{Y}),$$

subject to the constraints imposed by the biological context, such as the non-negativity of the parameters. This optimization problem is solved using numerical methods, given the potential non-linearity and complexity of the loss function.

Method limitations and Proposed Solutions. One significant constraint of Xu et al. (2019)'s method is the reliance on the assumption that the correlation structure adequately represents the underlying biological processes. If the relationships among the particle types change over time or under different environmental conditions, the model may fail to capture these dynamics accurately. Furthermore, while the method is designed to work with sparse data, extreme sparsity could still lead to unreliable estimates, particularly if the correlation between cell types is weak. A notable limitation of the work is that it is restricted to first-order reactions, significantly constraining the analysis and applicability to systems that exhibit higher-order dynamics.

2.3 Mathematical Tools and Methods

2.3.1 Exponential of matrices

The following result will be used in Chapter 5.
Given a squared matrix $P \in \mathbb{R}^{n\times n}$, its exponential is formally given by the power series

$$e^P = \sum_{k=0}^{\infty} \frac{P^k}{k!} = \mathbb{I}_p + P + \frac{1}{2}(P \cdot P) + \dots \tag{2.13}$$

The series converges for any matrix P, as P is a linear operator between Banach spaces, and the exponential series has a radius of convergence ∞. Although Eq. (2.13) is mathematically well-defined, it does not provide an explicit formula for practical computation. Thus, numerical schemes are typically employed. In particular, we will use the following approach (Al-Mohy and Higham, 2010):

1. *Scaling.* The matrix P is scaled by a factor of 2^{-s}, where s is chosen such that $\|P\|$ becomes sufficiently small to facilitate efficient computation of the matrix exponential. The scaled matrix is defined as:

$$P' = \frac{1}{2^s} P.$$

2. *Padé Approximation.* The scaled matrix exponential $e^{P'}$ is approximated using a rational function called the Padé approximant. For a matrix A, the Padé approximant of order m is given by:

$$e^A \approx \frac{\left(\mathbb{I}_n + \frac{A}{2m+1} + \frac{A^2}{(2m+1)(2m)} + \cdots\right)}{\left(\mathbb{I}_n - \frac{A}{2m+1} + \frac{A^2}{(2m+1)(2m)} - \cdots\right)},$$

where $\mathbb{I}_n$ denotes the identity matrix of dimension n and the number of terms depends on the chosen degree m of the approximation.

3. *Squaring.* After computing the matrix exponential of the scaled matrix P', the original scaling is recovered by successively squaring the result s times, as follows:

$$e^P = \left(e^{P'}\right)^{2^s}.$$

2.3.2 Analytical solutions to first-order ODE systems

The following result will be used in Chapter 5.

Consider the following non-homogeneous first-order ODE system

$$\frac{d}{dt}\mathbf{y}(t) = P \cdot \mathbf{y}(t) + \mathbf{b},$$

where $\mathbf{y}, \in \mathbb{R}^n$, $P \in \mathbb{R}^{n \times n}$, and $\mathbf{b} \in \mathbb{R}^n$. The system is solved by the method of variation of parameters. First, consider the homogeneous case:

$$\frac{d}{dt}\mathbf{y}(t) = P \cdot \mathbf{y}(t).$$

Using the notation of exponential of a matrix defined in (2.13), the solution to the homogeneous equation above is given by

$$\mathbf{y}_{\text{hom}}(t) = \exp(P(t - t_0)) \cdot \mathbf{y}(t_0),$$

where $\mathbf{y}(t_0)$ represents the initial condition. For the non-homogeneous equation, the general solution for the system is expressed as:

$$\mathbf{y}(t) = \exp(P(t - t_0)) \cdot \mathbf{y}(t_0) + \exp(P(t - t_0)) \cdot \int_{t_0}^{t} \exp(P(t - \tau)) \cdot \mathbf{b}\, d\tau.$$

If P is invertible, this expression simplifies to:

$$\mathbf{y}(t) = \exp(P(t - t_0)) \cdot \mathbf{y}(t_0) + P^{-1}(\exp(P(t - t_0)) - \mathbb{I}_n)\mathbf{b}.$$

.

2.3.3 Stiff problems

Stiffness is a common issue in solving ODEs systems when the solution contains components that evolve on very different time scales. This results in certain numerical methods, particularly explicit methods, being inefficient or unstable unless extremely small time steps are used. Since the notion of stiffness will be taken up in Chapter 5, as we will demonstrate that our proposed method is robust to this phenomenon, we here provide a semiformal definition, following Spijker (1996).

Definition 3. (Stiffness) Consider $\boldsymbol{U}(t) \in \mathbb{R}^n$, and a function $\boldsymbol{f} : [0, T] \times D \subset \mathbb{R}^n \to \mathbb{R}^n$. The following ODE system

$$\boldsymbol{U}'(t) = \boldsymbol{f}(t, \boldsymbol{U}(t)), \quad t \in [0, T], \quad \boldsymbol{U}(0) = \boldsymbol{U}_0,$$

is considered *stiff* if the largest step size h^* that ensures numerical stability for explicit methods is much smaller than the step size h_{acc} required to achieve the desired accuracy. In other words, stiffness occurs when

$$h^* \ll h_{\text{acc}}.$$

Numerical methods aim to approximate the solution $\boldsymbol{U}(t)$ at discrete time points t_k using a step size $h_k = t_k - t_{k-1}$. In the case of stiff problems, explicit methods such as

$$\boldsymbol{U}_{k+1} = \boldsymbol{U}_k + h_k \boldsymbol{f}(t_k, \boldsymbol{U}_k),$$

tend to become unstable unless the step size h_k is chosen to be extremely small. This instability arises because explicit methods impose stability constraints that are closely tied to the eigenvalues of the Jacobian matrix $J(t, \boldsymbol{U}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{U}}$, which governs the local behaviour of the system. Specifically, to quantify the degree of stiffness, one can introduce the concept of the *condition number* of the Jacobian matrix, defined as follows:

$$\kappa(J) = \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}.$$

In the definition above, $\sigma_{\text{max}}(J)$ and $\sigma_{\text{min}}(J)$ are the largest and smallest singular values of the Jacobian matrix J. A system is likely to be stiff if the condition number satisfies

$$\kappa(J) \gg 1.$$

The condition number provides a measure of how sensitive the system is to perturbations, with larger values indicating a greater degree of stiffness.

One of the most well-known examples demonstrating stiffness in quasi-reaction models is the Robertson (1966)'s problem. The model describes a set of three reactions involving the species A, B, and C, with the following

ODEs and quasi-reaction equations scheme

$$\begin{aligned}\frac{dY_1}{dt} &= -\theta_1 Y_1 + \theta_2 Y_2 \cdot Y_3\\ \frac{dY_2}{dt} &= \theta_1 Y_1 - \theta_2 Y_2 \cdot Y_3 - \theta_3 Y_2^2\\ \frac{dY_3}{dt} &= \theta_3 Y_2^2\end{aligned}$$

where $\boldsymbol{\theta} = (4 \cdot 10^{-2}, 3 \cdot 10^7, 1 \cdot 10^4)$, and initial value $\boldsymbol{y}_0 = (1, 0, 0)$.

In this problem, the reaction rates θ_1, θ_2, and θ_3 clearly exhibit significant differences in magnitude. Specifically, θ_1 is much smaller than θ_2 and θ_3, indicating that the first reaction occurs at a much slower rate compared to the other two, which are considerably faster. Given that the condition number calculated for a generic $\boldsymbol{y}_t = (1, 1, 1)$ is $k(J) = 2.23 \cdot 10^{19}$, an extremely small time step h is required when applying explicit numerical methods to solve the system in order to ensure stability for the fast reactions, even though the first reaction evolves on a much longer time scale and does not require such a fine resolution.

2.3.4 Optimization methods

Parametric inference techniques are fundamentally optimization problems, where the objective is to maximize a likelihood function or a similar criterion with respect to the parameters. In the thesis, we will deal with different types of optimization problems, including both constrained and unconstrained cases.

In Chapters 3 and 4, the reaction rates are modeled as follows

$$\boldsymbol{\theta} = \exp(X\boldsymbol{\beta}),$$

where $X \in \mathbb{R}^{N \times (q+1)r}$ represents the matrix of covariates (which is simply the identity matrix in the first Chapter), and $\boldsymbol{\beta} \in \mathbb{R}^{(q+1)r \times 1}$ is the vector of parameters to be estimated. The objective is the maximization of the likelihood function with respect to the parameters $\boldsymbol{\beta}$, which are indirectly related to the rates $\boldsymbol{\theta}$ through the exponential function, which means that the reaction rates

$\boldsymbol{\theta}$ are guaranteed to be positive. To solve this unconstrained optimization problem, we employ the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Shanno, 1970), a quasi-Newton method that iteratively improves an estimate of the inverse Hessian matrix without explicitly calculating it. Here we briefly present an outline of the algorithm.

Given a continuously differentiable objective function $f(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \mathbb{R}^r$, the BFGS algorithm seeks to find its minimum by iteratively updating the estimate of $\boldsymbol{\theta}$ using a gradient descent approach. At each iteration k, the parameter vector $\boldsymbol{\theta}_k$ is updated based on the inverse Hessian approximation B_k and the gradient of f evaluated at $\boldsymbol{\theta}_k$. The update rule for $\boldsymbol{\theta}$ is:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - B_k \nabla f(\boldsymbol{\theta}_k),$$

where $\nabla f(\boldsymbol{\theta}_k)$ represents the gradient of the objective function at iteration k. The updates B_k are defined with the following iterative rule:

$$B_{k+1} = B_k + \frac{\boldsymbol{J}_k \boldsymbol{J}_k^T}{\boldsymbol{J}_k^T \Delta\boldsymbol{\theta}_k} - \frac{B_k \Delta\boldsymbol{\theta}_k \Delta\boldsymbol{\theta}_k^T B_k}{\Delta\boldsymbol{\theta}_k^T B_k \Delta\boldsymbol{\theta}_k}, \tag{2.14}$$

where

$$\boldsymbol{J}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$$

is the change in the gradient between iterations k and $k+1$, and

$$\Delta\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$$

is the change in the parameter vector. The two additive terms in the update formula (2.14) have different goals: the first term increases the curvature in the direction of the gradient change $\boldsymbol{J}_k$, while the second term corrects the curvature in the direction of the parameter update $\Delta\boldsymbol{\theta}_k$.

In Chapter 5, the reaction rates are constant parameters rather than in exponential form. In this case, both our method and the Local Linear Approximation, which we use both for comparison and to obtain initial estimates for the optimization, involve a constrained minimization. As the reaction rates must remain positive, we impose box constraints to ensure this property during the optimization process. To this end, we employ the

L-BFGS-B algorithm (Byrd et al., 1995), which adapts the standard BFGS method to constrained problems. The parameter update is modified to include a projection step

$$\boldsymbol{\theta}_{k+1} = \mathrm{Proj}(\boldsymbol{\theta}_k - B_k \nabla f(\boldsymbol{\theta}_k)),$$

where $\mathrm{Proj}(\cdot)$ projects each element of $\boldsymbol{\theta}_k$ onto the feasible set defined by $\boldsymbol{\theta}^{MIN}$ and $\boldsymbol{\theta}^{MAX}$,

$$\mathrm{Proj}(\boldsymbol{\theta}_k) = \min\left(\boldsymbol{\theta}^{MAX}, \max\left(\boldsymbol{\theta}^{MIN}, \boldsymbol{\theta}_k\right)\right).$$

The L-BFGS-B algorithm effectively saves memory by storing only the last m updates to $\Delta\boldsymbol{\theta}_k$ and $\boldsymbol{J}_k$,and using only these values to update the inverse Hessian approximation. This allows the handling of large-scale problems, minimizing computational and memory usage. The update of B_k follows the rule in (2.14), with the gradient step constrained by the projection operator.

2.3.5 Useful probability results

Lemma 1 is used in the Kalman Filter update step, in Chapter 3. In the same section, Lemma 2 is applied to derive the expected log-likelihood of the complete data for the Expectation-Maximization algorithm.

Lemma 1. *Let $\boldsymbol{Z}$ and $\boldsymbol{X}$ be random vectors of size n and m, respectively. Suppose $\boldsymbol{Z}$ and $\boldsymbol{X}$ are jointly distributed as:*

$$\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{X} \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu_Z} \\ \boldsymbol{\mu_X} \end{pmatrix}, \begin{pmatrix} \Sigma_{\boldsymbol{Z}} & \Sigma_{\boldsymbol{ZX}} \\ \Sigma_{\boldsymbol{ZX}}^T & \Sigma_{\boldsymbol{X}} \end{pmatrix}\right).$$

Then, the marginal distribution of $\boldsymbol{Z}$ and the conditional distribution of $\boldsymbol{Z} \mid (\boldsymbol{X} = \boldsymbol{x})$ are:

$$\begin{aligned} \boldsymbol{Z} &\sim \mathcal{N}_n(\boldsymbol{\mu_Z}, \Sigma_{\boldsymbol{Z}}) \\ \boldsymbol{Z} \mid (\boldsymbol{X} = \boldsymbol{x}) &\sim \mathcal{N}_n\left(\boldsymbol{\mu_Z} + \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}(x - \boldsymbol{\mu_X}), \Sigma_{\boldsymbol{Z}} - \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{ZX}}^T\right) \end{aligned}$$

Proof. Define $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{Z}} \\ \boldsymbol{\mu}_{\boldsymbol{X}} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{\boldsymbol{Z}} & \Sigma_{\boldsymbol{ZX}} \\ \Sigma_{\boldsymbol{ZX}}^T & \Sigma_{\boldsymbol{X}} \end{pmatrix}$. We first compute the marginal distribution of $\boldsymbol{Z}$, then the conditional distribution $\boldsymbol{Z} \mid (\boldsymbol{X} = \boldsymbol{x})$.

Consider the following affine transformation,

$$\boldsymbol{Z} = A \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{X} \end{pmatrix},$$

where $A = \begin{pmatrix} \mathbb{I}_n & 0_{m\times m} \end{pmatrix}$, and $0_{m\times m}$ is a m-by-m matrix of zeros. Given that the Gaussian distribution is invariant under linear transformations, $\boldsymbol{Z}$ also follows a Gaussian distribution, with the following mean and variance:

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}] &= \mathbb{E}\left(A \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{X} \end{pmatrix} \right) = A\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{Z}} \\ \operatorname{Var}(\boldsymbol{Z}) &= A\Sigma A^T = \Sigma_{\boldsymbol{Z}} \end{aligned}$$

Thus, we find that $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{\mu}_{\boldsymbol{Z}}, \Sigma_{\boldsymbol{Z}})$. Using the transformation $A = \begin{pmatrix} 0_{n\times n} & \mathbb{I}_m \end{pmatrix}$, we derive that $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}_{\boldsymbol{X}}, \Sigma_{\boldsymbol{X}})$. For the conditional distribution $\boldsymbol{Z} \mid (\boldsymbol{X} = \boldsymbol{x})$, we use the following block matrix factorization of Σ:

$$\Sigma = \begin{pmatrix} \mathbb{I}_n & \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1} \\ 0_{m\times n} & \mathbb{I}_m \end{pmatrix} \begin{pmatrix} \Sigma_{\boldsymbol{Z}} - \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{ZX}}^T & 0_{n\times m} \\ 0_{m\times n} & \Sigma_{\boldsymbol{X}} \end{pmatrix} \begin{pmatrix} \mathbb{I}_n & 0_{m\times n} \\ \Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{ZX}}^T & \mathbb{I}_m \end{pmatrix}.$$

The inverse of Σ is:

$$\Sigma^{-1} = \begin{pmatrix} \mathbb{I}_n & 0_{n\times m} \\ -\Sigma_{\boldsymbol{Z}}^{-1}\Sigma_{\boldsymbol{ZX}}^T & \mathbb{I}_m \end{pmatrix} \begin{pmatrix} \Sigma_{\boldsymbol{Z}} - \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{ZX}}^T & 0_{n\times m} \\ 0_{m\times n} & \Sigma_{\boldsymbol{X}} \end{pmatrix} \begin{pmatrix} \mathbb{I}_n & -\Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1} \\ 0_{m\times n} & \mathbb{I}_m \end{pmatrix}. \tag{2.15}$$

Now, let the joint probability density functions of $(\boldsymbol{Z}, \boldsymbol{X})^T$, $\boldsymbol{Z}$, and $\boldsymbol{Z} \mid (\boldsymbol{X} = \boldsymbol{x})$ be denoted by $f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$, $f_{\boldsymbol{Z}}(\boldsymbol{z})$, and $f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x})$, respectively. By the definition of conditional probability:

$$f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})}.$$

From the properties of determinants, we obtain that

$$\begin{aligned}\left|\frac{\Sigma}{\Sigma_{\boldsymbol{X}}}\right| &= \frac{|\Sigma|}{\left|\begin{pmatrix}\mathbb{I}_n & 0_{n\times m}\\ 0_{m\times n} & \Sigma_{\boldsymbol{X}}\end{pmatrix}\right|}\\ &= \left|\Sigma\begin{pmatrix}\mathbb{I}_n & 0_{n\times m}\\ 0_{m\times n} & \Sigma_{\boldsymbol{X}}^{-1}\end{pmatrix}\right|\\ &= \left|\begin{pmatrix}\Sigma_{\boldsymbol{Z}} & \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}\\ \Sigma_{\boldsymbol{ZX}}^T & \mathbb{I}_m\end{pmatrix}\right|\\ &= \left|\Sigma_{\boldsymbol{Z}} - \underbrace{\Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}}_{B}\Sigma_{\boldsymbol{ZX}}^{\top}\right|.\end{aligned}$$

Using equation (2.15), we can derive that

$$\left(\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{pmatrix} - \boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{x}\end{pmatrix} - \boldsymbol{\mu}\right) - (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^T\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) =$$
$$(\boldsymbol{z} - (\boldsymbol{\mu}_{\boldsymbol{Z}} + B(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})))^T(\Sigma_{\boldsymbol{Z}} - B\Sigma_{\boldsymbol{ZX}}^T)^{-1}(\boldsymbol{z} - (\boldsymbol{\mu}_{\boldsymbol{Z}} + B(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}))).$$

Thus, we obtain

$$\begin{aligned}&f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) = (2\pi)^{-n/2}|\Sigma_{\boldsymbol{Z}} - B\boldsymbol{\Sigma}_{\boldsymbol{ZX}}^T|^{-1/2}\times\\ &\times \exp\left(-\frac{1}{2}(\boldsymbol{z} - (\boldsymbol{\mu}_{\boldsymbol{Z}} + B(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})))^T(\Sigma_{\boldsymbol{Z}} - B\Sigma_{\boldsymbol{ZX}}^T)^{-1}(\boldsymbol{z} - (\boldsymbol{\mu}_{\boldsymbol{Z}} + B(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})))\right),\end{aligned}$$

which defines the distribution $\mathcal{N}_n(\boldsymbol{\mu}_{\boldsymbol{Z}} + B(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}), \Sigma_{\boldsymbol{Z}} - B\Sigma_{\boldsymbol{ZX}}^T)$. The proof is completed as

$$\boldsymbol{Z}|(\boldsymbol{X} = \boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Z}} + \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}), \Sigma_{\boldsymbol{Z}} - \Sigma_{\boldsymbol{ZX}}\Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{ZX}}^T).$$

□

Lemma 2 (Expectation of Quadratic Forms)**.** *Let* $\boldsymbol{Z}$ *and* $\boldsymbol{X}$ *be random vectors of size* n*, and let* $A \in \mathbb{R}^{n\times n}$*. Then,*

$$\mathbb{E}[\boldsymbol{Z}^T A \cdot \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Z}]^T A\mathbb{E}[\boldsymbol{X}] + \mathit{Tr}(A \cdot \mathit{Cov}(\boldsymbol{X}, \boldsymbol{Z})).$$

Proof. We compute

$$\mathbb{E}[\boldsymbol{Z}^T A\boldsymbol{X}] = \mathbb{E}[\mathrm{Tr}(\boldsymbol{Z}^T A\boldsymbol{X})] = \mathbb{E}[\mathrm{Tr}(A\boldsymbol{X}\boldsymbol{Z}^T)] = \mathrm{Tr}(\mathbb{E}[A\boldsymbol{X}\boldsymbol{Z}^T]),$$

where the cyclic property of the trace operator has been used. Decomposing the last addend of the RHS of the equation above, we have that

$$\begin{aligned}\mathrm{Tr}(\mathbb{E}[A\boldsymbol{X}\boldsymbol{Z}^T]) &= \mathrm{Tr}(A \cdot \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z}) + \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Z}]^T) \\ &= \mathrm{Tr}(A \cdot \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z})) + \mathrm{Tr}(A \cdot \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Z}]^T) \\ &= \mathrm{Tr}(A \cdot \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z})) + \mathrm{Tr}(\mathbb{E}[\boldsymbol{Z}]^T A \cdot \mathbb{E}[\boldsymbol{X}]),\end{aligned}$$

where we have used the fact that both the trace and the expectation are linear operators. We conclude that

$$\mathbb{E}[\boldsymbol{Z}^T A \cdot \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Z}]^T A\mathbb{E}[\boldsymbol{X}] + \mathrm{Tr}(A \cdot \mathrm{Cov}(\boldsymbol{X},\boldsymbol{Z})).$$

□

2.4 Empirical Data used in the thesis

In this section, we present the datasets used to illustrate our methodologies, detailing their source, structure, and the manipulations required for the analysis. The first and second analysis concern COVID-19 transmission in Italy, characterized by closely spaced observations, as daily data was collected throughout the pandemic. The third analysis focuses on clonal hematopoietic tracking, where observations were collected at wider intervals, with responses recorded on a monthly timescale.

2.4.1 COVID-19 Transmission in Italy

As of 20 February 2020, the Italian Department of Protezione Civile and the National Institute of Statistics (ISTAT) have made COVID-19 data available at the provincial and regional levels through the platform http://dati.istat.it/Index.aspx?QueryId=18460 and the GitHub repository https://github.com/pcm-dpc/COVID-19. In Chapter 3 we use these datasets directly from the source, while in Chapter 4 the same data are imported using the *COVID-19 package* by Guidotti and Ardia (2020). The collected data included different stages of disease severity, cases of recovery, and deaths. Data are reported in both daily variation and cumulative form since the beginning of the pandemic.

We use the *basic reproduction number* R_0 as the principal metric to assess the severity of the disease's spread. R_0 is the expected number of secondary infections produced by a single infected individual in a completely susceptible population (Dietz, 1993). This epidemiological parameter serves as a threshold indicator: $R_0 > 1$ suggests that the disease will spread in the population, while $R_0 < 1$ implies a decline in transmission and eventual disease elimination. Mathematically, R_0 is determined by the ratio of the rate at which individuals enter into the state of being infected to the rate at which they exit it, either through recovery or death. Specifically, it aggregates the rates of infection, recovery, and death, providing a holistic view of the disease dynamics within a population. By incorporating these multiple transmission and recovery rates into a single measure, R_0 simplifies the interpretation of complex epidemiological data, giving a full comparison of various interventions and policy outcomes, imposed during different phases of the pandemic.

COVID-19 in Italy (2020)

In Chapter 3, we consider COVID-19 data from March 9, 2020 (the beginning of the first national lockdown) to January 13, 2021 (the conclusion of the second nationwide imposed lockdown). We choose this timeframe both for the expectation of interesting dynamics, as the lockdown imposed complete isolation and therefore non-interaction between regions, and for the copious set of results with which to compare the validity of the estimated results (Mingliang et al., 2022; Remuzzi and Remuzzi, 2020). Spatial granularity is considered at the regional level by separating the region of Trentino-Alto Adige into the two provinces of Trento and Bolzano, due to their distinct healthcare systems and independent management of COVID-19 directives (Signorelli, 2019). In this way, the analysis accounts for the specific public health responses and epidemiological conditions in each province. As the

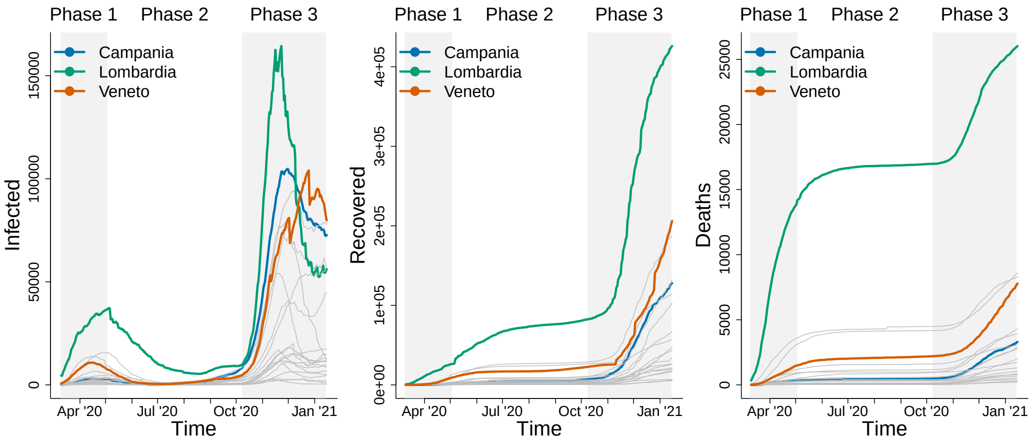


FIG. 2.2 *Dataset used in Chapter 3: evolution of the number of confirmed infections (left), recovered cases (centre), and deaths (right) across Italian regions.* *The three most affected regions—Campania, Lombardy, and Veneto—are indicated with distinct colours, while the remaining regions are shown in dark gray. Lombardy consistently exhibits the highest numbers of infections, recoveries, and deaths. The background shading indicates the three distinct phases of the analysis, with the darker sections corresponding to the lockdown periods. From the post-first-lockdown phase (white area), a shift in dynamics is observed: infections decrease while recoveries and deaths stabilize. In the subsequent non-lockdown phase (right-grey area), there is a marked resurgence in infections and deaths.*

daily cumulative number of infected shows a clear difference during the lockdown period from the non-lockdown period. We split the data into three distinct phases, as visually depicted in Figure 2.2 and assume constant reaction rates within each of these. The three phases are associated to different public health policies and restrictions. In particular,

1. The first phase corresponds to the initial national lockdown, implemented on March 9, 2020, which imposed strict restrictions on travel and gatherings (Conte, 2020c). During this period, as seen in Figure 2.2 (left), the number of infections rose sharply in Lombardy, which consistently exhibited the highest numbers throughout all phases. Campania and Veneto followed similar upward trends but at lower levels.

2. The second phase marks the easing of these restrictions, beginning on May 4, 2020, which allowed limited movement and the resumption of

some economic activities (Conte, 2020b). During this phase, infection rates appear relatively stable in all three regions, with no dramatic increases, indicating a temporary control of the spread. However, the cumulative numbers of recoveries and deaths continued to rise steadily, as shown in the centre and right panels of Figure 2.2.

3. The third phase, starting on October 8, 2020, saw the reintroduction of stricter measures, including mandatory mask-wearing and renewed limits on gatherings (Conte, 2020a). This phase coincides with a steep surge in infections, particularly in Lombardy, which reached its highest levels of the entire pandemic. The number of recoveries also spiked during this phase, suggesting that healthcare systems were actively managing more cases. The increase in deaths underscores the severity of the pandemic during this period.

Overall, the shaded regions in Figure 2.2 indicate the three distinct phases, with the darkest shading representing periods of strict lockdown, highlighting the effectiveness of public health interventions and their influence on the pandemic's trajectory, with distinct changes observed at the onset and relaxation of each phase.

COVID-19 in Lombardy (2021)

The first COVID-19 vaccine in Italy was administered on December 27, 2020, marking the beginning of the national vaccination campaign (Gozzi et al., 2022). To evaluate the effect of the vaccination rollout on the population, it was necessary to analyze data from the subsequent year when the impact of immunization could be meaningfully observed. For these reasons, the temporal dataset used in Chapter 4 spans the entire calendar year of 2021. From a spatial perspective, the study focuses on Lombardy, the region that recorded the highest number of infections during the first year of the pandemic (see Fig.2.2), and later emerged as the most heavily affected area in terms of overall cases and deaths. Lombardy was also the site where some of the strictest prevention and isolation measures were implemented, making it a pivotal case for assessing the efficacy of public health policies (Biology, 2022; Privitera, 2020).

We divided the infection process into three distinct stages: individuals infected at home, those hospitalized, and those critically ill. The first state is defined by subtracting the number of recovered cases, deaths, and hospitalizations from the total confirmed cases. This represents individuals who are infected but do not require hospitalization. The second state is calculated as the difference between the number of hospitalizations and ICU admissions, capturing patients who are hospitalized but not critically ill. The third state is directly obtained from the original dataset as well as the numbers of recovered individuals and deaths. To refine the analysis, the daily differences in confirmed cases are computed, reflecting changes over time.

Covariates Description. The period considered for the analysis encompasses a series of significant government interventions characterized by more targeted public health strategies and refined isolation protocols. These measures included restricted mobility, mandatory mask-wearing, limitations on access to public venues, and the enforcement of curfews. A key element that unifies the stringency of these interventions is the Government Response Index (GRI) (Hale et al., 2020), which quantifies the daily intensity of various public health policies. Since the original index differentiates between national and regional measures, assigning a positive or negative sign to each, we transform the GRI using absolute values to eliminate such distinction.

Environmental variables have been shown to influence the transmission, severity, and fatality rates associated with COVID-19 (Kifer et al., 2021), thus the average temperature in Lombardy has been included in our analysis (ARPA Lombardia, 2022). We expanded the monthly temperatures into a daily temperature sequence by repeating each average according to the number of days in the respective month. Then we smoothed the discrete values, appling a rolling mean with a window size of 5 days, centred on each day. The first and last two values, which were undefined due to the windowing, are set to the third and fourth values from the start and end of the sequence, respectively, to avoid edge effects.

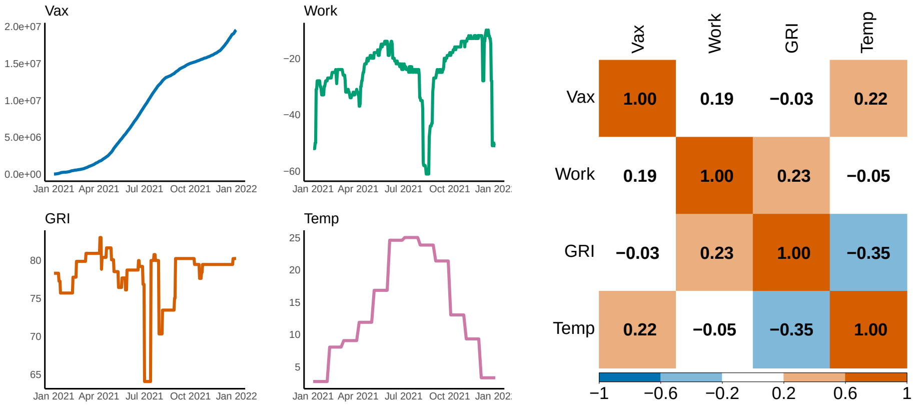


FIG. 2.3 ***Covariates used in Chapter 4.*** *The panels on the left display the evolution of four covariates, from January 2021 to January 2022: Vax (vaccination doses), Work (mobility trends for workplaces), GRI (Government Response Index), and Temp (smoothed monthly temperature). The right panel shows the correlation matrix between these covariates. The correlations suggest moderate relationships between the variables, with no strong multicollinearity present.*

Time spent at work significantly influences the probability of disease transmission, as individuals in workplace settings are often in closer proximity to one another, facilitating the spread of infectious agents (Guidotti and Ardia, 2020). The % time-at-work covariate is derived from the Google COVID-19 Mobility Reports, which present movement trends categorized by location (Google, 2021). This variable reflects the smoothed percentage of time individuals spent at work, relative to a baseline calculated as the median value from the five weeks preceding the pandemic (January 3 – February 6, 2020). In our analysis, we apply a smoothing technique to mitigate the impact of weekly patterns or artificial fluctuations caused by reduced activity during weekends.

Figure 2.3 illustrates the evolution of the four covariates from January 2021 to January 2022. The left panels depict the time series trends for each of these factors: The vaccinations variable shows a steady increase, reflecting the ongoing vaccination campaign, while the Work variable exhibits fluctuations that correspond to changes in mobility restrictions and work-from-home policies. In particular, the change in summer mobility compared to the pre-COVID period is visible. The GRI reflects the intensity of pub-

Covariate	Description	Time	Reference
Vax	Population with at least one dose of vaccine	Daily	Ministry of Health (2021) Guidotti and Ardia (2020)
GRI	Government Response Index	Daily	Hale et al. (2020)
Work	Mobility trends for places of work.	Daily	Google (2021)
Temp	Temperature.	Monthly	ARPA Lombardia (2022)

Table 2.2: **Covariates used in Chapter 4**. From left to right: variable names used in the text, along with their descriptions, time intervals, and data sources.

lic health interventions, showing pronounced peaks during critical periods in the spring and autumn, indicating more stringent measures in response to rising infection rates. In contrast, the GRI is lower during the summer months when the incidence of cases typically declines. Meanwhile, the average temperature fluctuates between 0 and 30 degrees Celsius, highlighting the seasonal variations that may influence both public health responses and the dynamics of the virus transmission. The right panel presents the correlation matrix among these covariates. The moderate values suggest that the selected variables were intentionally chosen to minimize multicollinearity, ensuring that each covariate contributes uniquely to the analysis. The covariates utilized in this analysis, along with their notation and sources, are detailed in Table 2.2.

2.4.2 Clonal Hematopoietic Data

The third analysis aims to analyze data recorded at widely spaced time intervals, a scenario that poses significant challenges for parameter estimation. In this context, we analyzed clonal tracking data collected from Rhesus Macaques, originally reported in (Wu et al., 2014), using the R package `Karen` by Del Core et al. (2022), which facilitated direct comparison of our results with those obtained by the original authors.

The study involves the mobilization of peripheral blood (MPB) CD34+

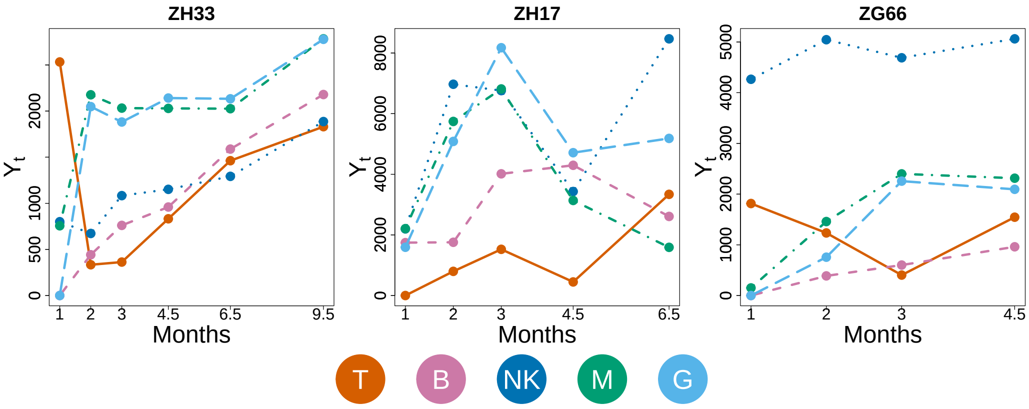


FIG. 2.4 ***Clonal tracking data used in Chapter 5.*** *Mean concentration over time of each Rhesus Macaque specimen following transplantation. The $p = 5$ cell types are reported with different colours and line styles.*

cells from three Rhesus Macaques, which are transduced with barcoded vectors to track the clonal dynamics. Following engraftment, granulocyte (G), monocyte (M), lymphoid T, B, and natural killer (NK) cells are flow sorted over periods of 9.5 months (ZH33), 6.5 months (ZH17), and 4.5 months (ZG66). The total number of clones collected amounts to 1165 (ZH33), 1280 (ZH17), and 1291 (ZG66). The dataset consists of multiple matrices representing time-series data, with each matrix corresponding to the lineage of a clone identifiable by its unique barcode sequence. We systematically remove any rows that contained only zero values, filtering out observations that lacked meaningful data. As a pre-processing step, we excluded time points where no barcodes are detected and removed all clones with fewer than three temporal observations, ensuring sufficient data for robust analysis and mitigating potential matrix inversion issues. This process yields a total of 555 unique barcode IDs, distributed across 434, 50, and 19 different clone types in specimens ZG66, ZH17, and ZH33, respectively. We then merged these individual datasets into a single comprehensive dataset for further analysis. By doing so, we assume that the chemical rates of hematopoietic differentiation are the same across all individuals of the Rhesus macaque species. Figure 2.4 shows the mean concentration over time of T, B, NK, M, and G for the three species (ZH33, ZH17, ZG66).

The observation times, originally recorded in months, are rescaled to a daily scale to allow comparison of the estimated parameters with other state-of-the-art studies, such as Del Core et al. (2023). This rescaling does not alter the dataset itself, as it only involves transforming the time parameters from months to days, which can be easily reverted to their original scale if necessary.

Chapter 3

Latent Event History Models for Quasi-Reaction Systems

[1] Published in **Computational Statistics & Data Analysis**, 2024 May 31. doi: 10.1016/j.csda.2024.107996

[2] Oral presentation at **SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation**, June 21-23, 2023, University of Ancona (IT)

3.1 Introduction

An increasing number of natural phenomena can be described by quasi-reaction systems of stochastic differential equations, as these are able to capture the inherent stochasticity of many processes. Examples include the stem cell differentiation process (Pellin et al., 2019, 2023), the dynamics of a biological system (Wilkinson, 2018) or of an infectious disease spreading (Britton et al., 2019), and the diverse applications of diffusion processes (Craigmile et al., 2023). The dynamics of these systems depend critically on parameters which are often unknown. Estimating these parameters is therefore important for characterizing and predicting the evolution of a dynamic system.

The likelihood of the intermittently observed process has rarely an explicit form (Wilkinson, 2018). To overcome this problem, Local Linear Approximation (LLA) methods provide an explicit approximation of the likelihood function under some assumptions (Shoji and Ozaki, 1998). Nevertheless, both in the case when observations are too spaced out in time and when the inter-observations times are too close, estimates based on the LLA are biased. Komorowski et al. (2011) present an extensive study on the effects of correlation between molecule concentrations on statistical inference, in the specific case of stochastic chemical kinetics models. Various approaches for reducing the variance of parameter estimators in a generic multi-response, non-linear model are available and could be used also in the case of dynamic systems. In the context of D-optimal designs, the most commonly used criterion assumes knowledge of the variance-covariance matrix (Fedorov, 2013). Although alternatives exist that use only an estimate of this matrix (Cooray-Wijesinha and Khuri, 1987), recent studies have observed that minimising the determinant of the information matrix is computationally efficient but not very robust (Hatzis and Larntz, 1992). An alternative approach is the use of Tikhonov regularisation techniques (Engl et al., 1996). However, if the measurements are taken very close together in time, the concentrations can be constant, leading to zero standard deviations and making also regularisation infeasible.

An approach to overcome these limitations is proposed. Intuitively, when the particles in the system are observed very close in time, one may be able to reconstruct which events of the stochastic process have taken place in order to result in a change of the system from the current to the next time point. Thus, the core element of the proposed approach involves integrating event history analysis into the framework of quasi-reaction systems. Originally conceived for sociological studies, event history models have been used on a range of applications, from engineering to medicine, economics, political science and psychology (Box-Steffensmeier and Jones, 2004). As the rates governing the evolutions of the state of the system and of the underlying event counting process are clearly linked, and they depend on the previous state of the system, the first contribution of the paper will be to formalise a joint statistical model that couples the two processes.

The second contribution of the paper is to develop an inferential procedure for the proposed model. As the occurrence of the events is not observed, an Expectation-Maximation (EM) algorithm for parameter estimation is derived. In particular, at the E-step, Kalman filter (Kalman, 1960) is used for the prediction of the latent events from the dynamics of the system observed on the entire time interval. The most popular version of Kalman filtering is for the case of Gaussian linear systems. However, an extension is required. Firstly, since the latent event counts are not Gaussian, the Poisson distribution is approximated with a continuous Gamma distribution, which is then transformed to a Gaussian distribution via a marginal transformation. Secondly, since the resulting system is non-linear, an extended Kalman filtering procedure is proposed for estimating the latent state of the event count process (Anderson and Moore, 2012). This allows the evaluation of the Q-function, which is then maximised at the M-step of the EM algorithm. In this way, the approach relates to other implementations of EM algorithms with an embedded Kalman filter, such as (Shumway and Stoffer, 1982) for dynamic linear systems, and more recent extensions for non-linear systems, such as the EM extended Kalman filter of Bar-Shalom et al. (2001), the EM unscented Kalman filter of Wan and Van Der Merwe (2000) and the EM particle filter of Zia et al. (2008).

The rest of the paper is organized as follows. In Section 3.2, the latent event history model for quasi-reaction systems is formalized. In Section 3.3, the EM algorithm for parameter estimation is described. In Section 3.4, a simulation study demonstrates the method's performance and highlights the settings where it is particularly advantageous compared to the existing LLA approaches. In Section 3.5, an illustration of the method on the modelling of the COVID-19 transmission dynamics in Italy is presented. Finally, in Section 3.6, conclusions and directions for future work are discussed.

3.2 Modelling quasi-reaction systems

Consider a closed system in which p substrates interact, each denoted as Q_l with $l = 1, \ldots, p$. These substrates could represent the compartments of an infectious disease model, the cell types in a cell differentiation model, or the different molecules in a biochemical reaction system. The j-th chemical reaction can generally be described as

$$k_{1j}Q_1 + ... + k_{pj}Q_p \xrightarrow{\beta_j} s_{1j}Q_1 + ... + s_{pj}Q_p \qquad j \in 1, \ldots, r, \tag{3.1}$$

where r indicates the number of reactions describing the dynamic system. Let $\mathcal{R}$ denote the set of possible reactions. The *stoichiometric coefficients* k_{lj} and s_{lj} are fixed integer values that describe the amount of substrate l, as reactant and product, respectively, that is needed for reaction j to occur, while $\theta_j = \exp(\beta_j) \in \mathbb{R}^+$ is the rate at which reaction j occurs.

The log-reaction rates $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ characterize the evolution of the dynamic system. These are the parameters that need to be estimated, given realizations of the state of the system over time. Let then $Y_l(t)$ denote the amount of the l-th particle at time t, with $t \in [0, T]$. Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_p(t))^T \in \mathbb{N}_0^p$ denote the state of the system at time t. Even if reaction equations like (3.1) are often used to represent kinetic models, as these facilitate a qualitative understanding of the dynamics, from a mathematical point of view chemical reactions are modelled primarily as systems of stochastic differential equations (Wilkinson, 2018). This methodology enables a quantitative interpretation of the dynamics, as it allows to study the temporal variation of the counts $\mathbf{Y}$ from the dynamics at the unit level.

According to the underlying dynamic system, particles encounter resulting in an instantaneous firing of one of the reactions. As a result, the system moves to the next state. Two viewpoints can be taken in the characterization of the stochastic process that induces changes in the state $\mathbf{Y}$ over time. The first viewpoint, presented in Section 3.2.1, models the stochastic process by which reactions occur and, as a by-product, the state of the system $\mathbf{Y}$ moves to a new configuration, deterministically. A second viewpoint, presented in Section 3.2.2, directly describes the change of the system based on the amount of particles available at a certain point in time and the hazard rate of each reaction at that point in time. As the occurrence of reactions is not observed, the second viewpoint is the most direct approach for modelling dynamic systems. Indeed, this is the approach considered in the literature and the one that results in the traditional LLA approaches for parameter estimation.

Instead, when the system is observed at small time intervals, one may be able to reconstruct the underlying process of reactions, leading to a more accurate characterization of the dynamic system. The main reason is the high temporal correlation between the states at small time scales. Motivated by this, Section 3.2.3 shows how the two viewpoints can be unified into a joint statistical model.

3.2.1 Event process

Let $e_j := (t_j, r_j)$ be the *event* that reaction $j \in \mathcal{R}$ occurs at time t_j. Associated with this marked point process and with each reaction j, there is a multivariate counting process

$$N_j(t) = \#\{\text{Reactions of type } j \text{ occurring in time interval } [0,t], j \in \mathcal{R}\}.$$

$N_j(t)$ is assumed to follow a non-homogeneous Poisson process

$$N_j(t) \sim \text{Poisson}(\Lambda_j(t)),$$

with cumulative rate

$$\Lambda_j(t) = E[N_j(t) \mid \mathcal{F}_{t^-}] = \int_0^t \lambda_j(\boldsymbol{Y}(u); \boldsymbol{\beta}) du,$$

where $\mathcal{F}_{t^-}$ is the history of the process up to, but excluding, time t.

The hazard rate $\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\beta})$ depends on the state of the system at time t as well as on the amount of particles of each type that are needed for each reaction to occur, i.e., the stoichiometric coefficients k_{lj} in (3.1). In particular, it holds that (Wilkinson, 2018)

$$\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\beta}) = \exp(\beta_j)\prod_{l=1}^{p}\binom{Y_l(t)}{k_{lj}}, \tag{3.2}$$

where $\binom{Y_l(t)}{k_{lj}} = 0$, for all $Y_l(t) < k_{lj}$.

3.2.2 Particle count process

The state of the system $\boldsymbol{Y}(t)$ is itself also a continuous time discrete Markov process. In the particular setting of a quasi-reaction system, it is possible to establish the temporal evolution of the probability distribution $P(\mathbf{Y};t)$, i.e., the probability that $\mathbf{Y}$ is the state of the system at time t. This will again depend on the state of the system just before time t. In particular, the distribution can be shown to satisfy the chemical master equations (Wilkinson, 2018)

$$\frac{dP(\boldsymbol{Y};t)}{dt} = \sum_{j\in\mathcal{R}}\left[\lambda_j\left(\boldsymbol{Y}(t)-V_{\cdot,j};\boldsymbol{\beta}\right)P\left(\boldsymbol{Y}-V_{\cdot,j};t\right)-\lambda_j(\boldsymbol{Y}(t);\boldsymbol{\beta})P(\boldsymbol{Y};t)\right], \tag{3.3}$$

where V denotes the net effect matrix, with (l,j) entry given by $v_{lj} = s_{lj}-k_{lj}$.

A solution of (3.3) gives the full transition probability kernel for the system dynamics. The master equations, however, can be solved analytically only in a small number of cases, due to the vast spectrum of conceivable state configurations (McQuarrie, 1967). On the other hand, from the master equations, one can derive the conditional expectation and variance of the rate of changes of the system. These are given, respectively, by

$$\begin{aligned}\frac{\mathbb{E}[\mathbf{Y}(t+dt)-\mathbf{Y}(t)\mid\mathbf{Y}(t)]}{dt} &= V\boldsymbol{\lambda}(\boldsymbol{Y}(t);\boldsymbol{\beta}), \qquad (3.4)\\ \frac{\mathbb{V}[\mathbf{Y}(t+dt)-\mathbf{Y}(t)\mid\mathbf{Y}(t)]}{dt} &= V\mathrm{diag}(\boldsymbol{\lambda}(\boldsymbol{Y}(t);\boldsymbol{\beta}))V^T.\end{aligned}$$

These two moments form the basis of the LLA solution to the master equations via a generalised least-squares approach (Pellin et al., 2019). Alternative approximations based on the van Kampen expansion have been proposed within a Bayesian inferential approach (Capistrán et al., 2012).

3.2.3 Latent event history model

The two characterizations described above are now merged into one joint model based on realizations of the process at discrete time points. Let then $\mathbf{Y}_i = \mathbf{Y}(t_i)$, $i = 0, \ldots, N$, be the state of the process at $N+1$, not necessarily equispaced, time points. Under the non-homogeneous Poisson process described in Section 3.2.1 and assuming that the hazard rates remain constant within the N time intervals, the increments of event counts follow a Poisson distribution, conditional on the history of the process. In particular,

$$\Delta N_{ij} = N_j(t_i) - N_j(t_{i-1}) \mid \mathcal{F}_{t_{i-1}} \sim \text{Poisson}(\mu_{ij}(\mathbf{Y}_{i-1}; \boldsymbol{\beta})), \quad j = 1, \ldots, r, \tag{3.5}$$

where

$$\mu_{ij}(\mathbf{Y}_{i-1}; \boldsymbol{\beta}) = (t_i - t_{i-1})\lambda_j(\mathbf{Y}_{i-1}; \boldsymbol{\beta}), \tag{3.6}$$

with $\lambda_j(\mathbf{Y}_{i-1}; \boldsymbol{\beta})$ defined as in Equation (3.2). For the rest of the manuscript, $\Delta\mathbf{N}_i$ and $\boldsymbol{\mu}_i$ denote the vectors of reaction counts ΔN_{ij} and rates $\mu_{ij}(\mathbf{Y}_{i-1}; \boldsymbol{\beta})$, respectively, in the interval $(t_{i-1}, t_i]$ across the r reactions.

It is clear how knowledge of the increments $\Delta\mathbf{N}_i$ would allow for perfect prediction of the state of the system at time t_i, since $\mathbf{Y}_i - \mathbf{Y}_{i-1} = V\Delta\mathbf{N}_i$. Combined with (3.5), this implies that $\mathbf{Y}_i - \mathbf{Y}_{i-1}$ is a linear combination of Poisson random variables, conditional on the history of the process. However, this linear combination does not have an explicit distribution in itself, and, more importantly, the increments for different particle types are not independent, leading to a further complication in the likelihood. For this reason, an approximate state-space formulation of the process that circumvents a direct full likelihood approach is proposed.

To this end, an approximation of the Poisson distribution of $\Delta\mathbf{N}_i$ with a continuous distribution is proposed. In particular, a Gamma distribution with a mean and variance matching that of $\Delta\mathbf{N}_i$, and with a similar skewness, is considered. In this way, the process can be rewritten as a Gaussian state space model. More in detail, the discrete increments ΔN_{ij} in Equation (3.5) are associated to the continuous random variable $X_{ij} = F_{ij}^{-1}(\Phi_{ij}(Z_{ij}))$, where F_{ij} is the CDF of a Gamma distribution with scale parameter 1 and shape parameter $\mu_{ij}(\mathbf{Y}_{i-1};\boldsymbol{\beta})$ from Equation (3.6), and Φ_{ij} is the CDF of a Gaussian distribution with mean and variance both equal to $\mu_{ij}(\mathbf{Y}_{i-1};\boldsymbol{\beta})$. So Z_{ij} is the Gaussian random variable that is uniquely associated to the Gamma distributed random variable X_{ij}, and with the same conditional mean and variance as the original ΔN_{ij} variable. In the remaining of the paper, $\mathbf{Z}_i$ will denote the r-dimensional vector of Gaussian random variables associated to the event counts in the interval $(t_{i-1}, t_i]$, $\mathbf{X}_i$ the corresponding Gamma random variables and G the function that transforms $\mathbf{Z}_i$ into $\mathbf{X}_i$, namely

$$\mathbf{X}_i = G(\mathbf{Z}_i) = \big(F_{i1}^{-1}(\Phi_{i1}(Z_{i1})), \ldots, F_{ir}^{-1}(\Phi_{ir}(Z_{ir}))\big).$$

With the latent event counts $\Delta\mathbf{N}_i$ approximated by $\mathbf{X}_i$, it follows that, approximately, $\mathbf{Y}_i - \mathbf{Y}_{i-1} = V\mathbf{X}_i = VG(\mathbf{Z}_i)$. In the following, the state space model will be formulated more generally, so as to account also for possible measurement error in the observations $\mathbf{Y}_i$, which may be relevant in some applied settings. In particular, the following latent event history model is proposed:

$$\begin{cases} \mathbf{Z}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, & \boldsymbol{\varepsilon}_i \sim \mathcal{N}\big(0, \mathrm{diag}(\boldsymbol{\mu}_i)\big), \\ \mathbf{Y}_i = \mathbf{Y}_{i-1} + VG(\mathbf{Z}_i) + \boldsymbol{\psi}_i, & \boldsymbol{\psi}_i \sim \mathcal{N}(0, \Sigma), i = 1, \ldots, N, \end{cases} \tag{3.7}$$

where $\boldsymbol{\psi}_i$ is a Gaussian noise vector with mean zero and variance-covariance $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. The case of no measurement error in $\mathbf{Y}_i$, which will be considered in the simulations, will correspond to the special case of $\sigma_l^2 = 0$, $l = 1, \ldots, p$. Figure 3.1 summarizes the dependence structure associated to the proposed model.

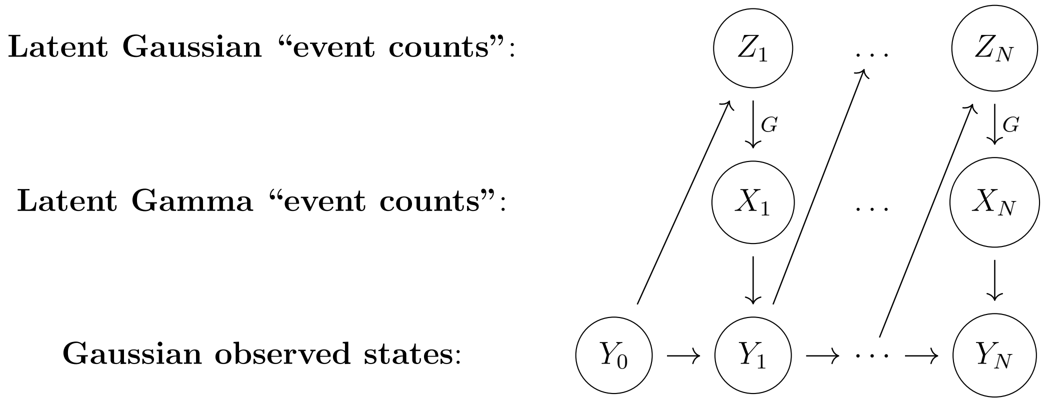

Latent Gaussian "event counts":
Latent Gamma "event counts":
Gaussian observed states:
Z_1
Z_N
G
G
X_1
X_N
Y_0
Y_1
Y_N

FIG. 3.1 ***Latent event history model DAG***. *The state of the system at time t_i, $\boldsymbol{Y}_i$, depends on the previous state, $\boldsymbol{Y}_{i-1}$, and on the number of reactions $\boldsymbol{X}_i$ that occur in the interval $(t_{i-1}, t_i]$. The latent $\boldsymbol{X}_i$ is approximated with a Gamma distribution and connected deterministically with a Gaussian random vector $\boldsymbol{Z}_i$, via a marginal transformation G. Notice how $\boldsymbol{Z}_i$ is independent of future states, $\boldsymbol{Y}_{(i+1):N}$, conditional on current and past states, $\boldsymbol{Y}_{0:i}$.*

3.3 Inference

This section discusses statistical inference of the latent event history model (3.7). Denoting with $\mathbf{Y}$ the $(N+1) \times p$ matrix of observations at the $N+1$ time points and $\mathbf{Z}$ the $(N+1) \times r$ matrix of latent variables, estimation of $\boldsymbol{\beta}$ and Σ requires the optimization of the marginal log-likelihood

$$\ell_{\mathbf{Y}}(\boldsymbol{\beta}, \Sigma) = \log \int_{\mathbf{Z}} L_{\mathbf{Z},\mathbf{Y}}(\boldsymbol{\beta}, \Sigma) d\mathbf{Z}. \tag{3.8}$$

As common in the presence of latent variables, an Expectation-Maximisation (EM) algorithm for parameter estimation is derived (Dempster et al., 1977). To this end, the complete log-likelihood, conditional on the initial state $\mathbf{Y}_0$ and assuming some measurement error in $\mathbf{Y}_i$ ($\Sigma \neq 0$), can be factorized into

$$\ell_{\mathbf{Z},\mathbf{Y}}(\boldsymbol{\beta}, \Sigma) = \sum_{i=1}^{N} \left[\ell_{\mathbf{Z}_i|\mathbf{Y}_{i-1}}(\boldsymbol{\beta}) + \ell_{\mathbf{Y}_i|\mathbf{Z}_i,\mathbf{Y}_{i-1}}(\Sigma) \right], \tag{3.9}$$

where

$$\ell_{\mathbf{Z}_i|\mathbf{Y}_{i-1}}(\boldsymbol{\beta}) = -\frac{1}{2}\Big\{r\log(2\pi)+\log(|\mathrm{diag}(\boldsymbol{\mu}_i)|)+\big[\mathbf{Z}_i-\boldsymbol{\mu}_i\big]^T(\mathrm{diag}(\boldsymbol{\mu}_i))^{-1}\big[\mathbf{Z}_i-\boldsymbol{\mu}_i\big]\Big\}$$

and

$$\begin{aligned}\ell_{\mathbf{Y}_i|\mathbf{Z}_i,\mathbf{Y}_{i-1}}(\Sigma) = &-\frac{1}{2}\Big\{p\log(2\pi)+\log(|\Sigma|)\\ &+\big[\mathbf{Y}_i-\mathbf{Y}_{i-1}-VG(\mathbf{Z}_i)\big]^T\Sigma^{-1}\big[\mathbf{Y}_i-\mathbf{Y}_{i-1}-VG(\mathbf{Z}_i)\big]\Big\}.\end{aligned}$$

The EM algorithm will then consist in the following two steps, which are iterated until convergence:

- *E-step*: Setting $\boldsymbol{\beta}$ and Σ to the current estimate of the parameters, $\boldsymbol{\beta}^*$ and Σ^*, respectively, compute the expected value of the complete log-likelihood (3.9) with respect to the distribution of the latent variables given the observations:

$$Q(\boldsymbol{\beta},\Sigma|\boldsymbol{\beta}^*,\Sigma^*) = \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\beta}^*,\Sigma^*}[\ell_{\mathbf{Z},\mathbf{Y}}(\boldsymbol{\beta},\Sigma)]. \tag{3.10}$$

- *M-step*: Find the optimal $\boldsymbol{\beta}$ and Σ by maximising the objective function (3.10) with respect to $\boldsymbol{\beta}$ and Σ.

In the next two sections, the computational aspects associated to the two steps, respectively, are discussed in detail.

3.3.1 E-step: Extended Kalman filtering

With the complete log-likelihood written as in (3.9), the Q-function (3.10) with slight abuse of notation is given by

$$Q(\boldsymbol{\beta},\Sigma|\boldsymbol{\beta}^*,\Sigma^*) = \mathbb{E}[\ell_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\beta})|\mathbf{Y}_{0:N}] + \mathbb{E}[\ell_{\mathbf{Y}|\mathbf{Z}\mathbf{Y}}(\Sigma)|\mathbf{Y}_{0:N}], \tag{3.11}$$

where $\mathbf{Y}_{0:N}$ denotes the data across all time points. Under model (3.7), the first term involves the following expectation

$$\begin{aligned}
\mathbb{E}[\ell_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\beta})|\mathbf{Y}_{0:N}] =& \mathbb{E}\Big[-\frac{1}{2}\sum_{i=1}^{N}\Big\{r\log(2\pi)+\log(|\mathrm{diag}(\boldsymbol{\mu}_i)|) \\
&+\big[\mathbf{Z}_i-\boldsymbol{\mu}_i\big]^T\mathrm{diag}(\boldsymbol{\mu}_i)^{-1}\big[\mathbf{Z}_i-\boldsymbol{\mu}_i\big]\Big\}|\mathbf{Y}_{0:N}\Big] \\
\propto& -\frac{1}{2}\sum_{i=1}^{N}\Big\{\mathbb{E}\big[\mathbf{Z}_i^T|\mathbf{Y}_{0:N}\big]\mathrm{diag}(\boldsymbol{\mu}_i)^{-1}\mathbb{E}\big[\mathbf{Z}_i|\mathbf{Y}_{0:N}\big] \\
&+\mathrm{Tr}\big[\mathrm{diag}(\boldsymbol{\mu}_i)^{-1}\mathbb{V}[\mathbf{Z}_i|\mathbf{Y}_{0:N}]\big]-2\mathbb{E}\big[\mathbf{Z}_i^T|\mathbf{Y}_{0:N}\big]\cdot\mathbf{1}+\boldsymbol{\mu}_i^T\cdot\mathbf{1}\Big\},
\end{aligned}$$

while expectation of the second term results in

$$\begin{aligned}
\mathbb{E}[\ell_{\mathbf{Y}|\mathbf{ZY}}(\Sigma)|\mathbf{Y}_{0:N}] =& \mathbb{E}\Big[-\frac{1}{2}\sum_{i=1}^{N}\Big\{p\log(2\pi)+\log(|\Sigma|) \\
&+\big[\Delta\mathbf{Y}_i-VG(\mathbf{Z}_i)\big]^T\Sigma^{-1}\big[\Delta\mathbf{Y}_i-VG(\mathbf{Z}_i)\big]\Big\}|\mathbf{Y}_{0:N}\Big] \\
\propto& -\frac{1}{2}\sum_{i=1}^{N}\Big\{-2\Delta\mathbf{Y}_i^T\Sigma^{-1}V\mathbb{E}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}\big] \\
&+\mathbb{E}\big[G(\mathbf{Z}_i)^T|\mathbf{Y}_{0:N}\big]V^T\Sigma^{-1}V\mathbb{E}\big[G(\mathbf{Z}_i)\mid\mathbf{Y}_{0:N}\big] \\
&+\mathrm{Tr}(\Sigma^{-1}V\mathbb{V}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}\big]V^T)\Big\},
\end{aligned}$$

with $\Delta\mathbf{Y}_i = \mathbf{Y}_i - \mathbf{Y}_{i-1}$ and keeping only the terms dependent on the latent variables.

In particular, the calculation of the Q-function requires the evaluation of the following first and second moments: $\mathbb{E}[\mathbf{Z}_i|\mathbf{Y}_{0:N}]$, $\mathbb{V}[\mathbf{Z}_i|\mathbf{Y}_{0:N}]$, $\mathbb{E}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}]$, and $\mathbb{V}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}]$. To this end, a Kalman filter approach is considered.

Firstly, notice how the dependences implied by model (3.7) are such that

$$\begin{aligned}\mathbb{E}[\mathbf{Z}_i|\mathbf{Y}_{0:N}] &= \mathbb{E}[\mathbf{Z}_i|\mathbf{Y}_{0:i}],\\ \mathbb{V}[\mathbf{Z}_i|\mathbf{Y}_{0:N}] &= \mathbb{V}[\mathbf{Z}_i|\mathbf{Y}_{0:i}],\\ \mathbb{E}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}] &= \mathbb{E}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}],\\ \mathbb{V}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:N}] &= \mathbb{V}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}],\end{aligned}$$

since $\mathbf{Z}_i$ is independent of future states, $\mathbf{Y}_{(i+1):N}$, conditional on current and past states, $\mathbf{Y}_{0:i}$ (Figure 3.1). In the following, the first two quantities are denoted with $\hat{\mathbf{z}}_{i|i}$ and $V_{i|i}$, respectively. This means that the smoothing step of a traditional Kalman filtering procedure is not needed, and only the prediction and update steps are. Secondly, the non-linearity in $\mathbf{Z}_i$ induced by the marginal transformation G means that a standard Kalman filter approach is not applicable. Thus, in order to calculate the first and second moments of $G(\mathbf{Z}_i)$, an extended Kalman filter is considered, where the function G is approximated with a second order Taylor expansion.

According to the derivations in A.1, the first two expectations are given by

$$\begin{aligned}\hat{\mathbf{z}}_{i|i} &= \mathbb{E}\left[\mathbf{Z}_i \mid \mathbf{Y}_{0:i}\right] = \hat{\mathbf{z}}_{i|i-1} + K_i\left[\mathbf{Y}_i - \mathbf{Y}_{i-1} - V\left(\mathbf{g}_{i|i-1} + \frac{1}{2}\text{vect}(V_{i|i-1}H_{i|i-1})\right)\right],\\ V_{i|i} &= \mathbb{E}\left[\left(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i}\right)\left(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i}\right)^T \mid \mathbf{Y}_{0:i}\right] = \left(\mathbb{I}_r - K_i V J_{i|i-1}\right) V_{i|i-1},\end{aligned}$$

where

$$K_i = (V V_{i|i-1} J_{i|i-1})^T (V J_{i|i-1} V_{i|i-1} J_{i|i-1}^T V^T + \Sigma)^{-1},$$

and where the various quantities predicted from data up to time t_{i-1}, which are formally defined in A.1, are dependent on a current estimate of parameters $\boldsymbol{\beta}^*$ and Σ^*. As for the moments of $G(\mathbf{Z}_i)$, these are approximated by

$$\begin{aligned}\mathbb{E}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}\big] &\approx \mathbf{g}_{i|i} + \frac{1}{2}\text{vect}(V_{i|i}H_{i|i}),\\ \mathbb{V}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}\big] &\approx J_{i|i}V_{i|i}J_{i|i}^T,\end{aligned}$$

```
Algorithm 2 Extended Kalman Filter (E-step)
Require: $\mathbf{Y}, \boldsymbol{\beta}^*, \Sigma^*, V$
  for $i = 1, \dots, N$ do
    1. Prediction step
         $\hat{\mathbf{z}}_{i|i-1} = \boldsymbol{\mu}_i$
         $V_{i|i-1} = \mathrm{diag}(\boldsymbol{\mu}_i)$
    2. Update step
         $\hat{\mathbf{z}}_{i|i} = \hat{\mathbf{z}}_{i|i-1} + K_i\Big[\mathbf{Y}_i - \mathbf{Y}_{i-1} - V\big(\mathbf{g}_{i|i-1} + \frac{1}{2}\mathrm{vect}(V_{i|i-1}H_{i|i-1})\big)\Big]$
         $V_{i|i} = \big(\mathbb{I} - K_i V J_{i|i-1}\big)\, V_{i|i-1}$
       with
         $K_i = (V V_{i|i-1} J_{i|i-1})^T (V J_{i|i-1} V_{i|i-1} J_{i|i-1}^T V^T + \Sigma)^{-1}$
         $\mu_{ij} = \exp(\beta_j) \prod_{l=1}^{p} \binom{Y_{lt_{i-1}}}{k_{lj}} (t_i - t_{i-1}) \qquad j = 1, \dots, r$
  end for
```

with

$$\mathbf{g}_{i|i} = G(\mathbf{Z})|_{\hat{\mathbf{z}}_{i|i}}, \qquad J_{i|i} = \frac{\partial G(\mathbf{Z})}{\partial \mathbf{Z}}|_{\hat{\mathbf{z}}_{i|i}}, \qquad H_{i|i} = \frac{\partial^2 G(\mathbf{Z})}{\partial \mathbf{Z}^2}|_{\hat{\mathbf{z}}_{i|i}}.$$

In particular, note how these moments depend on the moments of $\mathbf{Z}_i$ derived above, i.e., $\hat{\mathbf{z}}_{i|i}$ and $V_{i|i}$, so the latter are the main quantities that need to be calculated at the E-step.

Algorithm 2 summarizes the calculations required for the Kalman filter at the E-step of the algorithm, based on a current estimate of parameters, $\boldsymbol{\beta}^*$ and Σ^*. The Kalman filter predictions of the latent states are used in the evaluation of $Q(\boldsymbol{\beta}, \Sigma|\boldsymbol{\beta}^*, \Sigma^*) = \mathbb{E}[\ell_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\beta})|\mathbf{Y}_{0:N}] + \mathbb{E}[\ell_{\mathbf{Y}|\mathbf{ZY}}(\Sigma)|\mathbf{Y}_{0:N}]$, with

$$\begin{aligned}\mathbb{E}[\ell_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\beta})|\mathbf{Y}_{0:N}] = &-\frac{1}{2}Nr\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\Big\{\log(|\mathrm{diag}(\boldsymbol{\mu}_i)|)\\ &+ \hat{\mathbf{z}}_{i|i}^T \mathrm{diag}(\boldsymbol{\mu}_i)^{-1}\hat{\mathbf{z}}_{i|i}\\ &+ \mathrm{Tr}\big[\mathrm{diag}(\boldsymbol{\mu}_i)^{-1}\cdot V_{i|i}\big] - 2\hat{\mathbf{z}}_{i|i}^T\cdot\mathbf{1} + \boldsymbol{\mu}_i^T\cdot\mathbf{1}\Big\},\end{aligned} \qquad (3.12)$$

$$
\begin{aligned}
\mathbb{E}[\ell_{\mathbf{Y}|\mathbf{ZY}}(\Sigma)|\mathbf{Y}_{0:N}] = &-\frac{1}{2}N\log\left((2\pi)^p\prod_{l=1}^{p}\sigma_l^2\right)\\
&-\frac{1}{2}\sum_{i=1}^{N}\Big\{\Delta\mathbf{Y}_i^T\Sigma^{-1}\Delta\mathbf{Y}_i - 2\Delta\mathbf{Y}_i^T\Sigma^{-1}V\big(\mathbf{g}_{i|i}+\frac{1}{2}\mathrm{vect}(V_{i|i}H_{i|i})\big)\\
&+\big(\mathbf{g}_{i|i}+\frac{1}{2}\mathrm{vect}(V_{i|i}H_{i|i})\big)^TV^T\Sigma^{-1}V\big(\mathbf{g}_{i|i}+\frac{1}{2}\mathrm{vect}(V_{i|i}H_{i|i})\big)\\
&+\mathrm{Tr}\left(\Sigma^{-1}VJ_{i|i}V_{i|i}J^TV^T\right)\Big\},
\end{aligned}
\tag{3.13}
$$

and $\boldsymbol{\beta}^*$ and Σ^* the current values of the parameters used for the Kalman filter quantities $\hat{\mathbf{z}}_{i|i}$ and $V_{i|i}$.

3.3.2 M-step

The M-step maximizes the conditional expectation of the complete log-likelihood with respect to the parameters. Thus, the M-step involves the maximisation of the Q-function (3.11) with respect to $\boldsymbol{\beta}$ and Σ. Since the first term (3.12) does not depend on Σ, while the second term (3.13) does not depend directly on $\boldsymbol{\beta}$, the M-step results in the optimization of the first term (3.12) for the estimation of $\boldsymbol{\beta}$ and of the second term for the estimation of Σ. The latter is in fact available in closed form and is given by

$$
\begin{aligned}
\hat{\Sigma} =&\frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}(\Delta Y_i - VG(\mathbf{Z}_i))(\Delta Y_i - VG(\mathbf{Z}_i))^T\middle|\mathbf{Y}_{0:N}\right]\\
=&\frac{1}{N}\sum_{i=1}^{N}\Big\{\Delta Y_iY_i^T - 2\Delta Y_i(g_{i|i}+\frac{1}{2}V_{i|i}H_{i|i})^TV^T\\
&+V(g_{i|i}+\frac{1}{2}V_{i|i}H_{i|i})(g_{i|i}+\frac{1}{2}V_{i|i}H_{i|i})^TV^T+VJ_{i|i}V_{i|i}J_{i|i}^TV^T\Big\}.
\end{aligned}
\tag{3.14}
$$

The optimal values of $\boldsymbol{\beta}$ and Σ from the M-step are used as the new $\boldsymbol{\beta}^*$ and Σ^*, respectively, for computing a new expected log-likelihood at the E-step. This iterative procedure is repeated until convergence, e.g., until the estimates of $\boldsymbol{\beta}$ do not change significantly. Algorithm 3 summarizes the proposed EM algorithm.

Algorithm 3 EM algorithm

```
Require: $\mathbf{Y}, V, \boldsymbol{\beta}_{ini}, \Sigma_{ini}, \sigma^2, tol, maxit$
  while $err \geq tol$ & $it < maxit$ do
    for $i = 1, \ldots, N$ do
      1. E-step:
           Extended Kalman Filter: calculate $\hat{\mathbf{z}}_{i|i}$, $V_{i|i}$ from $\mathbf{Y}$, $V$, $\boldsymbol{\beta}_{old}$ $\Sigma_{old}$
      2. M-step:
           $\boldsymbol{\beta}_{new}, \Sigma_{new} \leftarrow \arg\max_{\boldsymbol{\beta},\Sigma} Q(\boldsymbol{\beta}, \Sigma | \boldsymbol{\beta}_{old}, \Sigma_{old}, \hat{\mathbf{z}}_{i|i}, V_{i|i})$
           $err \leftarrow \max ||\boldsymbol{\beta}_{new} - \boldsymbol{\beta}_{old}||_1^1$
           $\boldsymbol{\beta}_{old} \leftarrow \boldsymbol{\beta}_{new}$
           $\Sigma_{old} \leftarrow \Sigma_{new}$
           $it \leftarrow it + 1.$
    end for
  end while
```

3.3.3 Computational cost

The computational cost of the proposed EM algorithm is the combination of the computational cost of the E- and M-steps. At the E-step, the latent variables $\mathbf{Z}_i$ across the N time intervals are of dimension r, with r the number of reactions, and their covariance $V_{i|i}$ requires the inversion of a $p \times p$ matrix, where p is the number of substrates. Thus, the total complexity of the E-step is $\mathcal{O}(Nrp^3)$. On the other hand, the M-step concerns the optimization of an r-dimensional vector of parameters $\boldsymbol{\beta}$ and involves N inversions of an $r \times r$ matrix for the calculation of the objective function. Thus, the total complexity of the M-step is $\mathcal{O}(Nr^3p)$. This results in a computational cost of the full algorithm of the order $\mathcal{O}(Nr^3p^3)$, although this may vary depending on the speed of convergence of the numerical algorithm used for the optimization of the Q-function at the M-step.

3.3.4 Standard errors of reaction rates

Estimates of the reaction rates $\boldsymbol{\theta} = \exp(\boldsymbol{\beta})$ are the main output of the EM inference. Uncertainties on these point estimates can be summarised by their standard errors. Since the marginal log-likelihood in (3.8) is not a direct result of the EM algorithm, the standard errors are calculated from the Fisher information matrix associated to the Q-function, evaluated at the point estimates of $\boldsymbol{\theta}$ and Σ (Oakes, 1999). In particular, this is given by

$$I(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \widehat{\Sigma})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{3.15}$$

The variance of $\hat{\theta}_j$ is then given by $(I(\hat{\boldsymbol{\theta}})^{-1})_{jj}$.

If necessary, standard errors can be constructed also on $\boldsymbol{\beta}$. In particular, using the Delta method (Dorfman, 1938), the variances of $\boldsymbol{\beta}$ can be approximated by

$$\mathbb{V}(\hat{\beta}_j) = \frac{2}{\sum_{i=1}^{N} \left(\frac{2(\hat{z}_{i|i,j}^2 + V_{i|i,j})}{\hat{\mu}_{ij}} - 1 \right)},$$

with $\hat{\mu}_{ij}$ and $\hat{z}_{i|i,j}$ denoting the j-th elements of the vectors $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{z}}_{i|i}$, respectively, and $V_{i|i,j}$ the diagonal entry of $V_{i|i}$, all evaluated at the optimal point estimates of the parameters.

3.3.5 Model selection

In empirical settings, one may be interested in comparing different quasi-reaction systems of possibly varying complexity. Similarly to the derivation of the standard errors, a modified version of standard model selection criteria is considered, where the log-likelihood is replaced by the Q-function, which is instead a direct output of the EM algorithm (Ibrahim et al., 2008). In particular, the optimal model is taken as the one that minimizes the information criterion

$$IC = -2Q(\hat{\boldsymbol{\beta}}, \widehat{\Sigma}|\hat{\boldsymbol{\beta}}, \widehat{\Sigma}) + P(\hat{\boldsymbol{\beta}}), \tag{3.16}$$

with $Q(\hat{\boldsymbol{\beta}}, \widehat{\Sigma}|\hat{\boldsymbol{\beta}}, \widehat{\Sigma})$ the Q-function (3.11) evaluated upon convergence of the EM algorithm and $P(\hat{\boldsymbol{\beta}})$ a term penalizing model complexity. In the real application, the Bayesian Information Criterion (BIC) will be considered, where $P(\hat{\boldsymbol{\beta}}) = r \log(N)$, with r the number of reaction rates in the model and N the number of time intervals.

3.4 Simulation Study

In this section, a simulation study is provided to evaluate the performance of the proposed method under different settings and to highlight those where it is particularly advantageous compared to the existing LLA approaches. For the simulation, a dynamic system with a low number of particles ($p = 4$) and reactions ($r = 6$) is considered, in order to mimic a setting that is common in many applications, such as the cell differentiation process studied by Pellin et al. (2023). In the specifics, the 6 reactions contain one duplication, two death and three differentiation reactions. Figure 3.2a provides a graphical representation of the system, while Figure 3.2b reports the 6 reactions. These corresponds to the net effect matrix

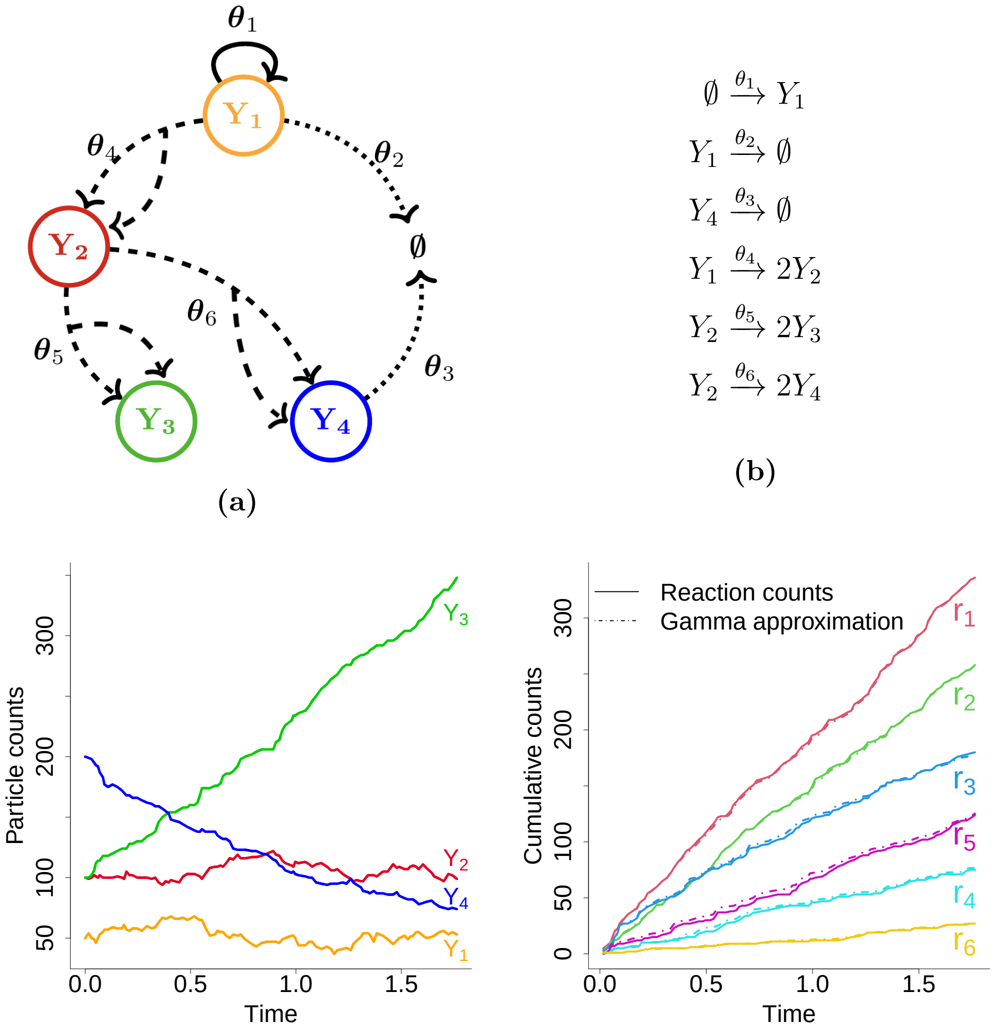
$$V = \begin{pmatrix} 1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 2 & -1 & -1 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & -1 & 0 & 0 & 2 \end{pmatrix}.$$

The parameters are set to

$$\boldsymbol{\beta}_{true} = \log(\boldsymbol{\theta}_{true}) = (5.30, 1.10, -0.11, -0.22, -0.22, -1.61)^T,$$

and no measurement error is considered ($\Sigma = 0$). Starting with initial particle counts set to $\mathbf{y}_0 = (50, 100, 100, 200)$, a Gillespie algorithm is used to generate the stochastic process over time (Gillespie, 1977). Figure 3.2c reports one run of the algorithm, while Figure 3.2d shows how the reaction counts $\mathbf{N}_i$ are close to those based on the Gamma approximation $\mathbf{X}_i$ of $\Delta\mathbf{N}_i$, with the Gamma distribution defined using the true parameters $\boldsymbol{\beta}$.

Improvement over local linear approximation approach In the first simulation study, the performance of the algorithm is compared against the existing LLA approach in terms of parameter estimation. Given the motivation behind the proposed methodology, one would expect an improvement when the interval between consecutive observations is particularly small, as this generates a strong temporal correlation among the particle counts. More-



(a)

(b)

(c)

(d)

FIG. 3.2 ***Specifications of the cell differentiation process used in the simulation study****. (a) Structure of the process with $p = 4$ particles. Each substrate is represented by a coloured node, whereas birth, death and differentiation reactions are denoted by full, dotted and dashed edges, respectively. (b) The corresponding quasi-reaction system. (c) An example of trajectories generated by means of a Gillespie algorithm. (d) Cumulative counts of the reactions increments $\Delta \boldsymbol{N}_i$ (full lines) and of their Gamma approximations $\boldsymbol{X}_i$ (dotted lines).*

over, one would expect the difference to be more pronounced at low sample sizes, i.e., a small number of time points, as this will make statistical inference more challenging in general and may amplify the effect of strong temporal correlations.

In order to test these hypotheses, a subset of the trajectories generated by the Gillespie algorithm is considered. In particular, observations are retained at every 10, 15, 20, 25 and 30 time points out of the originally sampled trajectories. These are referred to as *jumps*. The larger the jump is, the larger the gap between consecutive time points where the process is observed. This will generally translate into a large number of reactions that may have occurred between one time point and the next, although this will depend also on the dynamics of the process at the specific time interval. In order to test also the effect of sample size, in each of the settings, two scenarios are considered: one where the first $N = 5$ time intervals are considered and a second one where the first $N = 50$ time intervals are considered, generated as above.

Parameter estimation is conducted for each of the datasets using LLA and the proposed EM algorithm. LLA uses the moments in (3.4) as the basis of a generalised least-squares approach given the particle count data $\mathbf{Y}$. For the EM algorithm described in Algorithm 3, the LLA solution is set as starting value for $\boldsymbol{\beta}$ ($\boldsymbol{\beta}_{ini}$), $\Sigma = 0$, the net effect matrix is set to V defined as above, the tolerance for convergence to $tol = 0.002$ and the maximum number of iterations to $maxit = 300$. Upon convergence, the quality of the estimation is evaluated by calculating the Kullback-Leibler divergence between the estimated and the true parameters. In particular, this is defined by

$$KL(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_{true}) = \mathbb{E}_{\mathbf{y}^+}[\log p(\mathbf{y}^+|\boldsymbol{\beta}_{true}) - \log p(\mathbf{y}^+|\hat{\boldsymbol{\beta}})],$$

where $\mathbf{y}^+$ indicates an additional dataset with the same characteristics as the one used for inference, and generated from the same underling process defined by $\boldsymbol{\beta}_{true}$. The lower this value is, the closer the inferred process is to the true one.

Figure 3.3 reports the results in the form of boxplots across the 100 simulations for each of the settings.The results show how parameter estimation with the proposed EM algorithm is better than with the existing LLA approach, both in terms of the KL divergence (left panel) and estimation of one of the parameters (β_1, right panel). All plots show how the effects are more

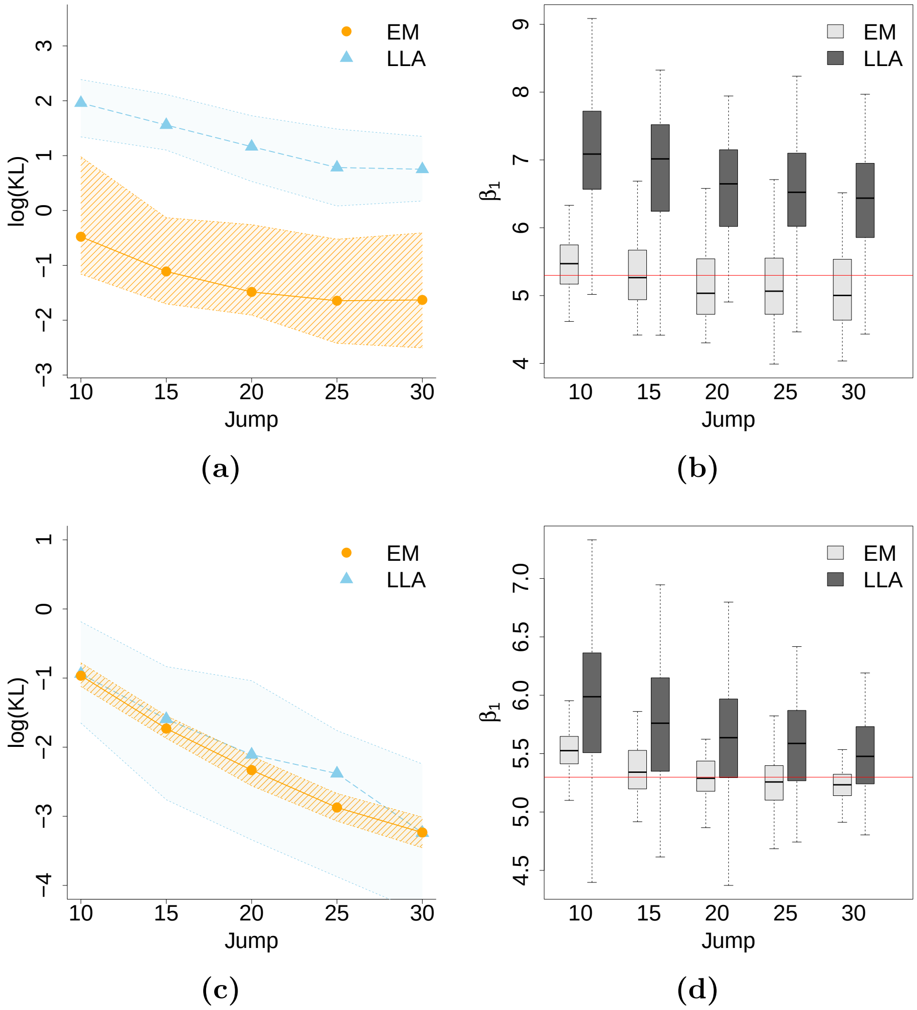
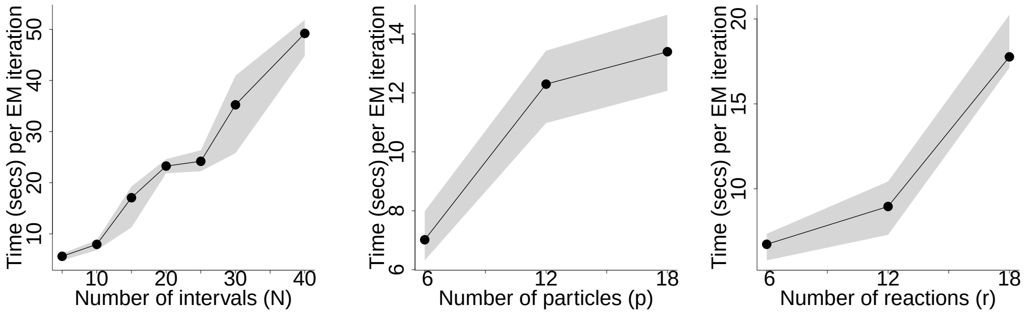


FIG. 3.3 ***Comparison between EM and LLA methods.*** *On the left, the KL measure, in the log scale, shows that parameter estimation with the EM algorithm is closer to the true parameters than with the LLA approach. On the right, the plots show how, for one of the parameters (β_1), estimates are more accurate with the EM than with the LLA approach. The true value is indicated by the horizontal red line. All plots show how the effects are more pronounced with $N = 5$ (3.3a, 3.3b) than with $N = 50$ (3.3c, 3.3d) time intervals. The boxplots are obtained across 100 simulations.*

pronounced with small sample sizes ($N = 5$, Figure 3.3a and 3.3b) than with larger sample sizes ($N = 50$, Figures 3.3c and 3.3d). Finally, Figure 3.3d in particular shows how the two approaches tend to converge to a similar



(a) **(b)** **(c)**

FIG. 3.4 ***Computational cost of EM algorithm.*** *Average computational time (in seconds) of one iteration of the EM algorithm in terms of (a) the number of time intervals (N), (b) the number of particles (p), (c) the number of reactions (r). Median, first and third quartiles are shown across 100 simulations.*

performance for larger time intervals (i.e., a large *jump*). This is to be expected, since temporal correlation will become less strong the larger the time interval. At the same time, the reconstruction of the reactions that have taken place within that time interval will also be less accurate. However, Figure 3.3d shows how, even in this case, estimation from the EM algorithm appears to be less biased and more accurate than with the LLA approach.

Computational cost in terms of number of time points, reactions, particles A second simulation study explores how the computational cost of the algorithm varies with respect to the number of time points (N), the number of reactions (r) and the number of particles (p). The results are shown in Figure 3.4. The first scenario (Figure 3.4a) considers the same generative process as before, fixing $jump = 30$ and letting the number of time intervals vary in $N = 5, 10, 15, 20, 25, 30, 40$. The plot shows how the average computational time of the EM algorithm is approximately linear in N.

The second scenario (Figure 3.4b) evaluates how the computational time varies with respect to the number of particles p. The three systems in Table 3.1 of A.2 are considered, with $jump = 40$ and $N = 10$. The systems are characterized by the same number of reactions as before ($r = 6$), but an increasing number of particles, namely $p = 6, 12, 18$, respectively. Figure 3.4b does not show the cubic dependence in p, that was anticipated. Given that the computational time is the combined time from the E- and the M-steps, this suggests a much slower M-step.

Finally, the third scenario (Figure 3.4c) evaluates how the computational time varies with respect to the number of reactions r. As before the parameters are set to $jump = 40$ and $N = 10$, but the three systems in Table 3.2 of A.2 are now considered. These are characterized by the same number of particles as before ($p = 6$), but an increasing number of reactions, namely $r = 6, 12, 18$, respectively. The plot shows a super-linear dependence in r.

3.5 Illustration on Italian COVID-19 data

This section provides a real data illustration focusing on the COVID-19 pandemic. Worldwide, more than 700 million infections and almost 7 million deaths were recorded as of August 16, 2023 (WHO, 2020). As a result, significant efforts have been made in order to understand the phenomenon and find strategies to control the spreading of the disease. Italy has been one of the countries of interest during the pandemic, being the first European country to experience a significant outbreak of the disease (Liao et al., 2020). The first case was confirmed on 31 January 2020. Since then, for more than two years, data were collected daily. The sufficiently close interval between observations is ideal for the application of the method, as it generates strong temporal correlations. The following analysis focusses on daily data within three specific time intervals, characterized by three different levels of contagion:

- **Phase 1**: $9^{th}March$ - $4^{th}May,$ 2020. Strong restrictions on travel throughout the country, banning all forms of gathering in private and public places (Conte, 2020c);
- **Phase 2**: $4^{th}May$ - $7^{th}October,$ 2020. Containment measures were relaxed, allowing the travelling for visits to relatives (within a region) and the restart of several production activities (Conte, 2020b);
- **Phase 3**: $8^{th}October,$ 2020 - $14^{th}January,$ 2021. Wearing of masks

became compulsory both outdoor and indoor, and assemblages were restricted (Conte, 2020a).

Within each of the three phases and using data from all 21 Italian regions, the proposed EM algorithm is used to fit the parameters of the following two dynamic systems:

Model A	**Model B**
$I_k \xrightarrow{\theta_{1k}} 2I_k$	$I_k \xrightarrow{\theta_{1k}} 2I_k$
$I_k \xrightarrow{\theta_2} R_k$	$I_k \xrightarrow{\theta_{2k}} R_k$
$I_k \xrightarrow{\theta_3} D_k$	$I_k \xrightarrow{\theta_{3k}} D_k$

The systems correspond to simple SIR compartmental models, where I is the number of infectious individuals, R the number of recovered individuals and D the number of deceased individuals (Simon, 2020). In particular, the first reaction models the creation of one infectious individual once a susceptible individual meets an infectious one. Note how the number of susceptible individuals S has been omitted, as it is almost constant throughout the observation period. The second and third reactions correspond to the cases of an infectious individual recovering and dying, respectively. The index k denotes the region. Thus, model B is characterized by region-specific rates for all three reactions. This results in a system with $p = 63$ particles and $r = 63$ rates. On the other hand, model A hypothesizes a simpler model where the recovery and death rates are assumed to be the same across Italy, under an assumption that these depend primarily on the specifics of the virus and are not as affected by the level of contagion in the population.

The proposed EM algorithm is used, with a tolerance $tol = 10^{-5}$, for the estimation of the reaction rates $\boldsymbol{\theta}$ and of the noise Σ. Using Equation (3.16) with a BIC penalty term, model A and model B result in $9.66{\cdot}10^5$ and $2.64{\cdot}10^5$ BIC values, respectively. This leads to the choice of the more complex model B, with region-specific recovery and death reaction rates. Since the BIC tends to select sparser models compared to other model selection criteria, this suggests that other model selection criteria would have led to the same conclusion. As for the parameter estimates, the error variances were generally

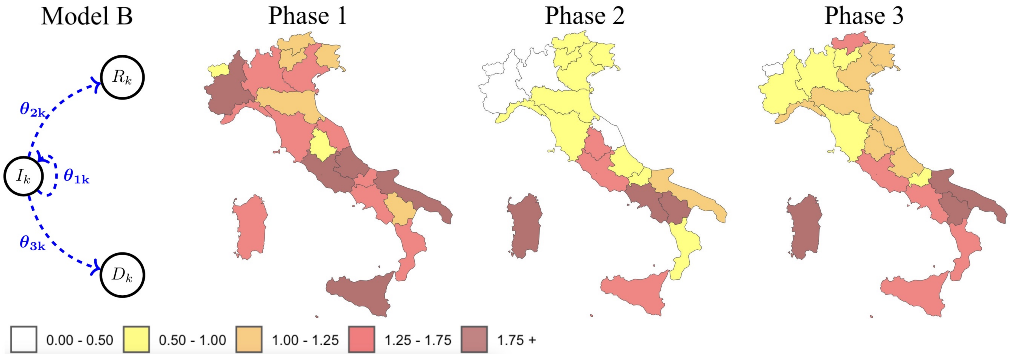


FIG. 3.5 ***Visualization of the estimated R_0 values in Italy.*** *On the left, the system of kinetic reactions for the k-th region. On the right, a visualization of the estimated values of the basic reproductive numbers for each Italian region, in the 3 phases of interest, categorised according to the colour coding used by the Italian government. An additional shade of dark red has been added to highlight the regions with the most critical R_0 values.*

far from zero, suggesting that some of the recorded cases were subject to a measurement error.

Figure 3.5 visualises the results in terms of the basic reproduction number R_0, which is the number of new infections that each infected individual produces on average (Wood and Wit, 2021). This can be estimated from the fitted models by calculating $\theta_{1k}/(\theta_{2k} + \theta_{3k})$ In Figure 3.5, these values are categorized according to the colour coding used by the Italian government to evaluate the severity of the disease spread. In particular, values below the boundary of $R_0 = 1$ are associated to a long-term decrease of the epidemic, while values above 1 indicate a long-term increase of the epidemic, with larger values (darker colours) associated to progressively more severe scenarios. The estimated R_0 values are in line with those from other studies (Giordano et al., 2020; Remuzzi and Remuzzi, 2020; Mingliang et al., 2022). The results in Figure 3.5 show how during Phase 2 the infection was limited, as a consequence of the containment measures implemented in Phase 1. During Phase 3, a revival of the disease spread is observed, in particular in the southern regions of Italy. The standard errors of R_0, calculated using the Delta method from the standard errors of the estimated reaction rates $\boldsymbol{\theta}$ given by (3.15), show significant differences in R_0 values between two consecutive phases at a 95% significance level, with the only exception of the

Autonomous Province of Bolzano between phase 1 and phase 2, and Molise and Sicily between phase 2 and phase 3. Moreover, although not assumed by the model, the R_0 estimates show some geographical clustering, which is to be expected given the movements of individuals between neighbouring regions.

3.6 Conclusions

A novel procedure for the statistical inference of quasi-reaction systems has been proposed. Local linear approximation methods tend to perform poorly when the system is observed at fine time intervals. This is due to numerical instability caused by strong correlations in the observations from one time point to the next. The proposed method focuses instead on reconstructing the underlying process of latent reactions. To this end, a latent event history model of the observed count process driven by a latent process of reactions is developed. A computationally efficient EM algorithm for parameter estimation is proposed, incorporating an extended Kalman filtering procedure for predicting the latent states. A simulation study demostrates how the proposed method performs better than the existing LLA approach, particularly when the time intervals between consecutive observations are small and the number of time points is low.

The method is illustrated by an application on the Italian Covid 19 data during the critical phase of the pandemic, between March 2020 and January 2021. The basic reproduction number R_0 of the 21 Italian regions estimated by the method in three consecutive phases of the pandemic shows higher values at the beginning and at the end of the time period. This is to be expected given the evolution of the disease and the societal restrictions that were imposed by the Italian government during this period.

The simple epidemic model considered is clearly a simplification of the pandemic process. The model does not consider inter-regional infections, nor effects from outside Italy or heterogeneity in the population. Most likely, ignoring this type of effects means that R_0 has been over-estimated by the models (Gomes et al., 2022). Future work will consider applying the same methodology to fit more complex models, such as the compartmental model of Wood and Wit (2021), which includes hospital infections and other types of interactions.

Data availability

The data used in this paper are available from the Dipartimento della Protezione Civile and the Istituto Nazionale di Statistica (ISTAT) via the webpage of the department (http://dati.istat.it/Index.aspx?QueryId=18460) and a Github repository (https://github.com/pcm-dpc/COVID-19).

APPENDIX

A.1 Kalman filtering (E-step)

This section discusses the extended Kalman filtering procedure that was developed for the evaluation of $\mathbb{E}[\mathbf{Z}_i|\mathbf{Y}_{0:i}]$, $\mathbb{V}[\mathbf{Z}_i|\mathbf{Y}_{0:i}]$, $\mathbb{E}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}]$, and $\mathbb{V}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}]$.

Prediction Step The prediction step calculates the first and second moments of $\mathbf{Z}_i$ conditional on $\mathbf{Y}_{0:i-1}$. According to model (3.7), these are in fact the conditional moments of $\mathbf{Z}_i$ given the state of the system at the previous time point, $\mathbf{Y}_{i-1}$. Thus

$$\begin{aligned} \hat{\mathbf{z}}_{i|i-1} &= \mathbb{E}\left[\mathbf{Z}_i \mid \mathbf{Y}_{i-1}\right] = \boldsymbol{\mu}_i, \\ V_{i|i-1} &= \mathbb{V}\left[\mathbf{Z}_i \mid \mathbf{Y}_{i-1}\right] = \operatorname{diag}(\boldsymbol{\mu}_i). \end{aligned} \tag{3.17}$$

Update Step Following from the prediction step, the update step refines these predictions by comparing them to the observed values at time i. In particular, the conditional distribution of $\mathbf{Z}_i$ is updated from the past with the information coming from $\mathbf{Y}_i$ by first deriving the joint distribution of $\mathbf{Y}_i$ and $\mathbf{Z}_i$ conditional on $\mathbf{Y}_{0:i-1}$. According to model (3.7), this is a multivariate Gaussian distribution, which can be written generically as

$$\left.\begin{matrix} \mathbf{Z}_i \\ \mathbf{Y}_i \end{matrix}\right| \mathbf{Y}_{0:i-1} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \right). \tag{3.18}$$

From the prediction step (3.17), $\mathbf{m}_1$ and S_{11} are already known. As regards to the other elements of the mean and covariance,

$$\begin{aligned}
\mathbf{m}_2 &= \mathbb{E}[\mathbf{Y}_i|\mathbf{Y}_{0:i-1}] = \mathbb{E}[\mathbf{Y}_{i-1} + VG(\mathbf{Z}_i) + \boldsymbol{\psi}_i|\mathbf{Y}_{0:i-1}] \\
&= \mathbf{Y}_{i-1} + V\mathbb{E}[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}], \\
S_{12} &= \mathbb{C}ov[\mathbf{Z}_i, \mathbf{Y}_i|\mathbf{Y}_{0:i-1}] \\
&= \mathbb{C}ov[\mathbf{Z}_i, \mathbf{Y}_{i-1} + VG(\mathbf{Z}_i) + \boldsymbol{\psi}_i|\mathbf{Y}_{0:i-1}] = V\mathbb{C}ov[\mathbf{Z}_i, G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}]V^T, \\
S_{22} &= \mathbb{V}[\mathbf{Y}_i|\mathbf{Y}_{0:i-1}] = \mathbb{V}[\mathbf{Y}_{i-1} + VG(\mathbf{Z}_i) + \boldsymbol{\psi}_i|\mathbf{Y}_{0:i-1}] \\
&= V\mathbb{V}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}\big]V^T + \Sigma.
\end{aligned} \tag{3.19}$$

In order to calculate the first and second moments of $G(\mathbf{Z}_i)$, the non-linear function G is approximated with its Taylor expansion of order 2 centered at $\hat{\mathbf{z}}_{i|i-1}$, i.e.,

$$G(\mathbf{Z}_i) \approx \mathbf{g}_{i|i-1} + J_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1}) + \frac{1}{2}\text{diag}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})H_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1}).$$

The first term of the expansion is the deterministic vector of size r

$$\mathbf{g}_{i|i-1} = G(\hat{\mathbf{z}}_{i|i-1}).$$

The other terms have a simplified form due to the fact that the j-th element of the function G is a function only of the j-th element of $\mathbf{Z}_i$. Thus, the $r \times r$ matrix of first derivatives is a diagonal matrix, with (j, j) element given by

$$\begin{aligned}
\big[J_{i|i-1}\big]_{jj} &= \big[\frac{\partial G(\mathbf{Z})}{\partial \mathbf{Z}}|_{\hat{\mathbf{z}}_{i|i}}\big]_{jj} = \frac{\partial G_j}{\partial z_{ij}}\bigg|_{\hat{z}_{ij|i-1}} \\
&= \frac{\partial(F_{ij}^{-1}(\Phi_{ij}(Z_{ij}))}{\partial z_{ij}}\bigg|_{\hat{z}_{ij|i-1}} = \left(\frac{\partial F_{ij}}{\partial x_{ij}}\bigg|_{G(\hat{z}_{ij|i-1})}\right)^{-1}\frac{\partial \Phi_{ij}}{\partial z_{ij}}\bigg|_{\hat{z}_{ij|i-1}},
\end{aligned}$$

where, using the functional form of the Normal and Gamma CDFs

$$\frac{\partial \Phi_{ij}}{\partial z_{ij}}(z) = \frac{e^{-\frac{(z-\mathbb{E}[Z_{ij}])^2}{2\mathbb{V}[Z_{ij}]}}}{\sqrt{2\pi\mathbb{V}[Z_{ij}]}}, \qquad \frac{\partial F_{ij}}{\partial x_{ij}}(x) = \frac{e^{-x}x^{\mathbb{E}[X_{ij}]-1}}{\Gamma(\mathbb{E}[X_{ij}])}\mathbb{1}_{[x>0]}.$$

Similarly, the $r \times r \times r$ Hessian matrix, can be written as an $r \times r$ diagonal matrix with second derivatives on the diagonal, namely

$$
\begin{aligned}
\left[H_{i|i-1}\right]_{jj} &= \frac{\partial}{\partial z_{ij}}\left[\left(\frac{\partial F_{ij}}{\partial x_{ij}}\Big|_{G(\hat{z}_{ij|i-1})}\right)^{-1}\frac{\partial \Phi_{ij}}{\partial z_{ij}}\bigg|_{\hat{z}_{ij|i-1}}\right] \\
&= \frac{-\frac{\partial^2 F_{ij}}{(\partial x_{ij})^2}\Big|_{G(\hat{z}_{ij|i-1})}\frac{\partial G_j}{\partial z_{ij}}\Big|_{\hat{z}_{ij|i-1}}\frac{\partial \Phi_{ij}}{\partial z_{ij}}\Big|_{\hat{z}_{ij|i-1}}+\frac{\partial F_{ij}}{\partial x_{ij}}\Big|_{G(\hat{z}_{ij|i-1})}\frac{\partial^2 \Phi_{ij}}{(\partial z_{ij})^2}\Big|_{\hat{z}_{ij|i-1}}}{\left(\frac{\partial F_{ij}}{\partial x_{ij}}\Big|_{G(\hat{z}_{ij|i-1})}\right)^2},
\end{aligned}
$$

where

$$
\frac{\partial^2 \Phi_{ij}}{\partial z_{ij}^2}(z) = \frac{(z-\mathbb{E}[Z_{ij}])e^{-\frac{(z-\mathbb{E}[Z_{ij}])^2}{2\mathbb{V}[Z_{ij}]}}}{\sqrt{2\pi}\mathbb{V}[Z_{ij}]^{3/2}}, \quad \frac{\partial^2 F_{ij}}{\partial x_{ij}^2}(x) = \frac{e^{-x}x^{\mathbb{E}[X_{ij}]-2}(\mathbb{E}[X_{ij}]-x-1)}{\Gamma(\mathbb{E}[X_{ij}])}\mathbb{1}_{[x>0]}.
$$

Going back to (3.19), the Taylor approximation can now be used to calculate the required conditional expectations. In particular,

$$
\begin{aligned}
\mathbb{E}\left[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}\right] \approx &\mathbb{E}\Big[\mathbf{g}_{i|i-1} + J_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1}) \\
&+ \frac{1}{2}\text{diag}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})H_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\Big] \\
=&\mathbf{g}_{i|i-1} + J_{i|i-1}\mathbb{E}\Big[\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1}|\mathbf{Y}_{0:i-1}\Big] \\
&+ \frac{1}{2}\mathbb{E}\Big[\text{diag}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})H_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\Big] \\
=&\mathbf{g}_{i|i-1} + \frac{1}{2}\text{vect}(V_{i|i-1}H_{i|i-1}),
\end{aligned}
$$

$$\begin{aligned}\mathbb{C}ov[\mathbf{Z}_i, G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}] \approx & \mathbb{C}ov\left[\mathbf{Z}_i, \mathbf{g}_{i|i-1}|\mathbf{Y}_{0:i-1}\right] \\ & + \mathbb{C}ov\left[\mathbf{Z}_i, J_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\right] \\ & + \mathbb{C}ov\left[\mathbf{Z}_i, \frac{1}{2}\text{diag}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})H_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\right] \\ = & \mathbb{C}ov\left[\mathbf{Z}_i, J_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\right] \\ = & \mathbb{V}ar[\mathbf{Z}_i|\mathbf{Y}_{0:i-1}]J_{i|i-1} = V_{i|i-1}J_{i|i-1},\end{aligned}$$

$$\begin{aligned}\mathbb{V}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i-1}\big] &\approx \mathbb{V}\left[\mathbf{g}_{i|i-1} + J_{i|i-1}(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1})|\mathbf{Y}_{0:i-1}\right] \\ &= J_{i|i-1}\mathbb{V}\left[\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i-1}|\mathbf{Y}_{0:i-1}\right]J_{i|i-1}^T = J_{i|i-1}V_{i|i-1}J_{i|i-1}^T.\end{aligned}$$

Finally, plugging these expressions into (3.17), it follows that

$$\begin{aligned}\mathbf{m}_2 &\approx \mathbf{Y}_{i-1} + V\left[\mathbf{g}_{i|i-1} + \frac{1}{2}\text{vect}(V_{i|i-1}H_{i|i-1})\right], \\ S_{22} &\approx V[J_{i|i-1}V_{i|i-1}J_{i|i-1}^T]V^T + \Sigma, \\ S_{12} &\approx VV_{i|i-1}J_{i|i-1},\end{aligned}$$

which, together with $\mathbf{m}_1$ and S_{11} derived previously, define the joint distribution (3.18) of $\mathbf{Z}_i$ and $\mathbf{Y}_i$ conditional on $\mathbf{Y}_{i-1}$. From this, using the formulae for the conditional distributions from a jointly Gaussian random vector, it follows that $\mathbf{Z}_i$, conditional on $\mathbf{Y}_{0:i}$, has a multivariate Gaussian distribution, with mean and covariance given, respectively, by

$$\begin{aligned}\hat{\mathbf{z}}_{i|i} &= \mathbb{E}\left[\mathbf{Z}_i \mid \mathbf{Y}_{0:i}\right] = \hat{\mathbf{z}}_{i|i-1} + K_i\left[\mathbf{Y}_i - \mathbf{Y}_{i-1} - V\left(\mathbf{g}_{i|i-1} + \frac{1}{2}\text{vect}(V_{i|i-1}H_{i|i-1})\right)\right], \\ V_{i|i} &= \mathbb{E}\left[\left(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i}\right)\left(\mathbf{Z}_i - \hat{\mathbf{z}}_{i|i}\right)^T \mid \mathbf{Y}_{0:i}\right] = \left(\mathbb{I}_r - K_iVJ_{i|i-1}\right)V_{i|i-1},\end{aligned}$$

where

$$K_i = (VV_{i|i-1}J_{i|i-1})^T(VJ_{i|i-1}V_{i|i-1}J_{i|i-1}^TV^T + \Sigma)^{-1}.$$

Note how the update step refines the conditional expectation found in the prediction step in proportion to the difference between the actual and estimated observations, i.e., the prediction error. Moreover, this is directly proportional to the magnitude of the Kalman *gain matrix* K_i, which captures the linear relationship between the noise and the variance of the latent variable (Kim and Bang, 2018).

Similarly to the earlier derivations,

$$\mathbb{E}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}\big] \approx \mathbf{g}_{i|i} + \frac{1}{2}\text{vect}(V_{i|i}H_{i|i}),$$
$$\mathbb{V}\big[G(\mathbf{Z}_i)|\mathbf{Y}_{0:i}\big] \approx J_{i|i}V_{i|i}J_{i|i}^T,$$

with

$$\mathbf{g}_{i|i} = G|_{\hat{\mathbf{z}}_{i|i}}, \qquad J_{i|i} = \frac{\partial G(\mathbf{Z})}{\partial \mathbf{Z}}|_{\hat{\mathbf{z}}_{i|i}}, \qquad H_{i|i} = \frac{\partial^2 G(\mathbf{Z})}{\partial \mathbf{Z}^2}|_{\hat{\mathbf{z}}_{i|i}}.$$

A.2 Dynamic systems used for the simulation study

This section reports the systems of reactions that were used in Section 3.4 for evaluating the computational complexity of the algorithm with respect to the number of particles p (Table 3.1) and the number of reactions r (Table 3.2).

$p = 6$

$$\begin{cases} Y_1 & \xrightarrow{\theta_1} Y_2 \\ Y_1 & \xrightarrow{\theta_2} Y_3 \\ Y_2 & \xrightarrow{\theta_3} Y_4 \\ Y_1 & \xrightarrow{\theta_4} Y_4 \\ Y_4 & \xrightarrow{\theta_5} Y_6 \\ Y_5 & \xrightarrow{\theta_6} Y_5 \end{cases}$$

$p = 12$

$$\begin{cases} Y_1 + Y_7 & \xrightarrow{\theta_1} Y_2 + Y_8 \\ Y_1 + Y_7 & \xrightarrow{\theta_2} Y_3 + Y_9 \\ Y_2 + Y_8 & \xrightarrow{\theta_3} Y_4 + Y_{10} \\ Y_1 + Y_7 & \xrightarrow{\theta_4} Y_4 + Y_{10} \\ Y_4 + Y_{10} & \xrightarrow{\theta_5} Y_6 + Y_{12} \\ Y_5 + Y_{11} & \xrightarrow{\theta_6} Y_5 + Y_{11} \end{cases}$$

$p = 18$

$$\begin{cases} Y_1 + Y_7 + Y_{13} & \xrightarrow{\theta_1} Y_2 + Y_8 + Y_{14} \\ Y_1 + Y_7 + Y_{13} & \xrightarrow{\theta_2} Y_3 + Y_9 + Y_{15} \\ Y_2 + Y_8 + Y_{14} & \xrightarrow{\theta_3} Y_4 + Y_{10} + Y_{16} \\ Y_1 + Y_7 + Y_{13} & \xrightarrow{\theta_4} Y_4 + Y_{10} + Y_{16} \\ Y_4 + Y_{10} + Y_{16} & \xrightarrow{\theta_5} Y_6 + Y_{12} + Y_{18} \\ Y_5 + Y_{11} + Y_{17} & \xrightarrow{\theta_6} Y_5 + Y_{11} + Y_{17} \end{cases}$$

Table 3.1: Three dynamic systems with $r = 6$ reactions, and an increasing number of particles ($p = 6, 12, 18$).

$r = 6$

$$\mathcal{R}_6 : \begin{cases} Y_2 & \xrightarrow{\theta_1} Y_1 + Y_3 \\ Y_3 & \xrightarrow{\theta_2} Y_2 + Y_4 \\ Y_4 & \xrightarrow{\theta_3} Y_3 + Y_5 \\ Y_5 & \xrightarrow{\theta_4} Y_4 + Y_6 \\ Y_5 & \xrightarrow{\theta_5} Y_6 \\ Y_6 & \xrightarrow{\theta_6} Y_1 \end{cases}$$

$r = 12$

$$\mathcal{R}_{12} : \mathcal{R}_6 \cup \begin{cases} Y_8 & \xrightarrow{\theta_1} Y_7 + Y_9 \\ Y_9 & \xrightarrow{\theta_2} Y_8 + Y_{10} \\ Y_{10} & \xrightarrow{\theta_3} Y_9 + Y_{11} \\ Y_{11} & \xrightarrow{\theta_4} Y_{10} + Y_{12} \\ Y_{12} & \xrightarrow{\theta_5} Y_1 1 \\ Y_7 & \xrightarrow{\theta_6} Y_1 2 \end{cases}$$

$r = 18$

$$\mathcal{R}_{18} : \mathcal{R}_{12} \cup \begin{cases} Y_{14} & \xrightarrow{\theta_1} Y_{13} + Y_{15} \\ Y_{15} & \xrightarrow{\theta_2} Y_{14} + Y_{16} \\ Y_{16} & \xrightarrow{\theta_3} Y_{15} + Y_{17} \\ Y_{17} & \xrightarrow{\theta_4} Y_{16} + Y_{18} \\ Y_{18} & \xrightarrow{\theta_5} Y_{17} \\ Y_{13} & \xrightarrow{\theta_6} Y_{18} \end{cases}$$

Table 3.2: Three dynamic systems with $p = 6$ particles, and an increasing number of reactions ($r = 6, 12, 18$).

Chapter 4

Inference for quasi-reaction models with covariate-dependent rates

[1] Published in **International Workshop on Statistical Modelling**, pp. 172-178, 2024 July 12. Springer Nature Switzerland.

[2] Oral presentation at **IWSM 2024-International Workshop on Statistical Modelling**, June 14-19, 2024, Durham University (UK).

4.1 Introduction

An increasing number of natural phenomena, such as infectious disease spreading (Britton et al., 2019), can be described by quasi-reaction systems of stochastic differential equations. A common modelling assumption is to associate a constant rate to each reaction. However, this assumption is too restrictive in many applications, leading to inferred dynamics that may be far from the true ones. Indeed, it is reasonable to assume that rates may vary dynamically or spatially, for example due to governmental imposed lockdowns, variations in the dynamics of contagion between geographical regions or between different groups of populations.

In the context of epidemic modelling, a certain level of heterogeneity can be achieved by adding new compartments and new reactions to the system. However, on the one hand, this is only a discrete approximation when the exogenous covariates are of a continuous nature, and, on the other hand, it may lead to overly complex models. As an alternative to this approach, in this paper, we propose to capture heterogeneity in the system dynamics by letting the rates of the quasi-reaction model depend directly on external covariates. In particular, we propose an extension of a recently developed latent event history model (Framba et al., 2024), by allowing log-reaction rates to be linearly dependent on a vector of covariates. In this way, the model is able to quantify the effect of covariates on the system dynamics. We adapt the Expectation-Maximization (EM) algorithm developed by Framba et al. (2024) to this new setting, and evaluate the effectiveness of this approach on simulated data and in the context of epidemic modelling.

The paper is organized as follows: Section 4.2 formalizes the proposed latent event history modelling approach. Section 4.3 evaluates the procedure on a simple SIR (Susceptible, Infected and Recovered) system, whose dynamics are affected by the start of a lockdown period. In Section 4.4, we show an illustration on COVID19 data from Italy, where the approach is able to assess the effect of environmental factors and public health interventions on the transmission and severity of the disease. Finally, in Section 4.5, we draw some conclusions.

4.2 Latent event history model

Consider a closed system in which p substrates, or compartments, interact via r reactions. Let k_{lj} be the stoichiometric coefficient, indicating the amount of substrate l needed for reaction j to occur, and let $\mathbf{Y}(t) = (Y_1(t), \dots, Y_p(t))$ be the state of the system at time t. The hazard of reaction j occurring instantaneously at time t is given by

$$\lambda_j(t) = \theta_j \prod_{l=1}^{p} \binom{Y_l(t-1)}{k_{lj}} \mathbb{1}_{Y_l(t-1) > k_{lj}},$$

with θ_j the non-negative rate associated to reaction j, for $j = 1, \dots, r$. These rates are typically assumed constant for each reaction (Framba et al., 2024). In contrast to this, in this paper, we propose to link the reaction rates to external covariates. In particular, given a vector of possibly time-dependent covariates $\mathbf{x} \in \mathbb{R}^q$, we assume

$$\theta_j = \exp(\mathbf{x}^t \boldsymbol{\beta}_j),$$

with $\boldsymbol{\beta}_j$ a q-dimensional vector of reaction-specific regression coefficients. In the following, we denote with $\boldsymbol{\beta}$ the vector of coefficients across all reactions. These are the parameters to be estimated.

The firing of reactions induces a change in the states of the system. In particular, if $\mathbf{Y}_i$, $i = 0, \dots, N$, is the state of the process at $N+1$, not necessarily equispaced, time points, then $\mathbf{Y}_i - \mathbf{Y}_{i-1} = V\mathbf{N}_i$, with V the $p \times r$ net-effect matrix , indicating the variation of the lth substrate due to the occurrence of the jth reaction, and $\mathbf{N}_i$ the number of each reaction occurring in the interval $(t_{i-1}, t_i]$. Since the number of reactions is not observed, we follow Framba et al. (2024) and propose to infer the dynamics of the system via the following state-space model

$$\begin{cases} \mathbf{Z}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, & \boldsymbol{\varepsilon}_i \sim \mathcal{N}\big(0, \operatorname{diag}(\boldsymbol{\mu}_i)\big), \\ \mathbf{Y}_i = \mathbf{Y}_{i-1} + VG(\mathbf{Z}_i) + \boldsymbol{\psi}_i, & \boldsymbol{\psi}_i \sim \mathcal{N}(0, \sigma_i^2 \cdot \mathbb{I}_p), \quad i = 1, \dots, N, \end{cases}$$

where $\boldsymbol{\psi}_i$ is a Gaussian measurement error, $\boldsymbol{\mu}_i = (t_i - t_{i-1})\boldsymbol{\lambda}(t_i; \boldsymbol{\beta})$ and $G(\mathbf{Z}_i)$, formally defined in Framba et al. (2024), is a continuous approximation to $\mathbf{N}_i$.

For the estimation of the parameters $\boldsymbol{\beta}$, we adapt the EM algorithm of Framba et al. (2024) to the presence of external covariates, making use of the extended Kalman filter for the prediction of the latent state variables at the E-step.

4.3 Simulation study: SIR system under lockdown

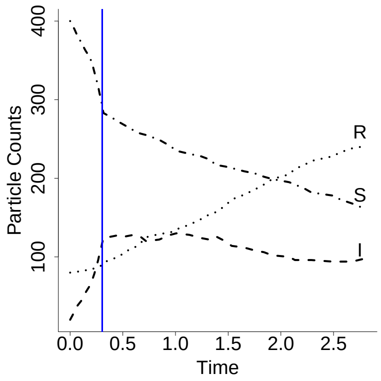
We evaluate the performance of the method on a simple SIR epidemiological model, consisting of $p = 3$ substrates (Susceptible, Infected, Recovered), and $r = 2$ reactions, given by

$$S + I \xrightarrow{\theta_1} 2I$$
$$I \xrightarrow{\theta_2} R$$

The first reaction represents the contagion from an infected person to a susceptible individual, while the second reaction denotes the recovery of an infected individual.

We introduce a binary covariate x to represent the impact of a lockdown. The covariate is initially set to 0 to signify a non-serious pandemic situation and changes to 1 when the susceptible-to-infected ratio exceeds a critical threshold, indicating the necessity of implementing lockdown measures to reduce the spread. The lockdown will have an impact on the reaction rates. In particular, letting $\theta_j = \exp(\beta_{0j} + \beta_{1j}x)$, we expect the infection rate θ_1 to reduce after the lockdown, so we set $\beta_{11} = -2.303$, while we expect the recovery rate θ_2 to not be affected by lockdown measures, so we set $\beta_{12} = 0$. For the other two parameters, we set $\beta_{01} = -4.094$ and $\beta_{02} = -0.693$, respectively. Figure 4.1a represents a trajectory simulated from this system and indeed shows how the curve of infection counts shifts from increasing to decreasing at the start of lockdown, represented by a vertical line.

We utilize the Gillespie algorithm to simulate 50 stochastic processes over



(a)

Rates	True	Median	SD
β_{01}	-4.094	-4.424	0.096
β_{11}	-2.303	-2.029	0.138
β_{02}	-0.693	-0.611	0.256
β_{12}	0.000	-0.227	0.252

(b)

FIG. 4.1 *(a) A trajectory of a stochastic SIR process. The vertical line denotes the beginning of lockdown, leading to a reduction in the infection rate; (b) Statistics of parameters across* 50 *simulations.*

time from this system, starting with an initial particle count configuration of $\mathbf{y}_0 = (400, 20, 80)$. Data are collected at intervals of 10 time points, selected from the originally sampled trajectories, for a total of 41 observations. The EM algorithm starts from $\beta_0 = (0, 0, 0, 0)$. We set no measurement error, i.e., $\sigma_i^2 = 0$ for all i, the tolerance for convergence to 0.002, and the maximum number of iterations to 100. Figure 4.1b reports the accuracy of parameter estimates, by comparing the true values with their median across the 50 simulations. Overall, the results indicate a close proximity between the estimated and the true values for all parameters.

4.4 Epidemic modelling of COVID-19

Italy was hit hard by the COVID-19 pandemic caused by the coronavirus SARS-CoV-2, and particularly so in the Lombardy region where the first cases were reported. The impact of the pandemic remains significant to these days: as of May 6th, 2024, the northern region has reported over 4.34 million confirmed cases and nearly $48,000$ deaths (Guidotti and Ardia, 2020). From

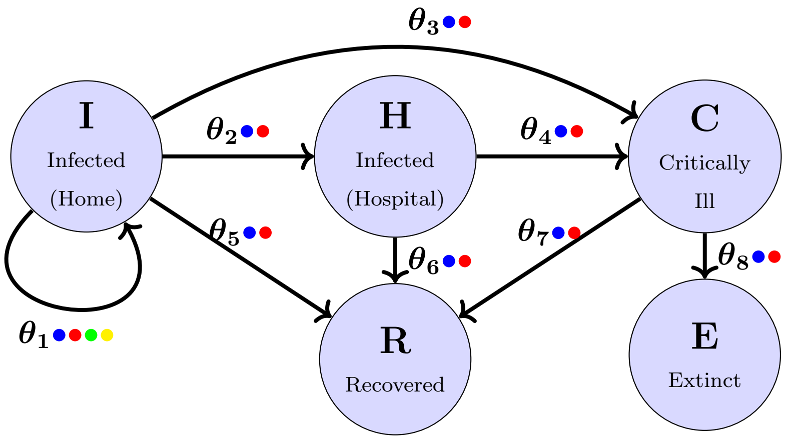


FIG. 4.2 *Graphical illustration of the epidemic model for COVID-19 spreading: arrows correspond to the 8 reactions, associated to different stages of the infection; dependence of reaction rates on time covariates is represented with the following colours: ● Vaccination Rate, ● Temperature, ● % Time at Work, ● GRI.*

day one of the pandemic, the Italian Institute of Health has started gathering a comprehensive dataset on the progression of the disease. In this section, we look closely at the data collected daily during the year 2021 (Guidotti and Ardia, 2020).

Figure 4.2 shows the system that we consider for modelling the dynamics of COVID-19 progression. In particular, the total population is partitioned into the following five compartments, which define the states of the dynamic system: **I**, infected but not in serious condition (homebound); **H**, infected and hospitalised; **C**, critically ill (requiring respiratory support); **R**, recovered (no longer infected); **E**, extinct (dead). The arrows correspond to the 8 reactions that define the system. It seems natural to assume that the implementation of mass vaccination campaigns and rigorous public health measures that were put in place to contain the pandemic had indeed an effect on the virus progression and spread. From a modelling point of view, this results in reaction rates that vary over time due to external time-dependent covariates.

For the analysis, we consider in particular the following covariates:

1. **Vaccination Rate**: cumulative sum of the vaccines delivered at time t. SARS-CoV-2 vaccination is proven to protect against both infection and manifestation of severe and fatal symptoms of the disease (Corrao

et al., 2022), so the dependence on this covariate has been included for all reactions.

2. **Temperature**: smoothed mean intensity of monthly temperatures retrieved from ARPA Lombardia (2022). Environmental factors have been found to have an impact on the transmissibility, severity, and mortality of COVID-19, so also this covariate has been considered for all reactions. (Kifer et al., 2021).

3. **% Time at Work**: smoothed variation of the percentage of time spent at work relative to its median value over the five weeks preceding the pandemic (Guidotti and Ardia, 2020). Viral transmission is clearly influenced by the time people spend at work, so this covariate has been included only for modelling the rate of infection.

4. **Government Response Index (GRI)**: a measure of the severity of government policies to reduce interactions, accounting for school and workplace closures, bans on public gatherings, suspension of public transport, stay-at-home mandates, public information campaigns, and international travel controls (Hale et al., 2020). Similarly to the previous covariate, this index has been included only for the reaction representing new infections.

All covariates are rescaled to zero mean and unit standard deviation.

We use the proposed EM algorithm to estimate the regression coefficients associated to each reaction rate. As initial value, we run the method of Framba et al. (2024) and set these estimated values as the intercepts to all models, and zero for all the other regression parameters. We set the tolerance for convergence to 10^{-5} and the maximum number of iterations to 200. The variances σ_i^2 associated with the increments of the i-th state are estimated offline. The results are presented in Table 4.1 in terms of estimated regression coefficients and standard errors. Since the marginal log-likelihood is not a direct output of the EM algorithm, we use the Fisher information matrix associated with the Q-function for the calculation of the standard errors (Oakes, 1999). The regression coefficients are all statistically significant and of a positive sign. As the variables are scaled, we deduce that temperature

Table 4.1: Estimated regression coefficients and standard errors for the COVID-19 model in Figure 4.2.

$\hat{\boldsymbol{\beta}}$	$I \to 2I$	$I \to H$	$I \to C$	$H \to C$	$I \to R$	$H \to R$	$C \to R$	$C \to E$
Intercept	-1.82	-2.15	-4.04	0.006	-4.65	0.657	1.547	1.378
	(0.001)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)	(0.002)	(0.002)
Vaccination Rate	0.657	0.473	1.296	1.690	1.201	0.605	1.232	1.443
	(0.001)	(0.001)	(0.003)	(0.002)	(0.004)	(0.001)	(0.002)	(0.003)
Temperature	0.725	0.824	1.354	1.018	1.628	0.905	0.673	1.370
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.002)	(0.002)
% Time at Work	0.003	-	-	-	-	-	-	-
	(0.001)	-	-	-	-	-	-	-
GRI	0.573	-	-	-	-	-	-	-
	(0.001)	-	-	-	-	-	-	-

and vaccination have the strongest effects on the system dynamics while percentage of time at work has the smallest effect.

From the estimates of the reaction rates, it is possible to obtain the basic reproduction number, expressing the number of new infections that each infected individual produces on average. In particular, this is defined by

$$R_0(t) = \frac{\theta_1(t)}{\theta_2(t) + \theta_3(t) + \theta_5(t)}$$

and is shown in Figure 4.3. The values estimated by our model (solid line) follow quite closely the ranges reported by the Italian government (Mattarella, 2021b), which are depicted with different colours. In particular, we correctly detect the peak of infection in the spring of 2021 and the period of lowest infection in the summer of 2021. In the winter of 2021, the results show how the rates were decreasing already prior to the strict measures coming into force (red column). This phenomenon was also observed in the very early stages of the pandemic, as reported in Wood and Wit (2021).

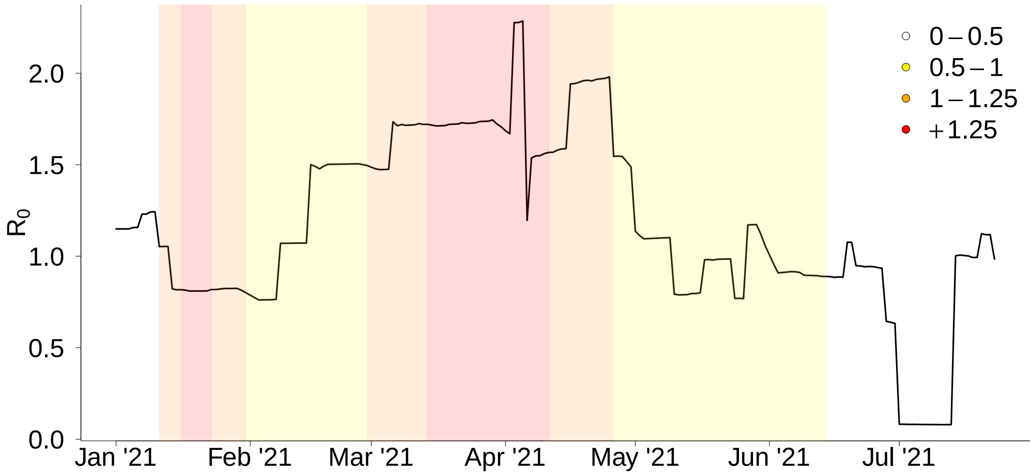


FIG. 4.3 *Comparison of basic reproduction number estimated by the model (solid line) and reported by the government (background colours) (Mattarella, 2021a).*

4.5 Conclusion

In this paper, we have proposed an extension of existing quasi-reaction models to account for covariate-dependent reaction rates. We have evaluated the effectiveness of the proposed approach in modelling the dynamics of disease spreading and their changes due to environmental factors and governmental and public health interventions.

Chapter 5

Inferring the dynamics of quasi-reaction systems via non-linear local mean-field approximations

[1] Submitted to **Computational Statistics & Data Analysis**, 2024 October 17. doi: 10.48550/arXiv.2410.13744

[2] Oral presentation at **SIS 2024- Statistics and Data Science Conference**, April 11-12, 2024, University of Palermo (IT).

5.1 Introduction

Reaction networks are an efficient 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, 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 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 moment of the process, but this requires considerable computational effort especially for large population sizes. The Bayesian inference approach in Boys et al. (2008) works well in data-poor scenarios, but is computationally inefficient. 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 process's nonlinearity. However, this approach is limited to unitary systems, where each reaction involves the transformation of a single element into one or more products. Such scenarios are 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). Xu et al. (2019) proposed a method-of-moments algorithm that matches second order moments. However, due to the statistical instability of the second moment, this approach fails in highly stochastic or nonlinear systems.

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 this paper, we propose an efficient method that extends the mean-field approach using a Taylor expansion not in time, as the above methods proposed, but in concentration. Starting from the chemical master equation and using the system of non-linear equations describing the dynamics of the process mean, we obtain a linear approximation of the rate function. This leads to an approximation of the system of ordinary differential equations (ODEs) with an explicit solution. By combining our method with a nonlinear least-squares method, it is possible to perform inference of the parameters governing the rate equations.

The paper is structured as follows. In section 5.2, we formalize the statistical modelling of quasi-reaction systems and introduce the generic mean-field approach. We define the proposed local mean-field approximation method and illustrate it in an example. We then study its resistance to stiffness, by comparing the performance of our method against several numerical approaches. The nonlinear least-squares procedure for parameter estimation is described in section 5.3, both from a methodological and computational point of view. Section 5.4 is reserved for simulation studies. By comparing the performance of the proposed algorithm with both the existing LLA method and other state-of-the-art approaches, we show its improved performance, particularly as the time interval between consecutive observations increases. In section 5.5, we illustrate the method in the analysis of a cell differentiation study by Wu et al. (2014). Via BIC model selection, we study the pattern of cell differentiation of 5 major blood cells, and estimate the parameters regulating the underlying reaction process.

5.2 Local Mean-field Approximation (LMA)

In this section, we describe the proposed method for estimation of parameters in a quasi-reaction system. To this end, in section 5.2.1, we first describe concisely a general quasi-reaction system, with the aim of deriving the general ODE formulation of the conditional mean of this process. In section 5.2.2, we show 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, we derive in section 5.2.3 a generic approximation for any quasi-reaction system.

5.2.1 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

$$k_{1j}Q_1 + ... + k_{pj}Q_p \xrightarrow{\theta_j} s_{1j}Q_1 + ... + s_{pj}Q_p \qquad j \in 1, \dots, r. \tag{5.1}$$

The occurrence of the j-th reaction leads to a change of $v_{lj} = s_{lj} - k_{lj}$ substrates for particle type l. Let V denote the net effect matrix having v_{lj} as lj-th element, and $K = \{k_{lj}\}$ the reactant matrix. Let $\boldsymbol{Y}(t) =$

$(Y_1(t), \ldots, Y_p(t))^T \in \mathbb{N}_0^p$ denote the continuous time counting process, with $Y_l(t)$ the number of l-th particles present in the system at time $t \in [0, T]$. Under standard assumptions, the conditional rate for the j-th reaction in such system is given as

$$\lambda_j(\boldsymbol{Y}(t); \boldsymbol{\theta}) = \theta_j \prod_{l=1}^{p} \binom{Y_l(t)}{k_{lj}}, \tag{5.2}$$

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 process dynamics. Let Θ 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 $\boldsymbol{\lambda}(\boldsymbol{Y}(t); \boldsymbol{\theta}) = \Theta \boldsymbol{\kappa}(t)$.

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 stochastic kinetic process, these are commonly referred to as the chemical master equation, which is given by

$$\frac{dp_t(\boldsymbol{y})}{dt} = \sum_{j=1}^{r} p_t(\boldsymbol{y} - \boldsymbol{v}_{\cdot j}) \lambda_j(\boldsymbol{y} - \boldsymbol{v}_{\cdot j}; \theta_j) - p_t(\boldsymbol{y}) \lambda_j(\boldsymbol{y}; \theta_j), \quad \forall\, \boldsymbol{y} \in \mathbb{N}_0^p. \tag{5.3}$$

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 (5.3).

In particular, one can obtain from the chemical master equation a set of ODEs describing the temporal evolution of lineage population concentration averages. To this end, let $\boldsymbol{m}(t + s|t)$ describe the evolution of $\mathbb{E}[\boldsymbol{Y}(t + s)|\boldsymbol{Y}(t) = \boldsymbol{y}(t)] = \sum_{\boldsymbol{y}} \boldsymbol{y} p_{t+s|t}(\boldsymbol{y})$. By taking the derivative, the following ODEs system describing the dynamics of the conditional mean of $\boldsymbol{Y}(t +$

$s)|\boldsymbol{Y}(t)$ can be obtained (Pellin et al., 2023):

$$\frac{dm_l(t+s|t)}{ds} = \sum_{j=1}^{r} v_{lj}\mathbb{E}[\lambda_j(\boldsymbol{Y}(t+s);\boldsymbol{\theta}) \mid \boldsymbol{Y}(t)], \quad l = 1, \ldots, p. \tag{5.4}$$

There have been several proposals for solving (5.4) 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 behaviour. In the next section, we show 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.

5.2.2 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 (5.2) is linear in $\boldsymbol{y}$, so

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

By adding the initial condition $\boldsymbol{m}(t|t) = \boldsymbol{y}(t)$, the system (5.4) is expressed by the following first order Cauchy differential equation,

$$\begin{cases} \dfrac{d\boldsymbol{m}(t+s|t)}{ds} = P_{\boldsymbol{\theta}}\boldsymbol{m}(t+s|t) + \boldsymbol{b}_{\boldsymbol{\theta}} \\ \boldsymbol{m}(t|t) = \boldsymbol{y}(t). \end{cases} \tag{5.5}$$

The coefficient matrix $P_{\boldsymbol{\theta}}$ and the inhomogeneous term $\boldsymbol{b}_{\boldsymbol{\theta}}$ 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. By employing the hazard function (5.2) and applying straightforward algebraic manipulations, they can be explicitly expressed in terms of the death and net effect matrices as $P_{\boldsymbol{\theta}} = V\Theta K^T$ and $b_{\boldsymbol{\theta},l} = \sum_j V_{lj}\theta_j \mathbb{1}_{\{K_{.j}=0\}}$. If the $P_{\boldsymbol{\theta}}$ matrix is invertible, the system (5.5) has the explicit solution

$$\boldsymbol{m}(t+s|t) = \exp\big(sP_{\boldsymbol{\theta}}\big)\boldsymbol{y}(t) + {P_{\boldsymbol{\theta}}}^{-1}\bigg(\exp\big(sP_{\boldsymbol{\theta}}\big) - \mathbb{I}_p\bigg)\boldsymbol{b}_{\boldsymbol{\theta}}. \tag{5.6}$$

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

5.2.3 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 (5.4) typically impossible. Our idea is to linearise 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 (5.6). Omitting the dependence of the hazard function on the parameters $\boldsymbol{\theta}$ for the sake of readability, we perform a first order Taylor expansion of $\boldsymbol{\lambda}(\boldsymbol{Y}(t+s))$ around $\boldsymbol{Y}(t)$,

$$\boldsymbol{\lambda}(\boldsymbol{Y}(t+s)) = \boldsymbol{\lambda}(\boldsymbol{Y}(t)) + \Lambda(\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)) + \boldsymbol{\eta}(t). \tag{5.7}$$

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

$$\eta_j = \frac{1}{2}\frac{\partial^2\lambda_j(\tilde{\boldsymbol{Y}})}{\partial Y_l(t)\partial Y_k(t)}(Y_l(t+s)-Y_l(t))(Y_k(t+s)-Y_k(t)), \quad l,k = 1,\dots,p \tag{5.8}$$

with $\tilde{\boldsymbol{Y}} \in (\boldsymbol{Y}(t), \boldsymbol{Y}(t+s))$. The Jacobian matrix $\Lambda = \dfrac{\partial\boldsymbol{\lambda}}{\partial\boldsymbol{y}}$ is explicitly defined by the proposition below, whose proof can be found in A.1.

Proposition 3. *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*

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

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 (5.7) in (5.4), we obtain the conditional mean ODEs system:

$$\begin{aligned}
\frac{d}{ds}\boldsymbol{m}(t+s|t) &= V\mathbb{E}\Big[\boldsymbol{\lambda}(\boldsymbol{Y}(t)) + \Lambda\big(\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)\big) \mid \boldsymbol{Y}(t) = \boldsymbol{y}(t)\Big] \\
&= V\Lambda\mathbb{E}[\boldsymbol{\lambda}(\boldsymbol{Y}(t+s))|\boldsymbol{Y}(t) = \boldsymbol{y}(t)] + V\boldsymbol{\lambda}(\boldsymbol{y}(t)) - V\Lambda\boldsymbol{y}(t) \\
&= \underbrace{V\,\Theta\,H}_{P_{\boldsymbol{\theta}}}\boldsymbol{m}(t+s|t) + \underbrace{V\,\Theta\,(\boldsymbol{\kappa}(t) - H\,\boldsymbol{y}(t))}_{\boldsymbol{b}_{\boldsymbol{\theta}}} \\
&= P_{\boldsymbol{\theta}}\boldsymbol{m}(t+s|t) + \boldsymbol{b}_{\boldsymbol{\theta}}. \qquad (5.9)
\end{aligned}$$

If $|P_{\boldsymbol{\theta}}| \neq 0$, the explicit solution of the ODEs system above is defined as in (5.6).

In terms of convergence to the solution of the original ODEs system (5.4), the approximation does not depend on the time step but on the difference between concentrations as indicated by the relation (5.8). Since we stop the expansion at first order, and if we consider at most second-order reactions, then the approximation error is determined by the norm of the Hessian matrix of the hazard function.

5.2.4 Example: cyclic chemical reaction network

In this section, we consider 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 (5.2), while the second, together with the hazard function, define the dynamics of the first moments of the concentrations in (5.4).

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 (5.4). The local mean-field approximation method described in section 5.2.3 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

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

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

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

We now plug in these quantities in (5.7), leading to the ODEs system approximation (5.9) with

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

By substituting the quantities above into equation (5.6), we obtain 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 $\boldsymbol{y} \neq (0.5, 0, 0.5)$.

5.2.5 Computational time and stiffness

The main advantage of the proposed method is that the ODEs system in (5.9), used to approximate the orginal system in (5.4), has an explicit solution. As we will show in this section, 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 Δt 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 (5.9) analytically, it is necessary to compute the exponential of a $p \times p$ matrix. For this, we utilize the Pade approximation with scaling and squaring method due to its efficiency with dense matrices. This approach has a computational complexity of $\mathcal{O}(p^3)$. Additionally, we need to compute the inversion of the same matrix, which incurs a similar computational cost. Consequently, the overall cost for calculating the analytical solution in (5.9) 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

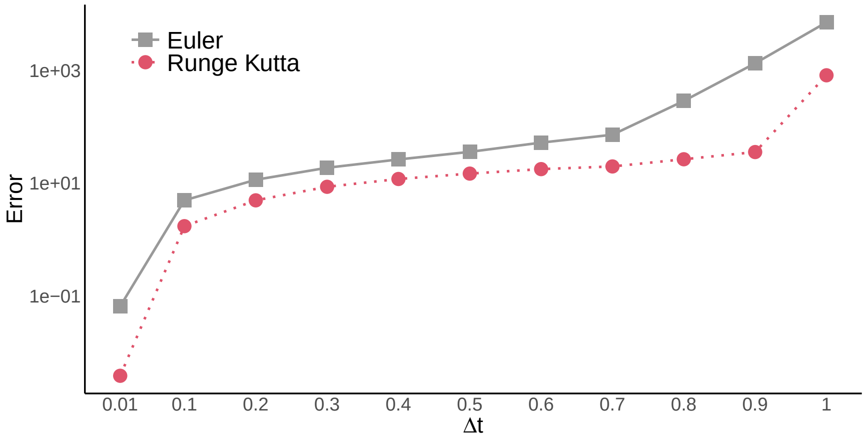


FIG. 5.1 *Mean absolute error between the numerical solution from each method and the analytical LMA solution in (5.6), as Δt increases: both Euler and Runge-Kutta methods experience a significant degradation in accuracy as Δt increases.*

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, 1966). However, as its associated net effect matrix is not full rank, $P_{\boldsymbol{\theta}} = V\Lambda$ is not invertible. Instead, we consider the cyclic chemical reaction system described in section 5.2.4, 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, we consider the ODEs system (5.9) associated to this network. The matrix associated to this system is invertible. Moreover, the eigenvalues of $P_{\boldsymbol{\theta}}$, given by $7.6 \cdot 10^{-7}$, $-4.2 \cdot 10^{-5}$, and $-3.8 \cdot 10^{0}$, are different from each other in 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.

Figure 5.1 shows the robustness of the Euler and Runge-Kutta methods when used to solve numerically the ODEs system (5.9) of this problem. The error is defined as the average of the absolute difference of the solution ob-

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

Table 5.1: Comparison of computational costs, error convergence, and performance in stiff problems for the explicit Euler, fourth-order Runge-Kutta, and the proposed LMA method.

tained by the proposed method and the analytical LMA solution in (5.6), 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 Δt increases. Table 5.1 summarises 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.

5.3 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.

5.3.1 Estimation of reaction rates

We consider a set of n replicates, 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 realisations 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. We define $\boldsymbol{m}(\boldsymbol{\theta}) = [\boldsymbol{m}_1(\boldsymbol{\theta}), \dots, \boldsymbol{m}_n(\boldsymbol{\theta})]$ 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 (5.9) 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,

$$\boldsymbol{Y}_{ci} = \boldsymbol{m}_{ci} + \boldsymbol{\varepsilon}_{ci}, \tag{5.10}$$

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

$$\hat{\boldsymbol{\theta}}_{LMA} = \arg\min_{\boldsymbol{\theta}\geq 0} \left\{ f(\boldsymbol{\theta}) = \sum_{c=1}^{n} \sum_{i=1}^{T_c} \left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}(\boldsymbol{\theta})\right]^T \left[\boldsymbol{Y}_{ci} - \boldsymbol{m}_{ci}(\boldsymbol{\theta})\right] \right\}. \tag{5.11}$$

To determine the optimum, we use an iterative approach which consists of two steps. First, we solve analytically the approximation of the ODE solution (5.9). Then, we apply the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box-constraints 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 A.2. The iterative procedure requires initial estimates $\hat{\boldsymbol{\theta}}_0$. Considering the potentially large number of parameters in the model, it is important to start the minimization of (5.11) with accurate initial values. A practical starting value is provided by the local linear approximation approach, which will also be used in the comparative study. A detailed description of the method can be found in A.3.

5.3.2 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). In this section, we do so 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

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial}{\partial \theta} f(\hat{\boldsymbol{\theta}})\right)\left(\frac{\partial}{\partial \theta} f(\hat{\boldsymbol{\theta}})\right)^T. \tag{5.12}$$

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 (5.9), i.e., the derivatives of

$$\boldsymbol{m}(t+s|t) = \exp\big(sP_{\boldsymbol{\theta}}\big)\boldsymbol{y}(t) + P_{\boldsymbol{\theta}}{}^{-1}\Big(\exp\big(sP_{\boldsymbol{\theta}}\big) - \mathbb{I}_p\Big)\boldsymbol{b}_{\boldsymbol{\theta}} \tag{5.13}$$

with respect to $\boldsymbol{\theta}$. For ease of notation, we consider a single-replicate scenario. We start by taking the derivatives of $P_{\boldsymbol{\theta}}$ and $\boldsymbol{b}_{\boldsymbol{\theta}}$ defined in (5.9). 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, we also obtain the derivative of the second term in (5.13),

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

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

$$\begin{aligned}\frac{\partial \boldsymbol{m}(t+s \mid t)}{\partial \theta_j} = &\left(s\int_0^1 \exp((1-u)sP_{\boldsymbol{\theta}})V\boldsymbol{e}_j H \exp(usP_{\boldsymbol{\theta}})\,du\right)\boldsymbol{y}(t)\\ &- P_{\boldsymbol{\theta}}^{-1}Ve_jHP_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}})-\mathbb{I}_p\right)b_{\boldsymbol{\theta}}\\ &+ P_{\boldsymbol{\theta}}^{-1}s\left(s\int_0^1 \exp((1-u)sP_{\boldsymbol{\theta}})V\boldsymbol{e}_j H \exp(usP_{\boldsymbol{\theta}})\,du\right)\boldsymbol{b}_{\boldsymbol{\theta}}\\ &+ P_{\boldsymbol{\theta}}^{-1}\left(\exp(sP_{\boldsymbol{\theta}})-\mathbb{I}_p\right)\left(Ve_j\boldsymbol{\kappa}(t) - Ve_jH\boldsymbol{y}(t)\right). \qquad (5.14)\end{aligned}$$

Taking into account that the minimization problem relies on T_c observations for each of n clone-type scenario, we define the $p \times r$-dimensional matrix $\boldsymbol{\xi}_{ci}$ such that the j-th column is the evaluation of (5.14) at $(t_{ci}, \boldsymbol{Y}_{ci})$. The gradient of the objective function with respect to $\boldsymbol{\theta}_j$ is the r-dimensional vector

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

Using this expression, the variances of the estimated rates $\hat{\boldsymbol{\theta}}$, which we approximate with the diagonal of $\mathcal{I}(\hat{\boldsymbol{\theta}})$ in (5.12), 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 derivations are checked numerically in A.4.

5.4 Simulation study

In this section, we present several simulation studies to assess the performance of the proposed method. The experimental setup, based on the system described in section 5.2.4, reflects conditions typical of numerous biological applications. In section 5.4.1, we evaluate the proposed method by varying the width between the observations Δt and the number of time points T. We compare the results with an alternative local linear approximation method. For short time steps, we expect the local linearization to be a serious competitor, whereas for large time steps the nonlinearity of the system will make our inferential scheme preferable. In section 5.4.2, we study how

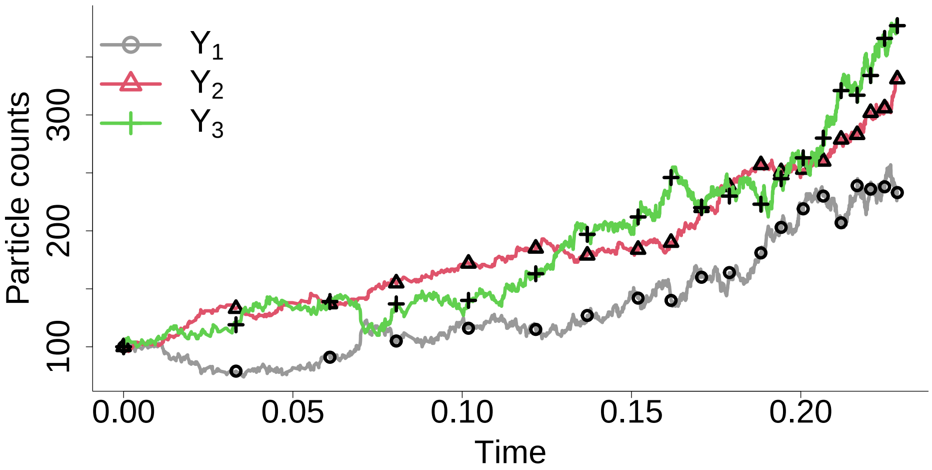


FIG. 5.2 *A trajectory generated using the Gillespie algorithm for the model described in section 5.2.4. The symbols indicate the $T = 20$ observations between every 100 simulated states (average $\Delta t \approx 0.012$).*

accuracy and computational time vary by increasing the number of nodes and reactions. Finally, in section 5.4.3 we compare our approach with a method-of-moments formulation introduced by Xu et al. (2019), which relies on matching model-derived and empirical correlations in cell type dynamics.

5.4.1 Performance as function of Δt and T

In this section, we compare the performance of the LMA inferential procedure across different time steps Δt and number of time points T. We use the Gillespie algorithm to simulate trajectories from the cyclic reaction system defined in section 5.2.4 with rates $\boldsymbol{\theta} = (0.2, 0.1, 0.2)$ and initial concentration $\mathbf{y}_0 = (100, 100, 100)$. Figure 5.2 shows an example of trajectories for one of the simulations.

In order to evaluate the dependence of the results on Δt, we extract measurements from the simulated trajectories. In order to get, on average, increasing Δt values, we consider 5 different cases where we retain every 10, 30 50, 70, 100 values, respectively, from each trajectory for a total of $T = 20$ measurements. See Figure 5.2 for an example of the last case. This leads to 5 different values for the average time steps $(0.003, 0.007, 0.009, .010, 0.012)$.

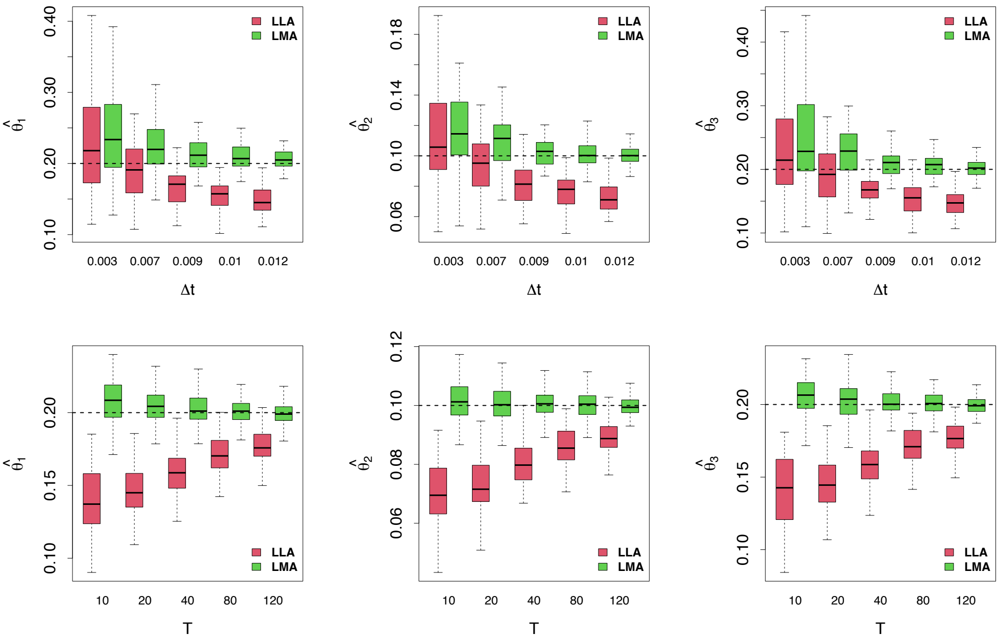


FIG. 5.3 *Distributions of the estimated rates from the local linear approximation (LLA, red) and local mean-field approximation (LMA, green) methods across* 100 *simulations. Top: Keeping* $T = 20$ *fixed, as the time step* Δt *increases, the LLA estimates, unlike LMA, show increased bias. Bottom: Keeping* $\Delta t = 0.012$ *fixed, increasing the number of time points* T *shrinks the standard error of the LMA method in a square-root fashion.*

In a second simulation, we fix the case $\Delta t = 0.012$ and consider 5 different values for the overall number of time points T. The simulations are repeated 100 times.

Figure 5.3 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 Δt 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.

5.4.2 Accuracy and computational time varying r and p

In this section, we examine the impact of varying the number of reactions p and the number of particle types p on the estimation accuracy and computational time required to estimate the parameters $\hat{\boldsymbol{\theta}}$. This analysis is conducted using the same experimental setup of section 5.4.1, 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, since the parameter values vary significantly in magnitude, we consider the Wasserstein 1-distance (Duy and Takeuchi, 2023) between the distribution of an estimated parameter and the corresponding true parameter. In particular, this is given by

$$W_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{r} \left| F_{\theta_j} - F_{\hat{\theta}_j} \right|,$$

where $F_{\hat{\theta}_j}$ is the empirical cumulative distribution of $\hat{\theta}_j$ from the 200 simulations, while F_{θ_j} is the degenerate distribution on the true value θ_j. The median and interquantiles of this measure are calculated by considering 1000 bootstrap versions of the 200 datasets and calculating the Wasserstein 1-distance from each of these.

In a first simulation, we vary the number of particle types $p \in \{3, 6, 9\}$. For $p = 3$, we consider the cyclic reaction system in section 5.2.4. In order to still have a $P_{\boldsymbol{\theta}}$ invertible, we generate the settings with a higher p by simply repeating the same system a number of times, as shown in Figure 5.9 of A.5. The results of the simulations are shown in Figure 5.4. 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, we fix the number of particles at $p = 3$ while progressively varying the number of reactions $r \in \{3, 6, 9, 12, 15\}$. We generate the different systems with varying r by following the scheme in Figure 5.10

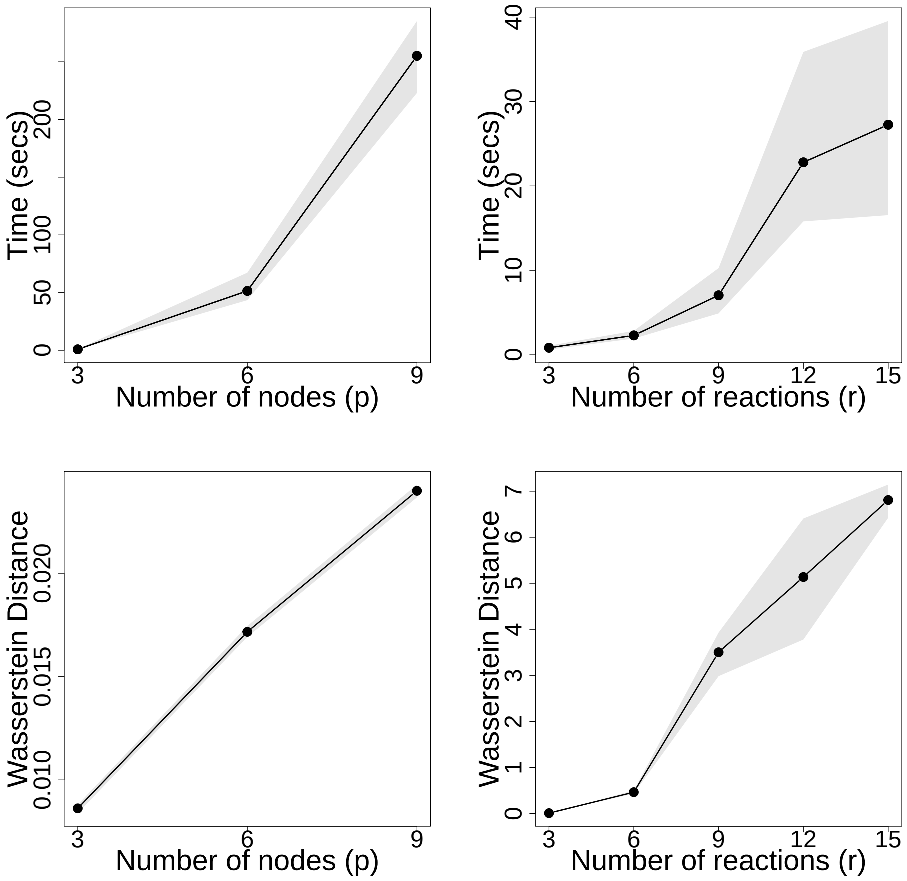


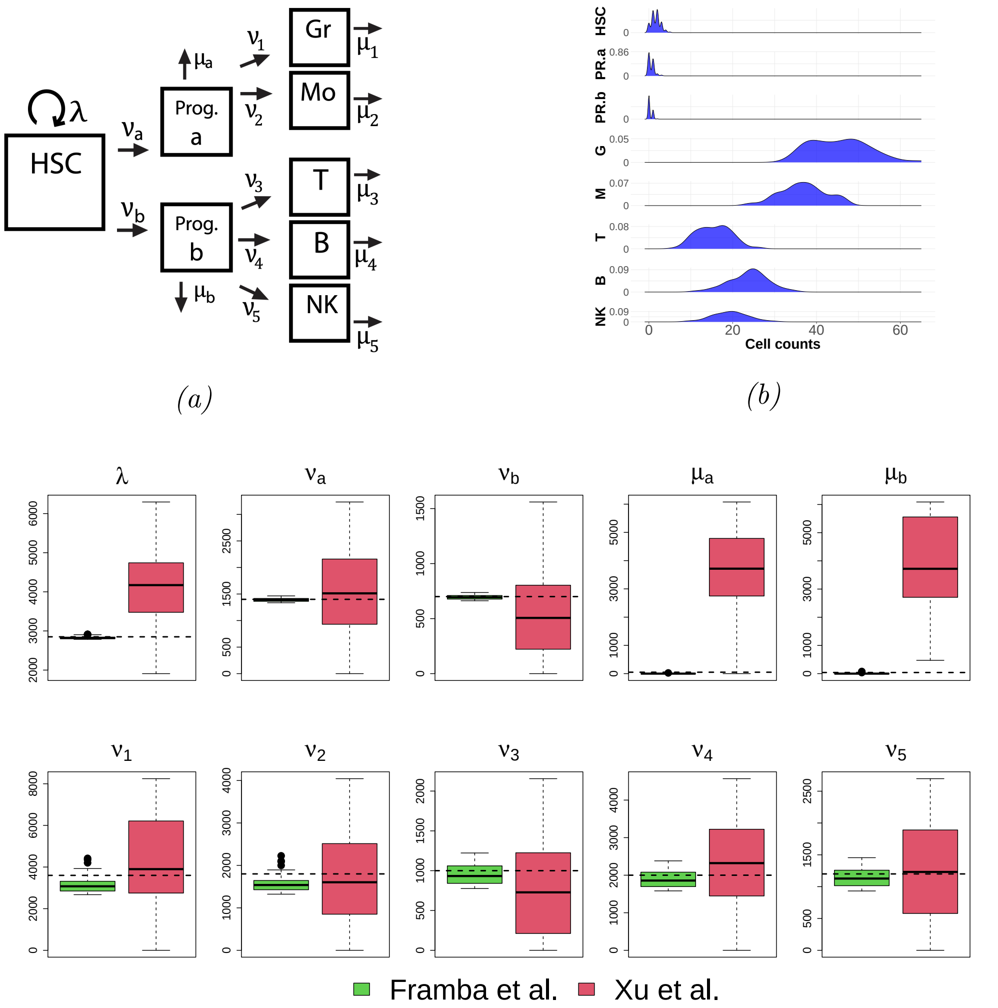
FIG. 5.4 *Median computational time (top) and Wasserstein 1-distance (bottom) of the LMA algorithm, as a function of the number of reactions (r, right) and states (p, left). The shaded area represents the interquartile range across 200 simulations.*

of A.5. The top-right panel of Figure 5.4 shows a quasi-cubic dependence of computational time on r, which aligns with the theoretical predictions discussed in section 5.2.5. The bottom right plot shows how the overall accuracy of the estimates improves approximately linearly with the increase in the number of reactions.

5.4.3 Comparison with M-estimator by Xu et al. (2019)

Finally, we present a simulation study comparing our proposed method with an alternative method proposed by Xu et al. (2019). The authors considered the empirical correlation between the dynamics of the particle types with the theoretical one defined by the chemical master equation. They applied their method to a human cell differentiation system involving hematopoietic stem cells (HSCs), two progenitors, and five mature cell types. The system is described in Figure 5.5. To obtain an analytical solution for the evolution of the moments, the authors assumed 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, we have kept the same assumption. Note that no changes to the method are necessary as it is a generalization of the case of linear hazard functions.

As HSCs and progenitor cells are latent in the reference method, we only identify the dataset with mature cells after generating the complete data, in order to maintain the same original setting. In terms of reaction rates, we multiply the original values defined in Xu et al. (2019) by 1000. This merely constitutes a time unit change. We set the HSC duplication parameter 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 $\mu_b = 40$. 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 $t = (0.00, 0.08, 0.11, 0.14, 0.17)$. We apply our proposed LMA approach and the method of Xu et al. (2019) to these simulated data. As in Xu et al. (2019), 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 Xu et al. (2019) algorithm relies on an initial value,



(a)

(b)

(c)

FIG. 5.5 *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 100 simulations. The 10 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).*

we perform a sensitivity checking by restarting the inference procedure 100 times for each simulation. The solution corresponding to the minimum cost function among the others is then selected as the final value.

The results of the simulation study, the steady-state distribution process and the parameter distributions with 300 simulations are shown in Figure 5.5.

The proposed LMA approach outperforms Xu et al. (2019) in several aspects. Firstly, most of the LMA parameter estimates, with the possible exception of $\hat{\nu}_1$ and $\hat{\nu}_2$, are unbiased, in contrast to the estimates obtained with the Xu et al. (2019) method. 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 method in Xu et al. (2019) is based on matching second order moments, which are inherently less stable than first order methods.

5.5 Cell lineage barcoding in rhesus macaques

Clonal tracing modeling is commonly used in genetic studies to enhance our 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 analysed cell production in non-human primates, which closely mirrors human physiology due to the similar lifespan and frequencies of hematopoietic stem progenitor cells (HSPCs) (Kim et al., 2000). In this article, we considered an in-vivo clonal tracking dataset on Rhesus Macaques (Wu et al., 2014) and aim to recover 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 utilising a high-diversity lentiviral barcode library. Then, such cells were reinfused into the three myeloableted animals (Shepherd et al., 2007). Once reinfused, the genetically modified HSPCs home to the bone marrow, where they engraft and begin to repopulate the

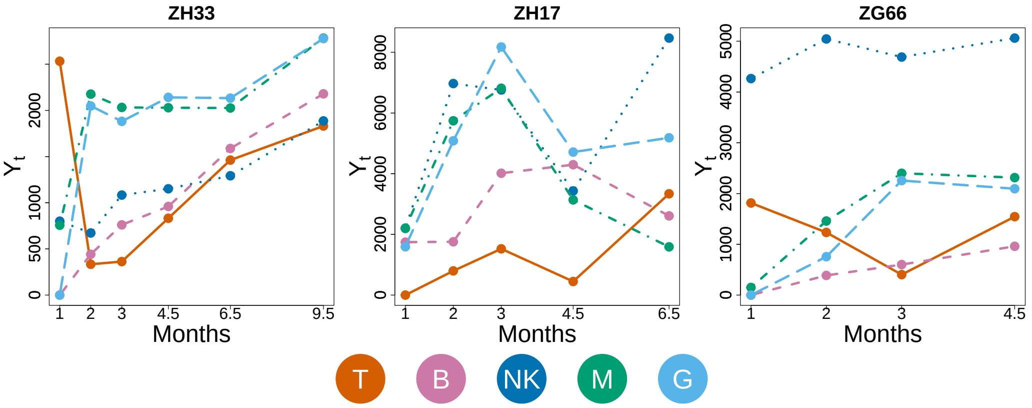


FIG. 5.6 *Average concentration over time of each Rhesus Macaque specimen following transplantation. The $p = 5$ cell types are reported with different colours and line styles.*

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, such Granulocytes (G), Monocytes (M), T, B, 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, we exclude the time points when no barcodes were detected 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 Figure 5.6 shows the average differentiation trajectories for each specimen. We posit that the dynamics of cell differentiation are universal and thus independent of individual subjects. Consequently, we assume that the differentiation process in each subject is governed by the same rate parameters. Following the same approach as in Pellin et al. (2023), we define the death reaction hazard rates as a quadratic function of the underlying abundances, representing a natural saturation effect.

Using the proposed LMA approach, we compare competing models of hematopoiesis from the available experimental data. To this end, we define the net matrix of the full model consisting of 30 reactions, of which 10 are birth and death reactions and 20 are all possible differentiations from one cell to another. We initialize the reaction rates 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), we keep such rates variable. We then search through the space of possible models, 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, we estimate the parameters using the LMA approach, i.e., by solving the optimisation problem (5.11). At each step of a stepwise procedure, we iteratively add and subtract the reaction that most reduces the Bayesian Information Criterion (BIC). This methodology leads to a sequence of models with increased complexity. The results are shown in Figure 5.7 (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 5.2 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 5.3.2. Due to the quadratic nature of the death rates and to particle counts in the order of 10^3 (Figure 5.6), we can see how reactions involving M and B tend to be the slowest (with rates in the order of 10 days), reactions involving G and T occurr 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.

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

Table 5.2: Estimated rates, and corresponding standard errors, from Rhesus Macaques gene therapy trial data, based on the optimal model of cell differentiation.

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, we determine the evidence that a specific reaction is included 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 θ_j, and w_j is the rescaled BIC-based weight. This is defined by

$$w_j = \frac{\exp\{-\frac{1}{2}(BIC_j - BIC_{\min})\}}{\sum_{h=1}^{r} \exp\{-\frac{1}{2}(BIC_h - BIC_{\min})\}}, \tag{5.15}$$

where $BIC_{\min}$ is the minimum BIC value among all the models considered. The cell differentiation network in Figure 5.7 (left) has edges whose thickness is proportional to the p_j value. From this we can see, how, although each element appears both as a reactant and a product in the optimal model (Table 5.2), NK cells do not tend to differentiate into other cell types. Moreover, the figure shows how 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.

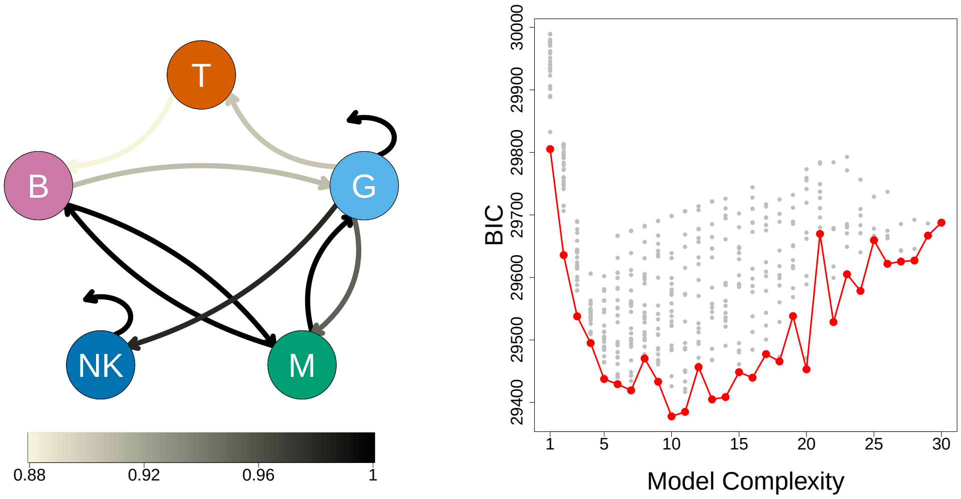


FIG. 5.7 *BIC model selection on Rhesus Macaques data. Right: Reaction systems ordered by complexity, from the simplest (single-reaction model) to the most complex (including all possible birth, death, and single-reactant differentiation events). For each complexity level, the BIC of the best model is 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* (5.15)*.*

5.6 Conclusion

In this study, we have developed and assessed a novel methodology for parameter inference in quasi-reaction systems, with a 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 is sparse or irregularly spaced. Our 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 other methods. The results demonstrate that our method significantly outperforms local linear approximation, particularly as the time interval between observations increases. Thanks to the availability of an explicit solution, the proposed method is shown to be robust to stiffness, a common occurrence in biological systems where processes operate at vastly different time scales. In addition, we illustrate the approach for the study of cell differentiation from gene therapy clonal tracking data. The approach returns an estimate of the cell populations dynamics and provides meaningful insights into the underlying biological processes.

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

APPENDIX

A.1 Jacobian of the hazard function

Proposition 4. *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*

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

Proof. First, we define the jl-th element of the Jacobian matrix $\Lambda(\boldsymbol{Y}(t);\boldsymbol{\theta})$ as

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

By applying the chain rule, we get

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

with δ_{il} an indicator function. Next, recalling the digamma function $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$, we differentiate the logarithm of the binomial coefficient:

$$\begin{aligned}\frac{\partial}{\partial Y_l}\log\binom{Y_l(t)}{k_{lj}} &= \frac{\partial}{\partial Y_l}\left[\log\Gamma(Y_l(t)+1) - \log\Gamma(k_{lj}+1) - \log\Gamma(Y_l(t)-k_{lj}+1)\right]\\ &= \psi(Y_l(t)+1) - \psi(Y_l(t)-k_{lj}+1).\end{aligned}$$

Applying logarithmic differentiation, we have:

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

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

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

□

A.2 LMA algorithm

Algorithm 4 gives the pseudo-code of the proposed nonlinear local mean-field (LMA) algorithm for parameter estimation in a quasi-reaction model.

Algorithm 4 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 (5.6)
    $\boldsymbol{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$
    $\boldsymbol{g}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k)$
    $H_{k+1} = \left(\mathbb{I}_r - \frac{\boldsymbol{s}_k \boldsymbol{g}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}\right) H_k \left(\mathbb{I}_r - \frac{\boldsymbol{g}_k \boldsymbol{s}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}\right) + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\boldsymbol{g}_k^T \boldsymbol{s}_k}$
    $k = k + 1$
end while
$\hat{\boldsymbol{\theta}}_{LMA} = \boldsymbol{\theta}_k$
```

A.3 Local Linear Approximation

In this section, we describe the Local Linear Approximation (LLA) approach for estimation of the rates $\boldsymbol{\theta}$. The LLA estimates are used in the comparative study in section 5.4.1 and as initial values for Algorithm 4 that iteratively solves the optimization problem (5.11).

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 (5.3), and assuming constant hazard rates between consecutive time points, the conditional moments are approximated as follows:

$$\mathbb{E}[\boldsymbol{Y}_{ci} \mid \boldsymbol{Y}_{c,i-1}] \simeq \boldsymbol{Y}_{c,i-1} + \sum_{j=1}^{r} v_{lj} \lambda_j(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta}) \Delta t_i \tag{5.16}$$

$$\begin{aligned}\mathbb{E}[Y_{c,i,l} Y_{c,i,k} \mid \boldsymbol{Y}_{c,i-1}] \simeq Y_{c,i-1,l} Y_{c,i-1,k} &+ \sum_{j=1}^{r} v_{lj} Y_{c,i-1,k} \lambda_j(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta}) \Delta t_i \\ &+ \sum_{j=1}^{r} v_{kj} Y_{c,i-1,l} \lambda_j(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta}) \Delta t_i + \sum_{j=1}^{r} v_{lj} v_{kj} \lambda_j(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta}) \Delta t_i.\end{aligned} \tag{5.17}$$

Since the hazard function $\lambda_j(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, the regression model (5.10) can be rewritten as:

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

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 $M_{ci}\boldsymbol{\theta} = V \operatorname{diag}(\boldsymbol{\lambda}_{ci}(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta})) \Delta t_i$ and $\Omega_{ci} = V \operatorname{diag}(\boldsymbol{\lambda}_{ci}(\boldsymbol{Y}_{c,i-1}; \boldsymbol{\theta})) V^T \Delta t_i$ represent the discretized drift function and the dispersion matrix of the concentration differentiation process, respectively.

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

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

A.4 Standard error approximation

In this section, we aim to validate the derivation of the standard error described in section 5.3.2. As explained in that section, under a maximum-likelihood estimation framework, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is asymptotically approximated by the inverse of the observed Fisher information matrix, for which we derive an explicit approximate formulation. We validate this derivation using the same model of section 5.2.4. In particular, we gen-

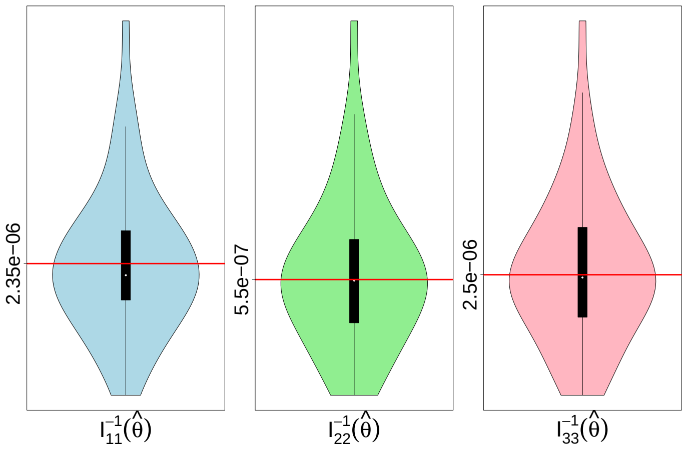


FIG. 5.8 *Violin plots showing the distribution across 50 simulations of the diagonal elements of the inverse of the observed Fisher information matrix for the 3 reaction rates of the system in section 5.2.4. The horizontal line represents the empirical variance of the estimates. The figure shows a good agreement between the theoretically and empirically derived standard errors of the estimated parameters.*

erate data from this model with 100 replicates for each of the 50 simulated datasets. Figure 5.8 presents violin plots corresponding to the diagonal elements of the inverse of the observed Fisher information matrix elements. 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.

A.5 Reaction systems used in the simulation study in section 5.4.2

This paragraph outlines the reaction systems used in section 5.4.2 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, we denote these with a subscript.

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

$$\mathcal{R}_3 : \begin{cases} 2\mathrm{A}_1 & \xrightarrow{0.2} 2\mathrm{B}_1 \\ \mathrm{A}_1 + \mathrm{B}_1 & \xrightarrow{0.1} 3\mathrm{C}_1 \\ 2\mathrm{C}_1 & \xrightarrow{0.2} 2\mathrm{A}_1 \end{cases} \qquad \mathcal{R}_6 : \mathcal{R}_3 \cup \begin{cases} 2\mathrm{A}_2 & \xrightarrow{0.2} 2\mathrm{B}_2 \\ \mathrm{A}_2 + \mathrm{B}_2 & \xrightarrow{0.1} 3\mathrm{C}_2 \\ 2\mathrm{C}_2 & \xrightarrow{0.2} 2\mathrm{A}_2 \end{cases}$$

$$p = 9$$

$$\mathcal{R}_9 : \mathcal{R}_6 \cup \begin{cases} 2\mathrm{A}_3 & \xrightarrow{0.2} 2\mathrm{B}_3 \\ \mathrm{A}_3 + \mathrm{B}_3 & \xrightarrow{0.1} 3\mathrm{C}_3 \\ 2\mathrm{C}_3 & \xrightarrow{0.2} 2\mathrm{A}_3 \end{cases}$$

FIG. 5.9 *Dynamic systems with an increasing number of particles ($p = 3, 6, 9$). Rates are reported above each reaction.*

$$\begin{array}{rrrrrr}
R_1 : & 2\mathrm{A} \xrightarrow{0.2} 2\mathrm{B} & R_4 : & 2\mathrm{A} \xrightarrow{0.01} \emptyset & R_7 : & 2\mathrm{A} \xrightarrow{0.1} 3\mathrm{B} \\
R_2 : & \mathrm{A} + \mathrm{B} \xrightarrow{0.1} 3\mathrm{C} & R_5 : & \mathrm{C} \xrightarrow{0.02} \emptyset & R_8 : & \mathrm{C} \xrightarrow{0.06} \mathrm{B} \\
R_3 : & 2\mathrm{C} \xrightarrow{0.2} 2\mathrm{A} & R_6 : & 2\mathrm{C} \xrightarrow{0.03} \emptyset & R_9 : & \mathrm{C} \xrightarrow{0.05} 2\mathrm{A}
\end{array}$$

$$\begin{array}{rrrr}
R_{10} : & \mathrm{A} \xrightarrow{0.1} \mathrm{C} & R_{13} : & \emptyset \xrightarrow{50} \mathrm{A} \\
R_{11} : & \mathrm{B} + \mathrm{C} \xrightarrow{0.09} \mathrm{A} & R_{14} : & \emptyset \xrightarrow{50} \mathrm{B} \\
R_{12} : & \mathrm{B} \xrightarrow{0.08} 2\mathrm{A} + \mathrm{C} & R_{15} : & \emptyset \xrightarrow{50} \mathrm{C}
\end{array}$$

FIG. 5.10 *Five 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 15 reactions. Reaction rates are reported above each reaction.*

Chapter 6

Conclusion

Quasi-reaction systems are widely used in many natural science studies due to their simplicity in modelling complex reaction networks and their broad applicability across multiple disciplines. The observation times of various natural phenomena modelled by these frameworks can vary considerably, with intervals ranging from frequent measurements, such as in daily COVID-19 monitoring, to more extended periods, as in hematopoietic clonal tracking studies. This variability can pose significant challenges to the accurate estimation of the parameters governing the stochastic differential equation that describe the system's dynamics. Our work introduces new inference algorithms that efficiently estimate reaction rates in such scenarios.

In the third chapter, we introduced a novel inferential approach based on event history models. This is particularly suited to systems with closely spaced observations. The method utilizes an EM algorithm that incorporates an extended Kalman filter for estimating latent reactions. Our approach is able to handle strong temporal correlations arising from dense data effectively, by providing accurate parameter estimates even when traditional methods encounter numerical instability. Furthermore, the proposed method offers a versatile framework that can be used in a wide range of applications involving discrete latent states. However, the approach has some limitations. Firstly, the Gamma distribution used to approximate Poisson-distributed events can be inaccurate for small means and may not match the skewness and kurtosis of the count variable. One potential solution is to employ mixtures of Gamma distributions so that higher-order moments can also be matched (Lindsay et al., 2000). Furthermore, the approximation errors introduced by the extended Kalman filter might impact the accuracy of latent event reconstruction, especially when large discrepancies arise between the estimated state trajectory and the nominal one. A more robust strategy involves incorporating additional input terms that account for higher-order effects while simultaneously minimizing the norm of the gain matrix, as suggested by Einicke and White (1999). Additionally, while comparisons with the LLA represent the typical approach in the context of quasi-reaction systems, future research could build on this by evaluating its performance against alternative Bayesian methods for state-space models, such as iterated filtering (Ionides et al., 2015) and stochastic approximation EM algorithms combined with particle methods (Lindsten, 2013), which may provide better uncertainty quantification and easier handling of missing values at the expense.

In the fourth chapter, we augment the latent event history model by integrating time-varying covariates in the modelling of the reaction rates. This modification offers several benefits. First, it increases flexibility in modelling dynamic reaction rates that may change in response to external influences. Second, it achieves this without adding unnecessary complications: unlike approaches that introduce additional compartments for each source of variability, our method efficiently incorporates covariates into the rate dynamics, making it applicable across a range of applications. Nonetheless, some limitations suggest opportunities for further refinement. One notable challenge is the potential misalignment between covariate effects and the observed data, particularly when covariate information is incomplete or imprecise. Addressing this could involve techniques like data imputation (Van Buuren, 2018). Additionally, we defined covariates as time-dependent functions, but they could also depend on spatial factors or be modelled as stochastic processes, as in (Zhang et al., 2022). Moreover, future work could focus on incorporating uncertainty quantification of the estimated basic reproduction number: this could be achieved by propagating the uncertainty from the covariate-dependent parameters, as done in the third chapter, or by employing standard approaches such as the bootstrap method (Tibshirani and Efron, 1993).

In the fifth chapter, we addressed the challenge of parameter estimation in quasi-reaction systems with sparsely observed data. We have developed a local mean-field approximation method that linearizes the nonlinear rate function using a first-order Taylor expansion. This method essentially simplifies the system to a unitary form, allowing for the explicit solution of the mean process of ordinary differential equations. One of the key advantages of this method is the higher accuracy of parameter estimation compared to state-of-the-art techniques. In addition, it has a low computational cost, as it is based on explicit solutions, which avoids the need for iterative numerical solvers. This characteristic also makes the method inherently resistant to stiffness, a common issue in natural processes. However, two main assumptions—the use of a first-order Taylor expansion and constant reaction rates over extended intervals—could be refined by adopting higher-order approximations, as demonstrated in (Lente et al., 2022), and by incorporating time-varying covariates, as introduced in Chapter 4. Additionally, to further improve the estimation of the parameters, the method could incorporate the variance-covariance matrix of the observations into the regression framework, similar to the approach of Pellin et al. (2023).

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Prof. Veronica Vinciotti, whose dedication to precision has inspired me to look beyond appearances and seek the true essence of things. I am especially grateful for her patience with my eagerness and her tolerance of my imperfect English. I would like to thank Prof. Ernst Wit, whose breadth of knowledge and simplicity in transmitting it I greatly admire. My renewed passion for scientific research is much to his credit. I want to thank both of the professors mentioned above for offering me the opportunity to go to both Lugano and Boston, as they were experiences that enriched me professionally and humanely.

I thank Prof. Danilo Pellin for unknowingly accompanying me throughout my PhD journey: from his articles I began my path and with his articles I conclude it. I thank him even more for the time he granted me at Harvard Medical School in his laboratory, and above all for sharing his time, space, and research spirit. All my life I have waited for such an experience, and I can only thank him for giving it to me. I also thank Giacomo and Alfonso for their valuable tips, debates, conversations.

My sincere thanks go to Simone Verzellesi and Giulia Cavicchioni, who were always with me if I needed advice on Analysis or Algebra, and most importantly, were there for me in my many moments of discouragement. I wish you all the best.

I would like to thank Martina Boschi, for being the first peer I could talk to about statistics, and being able to do so without prejudice or fear of confrontation.

I thank all the people in Ernst's lab at the Università Svizzera Italiana in Lugano, with whom I was able to share short but intense periods of time

and better understand what it means to do research in my own subjects.

I would like to thank my whole family from the bottom of my heart, with whom I have tried to share what work I was doing several times, and repeating it allowed me to understand it better myself. I thank you because you love me unconditionally, and we both know it.

You know how you begin a doctorate, not how (or if) you finish it. In these three years I have lived in 10 different houses, gone on 15 work missions, undergone 1 surgical operation and 68 visits to hospital centres. All this will never surpass the emotion of a wedding. I thank you, Cecilia, for being my eternal present.

Bibliography

Al-Mohy, A. H. and N. J. Higham (2010). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications 31*(3), 970–989.

Anderson, B. and J. Moore (2012). *Optimal filtering*. Courier Corporation.

ARPA Lombardia (2022). Temperatura minima media e massima mensili. https://www.arpalombardia.it/indicatori/2021/meteo-e-clima/temperatura-minima-media-e-massima-mensili/.

Baccelli, F., F. I. Karpelevich, M. Y. Kelbert, A. A. Puhalskii, A. N. Rybko, and Y. M. Suhov (1992). A mean-field limit for a class of queueing networks. *Journal of Statistical Physics 66*, 803–825.

Bar-Shalom, Y., X. Li, and T. Kirubarajan (2001). *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons.

Basile, R., R. Grima, and N. Popović (2013). A graph-based approach for the approximate solution of the chemical master equation. *Bulletin of Mathematical Biology 75*, 1653–1696.

Biology, P. C. (2022). Anatomy of the first six months of covid-19 vaccination campaign in italy. *PLOS Computational Biology 18*(5).

Box-Steffensmeier, J. and B. Jones (2004). *Event history modeling: A guide for social scientists*. Cambridge University Press.

Boyce, W. E., R. C. DiPrima, and D. B. Meade (2017). *Elementary differential equations*. John Wiley & Sons.

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

Britton, T., E. Pardoux, F. Ball, C. Laredo, D. Sirl, and V. C. Tran (2019). *Stochastic epidemic models with inference*, Volume 2255. Springer.

Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208.

Capistrán, M., J. Christen, and J. Velasco-Hernández (2012). Towards uncertainty quantification and inference in the stochastic SIR epidemic model. *Mathematical Biosciences 240*(2), 250–259.

Conte, G. (2020a). Dpcm 1 gennaio 2021. https://www.gazzettaufficiale.it/eli/id/2021/01/05/21G00001/sg.

Conte, G. (2020b). Dpcm 26 aprile 2020. https://www.gazzettaufficiale.it/eli/id/2020/04/27/20A02352/sg.

Conte, G. (2020c). Dpcm 9 marzo 2020. https://web.archive.org/web/20201001105941/http://www.governo.it/it/articolo/firmato-il-dpcm-9-marzo-2020/14276.

Cooray-Wijesinha, M. and A. Khuri (1987). The sequential generation of multiresponse d-optimal designs when the variance-covariance matrix is not known. *Communications in Statistics-Simulation and Computation 16*(1), 239–259.

Córdoba-Torres, P., F. Enriquez, and V. Fairén (1998). Optimal start of a taylor integrator by control of local error. *Computers in Physics 12*(2), 200–207.

Corrao, G., M. Franchi, D. Cereda, F. Bortolan, A. Zoli, et al. (2022). Persistence of protection against SARS-CoV-2 clinical outcomes up to 9 months since vaccine completion: a retrospective observational analysis in Lombardy, Italy. *The Lancet Infectious Diseases 22*(5), 649–656.

Craigmile, P., R. Herbei, G. Liu, and G. Schneider (2023). Statistical inference for stochastic differential equations. *WIREs Computational Statistics 15*(2), e1585.

Davis, M. E. and R. J. Davis (2012). *Fundamentals of chemical reaction engineering*. Courier Corporation.

Del Core, L., D. Pellin, M. Grzegorczyk, and E. Wit (2022). Stochastic modelling of cell differentiation networks from partially-observed clonal tracking data. *bioRxiv*, 2022–07.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research 2*(1), 23–41.

Dorfman, R. (1938). A note on the delta-method for finding variance formulae. *The Biometric Bulletin 1*(92), 129–137.

Duy, V. N. L. and I. Takeuchi (2023). Exact statistical inference for the wasserstein distance by selective inference: Selective inference for the wasserstein distance. *Annals of the Institute of Statistical Mathematics 75*(1), 127–157.

Einicke, G. A. and L. B. White (1999). Robust extended kalman filtering. *IEEE Transactions on Signal Processing 47*(9), 2596–2599.

Engl, H. W., M. Hanke, and A. Neubauer (1996). *Regularization of inverse problems*, Volume 375. Springer Science & Business Media.

Fedorov, V. (2013). *Theory of optimal experiments.* Elsevier.

Framba, M., V. Vinciotti, and E. C. Wit (2024). Latent event history models for quasi-reaction systems. *Computational Statistics & Data Analysis*, 107996.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry 81*(25), 2340–2361.

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications 188*(1-3), 404–425.

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

Giordano, G., F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine 26*(6), 855–860.

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

Golightly, A. and D. J. Wilkinson (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology 13*(3), 838–851.

Gomes, M., M. Ferreira, R. Corder, J. King, C. Souto-Maior, C. Penha-Gonçalves, G. Gonçalves, M. Chikina, W. Pegden, and R. Aguas (2022). Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. *Journal of Theoretical Biology 540*, 111063.

Google (2021). Google mobility data. https://www.google.com/covid19/mobility/.

Gozzi, N., M. Chinazzi, J. T. Davis, K. Mu, A. Pastore y Piontti, M. Ajelli, N. Perra, and A. Vespignani (2022). Anatomy of the first six months of covid-19 vaccination campaign in italy. *PLoS Computational Biology 18*(5), e1010146.

Grima, R. (2012). A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics 136*(15).

Guidotti, E. and D. Ardia (2020). Covid-19 data hub. *Journal of Open Source Software 5*(51), 2376.

Gupta, A., C. Schwab, and M. Khammash (2021). DeepCME: a deep learning framework for computing solution statistics of the chemical master equation. *PLoS Computational Biology 17*(12), e1009623.

Hale, J. K. and H. Koçak (2012). *Dynamics and bifurcations*, Volume 3. Springer Science & Business Media.

Hale, T., S. Webster, A. Petherick, T. Phillips, and B. Kira (2020). Oxford COVID-19 government response tracker (OxCGRT). *Last updated 8*, 30.

Hatzis, C. and K. Larntz (1992). Optimal design in nonlinear multiresponse estimation: Poisson model for filter feeding. *Biometrics*, 1235–1248.

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

Ibrahim, J., H. Zhu, and N. Tang (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association 103*(484), 1648–1658.

Ionides, E. L., D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King (2015). Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences 112*(3), 719–724.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*(1), 35–45.

Kennealy, J. P. and W. M. Moore (1977). A numerical method for chemical kinetics modeling based on the taylor series expansion. *The Journal of Physical Chemistry 81*(25), 2413–2419.

Kifer, D., D. Bugada, J. Villar-Garcia, I. Gudelj, C. Menni, et al. (2021). Effects of environmental factors on severity and mortality of COVID-19. *Frontiers in Medicine 7*, 607786.

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

Kim, Y. and H. Bang (2018). Introduction to Kalman filter and its applications. *Introduction and Implementations of the Kalman Filter 1*, 1–16.

Komorowski, M., B. Finkenstadt, D. Rand, C. Gillespie, and D. Wilkinson (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences 108*(21), 8645–8650.

Krijnen, W. P. and E. C. Wit (2022). *Computational and statistical methods for chemical engineering*. Chapman and Hall/CRC.

Lambert, J. (1974). Two unconventional classes of methods for stiff systems. *Stiff Differential Systems*, 171–186.

Lente, G., A. Fursenko, and R. Szabo (2022). Use of the taylor theorem to predict kinetic curves in an arbitrary mechanism. *Chemical Engineering Journal 445*, 136676.

Liao, Z., P. Lan, Z. Liao, Y. Zhang, and S. Liu (2020). TW-SIR: time-window based SIR for COVID-19 forecasts. *Scientific Reports 10*(1), 22454.

Lindsay, B. G., R. S. Pilla, and P. Basak (2000). Moment-based approximations of distributions using mixtures: Theory and applications. *Annals of the Institute of Statistical Mathematics 52*, 215–230.

Lindsten, F. (2013). An efficient stochastic approximation em algorithm using conditional particle filters. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6274–6278. IEEE.

Lotka, A. (1925). Elements of physical biology. *Williams and Wilkins*.

Mattarella, S. (2021a). Decreto del presidente del consiglio dei ministri 14 gennaio 2021. https://www.gazzettaufficiale.it/eli/id/2021/01/15/21A00221/sg.

Mattarella, S. (2021b). Decreto legge 23 luglio 2021, n. 105. https://www.gazzettaufficiale.it/eli/id/2021/07/23/21G00117/sg.

McQuarrie, D. A. (1967). Stochastic approach to chemical kinetics. *Journal of Applied Probability 4*(3), 413–478.

Milner, P., C. S. Gillespie, and D. J. Wilkinson (2011). Moment closure approximations for stochastic kinetic models with rational rate laws. *Mathematical Biosciences 231*(2), 99–104.

Milner, P., C. S. Gillespie, and D. J. Wilkinson (2013). Moment closure based parameter inference of stochastic kinetic models. *Statistics and Computing 23*, 287–295.

Mingliang, Z., T. E. Simos, and C. Tsitouras (2022). R0 estimation for COVID-19 pandemic through exponential fit. *Mathematical Methods in the Applied Sciences 45*(3), 1632–1639.

Ministry of Health (2021). Covid-19. https://github.com/pcm-dpc/COVID-19.

Mortimer, R. G. (2000). *Physical chemistry*. Academic Press.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology 61*(2), 479–482.

Pellin, D., L. Biasco, A. Aiuti, C. Di Serio, and E. Wit (2019). Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking. *Applied Network Science 4*(115).

Pellin, D., L. Biasco, S. Scala, C. Di Serio, and E. C. Wit (2023). Tracking hematopoietic stem cell evolution in a wiskott–aldrich clinical trial. *The Annals of Applied Statistics 17*(3), 1841–1860.

Privitera, G. (2020). First in, last out: Why lombardy is still italy's coronavirus hotspot. *Politico*.

Remuzzi, A. and G. Remuzzi (2020). COVID-19 and Italy: what next? *The Lancet 395*(10231), 1225–1228.

Robertson, H. (1966). The solution of a set of reaction rate equations. *Numerical Analysis: an Introduction*, 178–182.

Schnakenberg, J. (1976). Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern Physics 48*(4), 571–585.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation 24*(111), 647–656.

Shepherd, B. E., H.-P. Kiem, P. M. Lansdorp, C. E. Dunbar, G. Aubert, A. LaRochelle, R. Seggewiss, P. Guttorp, and J. L. Abkowitz (2007). Hematopoietic stem-cell behavior in nonhuman primates. *Blood, The Journal of the American Society of Hematology 110*(6), 1806–1813.

Shoji, I. (2013). Nonparametric estimation of nonlinear dynamics by metric-based local linear approximation. *Statistical Methods & Applications 22*, 341–353.

Shoji, I. and T. Ozaki (1998). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications 16*(4), 733–752.

Shumway, R. and D. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis 3*(4), 253–264.

Signorelli, C. (2019). Forty years (1978-2018) of vaccination policies in italy. *Acta bio-medica: Atenei Parmensis 90*(1), 127–133.

Simon, C. M. (2020). The SIR dynamic model of infectious disease transmission and its analogy with chemical kinetics. *PeerJ Physical Chemistry 2*, e14.

Sjöberg, P., P. Lötstedt, and J. Elf (2009). Fokker–planck approximation of the master equation in molecular biology. *Computing and Visualization in Science 12*, 37–50.

Spijker, M. N. (1996). Stiffness in numerical initial-value problems. *Journal of Computational and Applied Mathematics 72*(2), 393–406.

Swain, P. S., M. B. Elowitz, and E. D. Siggia (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences 99*(20), 12795–12800.

Teugels, J. L. (2008). Markov chains: Models, algorithms and applications.

Tibshirani, R. J. and B. Efron (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability 57*(1), 1–436.

Tsoularis, A. and J. Wallace (2002). Analysis of logistic growth models. *Mathematical Biosciences 179*(1), 21–55.

Tsugé, S. (2001). Rate increase in chemical reaction and its variance under turbulent equilibrium. *Combustion Science and Technology 162*(1), 303–330.

Ullah, M. and O. Wolkenhauer (2011). *Stochastic approaches for systems biology*, pp. 32. Springer Science & Business Media.

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*, Volume 1. Elsevier.

Wan, E. A. and R. Van Der Merwe (2000). The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pp. 153–158.

WHO (2020). WHO COVID-19 Dashboard. *Geneva: World Health Organization 2020.*

Wilkinson, D. J. (2018). *Stochastic modelling for systems biology.* Chapman and Hall/CRC.

Wood, S. and E. Wit (2021). Was R $<$ 1 before the English lockdowns? On modelling mechanistic detail, causality and inference about Covid-19. *Plos One 16*(9), e0257455.

Wu, C., B. Li, R. Lu, S. J. Koelle, Y. Yang, A. Jares, A. E. Krouse, M. Metzger, F. Liang, K. Loré, 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., S. Koelle, P. Guttorp, C. Wu, C. Dunbar, J. L. Abkowitz, and V. N. Minin (2019). Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *The Annals of Applied Statistics 13*(4), 2091–2119.

Ye, Y. and S.-l. Cai (1986). *Theory of limit cycles*, Volume 66. American Mathematical Soc.

Zhang, S., Q. Zhai, X. Shi, and X. Liu (2022). A wiener process model with dynamic covariate for degradation modeling and remaining useful life prediction. *IEEE Transactions on Reliability 72*(1), 214–223.

Zia, A., T. Kirubarajan, J. Reilly, D. Yee, K. Punithakumar, and S. Shirani (2008). An EM algorithm for nonlinear state estimation with model uncertainties. *IEEE Transactions on Signal Processing 56*(3), 921–936.