Author:
Whitehouse, Michael C; Whitehouse, Michael C

Title:
Fast and Consistent Inference in Compartmental Models

Introducing Poisson Approximate Likelihood Methods

Fast and Consistent Inference in Compartmental Models

Introducing Poisson Approximate Likelihood Methods

By

MICHAEL WHITEHOUSE

School of Mathematics
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Science.

JULY 2023

Abstract

This thesis introduces Poisson Approximate Likelihood (PAL) methods to address the challenge of scaling-up epidemiological inference to complex and heterogeneous models,. In contrast to the popular ODE approach to compartmental modelling, in which a large population limit is used to motivate a deterministic model, PALs are derived from approximate filtering equations for finite-population, stochastic compartmental models, and the large population limit drives consistency of maximum PAL estimators. The theoretical results contained within appear to be the first likelihood-based parameter estimation consistency results which apply to a broad class of partially observed stochastic compartmental models and address the large population limit. PALs are simple to implement, involving only elementary arithmetic operations and no tuning parameters, and fast to evaluate, requiring no simulation from the model and having computational cost independent of population size. Through examples we demonstrate how PALs can be used to: facilitate fast exact Bayesian inference within a Delayed Acceptance Particle Markov Chain Monte Carlo scheme; fit an age-structured model of influenza, taking advantage of automatic differentiation in Stan; compare over-dispersion mechanisms in a model of rotavirus by embedding PALs within sequential Monte Carlo; and evaluate the role of unit-specific parameters in a meta-population model of measles.

ACKNOWLEDGEMENTS

Thank you to Mum, Dad, and Caty for the support and for celebrating my successes.

Thank you to Professor Nick Whiteley, whose guidance, encouragement, and high standards have ensured this work is of a quality to be proud of.

Thank you to Dr Lorenzo Rimella, who has essentially been a second supervisor, alongside being a great friend.

Thank you to all my friends on the Compass CDT programme, I'm so lucky to have spent the past four years working alongside such an amazing group of people - you made it all the more worthwhile.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: MICHAEL WHITEHOUSE DATE: 31/07/23

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Introduction

This thesis is based on the paper 'Consistent and fast inference in compartmental models of epidemics using Poisson Approximate Likelihoods' (Whitehouse et al., 2023), which was written with Professor Nick Whiteley and Dr Lorenzo Rimella. This work was produced over the course of my PhD, funded by the EPSRC as part of the Compass centre for doctoral training at the University of Bristol. After a round of major, and then minor, revisions it was accepted to be published in the Journal of the Royal Statistical Society series B (Statistical Methodology).

Aims and Contributions

The quantification and characterisation of infectious disease dynamics are essential for informing official decision makers in their response to emerging epidemics; they are also crucial in understanding previous outbreaks in order to better prepare for the future. The most popular paradigm for modelling the spread of a disease through a population is that of compartmental models. The likelihood for such models is inaccessible in all but the simplest cases, therefore, in order to perform inference one needs to make approximations. Over the past few decades, computational advances have led to the development of many sophisticated and expensive simulation algorithms for inference in stochastic compartmental models. The aims and contributions of this thesis are to propose a class of computationally cheap and simple alternatives to these methods which are justified by rigorous consistency theory and demonstrated to be practically useful when applied to real world data. Section 2.6 provides a detailed breakdown of these contributions.

Structure

This thesis is organised as follows. Chapter 2 introduces the concept of compartmental modelling and reviews the literature on related inference algorithms. Chapter 3 introduces the Poisson approximate likelihood (PAL) and supporting derivations. Chapter 4 presents an argument outline for the consistency result which supports the PAL methodology. Chapter 5 proposes an extension of the PAL methodology to tackle over-dispersed models. Chapter 6 demonstrates the practical implementation of PAL methods through examples. Chapter 7 concludes the thesis with a discussion on the limitations of the methodology and avenues for future research. Appendix A presents the full proof for consistency. Appendix B contains supporting material for chapter 6.

CHAPTER 2

BACKGROUND

In this chapter we introduce compartmental models, discuss existing methods for inference, and introduce the ideas behind the PAL methodology.

2.1 Compartmental Models

Compartmental modelling is one of the most widespread methods for quantifying the dynamics of infectious diseases in populations, rooted in the works of McKendrick and Kermack in the 1920's (Kermack and McKendrick, 1927; McKendrick, 1925), Bartlett (1949, 1966) and Kendall (1956), see (Isham, 2005) for an overview. In this modelling paradigm individuals in a population transition between a collection of discrete compartments, usually representing disease states, where the rates of transition may depend on the current state of the population as a whole as well as possibly unknown parameters. This provides an interpretable, mechanistic framework in which to infer epidemic characteristics such as reproduction numbers, forecast disease dynamics and explore the possible impacts of public health interventions. Compartmental models are also popular in ecology and biochemistry, for example, (Fearnhead et al., 2014; Komorowski et al., 2009), but that is beyond the scope of the present work.

The earliest formulated compartmental models of epidemics consist of a small number of compartments, just three in the standard Susceptible-Infected-Recovered (SIR) model. Modern compartmental models often feature many more compartments, each corresponding to some combination of disease state and other covariates. By increasing the number of compartments, the modeller can specify a more precise representation of complex diseases and populations, such as multi-strain dynamics (Worden and Porco, 2017), subpopulations associated with, e.g., households or age-groups (Andrade and Duggan, 2020), and spatial information (Xia et al., 2004).

Modelling such features is considered a key challenge by epidemiologists (Ball et al., 2015; Funk et al., 2015; Riley et al., 2015; Wikramaratna et al., 2015).

However, the computational cost of fitting compartmental models to data, in general, grows with the number of compartments and also, in the cases of some methods, with the population size. Exact likelihood-based inference is intractable in general and approximate inference typically either involves deleterious model simplifications or involves highly sophisticated algorithms which incur a substantial computational cost. Thus scaling-up inference to complex models is an important and open challenge – this is the motivation for the present work.

Compartmental models come in various forms, some stochastic, some deterministic; some in continuous time, some in discrete time; some modelling finite populations, some motivated by large population asymptotics. Deterministic, ODE-based compartmental models are very popular in practice and often motivated by the fact they can be obtained from finite-population stochastic models in the large population limit. As a very simple example, consider the continuous-time, stochastic version of the SEIR model, with fixed population size n and numbers of susceptible, exposed, infective and removed individuals denoted $X_t^{(n)} := [S_t^{(n)} E_t^{(n)} I_t^{(n)} R_t^{(n)}]^\top$. Each susceptible individual becomes exposed at instantaneous rate $\beta n^{-1} I_t^{(n)}$, each exposed individual becomes infective at rate ρ, each infective individual is "removed" at rate γ, and $(X_t^{(n)})_{t\geq 0}$ is a jump-Markov process. General results concerning the convergence of jump-Markov processes to the solutions of ODE's (Kurtz, 1970, 1971) can be applied to show that, if $n^{-1}X_0^{(n)} \to x_0$ in probability, then for any $T > 0$ and $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} \|n^{-1}X_t^{(n)} - x_t\| > \delta \right) = 0, \tag{2.1}$$

where $(x_t)_{t\geq 0}$, $x_t \equiv [s_t\, e_t\, i_t\, r_t]^\top$, solves:

$$\frac{\mathrm{d}s_t}{\mathrm{d}t} = -\beta s_t i_t, \qquad \frac{\mathrm{d}e_t}{\mathrm{d}t} = \beta s_t i_t - \rho e_t, \qquad \frac{\mathrm{d}i_t}{\mathrm{d}t} = \rho e_t - \gamma i_t, \qquad \frac{\mathrm{d}r_t}{\mathrm{d}t} = \gamma i_t. \tag{2.2}$$

It follows from $S_0^{(n)} + E_0^{(n)} + I_0^{(n)} + R_0^{(n)} = n$ together with $n^{-1}X_0^{(n)} \to x_0$ and (2.2), that $s_t + e_t + i_t + r_t = 1$ for all $t \geq 0$. In order to use this ODE to model a population of size n, x_t is scaled back up by a factor of n, $x_t^{(n)} \equiv [s_t^{(n)}\, e_t^{(n)}\, i_t^{(n)}\, r_t^{(n)}]^\top := n[s_t\, e_t\, i_t\, r_t]^\top$, which satisfies the form of SEIR ODE usually encountered in practice:

$$\frac{\mathrm{d}s_t^{(n)}}{\mathrm{d}t} = -\beta s_t^{(n)} \frac{i_t^{(n)}}{n}, \quad \frac{\mathrm{d}e_t^{(n)}}{\mathrm{d}t} = \beta s_t^{(n)} \frac{i_t^{(n)}}{n} - \rho e_t^{(n)}, \quad \frac{\mathrm{d}i_t^{(n)}}{\mathrm{d}t} = \rho e_t^{(n)} - \gamma i_t^{(n)}, \quad \frac{\mathrm{d}r_t^{(n)}}{\mathrm{d}t} = \gamma i_t^{(n)}. \tag{2.3}$$

To relate $(X_t^{(n)})_{t\geq 0}$ or $(x_t^{(n)})_{t\geq 0}$ to data, for example, error-prone measurements of the number of newly infective individuals in given time periods, one usually postulates a probabilistic observation model, and evaluation of the likelihood function for the parameters (β, ρ, γ) then involves marginalising out $(X_t^{(n)})_{t\geq 0}$ in the case of the finite population stochastic model, which is intractable, or numerical approximation to $(x_t^{(n)})_{t\geq 0}$ in the case of the ODE.

Note here that the only way that $x_t^{(n)}$ depends on n is through the scaling factor $x_t^{(n)} = nx_t$. This, along with the lack of stochasticity, illustrates the simplicity but inflexibility of the ODE

approach to compartmental modelling. Indeed it has been recognised that ODE models cannot capture important epidemiological phenomena such as fade-out, extinction, lack of synchrony, or deviations from stable behaviour (Roberts et al., 2015, Sec. 8) and, somewhat more obviously, may under-represent uncertainty (King et al., 2015).

To summarise the above, consider the following conceptual workflow:

ODE 1. specify a finite population, stochastic, continuous-time compartmental model $(X_t^{(n)})_{t\geq 0}$;

ODE 2. scale $X_t^{(n)}$ by n^{-1} and take the large population limit $n \to \infty$ to obtain $(x_t)_{t\geq 0}$;

ODE 3. re-scale $(x_t)_{t\geq 0}$ by n to obtain $(x_t^{(n)})_{t\geq 0}$, on the appropriate scale for a population of size n;

ODE 4. numerically approximate $(x_t^{(n)})_{t\geq 0}$ and combine with an observation model to evaluate the likelihood function.

Of course in practice, someone can use the ODE model (2.3) without knowing anything about steps ODE 1.-3. We write out these steps in order to emphasise how the ODE approach differs to the PAL methods proposed in the present work, where crucially the limit $n \to \infty$ is taken later in the conceptual workflow:

PAL 1. specify a finite population, stochastic, discrete-time compartmental model;

PAL 2. combine this model with an observation model to obtain discrete-time filtering equations;

PAL 3. recursively approximate the filtering equations using Poisson distributions, thus defining the PAL;

PAL 4. take the large population limit, $n \to \infty$, to establish the consistency of the parameter estimator obtained by maximising the PAL.

The Latent Compartmental Model we work with is introduced in section 3.1. It allows the probabilities of individuals transitioning between compartments to depend on the state of the population as a whole in a quite general way, as well as allowing for immigration and emigration, constant or random and dynamic population size. Due to the general form of this compartmental model, we can treat classical disease states, such as SEIR, as well as discrete covariates or subpopulations such as spatial locations or age-groups, in a single framework.

In the next section we explore the connections of the present work to the literature. We survey the state of the art methods for inference in stochastic epidemic models and discuss the use of Poisson process approximations for inference.

2.2 Inference Algorithms for Stochastic Compartmental Models

Compartmental models concern case data, be it prevalence data - the number of individuals infected at a given time, or incidence data - the number of *newly* infected individuals over a given period. When applied to real data they are often combined with an observation model to capture an imperfect reporting mechanism, for example, due to asymptomatic cases or testing errors. This observation model results in a latent variable model.

SEIR example

As a simple running example of a Latent Compartmental Model we will consider the discrete-time susceptible-exposed-infective-removed (SEIR) model:

$$S_{t+1} = S_t - B_{t+1}, \quad E_{t+1} = E_t + B_{t+1} - C_{t+1}, \quad I_{t+1} = I_t + C_{t+1} - D_{t+1}, \qquad R_{t+1} = R_t + D_{t+1}.$$

With conditionally independent, binomially distributed random variables:

$$B_{t+1} \sim \text{Bin}(S_t, 1 - e^{-h\beta\frac{I_t}{n_t}}), \quad C_{t+1} \sim \text{Bin}(E_t, 1 - e^{-h\rho}), \quad D_{t+1} \sim \text{Bin}(I_t, 1 - e^{-h\gamma}), \tag{2.4}$$

where $h > 0$ is a time-step size. The observation at time t, say y_t, is a binomial under-reporting of the current number of infective individuals $y_t \sim \text{Bin}(I_t, q)$ for some reporting rate $q \in [0,1]$. Denote the parameter $\theta = [\beta\ \rho\ \gamma\ q]$. If one identifies $x_t := [S_t\, E_t\, I_t\, R_t]$ for $t \geq 0$, then the pair $(x_t)_{t\geq 0}$ and $(y_t)_{t\geq 1}$ defines a hidden Markov model with transition kernel implied by equation (2.4), emission distribution $y_t \sim \text{Bin}(I_t, q)$, and some initial distribution, say $x_0 \sim p_0$.

Evaluating the likelihood function for such a latent compartmental models requires the integrating out of all configurations of the population amongst the compartments, in the context of the simple SEIR example, suppressing dependence on θ:

$$p(y_{1:t}) = \sum_{x_{0:t}} p(x_{0:t}, y_{1:t}) = \sum_{x_{0:t}} p_0(x_0) \prod_{s=1}^{t} p(x_s|x_{s-1}) p(y_s|x_s). \tag{2.5}$$

The cost of this marginalisation explodes as the number of compartments and the population size grows far beyond those of the SEIR model. Computational advances over the past 25 years have allowed for the implementation of sophisticated and expensive algorithms, these advances have prompted the development of a variety of simulation-based inference methods. We now present some existing methods for approximating intractable marginal likelihoods and discuss their applications to stochastic compartmental models.

Sequential Monte Carlo

For a thorough introduction to this family of algorithms, see Chopin et al. (2020). For pedagogical purposes, we will consider sequential Monte Carlo methods in the context of hidden Markov Models, in particular the SEIR example - though these methods are applicable to a much broader range of scenarios. The general idea of Monte Carlo methods is to represent and approximate distributions with a discrete set of points, termed a sample of 'particles'. The objective of the particle filter family of sequential Monte Carlo algorithms is to produce finite sample approximations to sequences of filtering distributions, for example $p(S_t, E_t, I_t, R_t|y_{1:t})$ for $t \geq 1$ in the case of the SEIR model. One may also take as output an approximation to the marginal likelihood (2.5); a simple bootstrap filter targeting (2.5) is given by algorithm 1.

Algorithm 1 Bootstrap particle filter for the simple SEIR model

input: Number of particles n_{part}, parameter $\theta = [\beta\ \rho\ \gamma\ q]$.

initialise: $S_0^{(i)} \leftarrow S_0, \quad E_0^{(i)} \leftarrow E_0, \quad I_0^{(i)} \leftarrow I_0, \quad R_0^{(i)} \leftarrow R_0$ for $i = 1,\dots,n_{part}$

for $t \geq 1$:

  for $i = 1,\dots,n_{part}$:

    $B_t^{(i)} \sim \text{Bin}\left(S_{t-1}^{(i)}, 1 - e^{-h\beta\frac{I_{t-1}^{(i)}}{n_t}}\right), \quad C_t^{(i)} \sim \text{Bin}\left(E_{t-1}^{(i)}, 1-e^{-h\rho}\right), \quad D_t^{(i)} \sim \text{Bin}\left(I_{t-1}^{(i)}, 1-e^{-h\gamma}\right)$

    Set $S_t^{(i)} = S_{t-1}^{(i)} - B_t^{(i)}, \quad E_t^{(i)} = E_{t-1}^{(i)} + B_t^{(i)} - C_t^{(i)}, \quad I_t^{(i)} = I_{t-1}^{(i)} + C_t^{(i)} - D_t^{(i)}, \quad R_t^{(i)} = R_{t-1}^{(i)} + D_t^{(i)}$

    $\log w_t^{(i)} \leftarrow \log \text{Bin}(y_t | I_t^{(i)}, q)$

  end for

  $\log \hat{p}(y_t | y_{1:t-1}, \theta) \leftarrow \log\left(\frac{1}{n_{part}} \sum_{i=1}^{n_{part}} w_t^{(i)}\right)$

  $\bar{w}_t^{(i)} \leftarrow w_t^{(i)} / \sum_{j=1}^{n_{part}} w_t^{(j)}$

  resample $\left\{S_t^{(i)}, E_t^{(i)}, I_t^{(i)}, R_t^{(i)}\right\}_{i=1}^{n_{part}}$ according to the weights $\left\{\bar{w}_t^{(i)}\right\}_{i=1}^{n_{part}}$

end for

The output of this algorithm is the approximation:

$$\log p(y_{1:t}|\theta) \approx \sum_{s=1}^{t} \log \hat{p}(y_s | y_{1:s-1}, \theta).$$

For this simple model one can easily propose new particles from the transition kernel. In more complicated models using the model transition to propose new particles can lead to poor performance, such as particle collapse and high variance in the likelihood estimate, in some cases it may not even be possible. In this case one must choose an importance proposal. Consider a general hidden Markov model, $(x_t)_{t\geq 0}$ and $(y_t)_{t\geq 1}$, with transition kernel $f(x_t|x_{t-1})$ and emission $g(y_t|x_t)$. We would like to approximate:

$$p(y_{1:t}|\theta) = \int p(y_{1:t}, x_{0:t}|\theta) dx_{0:t} = \int p_0(x_0|\theta) \prod_{s=1}^{t} p(x_s|x_{s-1},\theta) p(y_s|x_s,\theta) dx_{0:t}. \tag{2.6}$$

A generic simple bootstrap particle filter which yields an approximation to (2.6) is presented in algorithm 2.

Algorithm 2 Bootstrap particle filter

input: proposal distribution $\pi(\cdot|\cdot)$, number of particles n_{part}, parameter θ.

initialise: $x_0^{(i)} \leftarrow x_0$ for $i = 1,\dots,n_{part}$

for $t \geq 1$:

  for $i = 1,\dots,n_{part}$:

    $x_t^{(i)} \sim \pi(\cdot | x_{0:t-1}^{(i)}, y_{1:t}, \theta)$

    $\log w_t^{(i)} \leftarrow \log g(y_t|x_t^{(i)},\theta) + \log f(x_t^{(i)}|x_{t-1}^{(i)},\theta) - \log \pi(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t}, \theta)$

  end for

  $\log \hat{p}(y_t | y_{1:t-1}, \theta) \leftarrow \log\left(\frac{1}{n_{part}} \sum_{i=1}^{n_{part}} w_t^{(i)}\right)$

  $\bar{w}_t^{(i)} \leftarrow w_t^{(i)} / \sum_{j=1}^{n_{part}} w_t^{(j)}$

  resample $\left\{x_t^{(i)}\right\}_{i=1}^{n_{part}}$ according to the weights $\left\{\bar{w}_t^{(i)}\right\}_{i=1}^{n_{part}}$

end for

Again, we make the approximation:

$$\log p(y_{1:t}|\theta) \approx \sum_{s=1}^{t} \log \hat{p}(y_s|y_{1:s-1},\theta). \tag{2.7}$$

It is well known that the efficiency of approximation (2.7) is heavily dependent on the choice of proposal π. For example, if π results in proposals such that $g(y_t|x_t^{(i)},\theta) = 0$ for all i, then the algorithm will fail due to particle collapse. The so-called 'optimal proposal' which minimises the variance of the importance weights is given by $\pi(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t}, \theta) = p(x_t^{(i)}|x_{t-1}^{(i)}, y_t, \theta)$ (Doucet et al., 2000), although this is often inaccessible aside from very convenient cases. Indeed, within epidemiology, methods for choosing π is an active field of research (Ju et al., 2021; Park and Ionides, 2020; Rimella et al., 2022). Furthermore, as the dimension of the latent process $(x_t)_{t\geq 0}$ grows, SMC algorithms suffer from the curse of dimensionality - in particular the bootstrap algorithm requires that n_{part} scales exponentially with the dimension of $(x_t)_{t\geq 0}$ (Snyder et al., 2008). This is a particular issue within epidemiology as it prohibits SMC from scaling up to perform inference on models with a large number of compartments. Work exists on tackling this specific issue (Ionides et al., 2022; Park and Ionides, 2020), but do not achieve the same level of scalability as PAL methods, as demonstrated in the example of section 6.4.

Given access to the output of algorithm 2 one can perform likelihood based inference by embedding it within, e.g., particle Markov chain Monte Carlo (Andrieu et al., 2010; Fasiolo et al., 2016) for a Bayesian approach, or an iterated filtering scheme (Ionides et al., 2011) for frequentist inference.

Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) methods encompass a family of algorithms designed to perform inference on Bayesian models for which the likelihood is inaccessible, but for which simulation is straight-forward. This poses ABC methods as an attractive candidate to fit easily simulated mechanistic models such as compartmental models. The general idea is to replace likelihood evaluation with a direct comparison between observed data and data simulated from the model, using some distance metric $d(\cdot,\cdot)$. The simplest instance of such an algorithm is given by the ABC rejection scheme (Rubin, 1984). Let $\pi(\theta)$ be a prior over the parameters of the SEIR model, algorithm 3 presents an ABC rejection scheme for this example.

Algorithm 3 ABC rejection sampler.

1: **input** data $y_{1:t}$, distance metric $d(\cdot,\cdot)$, threshold ε.
2: **for** $n \geq 1$:
3: $\quad\theta^* \sim \pi(\cdot)$
4: $\quad$simulate $y_{1:t}^*$ from the SEIR model with parameter θ^*
5: $\quad$If $d(y_{1:t}^*, y_{1:t}) < \varepsilon$ accept θ^*, otherwise reject
6: **end for**

Algorithm 3 results in a sample from the distribution $\pi(\theta|d(y^*_{1:t}, y_{1:t}) < \varepsilon)$ (Rubin, 1984).

In practice it is advised to replace the direct comparison between observation and simulation with a comparison between summary statistics, that is in line 5 we replace $d(y^*_{1:t}, y_{1:t})$ with $d(S(y^*_{1:t}), S(y_{1:t}))$ for some summary statistic function S. One intuitive and simple choice of comparison is the sum-of-squared differences between observed and simulated case counts, i.e.:

$$d(S(y^*_{1:t}), S(y_{1:t})) = \sum_{s=1}^{t} (y^*_s - y_s)^2$$

An alternative, suggested by (McKinley et al., 2009), is based on a chi-squared goodness-of-fit criterion which scales the contribution of each time point by the observed data to reflect that the variation changes as the epidemic progresses.

When the prior distribution is very different from the posterior, algorithm 3 has a high rejection rate. To tackle this one can embed ABC within an MCMC sampler to explore the space more efficiently, See algorithm 4.

Algorithm 4 ABC Markov Chain Monte Carlo

1: **input** data $y_{1:t}$, distance metric $d(\cdot,\cdot)$ and summary statistic function S, threshold ε, proposal distribution $q(\cdot|\cdot)$, initial θ_0.
2: **for** $n \geq 1$**:**
3: $\theta^* \sim q(\cdot|\theta_{n-1})$
4: Simulate $y^*_{1:t}$ from the SEIR model with parameter θ^*
5: If $d\left(S(y^*_{1:t}), S(y_{1:t})\right) < \varepsilon$:
6: With probability $\min\left\{1, \frac{\pi(\theta^*)q(\theta_{n-1}|\theta^*)}{\pi(\theta_{n-1})q(\theta^*|\theta_{n-1})}\right\}$ set $\theta_n = \theta^*$
7: Else set $\theta_n = \theta_{n-1}$

Algorithm 4 produces a Markov chain with stationary distribution $\pi\left(\theta|d\left(S(y^*_{1:t}), S(y_{1:t})\right) < \varepsilon\right)$ (Marjoram et al., 2003). The efficiency of ABC algorithms rely crucially on appropriate choices of distance metric and summary statistics which can be a barrier for practitioners, though there is an extensive literature advising on this issue (Fearnhead and Prangle, 2012; Prangle et al., 2014; Saulnier et al., 2017).

Data Augmentation Markov Chain Monte Carlo

We now describe the broad framework of data augmentation Markov chain Monte Carlo (DAMCMC), we take a general approach since the flavour of epidemiological model it is employed to fit often deviates somewhat from that of the SEIR model we have considered thus far. Consider a simple Bayesian model with:

1. prior $\theta \sim \pi(\cdot)$,

2. likelihood $y \sim p(\cdot|\theta)$,

and suppose that the likelihood p is intractable so that the posterior $\pi(\theta|y) \propto p(y|\theta)\pi(\theta)$ is inaccessible. DAMCMC consists of: augmenting the parameter space with some variable, say $\phi \sim \pi_{aug}(\cdot|\theta)$, such that one can access $p(y|\theta,\phi)$; designing an MCMC scheme to target the joint posterior for (θ,ϕ); discarding samples for ϕ and considering only samples from the marginal posterior for θ. A simple Metropolis Hastings algorithm targeting the posterior for (θ,ϕ) is given by algorithm 5, one obtains an approximate sample from the marginal posterior for θ by discarding samples for ϕ.

In the context of epidemiological models the case related observation data y is augmented with, for example, ϕ representing: infection event times (Walker et al., 2017); epidemic final severity (Demiris and O'Neill, 2005); a latent compartmental process (Morsomme and Xu, 2022) i.e. in the context of the SEIR model $\phi = ([S_t\, E_t\, I_t\, R_t])_{t\geq 0}$.

Algorithm 5 Data Augmentation Markov Chain Monte Carlo

input: θ_0, ϕ_0
1: **for** $n \geq 1$:
2: $\theta^* \sim \pi(\cdot)$
3: $\phi^* \sim \pi_{aug}(\cdot|\theta^*)$
4: with probability $\min\left\{1, \frac{p(y|\theta^*,\phi^*)}{p(y|\theta_{n-1},\phi_{n-1})}\right\}$ set $\theta_n = \theta^*$ and $\phi_n = \phi*$, else set $\theta_n = \theta_{n-1}$ and $\phi_n = \phi_{n-1}$.
5: **end for**

Data Augmentation methods are limited by computational overheads in their application to epidemiology. In our context data are naturally temporal; DA methods which sample subject histories require extensive book-keeping as the number of epidemiological events grow large and suffer in large population settings (Fintzi et al., 2017). Indeed, these methods can degrade substantially as the number of individuals grows over a few thousand (Fintzi et al., 2017), this is magnitudes smaller than the population sizes of the examples we consider in sections 6.3 and 6.4.

Simulation vs PAL Methods

In principle, if one can simulate from the model, say using Gillespie's algorithm (Gillespie, 1976), then one may perform inference using one of these algorithms. This has led to these methods being described as 'plug-and-play' and 'simulation based'. Clearly, this means that this family of algorithms are more versatile than the PAL methods we introduce, and are not restricted to the latent compartmental model we introduce in section 3.1.2. Indeed, their application remit extends far beyond epidemiology - from finance (Jasra and Del Moral, 2011; Jasra et al., 2011), to ecology (Beaumont, 2010; Fasiolo et al., 2016) , to population genetics (Beaumont et al., 2002).

This flexibility, however comes at a cost both in terms of computational complexity and difficulty of implementation. In practice, simulation based methods require fine tuning and careful adaptation to suit a specific target application. The need to choose algorithmic parameters places a burden on the practitioner on top of the computational costs involved. In contrast, the

vanilla PAL algorithm introduced in chapter 3 requires no tuning and relies solely on simple linear algebraic expressions, the computational benefits of the PAL in comparison to SMC methods are explored in section 6.1.

In chapter 5 we introduce methodology which extends the remit of PALs to models which include over-dispersion (Bretó and Ionides, 2011). This involves the introduction of latent variables which we integrate out by embedding PALs within sequential Monte Carlo scheme, introducing an element of simulation to the algorithm. We demonstrate, however, that the dimension of the integral being estimated is far smaller than that of the pure SMC alternative. Further to this, we recommend a choice of proposal which ensures efficient performance in terms of effective sample size and log-likelihood estimate variance.

Linear Noise Approximation

A functional central limit theorem associated with (2.1) due to Kurtz (1971) gives rise to an SDE known as the Linear Noise Approximation (LNA), see e.g., (Fearnhead et al., 2014; Komorowski et al., 2009) for practical details in a range of contexts. The LNA SDE has a multivariate Gaussian transition density which if combined with a suitable Gaussian observation model allows disease states to be marginalised out in closed form. The approximate likelihood function which thus arises may be used for inference directly, or as a surrogate for the exact likelihood function if a suitable correction can be applied, e.g. using Delayed Acceptance MCMC (Golightly et al., 2015).

The LNA may be described with a similar workflow to the ODE workflow in section 2.1 but with $(x_t)_{t\geq 0}$ and $(x_t)^{(n)}_{t\geq 0}$ replaced by the solutions of SDE's arising from the associated function CLT, and numerical approximations to $(x_t)^{(n)}_{t\geq 0}$ replaced by marginalising out. . The LNA involves computing covariance matrices associated with the set of compartments, and hence the computational cost of applying the LNA can scale with the third power of the number of compartments in general. In section 6.1 we make comparisons between the LNA and the proposed PAL methods, including time comparisons and a qualitative model comparison.

Other varieties of SDE-based approximations to finite-population stochastic compartmental models have been proposed (Allen, 2017), but their transition probabilities are usually not available in closed form and generally costly simulation-based methods are relied upon to fit these models to data (Cauchemez and Ferguson, 2008; Roberts and Stramer, 2001).

2.3 Poisson process approximations

Recursive approximation of filtering distributions using Poisson processes underlies the so-called Probability Hypothesis Density (PHD) filter of Mahler (2003), subsequently re-derived and generalised by Caron et al. (2011); Singh et al. (2009). These approximate methods pertain to models used for tracking targets in discrete-time moving on on a continuous space, as opposed to 'tracking' individuals moving through discrete disease states.

An notable connection here is that a specific, but epidemiologically uninteresting, case of one model we consider in section 3.2.1 coincides with a discrete-state version of the model considered in these works and the corresponding special case of our algorithm 6 would coincide with a discrete-state version of the PHD filter. The incidence data model we define in sections 3.1.3.2 and 3.1.3.3 is however notably different to the model of Mahler (2003); Singh et al. (2009); Caron et al. (2011), and particularly important for epidemiological data.

Whilst these methods were derived primarily for filtering purposes, parameter estimation using the PHD filter in spatial multi-target models was suggested by Singh et al. (2011) but without any rigorous justification. Parameter estimation is further explored by Mahler et al. (2011), however the focus lies with estimating clutter observation intensities and detection profiles, rather than parameters of the transition kernel. Such insights are vital in epidemiological applications since they allow one to make important inferences about transition rates and facilitate estimation of important epidemiological parameters, such as the reproduction number. Indeed, across the broad and substantial literature on the Probability Hypothesis Density there do not appear to be any theoretical results concerning parameter estimation consistency using the PHD filter. Extending our consistency results to the non-discrete setting of the PHD filter may be of considerable interest to the engineering community, but is beyond the scope of this thesis.

Approximate filtering for a limited class of epidemic models using multinomial rather than Poisson approximations was proposed by Whiteley and Rimella (2021), but without any consistency theory. In contrast, this thesis considers a far broader class of models and introduces rigorous justification for the methodology with a novel consistency result.

2.4 Parameter estimation consistency results for compartmental models

As surveyed above, in recent years much research on inference in compartmental models has focused on computational issues. The literature on consistency of parameter estimators is generally older and much more focused on specific instances of compartmental models for which inferential calculations can be made in closed form.

In the case of a fully-observed, continuous-time Susceptible-Infective-Removed (SIR) model, i.e. for which all infection and removal times are observed, maximum likelihood estimators of the infection and removal rate parameters are available in closed form and are consistent with asymptotically normal estimation error, established in the regime where the population size tends to infinity using martingale limit theorems, see (Becker, 1993), (Andersson and Britton, 2012, Ch. 9), and references therein. It is very unrealistic to assume that all infection and removal times are observed. For some restrictive cases of specific partial observations from the SIR model, such as when only the initial and final state of the population are observed, maximum

likelihood estimators are available and are consistent (Andersson and Britton, 2012, Ch. 10) However, perhaps owing to the specificity of the mathematics involved, the range of applications of martingale methods seems limited (Becker, 1993).

Many stochastic compartmental models of epidemics are transient, in the sense that with probability one the entire population eventually ends up in one compartment and stays there, such as the R compartment in SIR and SEIR. For this reason, it seems that the asymptotic regime of a finite, fixed population size and increasingly long time horizon is not a fruitful regime in which to study consistency of parameter estimators for many epidemic models. One exceptional non-transient case, at least in the infinite population setting, is the Susceptible-Infective-Susceptible model, see (Gourieroux and Jasiak, 2021) for a likelihood-based analysis in the regime where the time horizon tends to infinity. In this work, the authors establish consistency of transition rate estimators, but these are associated with a convenient observation model which does not consider imperfect observations.

It should further be noted that consistent point estimators of specific parameters within specific models, such as the Malthusian (the initial exponential growth rate in the number of infected individuals) parameter in an SEIR model (Lindenstrand and Svensson, 2013), or R_0 in an SIR model are available (Britton, 2010), but their derivations seem to be also very tied to the specifics of these models.

There appears to be a lack of consistency results for likelihood-based estimators for more general classes of compartmental models with a, theoretically, unbounded number of compartments.

2.5 Data-driven model selection for compartmental models

In the vast literature of compartmental models applications to real data, the choice of model specifics, such as deterministic vs stochastic, are almost always determined in a subjective manner, based on domain expert knowledge and computational convenience (Sun et al., 2015). Since, in most cases, the true likelihood of stochastic compartmental models is inaccessible, research has mainly focused on ABC approaches to model selection, usually based on approximating the Bayes factor (Toni et al., 2009). There are few instances of purely likelihood-based 'frequentist' model selection exercises. One example based on an Akaike information criterion (AIC) selection procedure, which performs a trade-off between model fit and complexity, is given by Stocks et al. (2020). Another, which assesses fit purely in terms of likelihood value is given by Ionides et al. (2022). In sections 6.3 and 6.4 we repeat the analyses of Stocks et al. (2020) and Ionides et al. (2022), respectively, within the PAL framework; in both cases, we report an overwhelming improvement in AIC and computational efficiency.

2.6 Motivating PAL methods

In this chapter, we have surveyed and discussed a variety of existing methods for inference in stochastic epidemic models. It is apparent that there are shortfalls in the currently available methods and gaps in the theoretical literature. In this thesis we will tackle these issues – in particular, we will address:

- **Computationally efficient and scalable methodology:** In chapter 3 we derive the PAL recursions, demonstrating their dependence on simple linear algebraic expressions. The computational advantages are explored in chapter 6. This addresses the lack of fast and efficient algorithms available for inference across the large range of models we consider, as discussed in section 2.2.

- **Consistency theory:** in chapter 4 we present the outline of the proof for the consistency of maximum PAL estimators. This addresses the lack of consistency results for parameter estimation pertaining to a generalised class of compartmental models of epidemics.

- **Practical utility:** chapter 6 extensively demonstrates the application of PAL methodology to real data, with strong results in terms of computation time and model goodness-of-fit. Furthermore, In sections 6.3 and 6.4 we demonstrate how one can embed models of increasing complexity within our highly flexible latent compartmental model class. Thus, we address the lack of computationally convenient likelihood based model selection procedures for compartmental models.

CHAPTER 3

DERIVING POISSON APPROXIMATE LIKELIHOODS

This chapter is organised as follows. Section 3.1 introduces the notation and models used throughout this thesis. In section 3.2 we propose and derive the PAL along with the supporting lemmas and proofs.

3.1 Models

3.1.1 Notation

The set of natural numbers, including 0, is denoted $\mathbb{N}_0$. The set of non-negative real numbers is denoted $\mathbb{R}_{\geq 0}$. For an integer $m \geq 1$, $[m] \coloneqq \{1, \ldots, m\}$. Matrices and vectors are denoted by bold upper-case and bold lower-case letters, respectively, e.g., $\mathbf{A}$ and $\mathbf{b}$, with non-bold upper-case and lower case used for their respective elements $A^{(i,j)}$, $b^{(i)}$.

All vectors are column vectors unless stated otherwise. We use $\mathbf{1}_m$ to denote the vector of m 1's and $\mathbf{0}_m$ to denote the vector of m 0's. The indicator function is denoted $\mathbb{I}[\cdot]$. The element-wise product of matrices and vectors are denoted $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{a} \odot \mathbf{b}$ respectively, the element-wise division of matrices and vectors are denoted $\mathbf{A} \oslash \mathbf{B}$ and $\mathbf{a} \oslash \mathbf{b}$ respectively, the outer product of vectors is denoted $\mathbf{a} \otimes \mathbf{b}$. The logarithm $\log \mathbf{A}$, factorial $\mathbf{A}!$, and exponential $\exp(\mathbf{A})$ are taken element-wise. For $\mathbf{x} \in \mathbb{N}_0^m$ we define $\boldsymbol{\eta}(\mathbf{x}) = [x^{(1)}/\mathbf{1}_m^\top \mathbf{x} \cdots x^{(m)}/\mathbf{1}_m^\top \mathbf{x}]^\top$ if $\mathbf{1}_m^\top \mathbf{x} > 0$, i.e. $\boldsymbol{\eta}(\mathbf{x})$ normalises $\mathbf{x}$ to yield a probability vector; and $\boldsymbol{\eta}(\mathbf{x}) = \mathbf{0}_m$ if $\mathbf{1}_m^\top \mathbf{x} = 0$. For $\mathbf{x} \in \mathbb{N}_0^m$ and $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^m$ we write $\mathbf{x} \sim \text{Pois}(\boldsymbol{\lambda})$ to denote that the elements of $\mathbf{x}$ are independent and element $x^{(i)}$ is Poisson distributed with parameter $\lambda^{(i)}$. We shall say that such a random vector $\mathbf{x}$ has a "vector-Poisson distribution". For a probability vector $\boldsymbol{\pi}$ we write $\text{Mult}(n, \boldsymbol{\pi})$ for the associated multinomial distribution. Similarly, for a random matrix $\mathbf{X} \in \mathbb{N}_0^{m \times l}$ and a matrix $\boldsymbol{\Lambda} \in \mathbb{R}_{\geq 0}^{m \times l}$, we write $\mathbf{X} \sim \text{Pois}(\boldsymbol{\Lambda})$ when the elements of $\mathbf{X}$ are independent with $X^{(i,j)}$ being Poisson distributed with parameter $\Lambda^{(i,j)}$. We call $\boldsymbol{\lambda}$ (resp. $\boldsymbol{\Lambda}$)

the intensity vector (resp. matrix). For a length-m vector $\mathbf{b}$ with non-negative elements, we call $\text{supp}(\mathbf{b}) \coloneqq \{i \in [m] : b^{(i)} > 0\}$ the support of $\mathbf{b}$. By convention, we take a sum over an empty set to be equal to 0, i.e. a sum of 0 terms. We write $\boldsymbol{e}_i$ for the vector of zeros except for a 1 in the ith entry.

3.1.2 Latent Compartmental Model

The model we consider is defined by: m, the number of compartments; n the expected initial population size; $\mathbb{P}_{0,n}$ an initial distribution on $\mathbb{N}_0^m$ such that $\mathbb{E}_{\mathbf{x}_0 \sim \mathbb{P}_{0,n}}[\mathbf{1}_m^\top \mathbf{x}_0] = n$, e.g., $\text{Pois}(\boldsymbol{\lambda}_0)$ for some $\boldsymbol{\lambda}_0 \in \mathbb{R}_{\geq 0}^m$ such that $\mathbf{1}_m^\top \boldsymbol{\lambda}_0 = n$, or $\text{Mult}(n, \boldsymbol{\pi}_0)$ for some length-m probability vector $\boldsymbol{\pi}_0$; a sequence, $\{\boldsymbol{\alpha}_t\}_{t \geq 1}$ with $\boldsymbol{\alpha}_t \in \mathbb{R}_{\geq 0}^m$ for all $t \geq 1$, of immigration intensity vectors; a sequence, $\{\boldsymbol{\delta}_t\}_{t \geq 0}$ with $\boldsymbol{\delta}_t \in [0,1]^m$ for all $t \geq 0$; and for each $t \geq 0$ a mapping from length-m probability vectors to size-$m \times m$ row-stochastic matrices, $\boldsymbol{\eta} \mapsto \mathbf{K}_{t,\boldsymbol{\eta}}$.

The population at time $t \in \mathbb{N}_0$ is a set of a random number n_t of random variables $\{\xi_t^{(1)}, \ldots, \xi_t^{(n_t)}\}$, each valued in $[m]$. The counts of individuals in each of the m compartments at time t are collected in $\mathbf{x}_t = [x_t^{(1)} \cdots x_t^{(m)}]^\top$, where $x_t^{(i)} = \sum_{j=1}^{n_t} \mathbb{I}[\xi_t^{(j)} = i]$. The population is initialised as a draw $\mathbf{x}_0 \sim \mathbb{P}_{0,n}$. The members of the population are exchangeable, labelled by, e.g., a uniformly random assignment of indices $\{\xi_0^{(1)}, \ldots, \xi_0^{(n_0)}\}$ subject to $x_0^{(j)} \coloneqq \sum_{i=1}^{n_0} \mathbb{I}[\xi_0^{(i)} = j]$. For $t \geq 1$, given $\{\xi_{t-1}^{(1)}, \ldots, \xi_{t-1}^{(n_{t-1})}\}$, we obtain n_t and $\{\xi_t^{(1)}, \ldots, \xi_t^{(n_t)}\}$ as follows. For $i = 1, \ldots n_{t-1}$, with probability $1 - \delta_t^{(\xi_{t-1}^{(i)})}$ the individual $\xi_{t-1}^{(i)}$ emigrates from $[m]$ to a state $0 \notin [m]$ from which it does not return. The counts of remaining individuals are collected in the vector $\bar{\mathbf{x}}_{t-1}$, where $\bar{x}_{t-1}^{(j)} \coloneqq \sum_{i=1}^{n_{t-1}} \mathbb{I}[\xi_{t-1}^{(i)} = j] \mathbb{I}[\phi_t^{(i)} = 1]$ and $\phi_t^{(i)} \sim \text{Bernoulli}(\delta_t^{(\xi_{t-1}^{(i)})})$.

For each i such that $\mathbb{I}[\phi_t^{(i)} = 1] = 1$, i.e. a remaining individual, $\xi_t^{(i)}$ is then drawn from the $\xi_{t-1}^{(i)}$'th row of $\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}$ and the resulting counts of individuals in the compartments $[m]$ are denoted $\tilde{\mathbf{x}}_t$ where $\tilde{x}_t^{(j)} \coloneqq \sum_{i=1}^{n_{t-1}} \mathbb{I}[\phi_t^{(i)} = 1] \mathbb{I}[\xi_t^{(i)} = j]$, if $\bar{\mathbf{x}}_{t-1} = \mathbf{0}_m$ then $\tilde{\mathbf{x}}_t = \mathbf{0}_m$. Let $\mathbf{Z}_t$ be the $m \times m$ matrix with elements $Z_t^{(i,j)} \coloneqq \sum_{k=1}^{n_{t-1}} \mathbb{I}[\xi_{t-1}^{(k)} = i, \xi_t^{(k)} = j]$, which counts the individuals transitioning from compartment i at $t-1$ to compartment j at time t. New individuals then immigrate into the compartments $[m]$ according to a vector-Poisson distribution $\hat{\mathbf{x}}_t \sim \text{Pois}(\boldsymbol{\alpha}_t)$ and the resulting combined counts of individuals are $\mathbf{x}_t \coloneqq \tilde{\mathbf{x}}_t + \hat{\mathbf{x}}_t$ with $n_t \coloneqq \mathbf{1}_m^\top (\tilde{\mathbf{x}}_t + \hat{\mathbf{x}}_t)$. The population $\{\xi_t^{(1)}, \ldots, \xi_t^{(n_t)}\}$ is then obtained by uniformly random assignment of indices subject to $x_t^{(j)} \coloneqq \sum_{i=1}^{n_t} \mathbb{I}[\xi_t^{(i)} = j]$. Note that under this model, the processes $(\mathbf{x}_t)_{t \geq 0}$ and $(\mathbf{Z}_t)_{t \geq 1}$ are Markov chains, although we shall not need explicit expressions for their transition probabilities.

If the matrix $\mathbf{K}_{t,\boldsymbol{\eta}}$ were to have no dependence on $\boldsymbol{\eta}$, then the Latent Compartmental Model is a discrete-state version of the dynamic spatial Poisson-process model underlying the PHD filter (Caron et al., 2011; Mahler, 2003; Singh et al., 2009). However, for epidemiological modelling it is critical that $\mathbf{K}_{t,\boldsymbol{\eta}}$ does depend on $\boldsymbol{\eta}$; for example in the case of SEIR as we shall now state, it is this dependence which models the mechanism of infection amongst the population.

SEIR example

As a very simple example of the Latent Compartmental Model consider the SEIR model:

$$S_{t+1} = S_t - B_t, \quad E_{t+1} = E_t + B_t - C_t, \quad I_{t+1} = I_t + C_t - D_t, \qquad R_{t+1} = R_t + D_t.$$

With conditionally independent, binomially distributed random variables:

$$B_t \sim \mathrm{Bin}(S_t, 1 - e^{-h\beta\frac{I_t}{n_t}}), \quad C_t \sim \mathrm{Bin}(E_t, 1 - e^{-h\rho}), \quad D_t \sim \mathrm{Bin}(I_t, 1 - e^{-h\gamma}),$$

where $h > 0$ is a time-step size. With no immigration or emigration, this model is cast as an instance of the model from section 3.1.2 by taking $m = 4$, identifying $\mathbf{x}_t \equiv [S_t\, E_t\, I_t\, R_t]^\top$ and:

$$\mathbf{K}_{t,\boldsymbol{\eta}} = \begin{bmatrix} e^{-h\beta\eta^{(3)}} & 1 - e^{-h\beta\eta^{(3)}} & 0 & 0 \\ 0 & e^{-h\rho} & 1 - e^{-h\rho} & 0 \\ 0 & 0 & e^{-h\gamma} & 1 - e^{-h\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.1}$$

3.1.3 Observation Models

3.1.3.1 Prevalence data

Epidemiological *prevalence data* pertain to the overall levels of susceptibility, exposure and infectivity in the population. In the context of the Latent Compartmental Model, such data are related to the counts of individuals in each compartment at given points in time, i.e., $(\mathbf{x}_t)_{t\geq 1}$. The observation at time $t \geq 1$ is an m-length vector $\mathbf{y}_t$ distributed as follows. With a vector $\mathbf{q}_t \in [0,1]^m$, for each $j \in [m]$ each individual in compartment j is independently detected with probability $q_t^{(j)}$, and the counts of detected individuals are collected in a vector $\bar{\mathbf{y}}_t$, i.e.,

$$\bar{y}_t^{(i)} \sim \mathrm{Bin}(x_t^{(i)}, q_t^{(i)}), \quad i \in [m]. \tag{3.2}$$

With $\mathbf{G}_t$ a row-stochastic matrix of size $m \times m$, each individual detected in compartment j is independently reported in compartment k with probability $G_t^{(j,k)}$. The counts of these reported individuals are collected in an m-length vector $\tilde{\mathbf{y}}_t$. The off-diagonal elements of the matrix $\mathbf{G}_t$ can be interpreted as the probabilities of mis-reporting between compartments. Then the observation $\mathbf{y}_t$ is given by:

$$\mathbf{y}_t = \tilde{\mathbf{y}}_t + \hat{\mathbf{y}}_t,$$

where independently $\hat{\mathbf{y}}_t \sim \mathrm{Pois}(\boldsymbol{\kappa}_t)$ for $\boldsymbol{\kappa}_t \in \mathbb{R}^m_{\geq 0}$, which can be interpreted as additive error counts. In epidemiological data usually only individuals associated with some subset of compartments are detected, and only at certain times. If individuals in say compartment i are not observed at time t, then for inference we will set $y_t^{(i)} = 0$ and $q_t^{(i)} = 0$.

More detailed interpretation of this observation model, in terms of e.g. epidemiological testing of the population, probability of false positives, etc., will be specific to the context in which the Latent Compartmental Model is applied. We provide discussion of this point illustrated by example in section 4.4.

3.1.3.2 Incidence data

Epidemiological measurements often involve data related to the number of newly infective or recovered individuals over given time periods – known as *incidence* data. In order to model such data, generalised to allow for transitions from any compartment to any compartment, we consider an observation at time $t \geq 1$ which is an $m \times m$ matrix $\mathbf{Y}_t$. The elements of $\mathbf{Y}_t$ are conditionally independent given $\mathbf{Z}_t$, and with a matrix $\mathbf{Q}_t \in [0,1]^{m \times m}$,

$$Y_t^{(i,j)} \sim \text{Bin}(Z_t^{(i,j)}, Q_t^{(i,j)}), \quad (i,j) \in [m] \times [m].$$

Similarly to the case of prevalence data, if $Y_t^{(i,j)}$ are missing, then for inference we set $Y_t^{(i,j)} = 0$ and $Q_t^{(i,j)} = 0$. One could extend this model to incorporate mis-reporting and/or additive error counts in a similar manner to in section 3.1.3.1, but for simplicity of presentation we do not do so.

In the context of the SEIR model, for example, the variable $Y_t^{(2,3)}$ models the number of individuals which are newly infective at time t, i.e. the count of the number of individuals which have transitioned $E \to I$ from time $t-1$ to t, subject to random under-reporting parameterised by $Q_t^{(i,j)}$.

3.1.3.3 Aggregated incidence data

In some situations it is desirable to model observations as in section 3.1.3.2, but with transitions of individuals between compartments occurring on a finer time-scale than observations. For example, consider the SEIR model and suppose each discrete time step corresponds to one week. Then the model in (3.1) assigns zero probability to a transition $S \to I$ in one week: in order to transition between $S \to I$, an individual must transit $S \to E$ and then $E \to I$, but at least two discrete time steps are needed for that to occur with positive probability. Similarly, transitions $E \to R$ in one week happen with zero probability. To model incidence data as in section 3.1.3.2 but allowing for these sort of multi-step transitions between observation times, we introduce a sequence of increasing integer observation times $(\tau_r)_{r \geq 1} \subset \mathbb{N}_0$ where $\tau_0 \coloneqq 0$. We then define $\bar{\mathbf{Y}}_r \coloneqq \sum_{t=\tau_{r-1}+1}^{\tau_r} \mathbf{Y}_t$, where $(\mathbf{Y}_t)_{t \geq 1}$ are distributed as per section 3.1.3.2. This model coincides with the model from that section in the case that $\tau_k = k$, we present these two models separately in order to help present a step-by-step explanation in section 3.2 of the corresponding filtering recursions.

In the context of the SEIR model, $\bar{Y}_r^{(2,3)}$ models the total number of individuals which have become infective between times τ_{r-1} and τ_r, subject to random under-reporting. If $\tau_r - \tau_{r-1} \geq 2$, this allows for two-step transitions of the form $S \to E \to I$ or $E \to I \to R$ to occur with positive probability between observations times.

3.2 Filtering recursions and Poisson Approximate Likelihoods

Our next objective is to state and explain the filtering recursions which are used to compute PALs. In section 3.2.1 we give the filtering recursion and PAL for the Latent Compartmental Model combined with the prevalence data model from section 3.1.3.1, we refer to this combination as case (I). In section 3.2.2 we give filtering recursions for a simplified case of the Latent Compartmental Model in which $n_0 = n$ *a.s.* for $n \in \mathbb{N}$, $\boldsymbol{\delta}_t = \mathbf{1}_m$, and $\boldsymbol{\alpha}_t = \mathbf{0}_m$ for all t, i.e. no emigration or immigration, combined with the incidence data model from sections 3.1.3.2 and 3.1.3.3. We refer to this as case (II). We discuss the filtering recursions in case (II) with $\boldsymbol{\delta}_t = \mathbf{1}_m$ and $\boldsymbol{\alpha}_t = \mathbf{0}_m$ only for ease of exposition. By expanding on the derivations we give in the following sections, the reader could obtain without great difficulty the filtering recursions for case (II) in the full generality of the Latent Compartmental Model and in section 6.4 we consider an example involving immigration, emigration and incidence data as an illustration.

Below we state a collection of lemmas which formalise the derivations of the steps in filtering recursions.

3.2.1 Case (I)

In this case, the observations $(\mathbf{y}_t)_{t\geq 1}$ follow the model from section 3.1.3.1. The pair of processes $(\mathbf{x}_t)_{t\geq 0}$ and $(\mathbf{y}_t)_{t\geq 1}$ constitutes a hidden Markov model: $(\mathbf{x}_t)_{t\geq 0}$ is a Markov chain, and $(\mathbf{y}_t)_{t\geq 1}$ are conditionally independent given $(\mathbf{x}_t)_{t\geq 0}$ with the conditional distribution of $\mathbf{y}_t$ given $(\mathbf{x}_t)_{t\geq 0}$ depending only on $\mathbf{x}_t$. Therefore the filtering distributions $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, obey a two-step recursion, with steps canonically referred to as "prediction" and "update":

$$p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \overset{\text{prediction}}{\longrightarrow} p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \overset{\text{update}}{\longrightarrow} p(\mathbf{x}_t|\mathbf{y}_{1:t}),$$

where, for $t \geq 1$,

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \sum_{\mathbf{x}_{t-1}\in\mathbb{N}_0^m} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}), \tag{3.3}$$

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}, \tag{3.4}$$

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \sum_{\mathbf{x}_t\in\mathbb{N}_0^m} p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}),$$

and here and below, by convention, conditioning on $\mathbf{y}_{1:0}$ is understood to mean no conditioning, $p(\cdot|\mathbf{y}_{1:0}) := p(\cdot)$. The marginal likelihood of the observations $\mathbf{y}_1, \ldots, \mathbf{y}_t$ can be written:

$$p(\mathbf{y}_{1:t}) = \prod_{s=1}^{t} p(\mathbf{y}_s|\mathbf{y}_{1:s-1}). \tag{3.5}$$

The general idea of the PAL is to obtain vector-Poisson distribution approximation to each of the terms $p(\mathbf{y}_1)$ and $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$, $t \geq 1$, computed via vector-Poisson approximations to each of the filtering distributions $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ and $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

Approximating the prediction step

For time step $t = 0$ we take a vector-Poisson approximation $\text{Pois}(\boldsymbol{\lambda}_0)$ to the initial distribution $\mathbb{P}_{0,n}$ by setting $\boldsymbol{\lambda}_0 \coloneqq \mathbb{E}_{\mathbf{x}_0 \sim \mathbb{P}_{0,n}}[\mathbf{x}_0]$ and $\bar{\boldsymbol{\lambda}}_0 \coloneqq \boldsymbol{\lambda}_0$. For $t \geq 1$, suppose we have obtained $\bar{\boldsymbol{\lambda}}_{t-1}$ and so defined a vector-Poisson approximation $\text{Pois}(\bar{\boldsymbol{\lambda}}_{t-1})$ to $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. In order to derive a vector-Poisson approximation to $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$, we need to consider the operation (3.3) in more detail, in accordance with the definition of the Latent Compartmental Model. We shall not need an explicit formula for the transition probabilities $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, but rather work with the intermediate quantities $\bar{\mathbf{x}}_{t-1}, \tilde{\mathbf{x}}_t, \hat{\mathbf{x}}_t$ introduced in section 3.1.2.

For $\bar{\mathbf{x}} \in \mathbb{R}^m$ and a length-m probability vector $\boldsymbol{\eta}$, let $M_t(\bar{\mathbf{x}}, \boldsymbol{\eta}, \cdot)$ be the probability mass function of $(\mathbf{1}_m^\top \mathbf{Z})^\top$ where the ith row of $\mathbf{Z} \in \mathbb{N}_0^{m \times m}$ has distribution $\text{Mult}(\bar{x}^{(i)}, \mathbf{K}_{t,\boldsymbol{\eta}}^{(i,\cdot)})$. Then we have:

$$\begin{aligned} p(\tilde{\mathbf{x}}_t|\mathbf{y}_{1:t-1}) &= \sum_{\bar{\mathbf{x}}_{t-1} \in \mathbb{N}_0^m} p(\bar{\mathbf{x}}_{t-1}|\mathbf{y}_{1:t-1}) p(\tilde{\mathbf{x}}_t|\bar{\mathbf{x}}_{t-1}) \\ &= \sum_{\bar{\mathbf{x}}_{t-1} \in \mathbb{N}_0^m} p(\bar{\mathbf{x}}_{t-1}|\mathbf{y}_{1:t-1}) M_t(\bar{\mathbf{x}}_{t-1}, \boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1}), \tilde{\mathbf{x}}_t), \end{aligned} \tag{3.6}$$

where $\bar{\mathbf{x}}_{t-1}$ is related to $\mathbf{x}_{t-1}$ by $\bar{x}_{t-1}^{(i)} \sim \text{Bin}(x_{t-1}^{(i)}, \delta_t^{(i)})$. The summation in (3.6) is too expensive to compute in general. To define an approximation which circumvents this issue, in (3.6) we replace $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ by its approximation $\text{Pois}(\bar{\boldsymbol{\lambda}}_{t-1})$, and replace $\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})$ by $\boldsymbol{\eta}(\mathbb{E}[\bar{\mathbf{x}}_{t-1}])$ where this expectation is under $\bar{\mathbf{x}}_{t-1} \sim \text{Pois}(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)$. Lemma 1 explains the rationale for making the vector-Poisson approximation

$$p(\tilde{\mathbf{x}}_t|\mathbf{y}_{1:t-1}) \approx \text{Pois}\left((\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)}\right).$$

Lemma 1. *Suppose that* $\mathbf{x} \sim \text{Pois}(\boldsymbol{\lambda})$ *for* $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^m$ *and* $\bar{x}^{(i)} \sim \text{Bin}(x^{(i)}, \delta^{(i)})$ *for* $\boldsymbol{\delta} \in [0,1]^m$. *Then* $\bar{\mathbf{x}} \sim \text{Pois}(\boldsymbol{\lambda} \odot \boldsymbol{\delta})$. *Furthermore, if* $\mu(\cdot)$ *is the probability mass function associated with* $\text{Pois}(\boldsymbol{\lambda} \odot \boldsymbol{\delta})$ *and* $\mathbb{E}_\mu[\cdot]$ *is the expected value under* μ, *then* $\sum_{\bar{\mathbf{x}} \in \mathbb{N}_0^m} \mu(\bar{\mathbf{x}}) M_t(\bar{\mathbf{x}}, \boldsymbol{\eta}(\mathbb{E}_\mu[\bar{\mathbf{x}}]), \cdot)$ *is the probability mass function associated with* $\text{Pois}\left((\boldsymbol{\lambda} \odot \boldsymbol{\delta})^\top \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda} \odot \boldsymbol{\delta})}\right)$.

Proof of Lemma 1 For the first result, consider the probability mass function of $\mathbf{x}$:

$$p(\mathbf{x}) = \prod_{j=1}^m \frac{e^{-\lambda^{(j)}} (\lambda^{(j)})^{(x^{(j)})}}{x^{(j)}!},$$

and for $0 \leq \bar{x}^{(j)} \leq x^{(j)}$ $j = 1, \ldots m$,

$$p(\bar{\mathbf{x}} \mid \mathbf{x}) = \prod_{j=1}^m \frac{x^{(j)}!}{\bar{x}^{(j)}! (x^{(j)} - \bar{x}^{(j)})!} (\delta^{(j)})^{\bar{x}^{(j)}} (1 - \delta^{(j)})^{x^{(j)} - \bar{x}^{(j)}},$$

So that

$$p(\mathbf{x}, \bar{\mathbf{x}}) = \prod_{j=1}^m \frac{e^{-\lambda^{(j)}} (\lambda^{(j)})^{x^{(j)}} (\delta^{(j)})^{\bar{x}^{(j)}} (1 - \delta^{(j)})^{x^{(j)} - \bar{x}^{(j)}}}{\bar{x}^{(j)}! (x^{(j)} - \bar{x}^{(j)})!},$$

and

$$\begin{aligned}
p(\bar{\mathbf{x}}) &= \sum_{x^{(i)}\geq\bar{x}^{(i)};i\in[m]}\prod_{j=1}^{m}\frac{e^{-\lambda^{(j)}}(\lambda^{(j)})^{x^{(j)}}(\delta^{(j)})^{\bar{x}^{(j)}}(1-\delta^{(j)})^{x^{(j)}-\bar{x}^{(j)}}}{\bar{x}^{(j)}!(x^{(j)}-\bar{x}^{(j)})!}\\
&= \left(\prod_{j=1}^{m}\frac{e^{-\lambda^{(j)}}(\delta^{(j)}\lambda^{(j)})^{\bar{x}^{(j)}}}{\bar{x}^{(j)}!}\right)\sum_{x^{(i)}\geq\bar{x}^{(i)};i\in[m]}\prod_{j=1}^{m}\frac{(\lambda^{(j)})^{(x^{(j)}-\bar{x}^{(j)})}(1-\delta^{(j)})^{(x^{(j)}-\bar{x}^{(j)})}}{(x^{(j)}-\bar{x}^{(j)})!}\\
&= \left(\prod_{j=1}^{m}\frac{e^{-\lambda^{(j)}}(\delta^{(j)}\lambda^{(j)})^{\bar{x}^{(j)}}}{\bar{x}^{(j)}!}\right)e^{\lambda^{(j)}(1-\delta^{(j)})}\\
&= \prod_{j=1}^{m}\frac{e^{-\lambda^{(j)}\delta^{(j)}}(\delta^{(j)}\lambda^{(j)})^{\bar{x}^{(j)}}}{\bar{x}^{(j)}!},
\end{aligned}$$

which is the probability mass function associated with $\text{Pois}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})$.

Now consider $\mathbf{x}'\sim M_t(\bar{\mathbf{x}},\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta}),\cdot)$ where $\bar{\mathbf{x}}\sim\mu$, so that $\sum_{\bar{\mathbf{x}}\in\mathbb{N}_0^m}\mu(\bar{\mathbf{x}})M_t(\bar{\mathbf{x}},\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta}),\cdot)$ is the marginal probability mass function of $\mathbf{x}'$. By the definition of M_t, $\mathbf{x}'=(\mathbf{1}_m^\top\mathbf{Z})$, where the rows of $\mathbf{Z}$ are conditionally independent given $\bar{\mathbf{x}}$, and the ith row of $\mathbf{Z}$ is distributed $\text{Mult}(\bar{x}^{(i)},\mathbf{K}^{(i,\cdot)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})})$. Now we can write the moment generating function (m.g.f.) of $\mathbf{x}'$ as:

$$\begin{aligned}
\mathscr{M}_{\mathbf{x}'}(\mathbf{b}) &= \mathbb{E}\left[\exp\left(\mathbf{1}_m^\top\mathbf{Z}^\top\mathbf{b}\right)\right]\\
&= \mathbb{E}\left[\exp\left(\sum_{i,j=1}^{m}Z^{(i,j)}b^{(j)}\right)\right]\\
&= \mathbb{E}\left[\prod_{j=1}^{m}\exp\left(\sum_{i=1}^{m}Z^{(i,j)}b^{(j)}\right)\right]\\
&= \mathbb{E}\left\{\prod_{i=1}^{m}\mathbb{E}\left[\exp\left(\sum_{j=1}^{m}Z^{(i,j)}b^{(j)}\right)\middle|\bar{\mathbf{x}}\right]\right\}.
\end{aligned}$$

Now we notice that $\mathbb{E}\left[\exp\left(\sum_{i=1}^{m}Z^{(i,j)}b^{(i)}\right)\middle|\bar{\mathbf{x}}\right]$ is the m.g.f. of $\text{Mult}(\bar{x}^{(i)},\mathbf{K}^{(i,\cdot)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})})$ so that

$$\begin{aligned}
\mathscr{M}_{\mathbf{x}'}(\mathbf{b}) &= \mathbb{E}\left\{\prod_{i=1}^{m}\left[\sum_{j=1}^{m}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}e^{b^{(i)}}\right]^{\bar{x}^{(j)}}\right\}\\
&= \sum_{\bar{x}^{(1)},\ldots,\bar{x}^{(m)}\in\mathbb{N}_0^m}\prod_{i=1}^{m}\left[\sum_{j=1}^{m}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}e^{b^{(j)}}\right]^{\bar{x}^{(i)}}\frac{e^{-\lambda^{(i)}\delta^{(i)}}(\lambda^{(i)}\delta^{(i)})^{\bar{x}^{(i)}}}{\bar{x}^{(i)}!}\\
&= \left(\prod_{i=1}^{m}e^{-\lambda^{(i)}\delta^{(i)}}\right)\sum_{\bar{x}^{(1)},\ldots,\bar{x}^{(m)}\in\mathbb{N}_0^m}\prod_{i=1}^{m}\frac{1}{\bar{x}^{(i)}!}\left[\sum_{j=1}^{m}\lambda^{(i)}\delta^{(i)}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}e^{b^{(j)}}\right]^{\bar{x}^{(i)}}\\
&= \prod_{i=1}^{m}\exp\left(-\lambda^{(i)}\delta^{(i)}+\sum_{j=1}^{m}\lambda^{(i)}\delta^{(i)}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}e^{b^{(j)}}\right)\\
&= \exp\left\{\sum_{i=1}^{m}\left(-\lambda^{(i)}\delta^{i)}\sum_{j=1}^{m}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}+\sum_{j=1}^{m}\lambda^{(i)}\delta^{(i)}K^{(i,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}e^{b^{(i)}}\right)\right\}\\
&= \prod_{j=1}^{m}\exp\left\{\left((\boldsymbol{\lambda}\odot\boldsymbol{\delta})^\top\mathbf{K}^{(\cdot,j)}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}\right)\left(e^{b^{(j)}}-1\right)\right\}.
\end{aligned}$$

We recognise this is the moment generating function of a $\text{Pois}\left((\boldsymbol{\lambda}\odot\boldsymbol{\delta})^\top \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}\odot\boldsymbol{\delta})}\right)$ random vector. ■

As per the definition of the Latent Compartmental Model, $\mathbf{x}_t$ is obtained by summing $\tilde{\mathbf{x}}_t$ with $\hat{\mathbf{x}}_t$ where $\hat{\mathbf{x}}_t \sim \text{Pois}(\boldsymbol{\alpha}_t)$. Since the sum of independent Poisson random variables is also Poisson with intensity given by the sum of the intensities, we then take the approximation

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \approx \text{Pois}(\boldsymbol{\lambda}_t), \quad \text{with} \quad \boldsymbol{\lambda}_t := (\bar{\boldsymbol{\lambda}}_{t-1}\odot\boldsymbol{\delta}_t)^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1}\odot\boldsymbol{\delta}_t)} + \boldsymbol{\alpha}_t.$$

Approximating the update step

In order to obtain a vector-Poisson approximation to $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ we substitute $\text{Pois}(\boldsymbol{\lambda}_t)$ in place of $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ in (3.4), which can be viewed as an application of Bayes' rule, and we shall define $\bar{\boldsymbol{\lambda}}_t$ to be the mean vector of the resulting distribution. Lemma 2 can be applied to calculate $\bar{\boldsymbol{\lambda}}_t$ in accordance with this recipe, leading us to:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \approx \text{Pois}(\bar{\boldsymbol{\lambda}}_t), \quad \bar{\boldsymbol{\lambda}}_t := [\mathbf{1}_m - \mathbf{q}_t + (\{\mathbf{y}_t^\top \oslash [(\mathbf{q}_t\odot\boldsymbol{\lambda}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top]\}[(\mathbf{1}_m\otimes\mathbf{q}_t)\odot\mathbf{G}_t^\top])^\top]\odot\boldsymbol{\lambda}_t,$$

Lemma 2 also tells us how to obtain a vector-Poisson approximation to $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$.

Lemma 2. *Suppose that* $\mathbf{x} \sim \text{Pois}(\boldsymbol{\lambda})$ *for given* $\boldsymbol{\lambda} \in \mathbb{R}^m_{\geq 0}$ *and let* $\bar{\mathbf{y}}$ *be a vector with conditionally independent elements distributed* $\bar{y}^{(i)} \sim \text{Bin}(x^{(i)}, q^{(i)})$ *for given* $\mathbf{q} \in [0,1]^m$. *For* $\mathbf{G}$ *a row-stochastic* $m \times m$ *matrix and* $\mathbf{M}$ *an* $m \times m$ *matrix with rows distributed* $\mathbf{M}^{(i,\cdot)} \sim \text{Mult}(\bar{y}^{(i)}, \mathbf{G}^{(i,\cdot)})$, *let* $\tilde{\mathbf{y}} := \sum_{i=1}^m \mathbf{M}^{(i,\cdot)}$ *and* $\mathbf{y} := \tilde{\mathbf{y}} + \hat{\mathbf{y}}$ *where* $\hat{\mathbf{y}} \sim \text{Pois}(\boldsymbol{\kappa})$ *for a given* $\boldsymbol{\kappa} \in \mathbb{R}^m_{\geq 0}$. *Then:*

$$\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] = [\mathbf{1}_m - \mathbf{q} + (\{\mathbf{y}^\top \oslash [(\mathbf{q}\odot\boldsymbol{\lambda})^\top \mathbf{G} + \boldsymbol{\kappa}^\top]\}[(\mathbf{1}_m\otimes\mathbf{q})\odot\mathbf{G}^\top])^\top]\odot\boldsymbol{\lambda}. \tag{3.7}$$

and $\mathbf{y} \sim \text{Pois}([(\boldsymbol{\lambda}\odot\mathbf{q})^\top\mathbf{G}]^\top + \boldsymbol{\kappa})$, *i.e.,*

$$\log p(\mathbf{y}) = -[(\boldsymbol{\lambda}_t\odot\mathbf{q})^\top\mathbf{G} + \boldsymbol{\kappa}^\top]\mathbf{1}_m + \mathbf{y}^\top \log([(\boldsymbol{\lambda}\odot\mathbf{q})^\top\mathbf{G}]^\top + \boldsymbol{\kappa}) - \mathbf{1}_m^\top \log(\mathbf{y}!),$$

with the convention $0\log 0 := 0$.

Proof of Lemma 2 We have $\bar{\mathbf{y}} \sim \text{Pois}(\boldsymbol{\lambda}\odot\mathbf{q})$ by the same reasoning as lemma 1. By definition $\hat{\mathbf{y}} = \mathbf{1}_m^\top\mathbf{M}$, hence the moment generating function of $\hat{\mathbf{y}}$ is:

$$\begin{aligned}
\mathscr{M}_{\hat{\mathbf{y}}}(\mathbf{b}) &= \mathbb{E}\left[\exp\left(\mathbf{1}_m^\top\mathbf{M}^\top\mathbf{b}\right)\right] \\
&= \mathbb{E}\left\{\prod_{i=1}^m \mathbb{E}\left[\exp\left(\sum_{j=1}^m M^{(i,j)}b^{(j)}\right)\middle|\bar{\mathbf{y}}\right]\right\} \\
&= \mathbb{E}\left\{\prod_{i=1}^m \left[\sum_{j=1}^m G^{(i,j)}e^{b^{(i)}}\right]^{\bar{y}^{(j)}}\right\} \\
&= \sum_{\bar{x}^{(1)},\ldots,\bar{x}^{(m)}\in\mathbb{N}_0^m} \prod_{i=1}^m \left[\sum_{j=1}^m G^{(i,j)}e^{b^{(j)}}\right]^{\bar{x}^{(i)}} \frac{e^{-\lambda^{(i)}q^{(i)}}(\lambda^{(i)}q^{(i)})^{\bar{x}^{(i)}}}{\bar{x}^{(i)}!} \\
&= \prod_{j=1}^m \exp\left\{\left((\boldsymbol{\lambda}\odot\boldsymbol{q})^\top\mathbf{G}^{(\cdot,j)}\right)\left(e^{b^{(j)}}-1\right)\right\}.
\end{aligned}$$

Which we recognise as the moment generating function of the Pois$((\boldsymbol{\lambda} \odot \mathbf{q})^\top \mathbf{G})$, the first result of the lemma then follows from applying element-wise the fact that the intensity of the sum of two independent Poisson random variables is the sum of the intensities.

We start the proof of (3.7) by considering the decomposition of $\mathbf{x}$ into the sum of random variables $\bar{\mathbf{y}}$ and $\check{\mathbf{x}}$ where $\check{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{y}}$. Then, $\bar{\mathbf{y}}$ and $\check{\mathbf{x}}$ are independent Poisson with intensity vectors $\mathbf{q} \odot \boldsymbol{\lambda}$ and $(\mathbf{1}_m - \mathbf{q}) \odot \boldsymbol{\lambda}$ respectively, see Kingman (1992)[Sec. 1.2]. Since $\check{\mathbf{x}}$ is independent of $\mathbf{y}$, we have that:

$$\mathbb{E}\left[\mathbf{x} \mid \mathbf{y}\right] = \left[\mathbf{1}_m - \mathbf{q}\right] \odot \boldsymbol{\lambda} + \mathbb{E}\left[\bar{\mathbf{y}} \mid \mathbf{y}\right]. \tag{3.8}$$

So, we need to characterise the distribution of $\bar{\mathbf{y}}$ given $\mathbf{y}$. Construct the random variable $\Xi \in \mathbb{N}_0^{(m+1)\times m}$ such that for $i, j \in [m]$, $\Xi^{(i,j)} = M^{(i,j)}$ and row $m+1$ of Ξ are the counts $\hat{\mathbf{y}} \sim \text{Pois}(\boldsymbol{\kappa})$. By this construction, $\sum_{j=1}^{m} \Xi^{(i,j)} = \bar{y}^{(i)}$ for $i = 1, \ldots, m$ and $\sum_{i=1}^{m+1} \Xi^{(i,j)} = y^{(j)}$ for $j = 1, \ldots, m$. Furthermore, the elements of Ξ are independently Poisson, see Kingman (1992)[Sec. 1.2], with intensity matrix $\Lambda \in \mathbb{R}^{(m+1)\times m}$ defined as follows:

$$\begin{aligned} \Lambda^{(i,j)} &= \lambda^{(i)} q^{(i)} G^{(i,j)} \quad \text{for} i = 1, \ldots m, \quad j = 1, \ldots m \\ \Lambda^{(m+1,j)} &= \kappa^{(j)} \quad \text{for} j = 1, \ldots m. \end{aligned}$$

If, for some $j, k \in [m]$, $\sum_{i=1}^{m+1} \Lambda^{(i,j)} = 0$, then we must have that $\Lambda^{(i,j)} = 0$ for all $i = 1, \ldots, m+1$ so that $\Xi^{(i,j)} = 0 \quad a.s.$. Otherwise we have that for $i = 1, \ldots, m+1$ and $j \in [m]$, $\Xi^{(i,j)}$ conditioned on $\sum_{k=1}^{m+1} \Xi^{(k,j)} = y^{(j)}$ is distributed

$$\text{Bin}\left(y^{(j)}, \frac{\Lambda^{(i,j)}}{\sum_{k=1}^{m+1} \Lambda^{(k,j)}}\right).$$

Hence, given $\mathbf{y}$, $\bar{y}^{(i)}$ has a Poisson-Binomial distribution with mean:

$$\mathbb{E}\left[\bar{y}^{(i)} \mid \mathbf{y}\right] = \mathbb{E}\left[\sum_{j=1}^{m} \Xi^{(i,j)} \mid \mathbf{y}\right] = \sum_{j=1}^{m} y^{(j)} \frac{\lambda^{(i)} q^{(i)} G^{(i,j)}}{\sum_{k=1}^{m+1} \lambda^{(k)} q^{(k)} G^{(k,j)} + \kappa^{(j)}}, \tag{3.9}$$

for $i = 1, \ldots, m$, where we set the jth term of the outer sum on the r.h.s to 0 if $\sum_{k=1}^{m+1} \lambda^{(k)} q^{(k)} G^{(k,j)} + \kappa^{(j)} = 0$ since that achieves

$$\mathbb{E}\left[\Xi^{(i,j)} \mid \mathbf{y}\right] = 0.$$

Writing (3.9) in vector form and substituting into (3.8) completes the proof.

■

Computing the PAL

Gathering together the approximations discussed above we arrive at the following algorithm.

Algorithm 6 Filtering for case (I)

initialise: $\bar{\boldsymbol{\lambda}}_0 \leftarrow \boldsymbol{\lambda}_0$
1: **for** $t \geq 1$:
2: $\quad \boldsymbol{\lambda}_t \leftarrow [(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)}]^\top + \boldsymbol{\alpha}_t$
3: $\quad \bar{\boldsymbol{\lambda}}_t \leftarrow [\mathbf{1}_m - \mathbf{q}_t + (\{\mathbf{y}_t^\top \oslash [(\mathbf{q}_t \odot \boldsymbol{\lambda}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top]\}[(\mathbf{1}_m \otimes \mathbf{q}_t) \odot \mathbf{G}_t^\top])^\top] \odot \boldsymbol{\lambda}_t$
4: $\quad \boldsymbol{\mu}_t \leftarrow [(\boldsymbol{\lambda}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t]^\top + \boldsymbol{\kappa}_t$
5: $\quad \ell(\mathbf{y}_t | \mathbf{y}_{1:t-1}) \leftarrow -\boldsymbol{\mu}_t^\top \mathbf{1}_m + \mathbf{y}_t^\top \log(\boldsymbol{\mu}_t) - \mathbf{1}_m^\top \log(\mathbf{y}_t!)$
6: **end for**

If, at line 3 of algorithm 6, we encounter 0/0 in performing the element-wise division operation we set the vector element in question to 0, which is in accordance with $p(\mathbf{x}_t|\mathbf{y}_{1:t}) \approx \text{Pois}(\bar{\boldsymbol{\lambda}}_t)$. At line 5 of algorithm 6 we apply the convention $0 \log 0 := 0$, in accordance with $p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \approx \text{Pois}(\boldsymbol{\mu}_t)$.

Mimicking (3.5), the log PAL associated with algorithm 6 is:

$$\log p(\mathbf{y}_{1:t}) \approx \sum_{s=1}^{t} \ell(\mathbf{y}_s|\mathbf{y}_{1:s-1}), \tag{3.10}$$

It is important to note that the term $\mathbf{1}_m^\top \log(\mathbf{y}_t!)$ in $\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ calculated in algorithm 6 has no dependence on the ingredients of the model, i.e., $\mathbf{K}_{t,\boldsymbol{\eta}}$, $\boldsymbol{\kappa}_t$, etc. and so in practice if one is computing the PAL in order to maximise it with respect to parameters of the model, or evaluate PAL ratios for different parameter values, the term $\mathbf{1}_m^\top \log(\mathbf{y}_t!)$ never needs to be computed.

3.2.2 Case (II)

In this case we consider the Latent Compartmental Model with $n_0 = n$ with probability 1, $\boldsymbol{\delta}_t = \mathbf{1}_m$ and $\boldsymbol{\alpha}_t = \mathbf{0}_m$ for all t, i.e. no emigration or immigration, and with the observations $(\tilde{\mathbf{Y}}_r)_{r\geq 1}$ following the model from section 3.1.3.3. For ease of exposition we start with the special case that $(\tau_r)_{r\geq 1} = \mathbb{N}$, in which case $(\tilde{\mathbf{Y}}_r)_{r\geq 1} \equiv (\mathbf{Y}_t)_{t\geq 1}$ and the model from section 3.1.3.3 reduces to that from section 3.1.3.2.

To derive the filtering recursions we follow a similar programme to case (I), starting from the fact that the pair of processes $(\mathbf{Z}_t)_{t\geq 1}$ and $(\mathbf{Y}_t)_{t\geq 1}$ constitutes a hidden Markov model, and approximating the following prediction and update operations:

$$p(\mathbf{Z}_{t-1}|\mathbf{Y}_{1:t-1}) \overset{\text{prediction}}{\longrightarrow} p(\mathbf{Z}_t|\mathbf{Y}_{1:t-1}) \overset{\text{update}}{\longrightarrow} p(\mathbf{Z}_t|\mathbf{Y}_{1:t}).$$

Approximating the prediction step when $(\tau_r)_{r\geq 1} = \mathbb{N}$

For $\mathbf{Z} \in \mathbb{N}_0^{m\times m}$ and a length-m probability vector $\boldsymbol{\eta}$, let $\bar{M}_t(\mathbf{Z}, \boldsymbol{\eta}, \cdot)$ be the probability mass function of a random $m \times m$ matrix, say $\tilde{\mathbf{Z}}$, such that $\mathbf{1}_m^\top \mathbf{Z} = (\tilde{\mathbf{Z}}\mathbf{1}_m)^\top$ with probability 1 and such that given the row sums $\tilde{\mathbf{Z}}\mathbf{1}_m = \mathbf{x}$, the rows of $\tilde{\mathbf{Z}}$ are conditionally independent with the conditional distribution of the i^{th} row being $\text{Mult}(x^{(i)}, \mathbf{K}_{t,\boldsymbol{\eta}}^{(i,\cdot)})$. By construction $\bar{M}_t(\mathbf{Z}_{t-1}, \boldsymbol{\eta}(\mathbf{1}_m^\top \mathbf{Z}_{t-1}), \mathbf{Z}_t)$ is equal

to $p(\mathbf{Z}_t|\mathbf{Z}_{t-1})$ for case (II), hence

$$\begin{aligned} p(\mathbf{Z}_t|\mathbf{Y}_{1:t-1}) &= \sum_{\mathbf{Z}_{t-1}\in\mathbb{N}_0^{m\times m}} p(\mathbf{Z}_{t-1}|\mathbf{Y}_{1:t-1})p(\mathbf{Z}_t|\mathbf{Z}_{t-1}) \\ &= \sum_{\mathbf{Z}_{t-1}\in\mathbb{N}_0^{m\times m}} p(\mathbf{Z}_{t-1}|\mathbf{Y}_{1:t-1})M_t(\mathbf{Z}_{t-1},\boldsymbol{\eta}(\mathbf{1}_m^\top\mathbf{Z}_{t-1}),\mathbf{Z}_t). \end{aligned} \tag{3.11}$$

Assuming we have already computed $\bar{\boldsymbol{\Lambda}}_{t-1}$ such that $p(\mathbf{Z}_{t-1}|\mathbf{Y}_{1:t-1}) \approx \text{Pois}(\bar{\boldsymbol{\Lambda}}_{t-1})$, we substitute this approximation in to (3.11) and replace $\boldsymbol{\eta}(\mathbf{1}_m^\top\mathbf{Z}_{t-1})$ by $\boldsymbol{\eta}(\mathbb{E}[\mathbf{1}_m^\top\mathbf{Z}_{t-1}])$ where this expectation is under $\mathbf{Z}_{t-1} \sim \text{Pois}(\bar{\boldsymbol{\Lambda}}_{t-1})$. Lemma 3 explains the rationale for then making the approximation:

$$p(\mathbf{Z}_t|\mathbf{Y}_{1:t-1}) \approx \text{Pois}(\boldsymbol{\Lambda}_t), \quad \Lambda_t := (\bar{\boldsymbol{\lambda}}_{t-1}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1})}, \quad \bar{\boldsymbol{\lambda}}_{t-1}^\top := \mathbf{1}_m^\top\bar{\boldsymbol{\Lambda}}_{t-1}.$$

Lemma 3. *If for a given $m\times m$ matrix Λ, $\bar{\mu}$ is the probability mass function associated with* $\text{Pois}(\Lambda)$ *and $\mathbb{E}_{\bar{\mu}}[\mathbf{1}_m^\top\mathbf{Z}]$ is the expected value of $\mathbf{1}_m^\top\mathbf{Z}$ where $\mathbf{Z}\sim\bar{\mu}$, then $\sum_{\mathbf{Z}\in\mathbb{N}_0^{m\times m}}\bar{\mu}(\mathbf{Z})\bar{M}_t(\mathbf{Z},\boldsymbol{\eta}(\mathbb{E}_{\bar{\mu}}[\mathbf{1}_m^\top\mathbf{Z}]),\cdot)$ is the probability mass function associated with* $\text{Pois}((\boldsymbol{\lambda}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})})$*, where $\boldsymbol{\lambda}^\top := \mathbf{1}_m^\top\boldsymbol{\Lambda}$.*

Proof of Lemma 3 Note $\boldsymbol{\eta}(\mathbb{E}_{\bar{\mu}}[\mathbf{1}_m^\top\mathbf{Z}]) = \boldsymbol{\eta}(\mathbf{1}_m^\top\boldsymbol{\Lambda}) = \boldsymbol{\eta}(\boldsymbol{\lambda}^\top)$. Let $\tilde{\mathbf{Z}}\sim\bar{M}_t(\mathbf{Z},\boldsymbol{\eta}(\boldsymbol{\lambda}),\cdot)$, then the moment generating function for $\tilde{\mathbf{Z}}$ is:

$$\begin{aligned} \mathbb{E}\left[\exp(\mathbf{1}_m^\top(\tilde{\mathbf{Z}}\odot\mathbf{B})\mathbf{1}_m)\right] &= \mathbb{E}\left[\prod_{i=1}^m\exp\left(\sum_{j=1}^m\tilde{Z}^{(i,j)}b^{(i,j)}\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^m\exp\left(\sum_{j=1}^m\tilde{Z}^{(i,j)}b^{(i,j)}\right)\middle|\mathbf{Z}\right]\right] \\ &= \mathbb{E}\left[\prod_{i=1}^m\mathbb{E}\left[\exp\left(\sum_{j=1}^m\tilde{Z}^{(i,j)}b^{(i,j)}\right)\middle|x^{(i)}\right]\right] \\ &= \mathbb{E}\left[\prod_{i=1}^m\left(\sum_{j=1}^m K_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})}^{(i,j)}e^{b^{(i,j)}}\right)^{x^{(i)}}\right] \\ &= \left(\prod_{i=1}^m e^{-\lambda^{(i)}}\right)\sum_{(x^{(1)},\ldots,x^{(m)})\in\mathbb{N}_0}\left(\sum_{j=1}^m K_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})}^{(i,j)}e^{b^{(i,j)}}\lambda^{(i)}\right)^{x^{(i)}}\frac{1}{x^{(i)}!} \\ &= \prod_{i=1}^m\exp\{-\lambda^{(i)}\sum_{j=1}^m K_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})}^{(i,j)}(1-e^{b^{(i,j)}})\} \\ &= \prod_{i,j=1}^m\exp\{-\lambda^{(i)}K_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})}^{(i,j)}(1-e^{b^{(i,j)}})\}, \end{aligned}$$

which we recognise as the moment generating function of a $\text{Pois}((\boldsymbol{\lambda}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda})})$ random matrix. ■

Approximating the update step when $(\tau_r)_{r\geq1} = \mathbb{N}$

We now apply Bayes' rule to $\text{Pois}(\boldsymbol{\Lambda}_t)$ and shall define $\bar{\boldsymbol{\Lambda}}_t$ to be the mean vector of the resulting distribution. Lemma 4 shows how to do this, leading to:

$$p(\mathbf{Z}_t|\mathbf{Y}_{1:t}) \approx \text{Pois}(\bar{\boldsymbol{\Lambda}}_t), \quad \bar{\boldsymbol{\Lambda}}_t := \mathbf{Y}_t + \Lambda_t\odot(\mathbf{1}_m\otimes\mathbf{1}_m - \mathbf{Q}_t),$$

Lemma 4. *Suppose that* $\mathbf{Z} \sim \text{Pois}(\Lambda)$ *for some* $\Lambda \in \mathbb{R}_{\geq 0}^{m \times m}$*, and that for some* $\mathbf{Q} \in \mathbb{R}_{\geq 0}^{m \times m}$*, given* $\mathbf{Z}$*,* $\mathbf{Y}$ *is a matrix with conditionally independent entries distributed:* $y^{(i,j)} \sim \text{Bin}(Z^{(i,j)}, q^{(i,j)})$*, then the conditional distribution of* $\mathbf{Z}$ *given* $\mathbf{Y}$ *is that of* $\mathbf{Y} + \mathbf{Z}^*$ *where:*

$$\mathbf{Z}^* \sim \text{Pois}\left(\Lambda \odot (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q})\right),$$

i.e.,

$$\mathbb{E}[\mathbf{Z}|\mathbf{Y}] = \mathbf{Y} + \Lambda \odot (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}),$$

and $\mathbf{Y} \sim \text{Pois}(\Lambda \odot \mathbf{Q})$*, i.e,*

$$\log p(\mathbf{Y}) = \mathbf{1}_m^\top (\Lambda \odot \mathbf{Q}) \mathbf{1}_m + \mathbf{1}_m^\top [\mathbf{Y} \odot \log(\Lambda \odot \mathbf{Q})] \mathbf{1}_m - \mathbf{1}_m^\top \log(\mathbf{Y}!) \mathbf{1}_m,$$

with the convention $0 \log 0 := 0$*.*

Proof of Lemma 4 We have:

$$p(\mathbf{Z}) = \prod_{i,j=1}^{m} \frac{e^{-\Lambda^{(i,j)}} (\Lambda^{(i,j)})^{Z^{(i,j)}}}{Z^{(i,j)}!},$$

furthermore:

$$p(\mathbf{Y}|\mathbf{Z}) = \prod_{i,j=1}^{m} \frac{Z^{(i,j)}!}{Y^{(i,j)}!(Z^{(i,j)} - Y^{(i,j)})!} {Q^{(i,j)}}^{Y^{(i,j)}} (1 - Q^{(i,j)})^{Z^{(i,j)} - Y^{(i,j)}}.$$

So that:

$$p(\mathbf{Z}, \mathbf{Y}) = \prod_{i,j=1}^{m} \frac{{Q^{(i,j)}}^{Y^{(i,j)}} (1 - Q^{(i,j)})^{Z^{(i,j)} - Y^{(i,j)}} e^{-\Lambda^{(i,j)}} (\Lambda^{(i,j)})^{Z^{(i,j)}}}{Y^{(i,j)}!(Z^{(i,j)} - Y^{(i,j)})!},$$

and

$$\begin{aligned}
p(\mathbf{Y}) &= \sum_{\{Z^{(i,j)} : Z^{(i,j)} \geq Y^{(i,j)}\}} \prod_{i,j=1}^{m} \frac{{Q^{(i,j)}}^{Y^{(i,j)}} (1 - Q^{(i,j)})^{Z^{(i,j)} - Y^{(i,j)}} e^{-\Lambda^{(i,j)}} (\Lambda^{(i,j)})^{Z^{(i,j)}}}{Y^{(i,j)}!(Z^{(i,j)} - Y^{(i,j)})!} \\
&= \prod_{i,j=1}^{m} \frac{e^{-\Lambda^{(i,j)}} (Q^{(i,j)} \Lambda^{(i,j)})^{Y^{(i,j)}}}{Y^{(i,j)}!} \sum_{Z^{(i,j)} - Y^{(i,j)} \geq 0} \frac{(\Lambda^{(i,j)}(1 - Q^{(i,j)}))^{Z^{(i,j)} - Y^{(i,j)}}}{(Z^{(i,j)} - Y^{(i,j)})!} \\
&= \prod_{i,j=1}^{m} \frac{e^{-\Lambda^{(i,j)}} (Q^{(i,j)} \Lambda^{(i,j)})^{Y^{(i,j)}}}{Y^{(i,j)}!} e^{\Lambda^{(i,j)}(1 - Q^{(i,j)})} \\
&= \prod_{i,j=1}^{m} \frac{e^{-\Lambda^{(i,j)} Q^{(i,j)}} (Q^{(i,j)} \Lambda^{(i,j)})^{Y^{(i,j)}}}{Y^{(i,j)}!}.
\end{aligned}$$

Dividing $p(\mathbf{Z}, \mathbf{Y})$ by $p(\mathbf{Y})$ gives:

$$p(\mathbf{Z} \mid \mathbf{Y}) = \prod_{i,j=1}^{m} \frac{e^{-\Lambda^{(i,j)}(1 - Q^{(i,j)})}}{(Z^{(i,j)} - Y^{(i,j)})!} (\Lambda^{(i,j)}(1 - Q^{(i,j)}))^{Z^{(i,j)} - Y^{(i,j)}}.$$

Giving the desired probability mass function of $\mathbf{Y} + \mathbf{Z}^*$. ■

Computing the PAL when $(\tau_r)_{r\geq 1} = \mathbb{N}$

Combining the above prediction and update approximations we arrive at algorithm 7.

Algorithm 7 Filtering for case (II) when $(\tau_r)_{r\geq 1} = \mathbb{N}$

initialise: $\bar{\boldsymbol{\lambda}}_0 \leftarrow \boldsymbol{\lambda}_0$
1: **for** $t \geq 1$:
2: $\quad \Lambda_t \leftarrow (\bar{\boldsymbol{\lambda}}_t \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_t)}$
3: $\quad \bar{\Lambda}_t \leftarrow \mathbf{Y}_t + (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_t) \odot \boldsymbol{\Lambda}_t$
4: $\quad \mathscr{L}(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}) \leftarrow -\mathbf{1}_m^\top(\Lambda_t \odot \mathbf{Q}_t)\mathbf{1}_m + \mathbf{1}_m^\top[\mathbf{Y}_t \odot \log(\Lambda_t \odot \mathbf{Q}_t)]\mathbf{1}_m - \mathbf{1}_m^\top \log(\mathbf{Y}_t!)\mathbf{1}_m$
5: $\quad \bar{\boldsymbol{\lambda}}_t \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_t)^\top$
6: **end for**

In algorithm 7 we adopt the same convention $0 \log 0 := 0$ as in algorithm 6. The log PAL associated with algorithm 7 is:

$$\log p(\mathbf{Y}_{1:t}) \approx \sum_{s=1}^{t} \mathscr{L}(\mathbf{Y}_s|\mathbf{Y}_{1:s-1}).$$

We now consider general $(\tau_r)_{r\geq 1}$. The filtering recursion is:

$$\begin{aligned} p(\mathbf{Z}_{\tau_{r-1}}|\bar{\mathbf{Y}}_{1:r-1}) &\overset{\text{prediction}}{\longrightarrow} p(\mathbf{Z}_{\tau_{r-1}+1}|\bar{\mathbf{Y}}_{1:r-1}) \overset{\text{prediction}}{\longrightarrow} \dots \\ &\overset{\text{prediction}}{\longrightarrow} p(\mathbf{Z}_{\tau_r}|\bar{\mathbf{Y}}_{1:r-1}) \overset{\text{update}}{\longrightarrow} p(\mathbf{Z}_{\tau_r}|\bar{\mathbf{Y}}_{1:r}). \end{aligned} \tag{3.12}$$

Approximating the prediction and update steps for general $(\tau_r)_{r\geq 1}$

Assuming that we are given $\Lambda_{\tau_{r-1}}$ such that $p(\mathbf{Z}_{\tau_{r-1}}|\bar{\mathbf{Y}}_{1:r-1}) \approx \text{Pois}(\Lambda_{\tau_{r-1}})$, each of the prediction steps in (3.12) is approximated by applying lemma 3, leading to lines 2-6 of algorithm 8. To approximate the update step, applying lemma 5 leads to lines 7-10 of algorithm 8.

Lemma 5. *For* $\boldsymbol{\lambda}_0 \in \mathbb{R}_{\geq 0}^m$ *and* $\tau \in \mathbb{N}$*, define:*

$$\boldsymbol{\Lambda}_t := (\boldsymbol{\lambda}_{t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}_{t-1})}, \quad \boldsymbol{\lambda}_t := (\mathbf{1}_m^\top \boldsymbol{\Lambda}_t)^\top, \quad t = 1, \dots, \tau,$$

and let $(\mathbf{Z}_t)_{t=1}^\tau$ *be independent with* $\mathbf{Z}_t \sim \text{Pois}(\boldsymbol{\Lambda}_t)$*. Suppose that given* $\mathbf{Z}_t$*,* $\mathbf{Y}_t$ *is a matrix with conditionally independent entries distributed* $Y_t^{(i,j)} \sim \text{Bin}(Z_t^{(i,j)}, Q^{(i,j)})$*, and let* $\bar{\mathbf{Y}} := \sum_{s=1}^\tau \mathbf{Y}_s$*. Then:*

$$\mathbb{E}\left[\mathbf{Z}_\tau|\bar{\mathbf{Y}}\right] = (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_\tau) \odot \boldsymbol{\Lambda}_t + \bar{\mathbf{Y}} \odot \boldsymbol{\Lambda}_\tau \odot \mathbf{Q}_\tau \oslash \left(\sum_{t=1}^{\tau} \boldsymbol{\Lambda}_t \odot \mathbf{Q}_t\right),$$

and $\bar{\mathbf{Y}} \sim \text{Pois}(\sum_{t=1}^\tau \boldsymbol{\Lambda}_t \odot \mathbf{Q}_t)$*, i.e.,*

$$\log p(\bar{\mathbf{Y}}) = \mathbf{1}_m^\top \mathbf{M} \mathbf{1}_m + \mathbf{1}_m^\top (\bar{\mathbf{Y}} \odot \log \mathbf{M}) \mathbf{1}_m - \mathbf{1}_m^\top \log(\bar{\mathbf{Y}}!) \mathbf{1}_m,$$

where $\mathbf{M} := \sum_{t=1}^\tau \boldsymbol{\Lambda}_t \odot \mathbf{Q}_t$ *and by convention* $0 \log 0 := 0$*.*

Proof of Lemma 5 By lemma 4 we have that for each $t = 1, \ldots, \tau$, $\mathbf{Y}_t \sim \text{Pois}(\mathbf{\Lambda}_t \odot \mathbf{Q}_t)$ and:

$$\mathbb{E}[\mathbf{Z}_\tau \mid \mathbf{Y}_\tau] = (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_\tau) \odot \mathbf{\Lambda}_\tau + \mathbf{Y}_\tau.$$

Since $\bar{Y}^{(i,j)}$ is the sum of independent Poisson random variables $Y_t^{(i,j)}$, we have $\mathbf{Y} \sim \text{Pois}(\sum_{t=1}^{\tau} \mathbf{\Lambda}_t \odot \mathbf{Q}_t)$ and given $\bar{Y}^{(i,j)}$, $Y_\tau^{(i,j)}$ is distributed $\text{Bin}(\bar{Y}^{(i,j)}, \Lambda_\tau^{(i,j)} Q_t^{(i,j)} / \sum_{t=1}^{\tau} \Lambda_t^{(i,j)} Q_t^{(i,j)})$. Hence by the tower law:

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_\tau \mid \mathbf{Y}] &= \mathbb{E}[\mathbb{E}[\mathbf{Z}_\tau \mid \mathbf{Y}_\tau] \mid \mathbf{Y}] \\
&= (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_t) \odot \mathbf{\Lambda}_\tau + \mathbb{E}[\mathbf{Y}_\tau \mid \mathbf{Y}] \\
&= (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_t) \odot \mathbf{\Lambda}_\tau + \bar{\mathbf{Y}} \odot \mathbf{\Lambda}_\tau \odot \mathbf{Q}_\tau \oslash \left(\sum_{t=1}^{\tau} \mathbf{\Lambda}_t \odot \mathbf{Q}_t \right),
\end{aligned}$$

in the case that all elements of $\sum_{t=1}^{\tau} \mathbf{\Lambda}_t \odot \mathbf{Q}_t$ are strictly positive. Otherwise we have $\mathbb{E}[Z_\tau^{(i,j)} \mid \mathbf{Y}] = (1 - Q_t^{(i,j)}) \Lambda_\tau^{(i,j)}$ for any (i,j) such that $\left[\sum_{t=1}^{\tau} \mathbf{\Lambda}_t \odot \mathbf{Q}_t\right]^{(i,j)} = 0$, since the latter equality implies $[\mathbf{\Lambda}_\tau \odot \mathbf{Q}_\tau]^{(i,j)} = 0$, which in turn implies $Y_\tau^{(i,j)} = 0$ almost surely.

■

Computing the PAL for general $(\tau_r)_{r \geq 1}$

Algorithm 8 Filtering for case (II) with general $(\tau_r)_{r \geq 1}$

initialise: $\bar{\boldsymbol{\lambda}}_0 \leftarrow \boldsymbol{\lambda}_0$.
1: **for** $r \geq 1$:
2: **for** $t = \tau_{r-1} + 1, \ldots, \tau_r - 1$:
3: $\mathbf{\Lambda}_t \leftarrow (\bar{\boldsymbol{\lambda}}_{t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{t, \boldsymbol{\eta}(\bar{\lambda}_{t-1})}$
4: $\bar{\boldsymbol{\lambda}}_t \leftarrow (\mathbf{1}_m^\top \mathbf{\Lambda}_t)^\top$
5: **end for**
6: $\mathbf{\Lambda}_{\tau_r} \leftarrow (\boldsymbol{\lambda}_{\tau_r - 1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{\tau_r, \boldsymbol{\eta}(\boldsymbol{\lambda}_{\tau_r - 1})}$
7: $\mathbf{M}_r \leftarrow \sum_{t=\tau_{r-1}+1}^{\tau_r} \mathbf{\Lambda}_t \odot \mathbf{Q}_t$
8: $\bar{\mathbf{\Lambda}}_{\tau_r} \leftarrow (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{\tau_r}) \odot \mathbf{\Lambda}_{\tau_r} + \bar{\mathbf{Y}}_r \odot \mathbf{\Lambda}_{\tau_r} \odot \mathbf{Q}_{\tau_r} \oslash \mathbf{M}_r$
9: $\mathscr{L}(\bar{\mathbf{Y}}_r | \bar{\mathbf{Y}}_{1:r-1}) \leftarrow -\mathbf{1}_m^\top \mathbf{M}_r \mathbf{1}_m + \mathbf{1}_m^\top (\bar{\mathbf{Y}}_r \odot \log \mathbf{M}_r) \mathbf{1}_m - \mathbf{1}_m^\top \log(\bar{\mathbf{Y}}_r!) \mathbf{1}_m$
10: $\bar{\boldsymbol{\lambda}}_{\tau_r} \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_{\tau_r})^\top$
11: **end for**

In algorithm 8 we adopt the same conventions concerning 0/0 and $0 \log 0 := 0$ as in algorithm 6. The log PAL associated with algorithm 8 is:

$$\log p(\bar{\mathbf{Y}}_{1:r}) \approx \sum_{s=1}^{r} \mathscr{L}(\bar{\mathbf{Y}}_s | \bar{\mathbf{Y}}_{1:s-1}), \tag{3.13}$$

where, as per line of 9 of algorithm 8, each term $\mathscr{L}(\bar{\mathbf{Y}}_r | \bar{\mathbf{Y}}_{1:r-1})$ is the log probability mass function of $\text{Pois}(\mathbf{M}_r)$ evaluated at $\bar{\mathbf{Y}}_r$.

CHAPTER 4

CONSISTENCY THEORY

Whilst the results in section 3.2 explain how the steps in algorithms 6-8 and the associated PALs are motivated by recursive vector-Poisson approximations, so far nothing we have stated quantifies the quality of these approximations, nor the PALs. In this chapter we present consistency results for parameter estimators defined by maximising PALs, it is organised as follows. Section 4.1 introduces further notation and definitions necessary for our consistency proofs. Section 4.2 states our assumptions. Section 4.3 outlines the consistency result; the full argument and associated proofs are arduous and repetitive, hence they are presented in appendix A. Section 4.4 presents a simulated example to empirically illustrate the theoretical results.

4.1 Notation and definitions for the consistency results

We now introduce explicit notation for dependence of various quantities on a parameter vector $\boldsymbol{\theta}$; we allow $\mathbb{P}_{0,n}, \mathbf{K}_{t,\boldsymbol{\eta}}, \mathbf{q}_t, \mathbf{Q}_t, \mathbf{G}_t, \boldsymbol{\delta}_t$ to depend on $\boldsymbol{\theta}$, and reflect this throughout section 4.3 with notation $\mathbb{P}^{\boldsymbol{\theta}}_{0,n}, \mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta}), \mathbf{q}_t(\boldsymbol{\theta}), \mathbf{Q}_t(\boldsymbol{\theta}), \mathbf{G}_t(\boldsymbol{\theta}), \boldsymbol{\delta}_t(\boldsymbol{\theta})$. We allow $\boldsymbol{\kappa}_t$ and $\boldsymbol{\alpha}_t$ to depend on $\boldsymbol{\theta}$, as well as the expected initial population size n, with notation $\boldsymbol{\kappa}_{t,n}(\boldsymbol{\theta})$ and $\boldsymbol{\alpha}_{t,n}(\boldsymbol{\theta})$. We also need to make explicit the dependence on n and $\boldsymbol{\theta}$ of the quantities computed in algorithms 6 and 8; we write these as: $\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta})$, $\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}), \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})$; and $\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta})$, $\bar{\boldsymbol{\Lambda}}_{t,n}(\boldsymbol{\theta})$, $\mathbf{M}_{r,n}(\boldsymbol{\theta})$.

In either case (I) or (II), one can think of the expected initial population size n as a global model index. We write $(\Omega_n, \mathscr{F}_n, \mathbb{P}^{\boldsymbol{\theta}}_n)$ for a probability space underlying each of these cases with expected initial population size n; in the context of case (I), $\mathbb{P}^{\boldsymbol{\theta}}_n$ is the joint distribution of $(\mathbf{x}_t)_{t\geq 0}$ and $(\mathbf{y}_t)_{t\geq 1}$ (as formulated in section 3.1) whilst in the context of case (II), $\mathbb{P}^{\boldsymbol{\theta}}_n$ is the joint distribution of $(\mathbf{Z}_t)_{t\geq 1}$ and $(\bar{\mathbf{Y}}_r)_{r\geq 1}$. In either case the overall probability space we shall work with is $(\Omega, \mathscr{F}, \mathbb{P}^{\boldsymbol{\theta}}) := (\prod_{n\geq 1}\Omega_n, \bigotimes_{n\geq 1}\mathscr{F}_n, \bigotimes_{n\geq 1}\mathbb{P}^{\boldsymbol{\theta}}_n)$. From henceforth we denote by $\boldsymbol{\theta}^* \in \Theta$ an arbitrarily

chosen but then fixed data-generating parameter (DGP). Almost sure convergence under $\mathbb{P}^{\boldsymbol{\theta}^*}$ is denoted $\xrightarrow[a.s.]{\boldsymbol{\theta}^*}$.

We now fix a time horizon $T \geq 1$ where for case (I), T is any positive integer, whilst for case (II), we assume $T = \tau_R$ for some $R \geq 1$. Since this time horizon is fixed, it will not appear explicitly in some of the notation for our consistency results. However, in order to state and prove various results, we need to make the dependence on $\boldsymbol{\theta}$ and n of the PALs computed using algorithms 6 and 8 explicit. To do so we define

$$\ell_n(\boldsymbol{\theta}) \coloneqq \sum_{t=1}^{T} \ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}), \qquad \mathscr{L}_n(\boldsymbol{\theta}) \coloneqq \sum_{r=1}^{R} \mathscr{L}(\bar{\mathbf{Y}}_r|\bar{\mathbf{Y}}_{1:r-1}),$$

where it is to be understood that each of the terms $\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ and $\mathscr{L}(\bar{\mathbf{Y}}_r|\bar{\mathbf{Y}}_{1:r-1})$ are computed using respectively algorithms 6 and 8 with parameter value $\boldsymbol{\theta}$ and expected initial population size n, and where the distribution of the random variables $\mathbf{y}_{1:T}$ and $\bar{\mathbf{Y}}_{1:R}$ is specified by the DGP $\boldsymbol{\theta}^*$ and the expected initial population size n. The fact that $\ell_n(\boldsymbol{\theta})$ and $\mathscr{L}_n(\boldsymbol{\theta})$ are functions of respectively $\mathbf{y}_{1:T}$ and $\bar{\mathbf{Y}}_{1:R}$ is not shown in the notation.

4.2 Assumptions

Assumption 1. *The parameter space $\Theta \subset \mathbb{R}^d$ is compact.*

Assumption 2. *For all probability vectors $\boldsymbol{\eta}$, $t \geq 1$, and $n \geq 1$, $\mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta}), \mathbf{q}_t(\boldsymbol{\theta}), \mathbf{Q}_t(\boldsymbol{\theta})$, $\mathbf{G}_t(\boldsymbol{\theta})$, $\boldsymbol{\delta}_t(\boldsymbol{\theta})$, $\boldsymbol{\kappa}_{t,n}(\boldsymbol{\theta})$ and $\boldsymbol{\alpha}_{t,n}(\boldsymbol{\theta})$ are continuous functions of $\boldsymbol{\theta}$, and the supports of these vectors and the supports of each matrix row do not depend on $\boldsymbol{\theta}$ or n. For all $\boldsymbol{\theta} \in \Theta$ and $t \geq 1$, $\text{supp}(\boldsymbol{\delta}_t(\boldsymbol{\theta})) = [m]$, i.e. $\boldsymbol{\delta}_t(\boldsymbol{\theta})$ has no entries equal to 0. Furthermore, there exist continuous functions of $\boldsymbol{\theta}$ mapping $\Theta \to \mathbb{R}^m_{\geq 0}$, $\boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})$ and $\boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})$, such that $\text{supp}(\boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})) = \text{supp}(\boldsymbol{\kappa}_{t,n}(\boldsymbol{\theta}))$ and $\text{supp}(\boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})) = \text{supp}(\boldsymbol{\alpha}_{t,n}(\boldsymbol{\theta}))$ for all n, and for each $\boldsymbol{\theta} \in \Theta$ there exist $a_1 > 0$, $a_2 > 0$, $\gamma_1 > 0$, and $\gamma_2 > 0$ such that:*

$$\|n^{-1}\boldsymbol{\kappa}_{t,n}(\boldsymbol{\theta}) - \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})\|_\infty < a_1 n^{-(\frac{1}{4}+\gamma_1)},$$
$$\|n^{-1}\boldsymbol{\alpha}_{t,n}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})\|_\infty < a_2 n^{-(\frac{1}{4}+\gamma_2)}.$$

Assumption 3. *For all $\boldsymbol{\theta} \in \Theta$, there exists a constant $c > 0$ such that for all $t \geq 1$, all vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^m$, and all probability vectors $\boldsymbol{\eta}, \boldsymbol{\eta}'$:*

$$|\boldsymbol{f}_1^\top \mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta})\boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{K}_{t,\boldsymbol{\eta}'}(\boldsymbol{\theta})\boldsymbol{f}_2| \leq c\|\boldsymbol{f}_1\|_\infty \|\boldsymbol{f}_2\|_\infty \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_\infty.$$

Furthermore, if $\text{supp}(\boldsymbol{\eta}) \subseteq \text{supp}(\boldsymbol{\eta}')$ then $\text{supp}(\mathbf{K}^{i,\cdot}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta})) \subseteq \text{supp}(\mathbf{K}^{i,\cdot}_{t,\boldsymbol{\eta}'}(\boldsymbol{\theta}))$ for all $i \in [m]$.

Assumption 4. *Let $\boldsymbol{\theta} \in \Theta$, $n \in \mathbb{N}$, and $\mathbf{x}_0 \sim \mathbb{P}^{\boldsymbol{\theta}}_{0,n}$. There exists $\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta})$ which is a continuous mapping $\Theta \to \mathbb{R}^m_{\geq 0}$ such that the support of $\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta})$, which is not the empty set, does not depend on $\boldsymbol{\theta}$, and there exists $\gamma_0 > 0$ such that for any $\boldsymbol{f} \in \mathbb{R}^m$ there exists a $c_0 > 0$ such that:*

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}^\top \mathbf{x}_0 - \boldsymbol{f}^\top \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta})\right|^4\right]^{\frac{1}{4}} < c_0 n^{-(\frac{1}{4}+\gamma_0)}.$$

Furthermore, there exists some $c > 0$ *and* $\gamma > 0$ *such that:*

$$\|n^{-1}\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}) - \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta})\|_\infty < cn^{-(\frac{1}{4}+\gamma)},$$

and $\mathrm{supp}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta})) = \mathrm{supp}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}))$ *for all* $\boldsymbol{\theta} \in \Theta$ *and* $n \in \mathbb{N}$.

The compactness of Θ in assumption 1 and the continuity in $\boldsymbol{\theta}$ of various quantities in assumption 2 are fairly standard assumptions in proofs of consistency of maximum likelihood estimators. The conditions on the supports of various vectors in assumptions 2-4 are used to rule out the possibility that different parameter values may induce mutually singular distributions over observations, this helps us ensure well-defined contrast functions in our consistency proofs. Assumption 4 asserts that the scaled initial population configuration, $n^{-1}\mathbf{x}_0$, obeys a law of large numbers.

4.3 Main consistency theorem and outline of the proof

In order to state and explain our main consistency result, theorem 1, we now summarise some intermediate results concerning the asymptotic behaviour of the models and quantities calculated using algorithms 6 and 8. Precise statements and proofs of these intermediate results are in appendix A.

Laws of large numbers. The first step is to establish laws of large numbers for the Latent Compartmental Model, and hence for the observations, these results are stated and proved in section A.1. In case (I) we show that for certain deterministic vectors $\boldsymbol{\nu}_t(\boldsymbol{\theta}^*)$, $t \geq 1$,

$$\frac{1}{n}\mathbf{x}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\nu}_t(\boldsymbol{\theta}^*), \qquad \frac{1}{n}\mathbf{y}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} [(\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) \odot \mathbf{q}_t(\boldsymbol{\theta}^*))^\top \mathbf{G}_t(\boldsymbol{\theta}^*)]^\top + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}^*), \tag{4.1}$$

and in case (II), for certain deterministic matrices $\mathbf{N}_t(\boldsymbol{\theta}^*)$, $t \geq 1$,

$$\frac{1}{n}\mathbf{Z}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{N}_t(\boldsymbol{\theta}^*), \qquad \frac{1}{n}\bar{\mathbf{Y}}_r \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \sum_{t=\tau_{r-1}+1}^{\tau_r} \mathbf{N}_t(\boldsymbol{\theta}^*) \odot \mathbf{Q}_t(\boldsymbol{\theta}^*). \tag{4.2}$$

The vectors $\boldsymbol{\nu}_t(\boldsymbol{\theta}^*)$ and matrices $\mathbf{N}_t(\boldsymbol{\theta}^*)$ satisfy recursive (in time) formulae and the convergence of $\frac{1}{n}\mathbf{x}_t$ and $\frac{1}{n}\mathbf{Z}_t$ as $n \to \infty$ is a discrete time analogue of the convergence of the continuous time, stochastic model to the solution of the ODE in (2.1), i.e. a discrete-time counterpart of the results of (Kurtz, 1970).

Filtering intensity limits and asymptotic filtering accuracy. Making use of the laws of large numbers for the observations, the next step is to establish convergence to deterministic limits of intensity vectors and matrices computed using respectively algorithms 6 and 8 and which thus define the PALs (3.10) and (3.13). This is the subject of section A.2. In case (I) we find

deterministic vectors $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ and $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$, $t \geq 1$, $\boldsymbol{\theta} \in \Theta$, where $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ is a function of $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$, such that:

$$\frac{1}{n}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \qquad \frac{1}{n}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}).$$

In case (II) we find deterministic matrices $\boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ and $\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$, $t \geq 1$, $r \geq 1$, $\boldsymbol{\theta} \in \Theta$, where $\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ is a function of $\boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ for $t = \tau_{r-1}+1,\ldots,\tau_r$, such that:

$$\frac{1}{n}\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \qquad \frac{1}{n}\mathbf{M}_{r,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}).$$

A notable fact about the limiting filtering intensities $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ and $\boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ that we uncover (see remarks 1 and 2 in section A.2) is that:

$$\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = \boldsymbol{\nu}_t(\boldsymbol{\theta}^*), \qquad \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = \mathbf{N}_t(\boldsymbol{\theta}^*),$$

where $\boldsymbol{\nu}_t(\boldsymbol{\theta}^*)$ and $\mathbf{N}_t(\boldsymbol{\theta}^*)$ are as in (4.1) and (4.2). In this sense, running algorithms 6 and 8 with the model specified by the DGP $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^*$ is asymptotically accurate as $n \to \infty$, in spite of the recursive Poisson approximations involved in these procedures.

Contrast functions. We then construct contrast functions associated with the PALs. This is the subject of section A.3. The contrast functions turn out to be in the form of Kullback-Liebler divergences. In case (I),

$$\frac{1}{n}\ell_n(\boldsymbol{\theta}) - \frac{1}{n}\ell_n(\boldsymbol{\theta}^*) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\sum_{t=1}^{T} \mathrm{KL}\left(\mathrm{Pois}[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)] \,\|\, \mathrm{Pois}[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\right),$$

and in case (II),

$$\frac{1}{n}\mathscr{L}_n(\boldsymbol{\theta}) - \frac{1}{n}\mathscr{L}_n(\boldsymbol{\theta}^*) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\sum_{r=1}^{R} \mathrm{KL}\left(\mathrm{Pois}[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)] \,\|\, \mathrm{Pois}[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\right),$$

where in each case the convergence is established to be uniform in $\boldsymbol{\theta}$.

Convergence of the maximum PAL estimators. With:

$$\Theta^*_{(I)} \coloneqq \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \text{ for all } t = 1,\ldots T\},$$
$$\Theta^*_{(II)} \coloneqq \{\boldsymbol{\theta} \in \Theta : \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \text{ for all } r = 1,\ldots R\},$$

uniform convergence to the contrast functions as well as standard continuity and compactness arguments are used to complete the proof of our main consistency result:

Theorem 1. *Let assumptions 1-4 hold and let $\hat{\boldsymbol{\theta}}_n$ be a maximiser of $\ell_n(\boldsymbol{\theta})$ (resp. $\mathscr{L}_n(\boldsymbol{\theta})$). Then $\hat{\boldsymbol{\theta}}_n$ converges to $\Theta^*_{(I)}$ (resp. $\Theta^*_{(II)}$) as $n \to \infty$, $\mathbb{P}^{\boldsymbol{\theta}^*}$-almost surely.*

The proof is in section A.4.

Identifiability. We now provide some further insight into the sets $\Theta^*_{(I)}$ and $\Theta^*_{(II)}$ in order to explain in what sense the model is identified under theorem 1. In section A.5 we show that for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned} \boldsymbol{\theta} \in \Theta^*_{(I)} &\iff \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*), \quad \forall t = 1,\dots,T, \\ \boldsymbol{\theta} \in \Theta^*_{(II)} &\iff \mathbf{M}_{r,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*), \quad \forall r = 1,\dots,R. \end{aligned}$$

The vector $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ turns out (see remark 1) to be equal to the r.h.s. of the second $\mathbb{P}^{\boldsymbol{\theta}^*}$-almost sure limit in (4.1). Thus for case (I), the convergence to $\Theta^*_{(I)}$ in theorem 1 tells us that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ approaches the set of $\boldsymbol{\theta}$ such that the $\mathbb{P}^{\boldsymbol{\theta}}$-almost sure limit of $\frac{1}{n}\mathbf{y}_t$ is the same as the $\mathbb{P}^{\boldsymbol{\theta}^*}$-almost sure limit of $\frac{1}{n}\mathbf{y}_t$, for all $t = 1,\dots,T$. Similarly for case (II), $\mathbf{M}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ turns out (see remark 2) to be equal to the r.h.s. of the second limit in (4.2), and the convergence to $\Theta^*_{(II)}$ in theorem 1 tells us that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ approaches the set of $\boldsymbol{\theta}$ such that the $\mathbb{P}^{\boldsymbol{\theta}}$-almost sure limit of $\frac{1}{n}\bar{\mathbf{Y}}_r$ is the same as the $\mathbb{P}^{\boldsymbol{\theta}^*}$-almost sure limit of $\frac{1}{n}\bar{\mathbf{Y}}_r$, for all $r = 1,\dots,R$.

4.4 A simulated example

Consider a simple SEIR model with immigration and emigration: $\mathbb{P}_{0,n} = \text{Mult}\left(n, [0.99\ 0\ 0.01\ 0]^\top\right)$ with $\boldsymbol{\alpha}_{t,n} = \left[\frac{4}{100}n\ \frac{4}{100}n\ \frac{4}{100}n\ \frac{4}{100}n\right]^\top$, $\boldsymbol{\delta}_t = \left[\frac{98}{100}\ \frac{98}{100}\ \frac{98}{100}\ \frac{98}{100}\right]^\top$, $\boldsymbol{\kappa}_{t,n} = \left[\frac{1}{100}n\ \frac{1}{100}n\ \frac{1}{100}n\ \frac{1}{100}n\right]^\top$, $\mathbf{q}_t = [0.1\ 0.1\ 0.3\ 0.2]^\top$ for all t, and

$$\mathbf{K}_{t,\boldsymbol{\eta}} = \begin{bmatrix} e^{-\beta\eta^{(3)}} & 1-e^{-\beta\eta^{(3)}} & 0 & 0 \\ 0 & e^{-\rho} & 1-e^{-\rho} & 0 \\ 0 & 0 & e^{-\gamma} & 1-e^{-\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{G}_t = \begin{bmatrix} 0.95 & 0 & 0.05 & 0 \\ 0.3 & 0 & 0.7 & 0 \\ 0.15 & 0 & 0.85 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with DGP $\boldsymbol{\theta}^* = [\beta^*\ \rho^*\ \gamma^*]^\top = [0.5\ 0.05\ 0.1]^\top$.

This observation model can be interpreted as follows: with probability $q_t^{(i)}$ each individual in compartment i is tested for disease. Allowing $q_t^{(i)}$ to vary across i could model, for example, infective individuals being more likely to be tested. The above choice of $\mathbf{G}_t$ allows for false-positives (first row) and false-negatives (third row), where those testing positive are considered infective, and those testing negative are considered susceptible. Of course, other choices are possible.

The top two rows of plots in figure 4.1 show $n^{-1}\mathbf{x}_t$ and $n^{-1}\mathbf{y}_t$ simulated 50 times from the model with population sizes $n \in \{100, 1000, 10000, 100000\}$. Note that in the top row, the fact that trajectories for compartment S in $n^{-1}\mathbf{x}_t$ are valued above 1 in places is explained in terms of immigration into the S compartment exceeding the combined effect of emigration from S and individuals transitioning from S to E. With $n = 100$, the fact that some trajectories for the S compartment are roughly increasing over time corresponds to the lack of an outbreak; for other trajectories which rise and then fall, an outbreak does occur.

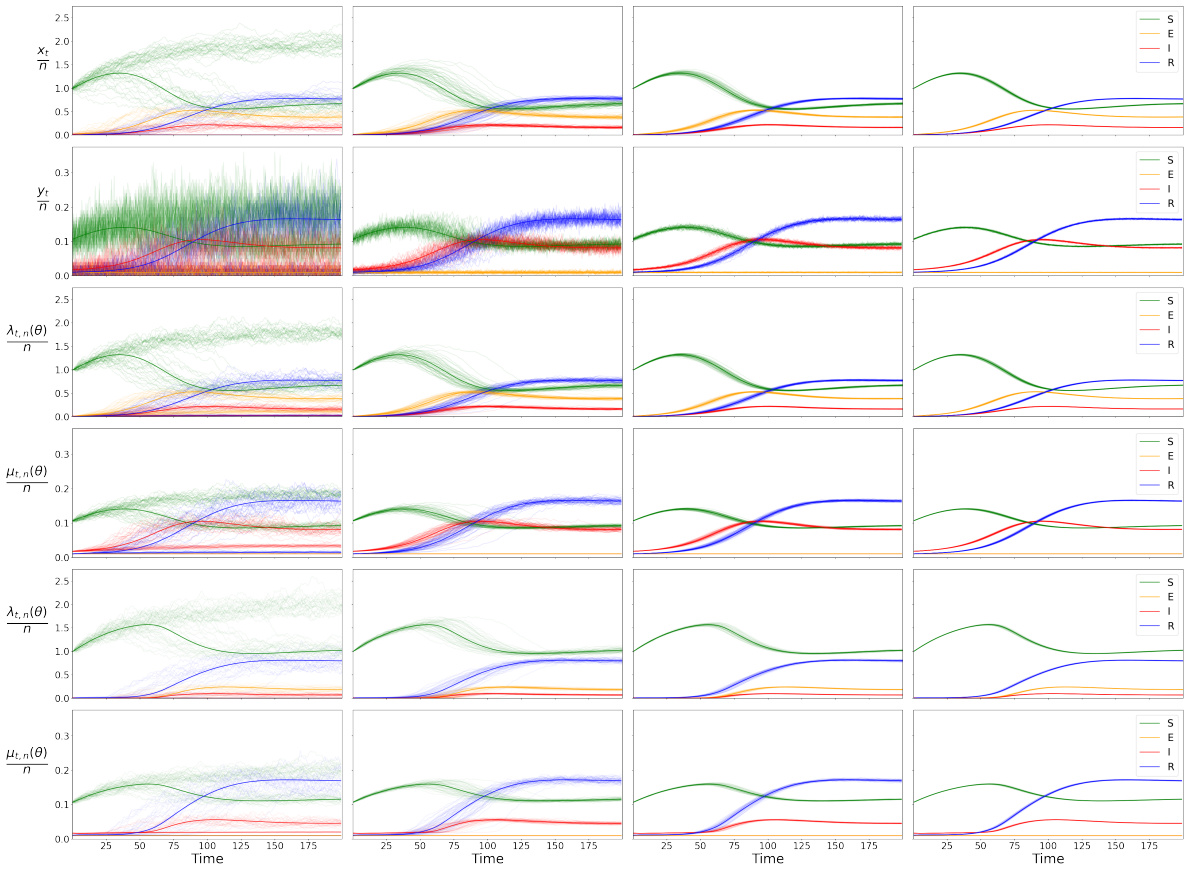

Figure 4.1: Simulation SEIR example. Top two rows: asymptotic behaviour of $\mathbf{x}_t/n$ and $\mathbf{y}_t/n$; 50 simulations from the model (light lines) and theoretical deterministic $n \to \infty$ limits (bold line) for each population size (left to right), $n \in \{100, 1000, 10000, 100000\}$. Middle two rows: filtering intensities associated with the 50 simulated data sets with $\boldsymbol{\theta}$ taken to be $\boldsymbol{\theta}^*$. Bottom two rows: filtering with $\boldsymbol{\theta}$ set erroneously $\beta = 0.1$, $\gamma = 0.3$, and all other parameters set as for the middle two rows.

Due to the choices of $\mathbb{P}_{0,n}$, $\boldsymbol{\alpha}_{t,n}$ and $\boldsymbol{\kappa}_{t,n}$ set out above, it is immediate that the vectors $\boldsymbol{\lambda}_{0,\infty}$, $\boldsymbol{\alpha}_{t,\infty}$ and $\boldsymbol{\kappa}_{t,\infty}$ appearing in assumptions 4 and 2 exist. The convergence of $n^{-1}\mathbf{x}_t$ and $n^{-1}\mathbf{y}_t$ as $n \to \infty$ to deterministic limits as discussed in section 4.3 is evident in figure 4.1.

The middle two rows of figure 4.1 show the behaviour of the scaled filtering intensities $n^{-1}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta})$ and $n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})$ obtained from algorithm 6 in the case of correctly specified parameters $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^*$. It is evident that, as per the discussion of asymptotic filtering accuracy in section 4.3, as $n \to \infty$ these quantities converge to the same deterministic limits as do $n^{-1}\mathbf{x}_t$ and $n^{-1}\mathbf{y}_t$, respectively. On the other hand, as illustrated in the bottom two rows of figure 4.1, when the model is not correctly specified, then $\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})$ converge to limits which are not equal to the limits of $n^{-1}\mathbf{x}_t$ and $n^{-1}\mathbf{y}_t$.

Figure 4.2 illustrates the behaviour of the scaled log-PAL $n^{-1}\ell_n(\boldsymbol{\theta})$ evaluated over a find grid of values fo $\boldsymbol{\theta} = [\beta\ \gamma]^\top$ (all other parameters held constant). Each purple surface in each plot corresponds to a different data set simulated from the model, as in the second row of figure 4.1. As n grows, figure 4.1 evidences convergence of the maximum PAL estimates to the true

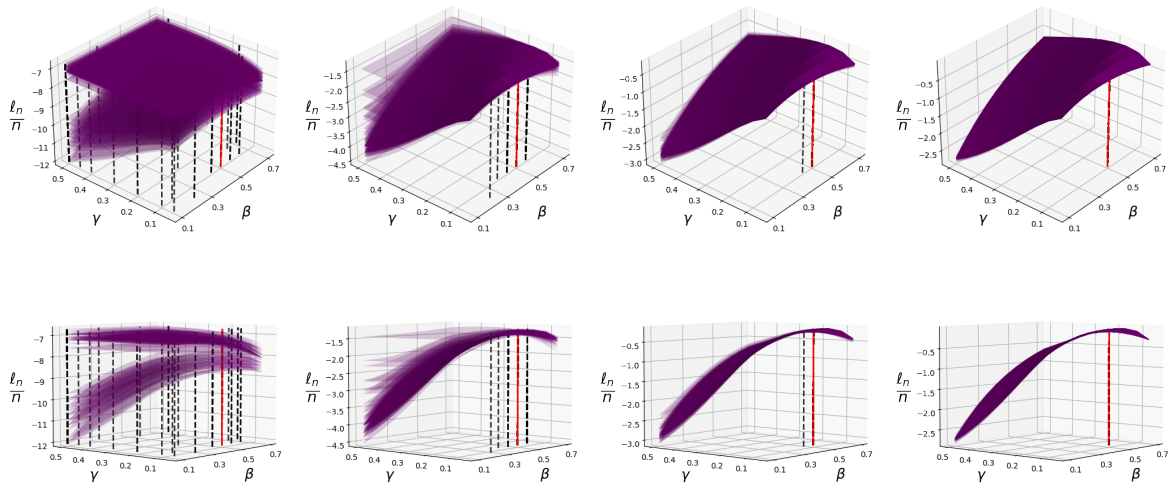

Figure 4.2: Simulation SEIR example. Purple surfaces within each plot are the scaled log-PAL surfaces associated with 50 data sets simulated from the model with the DGP. From left to right: $n = 100, 1000, 10000, 100000$. Vertical black dashed lines are the maximum PAL estimates for each surface, the vertical red line is the DGP. The two rows show the same 3-d plots from different viewing angles.

parameter value, as per theorem 1.

CHAPTER 5

OVER-DISPERSION

This chapter is organised as follows. Section 5.1 introduces the concept of over-dispersion and its importance in epidemiology. Section 5.2 proposes a methodology for fitting over-dispersed models within the PAL framework. Section 5.3 demonstrates the methodology and empirically validates its performance with a simple simulated example.

5.1 Introducing over-dispersion

Over-dispersion is an important modelling consideration in many epidemiological contexts and may have substantial implications for model fit and predictive uncertainty. The models we have considered so far are equi-dispersed in the sense of Bretó and Ionides (2011). That is, they are based on distributions, namely binomial and Poisson, for which the variance is less than or equal to the mean – this is an undesirable limitation, as we will show in chapter 6. For compartmental models in general, over-dispersion can be incorporated in either the transition or observation models, or both, see for example (Stocks et al., 2020). In the context of the models from section 3.1, a natural approach would be to replace the binomial and Poisson-distributed elements of the latent compartmental model (section 3.1.2) and/or observation models (sections 3.1.3.1-3.1.3.3) with over-dispersed counterparts, such as beta-binomial and negative binomial distributions. It appears that analytically tractable PAL-style approximations cannot be derived for such models. However, one can often construct over-dispersed distributions as compound distributions through introduction of latent variables, e.g. placing a beta prior on $q_t^{(i)}$ in (3.2) and then integrating out would result in a marginally beta-binomial observation model. Similarly, priors could be placed on parameters which specify the matrix $\mathbf{K}_{t,\boldsymbol{\eta}}$, the immigration and emigration parameters $\boldsymbol{\alpha}_t$, $\boldsymbol{\delta}_t$, the spurious observation intensity $\boldsymbol{\kappa}_t$, and so on. It is through this latent variable perspective that we extend the use of the PAL to deal with over-dispersion.

5.2 Dealing with over-dispersion in the PAL framework

Consider the latent compartmental model from section 3.1.2 combined with observation mechanism from section 3.1.3.1 with parameter $\boldsymbol{\theta}$ (the observation models from sections 3.1.3.2 and 3.1.3.3 can be handled in a very similar manner). We consider $\boldsymbol{\theta}$ to be partitioned into two components: $\boldsymbol{\theta} = [\boldsymbol{\vartheta}\, \bar{\boldsymbol{\theta}}_{1:T}]$, where $\boldsymbol{\vartheta}$ consists of parameters which are either fixed or to be estimated, and $\bar{\boldsymbol{\theta}}_{1:T} \sim f(\cdot|\boldsymbol{\varphi})$ are to be integrated out, for some density f and hyperparameter $\boldsymbol{\varphi}$. A default approach would be for $\bar{\boldsymbol{\theta}}_{1:T}$ to be independent under $f(\cdot|\boldsymbol{\varphi})$, but Markovian or other dependence could be incorporated.

We assume that the elements of the model are parameterised such that:

$$\begin{aligned}
\boldsymbol{\alpha}_t(\boldsymbol{\theta}) &= \boldsymbol{\alpha}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \quad \boldsymbol{\delta}_t(\boldsymbol{\theta}) = \boldsymbol{\delta}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \quad \mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta}) = \mathbf{K}_{\boldsymbol{\eta}}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \\
\boldsymbol{\kappa}_t(\boldsymbol{\theta}) &= \boldsymbol{\kappa}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \quad \mathbf{q}_t(\boldsymbol{\theta}) = \mathbf{q}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \quad \mathbf{G}_t(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t),
\end{aligned}$$

for some given functions $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, etc., which implies that:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\theta}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \qquad p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}) = p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t),$$

and in turn that $\bar{\boldsymbol{\theta}}_t$ is conditionally independent of $\mathbf{y}_{1:t-1}$ given $\bar{\boldsymbol{\theta}}_{1:t-1}$, $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}$.

Let us derive the marginal likelihood for the parameters $[\boldsymbol{\vartheta}\, \boldsymbol{\varphi}]$ with $\bar{\boldsymbol{\theta}}_{1:T}$ integrated out. Momentarily regarding $[\boldsymbol{\vartheta}\, \boldsymbol{\varphi}]$ as fixed and suppressing it from notation, consider the recursive relationship:

$$\begin{aligned}
p(\mathbf{y}_{1:t}, \bar{\boldsymbol{\theta}}_{1:t}) &= p(\mathbf{y}_t, \bar{\boldsymbol{\theta}}_t|\mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) p(\mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) \\
&= p(\mathbf{y}_t|\bar{\boldsymbol{\theta}}_t, \mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) p(\bar{\boldsymbol{\theta}}_t|\mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) p(\mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) \\
&= p(\mathbf{y}_t|\bar{\boldsymbol{\theta}}_t, \mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}) f(\bar{\boldsymbol{\theta}}_t|\bar{\boldsymbol{\theta}}_{1:t-1}) p(\mathbf{y}_{1:t-1}, \bar{\boldsymbol{\theta}}_{1:t-1}),
\end{aligned}$$

where the third equality holds due to the aforementioned conditional independence. Now, reintroducing $[\boldsymbol{\vartheta}\, \boldsymbol{\varphi}]$ to the notation, we have:

$$\begin{aligned}
p(\mathbf{y}_{1:T}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) &= \int p(\mathbf{y}_{1:T}, \bar{\boldsymbol{\theta}}_{1:T}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) d\bar{\boldsymbol{\theta}}_{1:T} \\
&= \int \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_{1:t}) f(\bar{\boldsymbol{\theta}}_t|\bar{\boldsymbol{\theta}}_{1:t-1}, \boldsymbol{\varphi}) d\bar{\boldsymbol{\theta}}_{1:T}.
\end{aligned}$$

We can approximate this using the PAL:

$$p(\mathbf{y}_{1:T}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \approx \int \prod_{t=1}^{T} \exp\left\{\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_{1:t})\right\} f(\bar{\boldsymbol{\theta}}_t|\bar{\boldsymbol{\theta}}_{1:t-1}, \boldsymbol{\varphi}) d\bar{\boldsymbol{\theta}}_{1:T}, \tag{5.1}$$

where ℓ is defined as per algorithm 6. The right-hand side of (5.1) can be efficiently numerically approximated by embedding PAL computations within sequential Monte Carlo – see (Chopin et al., 2020) for an introduction to this family of Monte Carlo algorithms. Such a scheme is given by algorithm 10 and its subroutine algorithm 9.

In line 3 of algorithm 10 we take the convention $\pi(\cdot|\bar{\boldsymbol{\theta}}_{1:0}^{(i)}, \bar{\boldsymbol{\lambda}}_{t-1}^{(i)}, \mathbf{y}_{1:1}) \coloneqq \pi(\cdot|\bar{\boldsymbol{\lambda}}_0^{(i)}, \mathbf{y}_1)$. Algorithm 10 yields a Monte Carlo approximation to the r.h.s. of (5.1), so overall we obtain:

$$\log p(\mathbf{y}_{1:t}|\boldsymbol{\vartheta}, \boldsymbol{\varphi}) \approx \sum_{s=1}^{t} \widehat{\ell}(\mathbf{y}_s|\mathbf{y}_{1:s-1}, \boldsymbol{\vartheta}, \boldsymbol{\varphi}).$$

We stress there are two ingredients to this approximation: the Monte Carlo approximation and the PAL approximation. Whilst the main emphasis above regarding $\bar{\boldsymbol{\theta}}_{1:t}$ is that they are to be integrated out, a benefit of algorithm 10 is that it also yields the approximation:

$$p(\bar{\boldsymbol{\theta}}_t|\mathbf{y}_{1:t}, \boldsymbol{\vartheta}, \boldsymbol{\varphi}) \approx \sum_{i=1}^{n_{part}} \bar{w}_t^{(i)} \delta_{\bar{\boldsymbol{\theta}}_t^{(i)}}, \tag{5.2}$$

which enables inference for $\bar{\boldsymbol{\theta}}_t$ on the basis of observations $\mathbf{y}_{1:t}$.

Algorithm 9 PALSMC subroutine

input: $\bar{\boldsymbol{\lambda}}_{t-1}$ and $[\boldsymbol{\vartheta}\ \bar{\boldsymbol{\theta}}_t]$

1: $\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \boldsymbol{\delta}_t \leftarrow \boldsymbol{\delta}_t(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \mathbf{K}_{t,\boldsymbol{\eta}} \leftarrow \mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \mathbf{q}_t \leftarrow \mathbf{q}_t(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t),$
$\boldsymbol{\kappa}_t \leftarrow \boldsymbol{\kappa}_t(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t), \mathbf{G}_t \leftarrow \mathbf{G}_t(\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t)$

2: $\boldsymbol{\lambda}_t \leftarrow [(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1} \odot \boldsymbol{\delta}_t)}]^\top + \boldsymbol{\alpha}_t$

3: $\bar{\boldsymbol{\lambda}}_t \leftarrow [\mathbf{1}_m - \mathbf{q}_t + (\{\mathbf{y}_t^\top \oslash [(\mathbf{q}_t \odot \boldsymbol{\lambda}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top]\}[(\mathbf{1}_m \otimes \mathbf{q}_t) \odot \mathbf{G}_t^\top])^\top] \odot \boldsymbol{\lambda}_t$

4: $\boldsymbol{\mu}_t \leftarrow [(\boldsymbol{\lambda}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t]^\top + \boldsymbol{\kappa}_t$

5: $\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \leftarrow -\boldsymbol{\mu}_t^\top \mathbf{1}_m + \mathbf{y}_t^\top \log(\boldsymbol{\mu}_t) - \mathbf{1}_m^\top \log(\mathbf{y}_t!)$

return $\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ and $\bar{\boldsymbol{\lambda}}_t$

Algorithm 10 PALSMC

input: proposal distribution $\pi(\cdot|\cdot)$, number of particles n_{part}, parameter $[\boldsymbol{\vartheta}\ \boldsymbol{\varphi}]$.

initialise: $\bar{\boldsymbol{\lambda}}_0^{(i)} \leftarrow \boldsymbol{\lambda}_0$ for $i = 1, \ldots, n_{part}$

1: **for** $t \geq 1$:
2: **for** $i = 1, \ldots, n_{part}$:
3: $\bar{\boldsymbol{\theta}}_t^{(i)} \sim \pi(\cdot|\bar{\boldsymbol{\theta}}_{1:t-1}^{(i)}, \bar{\boldsymbol{\lambda}}_{t-1}^{(i)}, \mathbf{y}_{1:t})$
4: Obtain $\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_{1:t}^{(i)})$ and $\bar{\boldsymbol{\lambda}}_t^{(i)}$ from algorithm 9 with input $\bar{\boldsymbol{\lambda}}_{t-1}^{(i)}$ and $\left[\boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_t^{(i)}\right]$
5: $\log w_t^{(i)} \leftarrow \ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}_{1:t}^{(i)}) + \log f(\bar{\boldsymbol{\theta}}_t^{(i)}|\bar{\boldsymbol{\theta}}_{1:t-1}^{(i)}, \boldsymbol{\varphi}) - \log \pi(\bar{\boldsymbol{\theta}}_t^{(i)}|\bar{\boldsymbol{\theta}}_{1:t-1}^{(i)}, \bar{\boldsymbol{\lambda}}_{t-1}^{(i)}, \mathbf{y}_{1:t})$
6: **end for**
7: $\widehat{\ell}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \boldsymbol{\varphi}) \leftarrow \log\left(\frac{1}{n_{part}} \sum_{i=1}^{n_{part}} w_t^{(i)}\right)$
8: $\bar{w}_t^{(i)} \leftarrow w_t^{(i)} / \sum_{j=1}^{n_{part}} w_t^{(j)}$
9: resample $\left\{\bar{\boldsymbol{\theta}}_{1:t}^{(i)}, \bar{\boldsymbol{\lambda}}_t^{(i)}\right\}_{i=1}^{n_{part}}$ according to the weights $\left\{\bar{w}_t^{(i)}\right\}_{i=1}^{n_{part}}$
10: **end for**

In section 5.3 we explore ways in which the large population theory from chapter 4 is relevant to the construction and behaviour of PALSMC algorithms for over-dispersed models:

- It is well known that the efficiency of sequential Monte Carlo methods can be highly sensitive to the choice of the proposal distribution, π in algorithm 10. If we could choose

$\pi(\bar{\boldsymbol{\theta}}_t|\bar{\boldsymbol{\theta}}^{(i)}_{1:t-1}, \bar{\boldsymbol{\lambda}}^{(i)}_{t-1}\mathbf{y}_{1:t})$ to be proportional (as a function of $\bar{\boldsymbol{\theta}}_t$) to:

$$\exp\left[\ell(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\vartheta}, \bar{\boldsymbol{\theta}}^{(i)}_{1:t-1}, \bar{\boldsymbol{\theta}}_t)\right] f(\bar{\boldsymbol{\theta}}_t|\bar{\boldsymbol{\theta}}^{(i)}_{1:t-1}, \boldsymbol{\varphi}), \tag{5.3}$$

then the weight $w^{(i)}_t$ would have no dependence on $\bar{\boldsymbol{\theta}}^{(i)}_t$. Consequently the variability of the weight would be reduced and the overall efficiency of the PALSMC algorithm likely improved. This "optimal" choice or π is often not analytically tractable, but inspired by our consistency theory we suggest Laplace approximation to it. We demonstrate such proposals in simulation-based and real data examples in sections 5.3 – 6.4 and find them to be very efficient in practice.

- Through a simulation example in section 5.3, we illustrate that even for our over-dispersed models, where one might expect estimation consistency to be ruled out (the additional hierarchical components lead to violation of the various continuity assumptions of chapter 4), increasing population size can in fact increase the accuracy of point estimates of $\bar{\boldsymbol{\theta}}_t$ obtained from the r.h.s. of (5.2). The explanation for this is that, whilst the model may be over-dispersed, once $\bar{\boldsymbol{\theta}}_{1:t}$ are integrated out, it is equi-dispersed *conditional* on $\bar{\boldsymbol{\theta}}_{1:t}$.

In the examples in chapter 6 we also expand on algorithm 10 to include sophisticated resampling schemes and block particle filtering techniques (Rebeschini and Van Handel, 2015).

5.3 Pedagogical over-dispersed SEIR example

To demonstrate inference for an over-dispersed model using PALSMC we consider a simple SEIR model for which the latent population $\mathbf{x}_t \equiv [S_t\, E_t\, I_t\, R_t]^\top$ evolves according to transition matrix (3.1), with immigration and emigration parameters, $\boldsymbol{\alpha}_t$ and $\boldsymbol{\delta}_t$, combined with the observation model $y_t \sim \text{Binom}(I_t, q_t)$. We assume $\boldsymbol{\alpha}_t$ and $\boldsymbol{\delta}_t$ are known. We can cast this model in the form discussed in section 5.2 by identifying $\boldsymbol{\vartheta} = [\beta\ \rho\ \gamma]$, $\bar{\boldsymbol{\theta}}_{1:T} = q_{1:T}$, and choosing $f(\cdot|\boldsymbol{\varphi})$ to make $q_{1:T}$ i.i.d. according to a truncated normal distribution $q_t \sim \mathcal{N}(\mu_q, \sigma^2_q)_{\geq 0, \leq 1}$, with $\boldsymbol{\varphi} = [\mu_q\ \sigma^2_q]$, $\mu_q \in [0,1]$ and $\sigma^2_q > 0$. We give the details of a PALSMC scheme for this model in algorithm 11, and include the derivation of efficient, data-informed proposals by Laplace approximation to (5.3), inspired by the theory from chapter 4. The specific PALSMC scheme used for this section is given by algorithm 11.

Filtering and parameter estimation simulation study

To assess the ability of the PALSMC scheme to recover ground truth quantities, we simulated data from the model with $[\beta\ \rho\ \gamma\ \mu_q\ \sigma^2_q] = [0.8\ 0.1\ 0.2\ 0.5\ 0.1]$, $\boldsymbol{\pi}_0 = [0.99\ 0\ 0.01\ 0]^\top$, $\boldsymbol{\alpha}_t = 0.05\boldsymbol{\pi}_0$ and $\boldsymbol{\delta}_t = [0.95\ 0.95\ 0.95\ 0.95]^\top$. The first two rows of figure 5.1 explore the performance of PALSMC with increasing population size n and using the data-generating values of $[\boldsymbol{\vartheta}\ \boldsymbol{\varphi}]$. This collection of plots was created by first sampling a single draw of latent variables $q_{1:100} \sim f(\cdot|\boldsymbol{\varphi})$, then for

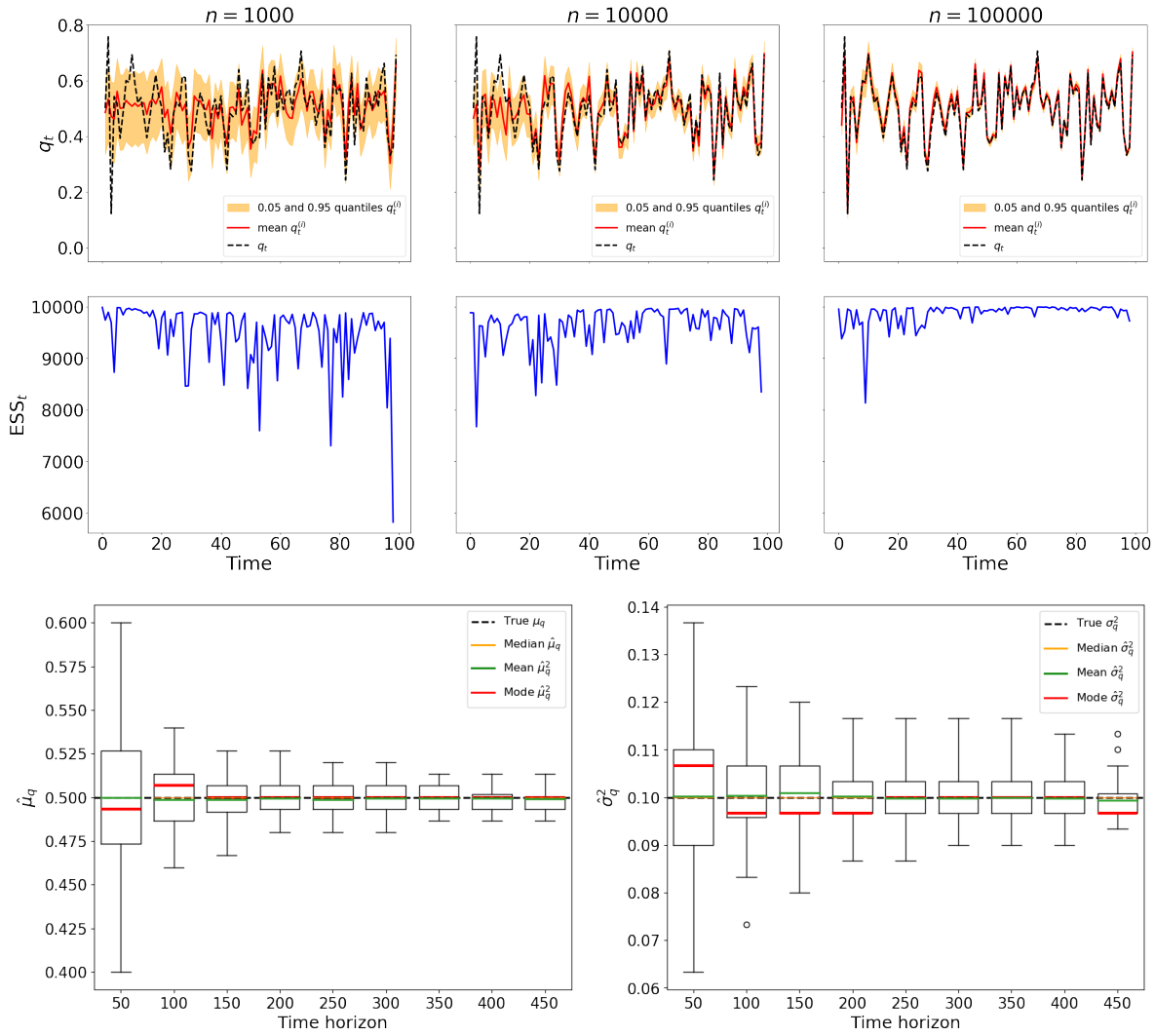

Figure 5.1: Pedagogical over-dispersed SEIR example. Top two rows: filtering distribution approximations and ESS obtained from PALSMC with $n_{part} = 10^4$ particles and increasing model population size n. Bottom row: maximum PALSMC estimation of hyper-parameters $\boldsymbol{\varphi} = [\mu_q \; \sigma_q^2]$ over increasing time horizons. Each boxplot summarises 100 hyper-parameter estimates.

each value of $n = 10^3, 10^4, 10^5$, generating data $\mathbf{y}_{1:100}$ from the model conditional on $q_{1:100}$, and running the PALSMC algorithm. We see that the effective sample size (ESS) is high across all values of population size n, indicating a good approximation to the r.h.s. of (5.1); this reflects the careful choice of proposal distribution. As in (5.2), for each $t \geq 1$, the PALSMC algorithm yields a Monte Carlo approximation $p(q_t|y_{1:t}) \approx \sum_{i=1}^{n_{part}} \bar{w}_t^{(i)} \delta_{q_t^{(i)}}$. The first row of plots in figure 5.1 demonstrates that these PALSMC filtering approximations concentrate on the true $q_{1:t}$ as the population size n grows. This is in keeping with the theory of chapter 4, which tells us that $\text{argmax}_{q_{1:t}, \boldsymbol{\vartheta}} \ell(y_{1:t}|q_{1:t}, \boldsymbol{\vartheta})$ converges to the data generating $[q_{1:t} \; \boldsymbol{\vartheta}]$ in the large population limit $n \to \infty$.

We also explored the ability of the procedure to recover the data generating hyperparameters

$\boldsymbol{\varphi} = [\mu_q \ \sigma_q^2]$; in the bottom plots of figure 5.1. Here each boxplot summarises 100 estimates, each estimate was obtained as follows: (1) simulate $q_{1:450} \sim f(\cdot|\boldsymbol{\varphi})$ and data $\mathbf{y}_{1:450}$ from the model with population size $n = 10^4$, (2) construct a 2-dimensional grid of candidate values for estimation of $[\mu_q \ \sigma_q^2]$, (3) run PALSMC with input $\mathbf{y}_{1:450}$ for each grid point, with $n_{part} = 10^4$ particles and $\boldsymbol{\vartheta}$ set to the DGP, (4) at time-steps $t = 50, 100, 150, ...$ report as an estimate of $[\mu_q \ \sigma_q^2]$ the value on the grid for which the largest value of $\widehat{\ell}(y_{1:t}|\mu_q, \sigma_q^2, \boldsymbol{\vartheta})$ was obtained across the PALSMC runs. We see from these boxplots that, for increasing time horizon T, the maximum PALSMC estimators obtained across 100 simulations converge towards the data generating $\boldsymbol{\varphi}$ with little bias.

Overall, these simulation results illustrate that, even in an over-dispersed setting, a large population can be useful in estimating $\bar{\boldsymbol{\theta}}_{1:t}$, whilst a large time horizon can be useful in recovering hyperparameters $\boldsymbol{\varphi}$.

Deriving a proposal informed by observations

Let $f(\cdot|\mu_q, \sigma_q^2)$ be the density associated with a $\mathcal{N}(\mu_q, \sigma_q^2)_{\geq 0, \leq 1}$ random variable. We would like to make proposals informed by observations, to that end we seek a Laplace approximation to:

$$\hat{p}(q_t \mid y_{1:t}, q_{1:t-1}) := \frac{\exp \ell(y_t \mid y_{1:t-1}, q_{1:t}) f(q_t|\mu_q, \sigma_q^2)}{\int \exp \ell(y_t \mid y_{1:t-1}, q_{1:t}) f(q_t|\mu_q, \sigma_q^2) dq_t}.$$

Suppressing dependence on the particle, let $\boldsymbol{\lambda}_t$ be calculated as per line 3 of algorithm 11. We have for some constant C_1 and C_2:

$$\begin{aligned} \log \hat{p}(q_t|y_t) &= \ell(y_t \mid y_{1:t-1}, q_{1:t}) + f(q_t|\mu_q, \sigma_q^2) + C_1 \\ &= y_t \log(q_t) + y_t \log(\lambda_t^{(3)}) - q_t \lambda_t^{(3)} - \log y_t! - \frac{1}{2}\left(\frac{q_t - \mu_q}{\sigma_q}\right)^2 + C_2. \end{aligned} \tag{5.4}$$

To get the mean for a Laplace approximation to (5.4) we must find it's maximum w.r.t. q_t, hence:

$$\begin{aligned} \frac{d \log \hat{p}(q_t \mid y_t)}{dq_t} = \frac{y_t}{q_t} - \lambda_t^{(3)} - \frac{q_t - \mu_q}{\sigma_q^2} &= 0 \\ \iff (q_t)^2 + (\lambda_t^{(3)}\sigma_q^2 - \mu_q) q_t - y_t \sigma_q^2 &= 0 \\ \implies q_t = \frac{1}{2}\left(\mu_q - \lambda_t^{(3)}\sigma_q^2 + \sqrt{(\lambda_t^{(3)}\sigma_q^2 - \mu_q)^2 + 4 y_t \sigma_q^2}\right) &=: \mu_{prop}. \end{aligned}$$

For the variance we find the second derivative and evaluate it at μ_{prop}:

$$\begin{aligned} \frac{d^2 \log \hat{p}(q_t \mid y_t)}{d(q_t)^2} &= -\frac{y_t}{(q_t)^2} - \frac{1}{\sigma_q^2} \\ \implies \sigma_{prop}^2 &= \left(\frac{y_t}{\mu_{prop}^2} + \frac{1}{\sigma_q^2}\right)^{-1}. \end{aligned}$$

To be congruent with the support of q_t we truncate the proposal to be:

$$\mathcal{N}(\mu_{prop}, \sigma_{prop}^2)_{\geq 0, \leq 1}, \tag{5.5}$$

denote its density as $\pi(\cdot|\mu_{prop}, \sigma^2_{prop})$.

Algorithm 11 PAL within SMC

initialise: $\bar{\boldsymbol{\lambda}}_{0,i} \leftarrow \boldsymbol{\lambda}_0$ for $i = 1$ to n_{part}.

1: **for** $t \geq 1$:
2: **for** $i = 1, \ldots, n_{part}$:
3: $\boldsymbol{\lambda}_t^{(i)} \leftarrow \left(\bar{\boldsymbol{\lambda}}_{t-1}^{(i)} \odot \boldsymbol{\delta}_t\right)^\top \mathbf{K}_{t,\eta\left(\bar{\boldsymbol{\lambda}}_t^{(i)}\right)} + \boldsymbol{\alpha}_t$
4: $q_t^{(i)} \sim \mathcal{N}\left(\mu_{prop}, \sigma^2_{prop}\right)_{\geq 0, \leq 1}$ as per (5.5)
5: $\mathbf{q}_t^{(i)} \leftarrow [0\ 0\ q_t^{(i)}\ 0]^\top$
6: $\log w_t^{(i)} \leftarrow \mathbf{y}_t^\top \log \boldsymbol{\lambda}_t^{(i)} \odot \mathbf{q}_t^{(i)} - \boldsymbol{\lambda}_t^\top \mathbf{q}_t^{(i)} - \log \mathbf{y}_t! + \log f(q_t^{(i)}|\mu_q, \sigma_q^2) - \log \pi(q_t^{(i)}|\mu_{prop}, \sigma^2_{prop})$
7: $\bar{\boldsymbol{\lambda}}_t^{(i)} \leftarrow \left(\mathbf{1}_m - \mathbf{q}_t^{(i)}\right) \odot \boldsymbol{\lambda}_t^{(i)} + \mathbf{y}_t$
8: **end for**
9: $\ell(y_t \mid y_{1:t-1}) \leftarrow \frac{1}{n_{part}} \sum_{i=1}^{n_{part}} w_t^{(i)}$
10: $\bar{w}_t^{(i)} \leftarrow w_t^{(i)} / \sum_{j=1}^{n_{part}} w_t^{(j)}$
11: **resample** $\left\{\bar{\boldsymbol{\lambda}}_t^{(i)}, q_t^{(i)}\right\}_{i=1}^{n_{part}}$ according to a systematic resampling scheme with weights $\left\{\bar{w}_t^{(i)}\right\}_{i=1}^{n_{part}}$
12: **end for**

CHAPTER 6

EXAMPLES

This chapter presents extensive examples of the PAL methodology being applied to real world data, it is organised as follows. Section 6.1 demonstrates how to use PALs within delayed acceptance particle Markov chain Monte Carlo to speed up exact Bayesian inference. Section 6.2 demonstrates how to use PALs within Stan to perform inference with Hamiltonian Monte Carlo. Section 6.3 demonstrates how to perform a model selection procedure within the PAL framework, with an application to rotavirus infections in Germany. Section 6.4 evaluates the role of unit-specific parameters in a large scale meta-population model of measles. Code for all examples is available at: https://github.com/Michael-Whitehouse/PAL. The algorithm and code for the measles example and the boarding school example LNA comparison were written in collaboration with Lorenzo Rimella.

6.1 Delayed Acceptance PMCMC for the boarding school influenza outbreak

This example illustrates the use of the PAL within delayed acceptance PMCMC, specifically the delayed acceptance Particle Marginal Metropolis Hastings (daPMMH) algorithm of Golightly et al. (2015).

Data and model

The data set is the well-known boarding school influenza outbreak data, recorded at a British boarding school in 1978 and reported in the British Medical Journal (Anon, 1978; Davies et al., 1982). The data are available in the R package “pomp” (King et al., 2016). On day one there was one infection and over the course of the 14 day epidemic a total of 512 students reported

symptoms from a population of $n = 763$. The observations are prevalence data: daily counts of the total number of symptomatic individuals. We cast this an instance of case (I), using a simple SIR model, where the initial state of the population is fixed to $[763\ 1\ 0]^\top$ and we define the matrix $\mathbf{K}_{t,\boldsymbol{\eta}}$ as follows:

$$\mathbf{K}_{t,\boldsymbol{\eta}} = \begin{bmatrix} e^{-\beta\eta^{(2)}} & 1 - e^{-\beta\eta^{(2)}} & 0 \\ 0 & e^{-\gamma} & 1 - e^{-\gamma} \\ 0 & 0 & 1 \end{bmatrix},$$

where β and γ are to be estimated. Observations y_t are modelled as binomially under-reported counts of infected individuals, that is, given $x_t^{(2)}$, $y_t \sim \text{Bin}(x_t^{(2)}, q)$ where $q \in [0,1]$ is unknown and to be estimated. To connect with the notation of algorithm 6 we have $\mathbf{y}_t \equiv [0\ y_t\ 0]^\top$ and $\mathbf{q}_t \equiv [0\ q\ 0]^\top$ for $t \geq 1$.

Delayed Acceptance Particle Marginal Metropolis Hastings

In the standard PMMH algorithm (Andrieu et al., 2010), one calculates a particle filter approximation to the likelihood for each proposed parameter value, which is typically a computationally intensive operation. The daPMMH algorithm introduces an additional 'pre-screening' acceptance step based on an approximate likelihood which is assumed to be cheap to evaluate. Only if the proposed parameter is accepted in this initial step is a particle filter approximation to the likelihood then evaluated; thus in performing this additional step, one seeks to avoid running a particle filter for proposals which are likely to be rejected. Details of the validity of the scheme, in the sense that it indeed targets the true posterior distribution over the parameters, can be found in Golightly et al. (2015). Algorithm 12 illustrates how to use a PAL within a daPMMH.

We stress that, although for the SIR model the number of compartments is small ($m = 3$), and for the data set in question the population size is fairly small ($n = 763$), this actually presents a stern relative speed test for PALs versus particle filters: the particle filter element of the daPMMH and PMMH algorithms involves simulating from the latent compartmental model, and the overall cost of the particle filter, therefore, grows with both the number of compartments and the size of the population, as well as the number of particles. By contrast, evaluating the PAL involves no random number generation and has a cost independent of population size. Thus, if a relative speed gain using PALs can be demonstrated with a small population size and small number of compartments, it is reasonable to expect an even greater relative speed gain for models with larger numbers of compartments and larger populations.

Results

We compare the performance of three algorithms: PALMH: a Metropolis-within-Gibbs algorithm with the PAL substituted in place of the exact likelihood, i.e., targeting an approximation to the exact posterior distribution; PMMH: a standard Particle Marginal Metropolis-Hastings within Gibbs; daPMMH: a delayed acceptance Particle Marginal Metropolis-Hastings within Gibbs, in

which we use the PAL for the delayed acceptance step. We apply these three methods to both a synthetic and a real dataset. For all three algorithms we use Gaussian random walk proposals independently for each element of $\boldsymbol{\theta}$. The random walk variances are tuned to ensure acceptance rates between 20% and 40%. The PMMH and daPMMH algorithms were each run with 1000 particles. All experiments were run on a single core of a 1.90 GHz i7-8650U CPU.

The parameters of the model are collected in the vector $\boldsymbol{\theta} = [\beta\ \gamma\ q]^\top$. We consider a fairly vague prior $p(\boldsymbol{\theta}) = p(\beta)p(\gamma)p(q)$, where $p(\beta)$ and $p(\gamma)$ are truncated Gaussian densities $\mathcal{N}(0,1)_{\geq 0}$ and $p(q)$ is a truncated Gaussian density $\mathcal{N}(0.5,0.5)_{\geq 0,\leq 1}$.

Simulated data. We simulated an epidemic for 14 days with the parameter regime $\boldsymbol{\theta}^* = [\beta^*\ \gamma^*\ q^*]^\top = [2\ 0.5\ 0.8\ ]^\top$. For each of the PALMH, PMMH, and daPMMH we ran a 5×10^5 length chain, discarded 10^5 for burn in and then thinned to a sample of 2.5×10^5. Trace plots, autocorrelation plots, and posterior sample histograms for each scheme are presented in section B.1 of the appendix, the rates of decay of the ACFs with respect to lag for the daPMMH and PMMH algorithms are similar, the rate of decay for the PALMH algorithm is faster. The Monte Carlo approximations of the posterior marginals are closely matched across the three algorithms, see table 6.1 for summary statistics, and are concentrated around the data generating parameters. A single evaluation of the PAL took a mean time of 9.4×10^{-6} seconds, the particle filter approximation to the likelihood took a mean time of 4.5×10^{-3} seconds, both algorithms were implemented with Rcpp.

Real data. On the real data we ran the PALMH, PMMH, and daPMMH for 5×10^5 iterations each, with run times of 12.2 minutes, 4.5 hours, and 2.8 hours respectively, exhibiting the speed benefits of the PAL approach. Trace plots, autocorrelation plots, and approximate posterior sample histograms for each scheme are presented in section B.1, the rate of decay of the ACF with lag is similar for the daPMMH and PMMH algorithms, the rate of decay for the PALMH algorithm is faster. The daPMMH and PMMH algorithms yield very similar approximate posterior marginals as expected – see table 6.2. The posterior marginals obtained from the PALMH scheme exhibit modes in different locations to those from PMMH/daPMMH, with the following epidemiological interpretation. The approximate posterior marginals obtained from PALMH correspond to a fast growing outbreak (large β), with individuals spending longer in the infected state (small γ) and a relatively lower reporting rate (relatively small q). By contrast, the PMMH/daPMMH marginals suggest a slower outbreak (smaller β) with less time spent in the infected compartment (larger γ), but with a higher case reporting rate (relatively high q). Posterior predictive checks (Gelman et al., 1995) show that, while having contrasting epidemiological interpretations (potentially due to model mis-specification), both PALMH and PMMH/daPMMH achieve good coverage of the data, see figure 6.1. The mean trajectories from these posterior predictive distributions reflect the above interpretations of posterior marginals. The posterior predictive means and credible regions were calculated from 10000 samples from the posterior predictive distributions produced by the

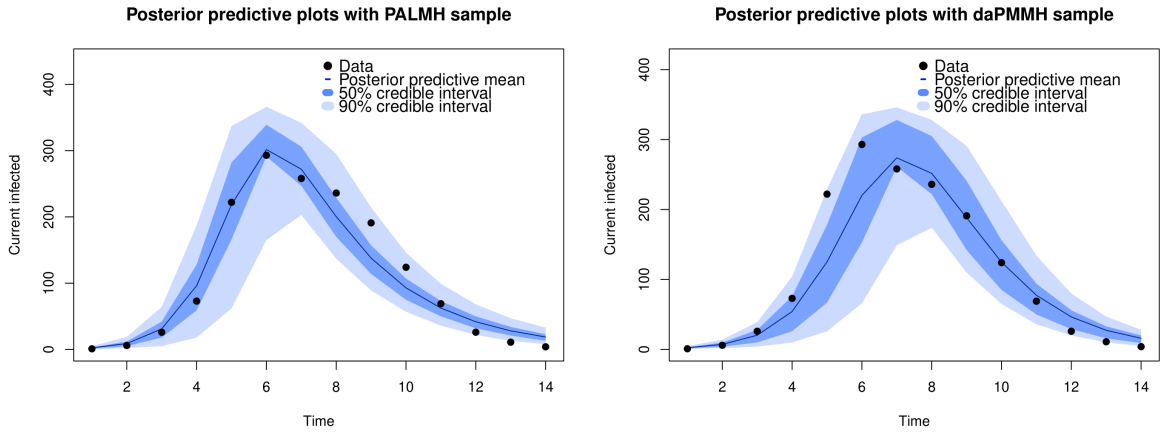

Figure 6.1: Boarding school influenza example. Means and credible intervals for posterior predictive distributions.

PALMH and daPMMH respectively; each sample from the posterior predictive distribution was generated by: sampling a parameter $\boldsymbol{\theta}'$ from the approximate posterior; then using $\boldsymbol{\theta}'$ to simulate an epidemic trajectory and data record from the model.

We performed inference on this data set using a Linear Noise Approximation to the likelihood as described in section 4 of Fearnhead et al. (2014) within a Metropolis Hastings scheme, the full results can be found in the appendix. We find that, whilst the LNA and the PAL perform similarly in terms of parameter inference, the Latent Compartmental Model from which the PAL is derived is more congruent with reality than the SDE model, since the latter allows non-integer and negative counts of individuals in compartments. Furthermore, a single evaluation of the PAL was approximately ~ 100 times faster than a single evaluation of the LNA marginal likelihood for this dataset.

Parameter	True value	PALMH	PMMH	daPMMH
β	2	2.10 (1.88, 2.34)	2.08 (1.85 2.35)	2.08 (1.85,2.35)
γ	0.5	0.51 (0.42, 0.63)	0.53 (0.44 0.65)	0.53 (0.43, 0.65)
q	0.8	0.81 (0.70, 0.94)	0.82 (0.71, 0.96)	0.82 (0.71, 0.96)

Table 6.1: Boarding school model posterior means and 95% credible interval, synthetic data.

Parameter	PALMH	PMMH	daPMMH
β	2.98 (2.60,3.30)	2.30 (2.00,2.68)	2.30 (2.00,2.68)
γ	0.406 (0.35,0.47)	0.58 (0.47,0.68)	0.58 (0.47,0.68)
q	0.69 (0.62,0.77)	0.90 (0.76,0.99)	0.90 (0.76,0.99)

Table 6.2: Boarding school model posterior means and 95% credible interval, real data.

PALMH: prior sensitivity analysis

Section 6.1 explores Bayesian analysis on real data under vague priors, with results in table 6.2. Whilst both the PALMH and PMMH schemes result in identical inferences on simulated data, there are discrepancies on the real-world boarding school data - which could be attributed to a mis-specified model. The estimated parameters under the PALMH suggest an R_0 of around 7.3 (the PMMH estimates suggest an R_0 of around 4) which is consistent with the entire population being infected at some point during the epidemic - one can question whether this is a realistic inference. Given the closed nature of this epidemic, along with the likelihood of close monitoring of the individuals in the system, one could afford to place stronger priors on q. In this section, we explore the inferences one can make using the PALMH scheme under more informative priors. We consider the following scenarios:

1. $\mathcal{N}(0.5, 0.5)_{\geq 0, \leq 1}$ - the vague prior used in the original analysis.
2. Beta$(9, 1)$ - an informative prior with mean 0.9 and variance 0.0082 and mode 1.
3. Beta$(95, 5)$ - a strongly informative prior with mean 0.95 and variance 0.00047 and mode 1.
4. $\delta_{0.9}$ - an atomic prior on 0.9 (the mean inferred q under the PMMH analysis).
5. $\delta_{0.95}$ - an atomic prior on 0.95.

For each of these we ran a 5×10^5 length chain, discarded 10^5 for burn in and then thinned to a sample of 2.5×10^5, we summarise our findings in table 6.3. We find that as the prior belief in a high reporting rate is strengthened, the resulting estimated R_0 lowers. If one places strong prior belief in a high reporting rate, see the Beta$(95, 5)$ and $\delta_{0.95}$ columns, then the inferred R_0 falls in line with our findings using the PMMH procedure.

Parameter	$\mathcal{N}(0.5, 0.5)_{\geq 0, \leq 1}$	Beta$(9, 1)$	Beta$(95, 5)$	$\delta_{0.9}$	$\delta_{0.95}$
β	2.98(2.60,3.30)	2.77(2.29,3.22)	2.39(2.01,2.84)	2.46(2.08,2.92)	2.35(1.98,2.77)
ρ	0.41(0.35,0.47)	0.44(0.35,0.57)	0.58(0.49,0.68)	0.55(0.48,0.63)	0.59(0.52,0.68)
q	0.69(0.62,0.77)	0.74(0.63,0.90)	0.94(0.86,0.97)	0.90	0.95
R_0	6.91	6.47	4.14	4.52	4.05

Table 6.3: Boarding school model PALMH prior sensitivity analysis. Posterior means and 95% credible interval, real data under various prior assumptions with R_0 posterior mean point estimates.

To investigate the disparity between inferences using the PALMH procedure vs the PMMH procedure when applied to real data, exhibited in table 6.2, we repeated the analysis with a fixed $q = 0.9$ (equivalent to the $\delta_{0.9}$ prior). The resulting posteriors for the PALMH and PMMH procedures still exhibited some differences, but were much more similar as a result of this stronger assumption:

- The posterior mean and 95% credible interval for β under the PMMH procedure was 2.14 (1.91,2.40), to be compared with 2.46(2.08,2.92) for PALMH.
- The posterior mean and 95% credible interval for γ under the PMMH procedure was 0.58 (0.53,0.64), to be compared with 0.55(0.48,0.63) for PALMH.
- The posterior mean estimates for R_0 under the PMMH and PALMH procedures were 4.52 and 3.70, respectively.

Algorithm details for section 6.1

The following algorithm describes how the PAL can be used within a delayed acceptance pmcmc scheme.

Algorithm 12 Delayed acceptance PMMH algorithm with PAL

Initialise: $i = 0$, set $\boldsymbol{\theta}_0$ arbitrarily.

1: Run a particle filter to produce an approximation to $p(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_0)$ and denote this as $\hat{p}(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_0)$.

2: Run algorithm 6 to produce a PAL approximation to $p(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_0)$ and denote this as $\hat{p}_a(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_0)$.

3: **for** $i \geq 1$:

4: sample $\boldsymbol{\theta}_* \sim q(\cdot \mid \boldsymbol{\theta}_{i-1})$.

5: **stage 1**

- Run algorithm 6 to produce a PAL approximation to $p(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_*)$ and denote this as $\hat{p}_a(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_*)$.
- With probability:

$$\alpha_1(\boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_*) = \min\left\{1, \frac{\hat{p}_a(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)}{\hat{p}_a(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_{i-1})p(\boldsymbol{\theta}_{i-1})} \frac{q(\boldsymbol{\theta}_{i-1} \mid \boldsymbol{\theta}_*)}{q(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{i-1})}\right\},$$

run a particle filter to produce an approximation to $p(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_*)$, denote this as $\hat{p}(\mathbf{y}_{1:t} \mid \boldsymbol{\theta}_*)$ and go to **Stage 2**. Otherwise, set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$, set $i = i + 1$ and return to 4.

6: **stage 2**

With probability

$$\alpha_2(\boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_*) = \min\left\{1, \frac{\hat{p}(\mathbf{y}_{1:t}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)}{\hat{p}(\mathbf{y}_{1:t}|\boldsymbol{\theta}_{i-1})p(\boldsymbol{\theta}_{i-1})} \frac{\hat{p}_a(\mathbf{y}_{1:t}|\boldsymbol{\theta}_{i-1})p(\boldsymbol{\theta}_{i-1})}{\hat{p}_a(\mathbf{y}_{1:t}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)}\right\},$$

set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_*$, otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$. Set $i = i + 1$ and return to 4.

7: **end for**

Time comparisons with the Linear Noise Approximation

For comparisons with the PAL we consider LNAMH: a Metropolis-within-Gibbs algorithm with a Linear noise approximation (LNA) to the likelihood of a stochastic differential equation model used in the accept/reject step, see Fearnhead et al. (2014) for details. We apply the LNAMH to the real boarding school dataset to compare and contrast to the PAL, for these comparisons

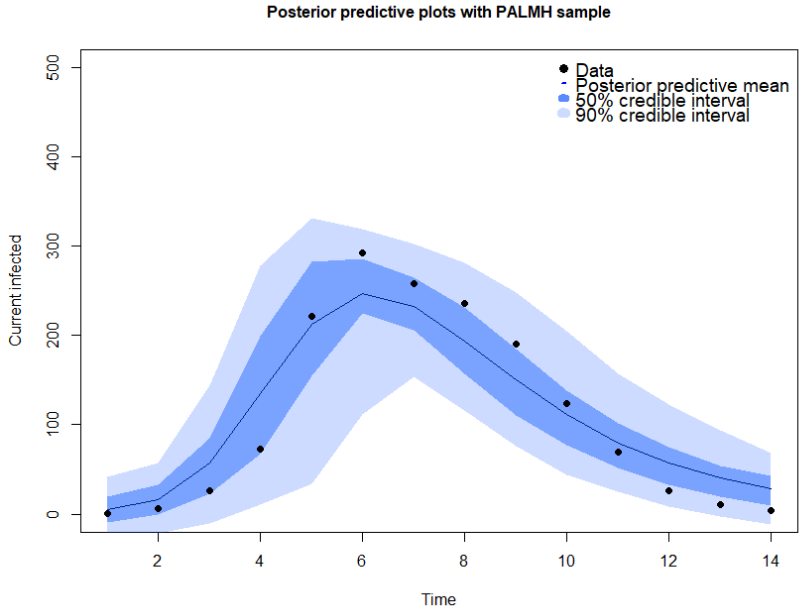

Figure 6.2: Posterior predictive distribution for LNAMH sample. To produce this plot we sampled a parameter from the approximate posterior and simulated from the SDE model 1000 times.

we implement the PAL in base R, whereas the LNA computations use base R interfaced with FORTRAN for cumbersome ODE solving calculations. The LNAMH implementation introduces an extra parameter in the variance of a Gaussian observation model, analogous to $V(\boldsymbol{\theta})$ in section 4.2 of Fearnhead et al. (2014), which we will denote as v; we consider a vague truncated Gaussian prior of $\mathcal{N}(400,300)_{\geq 0}$. We ran the chain for 100k iterations, discarded the first 20k and thinned to a sample of 25k to produce the posterior histograms.

The posterior predictive plot associated with the LNAMH sample, figure 6.2, demonstrates good coverage of the data, yet they help illustrate some of its shortfalls in comparison to the PAL approach: the Gaussian nature of the ingredients of the LNA permits non-integer and even allows negative valued observations, which is clearly not parsimonious with reality; further, modelling with a constant in time observation variance leads to underconfidence in the start and end of the data record. In order to circumvent these issues within the LNA framework, one would have to turn to sophisticated and expensive methods; alternatively, one could avoid each of them for free through the use of PALs.

Figure 6.3 reports the mean time ratio between a single evaluation of the LNA likelihood and a single evaluation of the PAL for varying ODE solver intermediate time step choices for the LNA and analogous choice of h for the PAL. The order of magnitude of the speed gains is around 100 for the PAL, demonstrating the significant speed benefits given by the simplicity of computations needed to compute the PAL in comparison to cumbersome ODE solution calculations. Experiments were performed on an Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz processor.

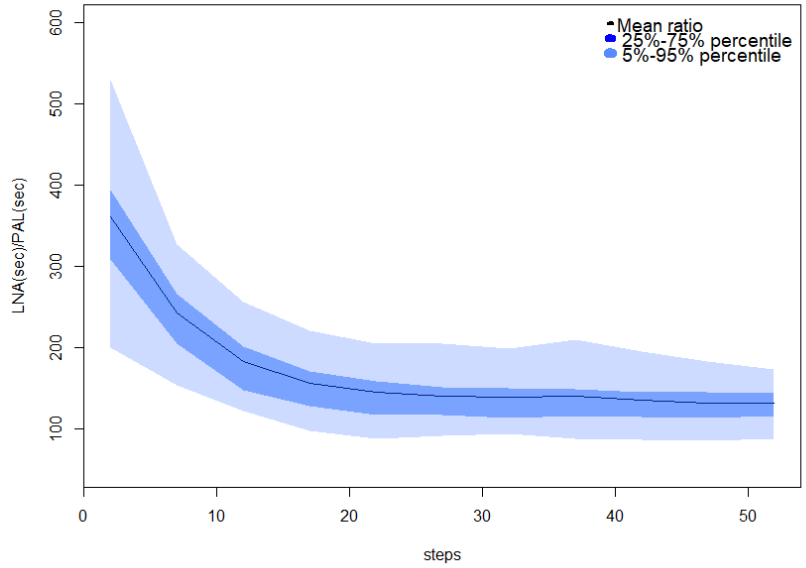

Figure 6.3: Time comparisons for the LNA. The ratio of one evaluation of the LNA likelihood to one evaluation of the PAL for varying ODE solver intermediate time steps, with the comparative PAL evaluation run with $h = 1$/number of timesteps. Percentiles are based on 1000 runs.

6.2 Inference using automatic differentiation and HMC for an age-structured model of 'flu

In this example, we demonstrate PALs for an age-structured model of a 1957 outbreak of influenza in Wales. Computation is performed using the probabilistic programming language Stan (Carpenter et al., 2017), taking advantage of automatic differentiation to implement Hamiltonian Monte Carlo (HMC). This example also highlights how the general Latent Compartmental Model can accommodate discrete or discretisable covariates associated with subpopulations: in this case the covariates are indicators of the age-group which individuals belong to and this is reflected in the compartment structure of the model.

Data and Model

The data consist of 19 weeks of incidence data in the form of GP symptom reports for a town with population size 8000 across 4 age groups: 00 – 04, 05 – 14, 15 – 44, and 45+. The data were analysed by Vynnycky and Edmunds (2008) and are available via the Github page associated with (Andrade and Duggan, 2020). For each age group $k = 1,\ldots,4$,

$$S_{k,t+1} = S_{k,t} - B_{k,t}, \qquad E_{k,t+1} = E_{k,t} + B_{k,t} - C_{k,t},$$
$$I_{k,t+1} = I_{k,t} + C_{k,t} - D_{k,t}, \qquad R_{k,t+1} = R_{k,t} - D_{k,t},$$

with conditionally independent increments: $B_{k,t} \sim \text{Bin}(S_{k,t}, 1-e^{-h\bar{\beta}_{k,t}})$, $C_{k,t} \sim \text{Bin}(E_{k,t}, 1-e^{-h\rho})$, $D_{k,t} \sim \text{Bin}(I_{k,t}, 1-e^{-h\gamma})$, where

$$\begin{bmatrix} \bar{\beta}_{1,t} \\ \vdots \\ \bar{\beta}_{4,t} \end{bmatrix} = \underbrace{\begin{bmatrix} \beta_{11} & \dots & \beta_{14} \\ \vdots & \ddots & \vdots \\ \beta_{14} & \dots & \beta_{44} \end{bmatrix}}_{=:\mathbf{B}} \begin{bmatrix} I_{1,t} \\ \vdots \\ I_{4,t} \end{bmatrix} \frac{1}{n}. \tag{6.1}$$

$\mathbf{B}$ is a symmetric matrix with element β_{ij} representing the rate at which two individuals, one from the susceptible compartment of the ith age group and the other from the infective compartment of the jth age group come into effective contact.

The mean time spent in the exposed compartment $1/\rho$ and the mean recovery time $1/\gamma$ are taken to be independent of age group and set to be 1.5 days, following Andrade and Duggan (2020). We assume that the model evolves daily with $h = 1/7$ and that observations consist of cumulative weekly transitions from the E to I compartments for each age group, that is we have observation times at times $\tau_r = 7r$ for $r = 1,\dots,R$ corresponding to the end of each week. In the setting of case (II) we denote observations $\bar{\mathbf{Y}}_{k,r} = \sum_{s=\tau_{r-1}+1}^{\tau_r} \mathbf{Y}_{k,t}$ where each element of each $\mathbf{Y}_{k,t}$ is equal to zero except the $(2,3)$th element corresponding to transitions from compartment E to I which, conditional on $C_{k,t}$, is distributed $Y_{k,t}^{(2,3)} \sim \text{Bin}(C_{k,t}, Q_{k,t}^{(2,3)})$, where $\mathbf{Q}_{k,t} \in [0,1]^{4\times 4}$ has elements equal to zero except for the $(2,3)$th entry which is equal to an age group dependant under reporting parameter $q_k \in (0,1)$ which is to be estimated. We give details of how this model is written as an instance of the Latent Compartmental Model and the algorithm used to calculate the PAL in the appendix.

Hamiltonian Monte Carlo with automatic differentiation in Stan

We now consider MCMC sampling to approximate the posterior $p(\boldsymbol{\theta}|\bar{\mathbf{Y}}_{1:R})$. The probabilistic programming language Stan (Carpenter et al., 2017) provides a framework for implementing HMC – a type of MCMC algorithm which uses auxiliary "momentum" variables to help explore the posterior – in which the user only needs to specify priors and provide a function which evaluates the likelihood for the model. Stan uses Automatic Differentiation (AD) to compute gradients and update the auxiliary HMC variables without the need for user input. Since the PAL consists of recursive compositions of elementary linear algebra operations, it is a natural candidate for AD.

Results

We implemented a Stan program incorporating the PAL, details of which are given in section B.2 of the appendix. We stress that here we do not correct for the fact that the PAL is only an approximation to the true likelihood, so Stan is targeting an approximation to the true

intractable posterior, although in a separate example in the appendix we explore corrections using Delayed-Acceptance MCMC methods.

The parameters to be estimated are $\boldsymbol{\theta} = [\beta_{11} \cdots \beta_{44}\ q_1 \cdots q_4]^\top$, the initial state for each age group is assumed known as $\mathbf{x}_{1,0} = [948\ 0\ 1\ 0]^\top$, $\mathbf{x}_{2,0} = [1689\ 0\ 1\ 0]^\top$, $\mathbf{x}_{3,0} = [3466\ 0\ 1\ 0]^\top$, $\mathbf{x}_{4,0} = [1894\ 0\ 1\ 0]^\top$. We used vague gamma priors $\beta_{ij} \sim \text{Gamma}(5,1)$ for $i,j = 1,\ldots,4$ and a vague truncated normal prior $q_k \sim \mathcal{N}(0.5, 0.5)_{\geq 0, \leq 1}$ for $k = 1,\ldots,4$. The HMC sampler was run to produce a chain of length 5×10^5 iterations, a burn-in period of size 10^5 was discarded and the remaining was thinned to produce a sample of 2.5×10^4. We report approximate posterior distributions and trace plots in section B.2 of the appendix, these show no signs of unsatisfactory mixing. Figure 6.5 reports the posterior predictive distributions and credible intervals, we see good coverage of observed data.

We repeated the analysis using an ODE version of the same age-structured SEIR model, from Andrade and Duggan (2020), with a Poisson reporting model: we use as emission distribution a Poisson distribution with rate given by the ODE solution scaled by an under-reporting parameter. This was implemented in the Stan framework using the code available in Andrade and Duggan (2020), we again sampled a chain of length 5×10^5 iterations, discarded a burn-in period of size 10^5, and thinned the remaining to produce a sample of 2.5×10^4. To calculate the reproduction number R_0 for stratified models such as this, one must calculate the so called *next generation matrix* (Van den Driessche and Watmough, 2002) which has elements given by $\frac{n_i \beta_{ij}}{n\gamma}$ where n_i is the population size of the ith age group. R_0 is then given by the largest modulus of the eigenvalues of the next generation matrix (Diekmann et al., 1990). Using this definition, we can produce approximate posterior distributions of R_0 using each of the PAL and ODE procedures, which we report in figure 6.4. The approximate posteriors concentrate around 1.42 using the PAL and 1.82 using the ODE model. This disparity in estimates can be related to the features of the posterior predictive distributions reported in figures 6.5 and 6.6: the distribution of trajectories in figure 6.6 appears to 'overshoot' the data in comparison to those in figure 6.5, reflecting the higher force of infection implied by the ODE procedure in contrast to the PAL procedure. These posterior predictive plots also exhibit the inherent inflexibility of the ODE model: since the latent process is deterministic, random variations in the data away from the ODE trajectory must be explained as observation error. As is apparent in the 45+ age group, this rigidity in modelling results in overconfidence and a poor fit compared to that of the stochastic model combined with the PAL procedure.

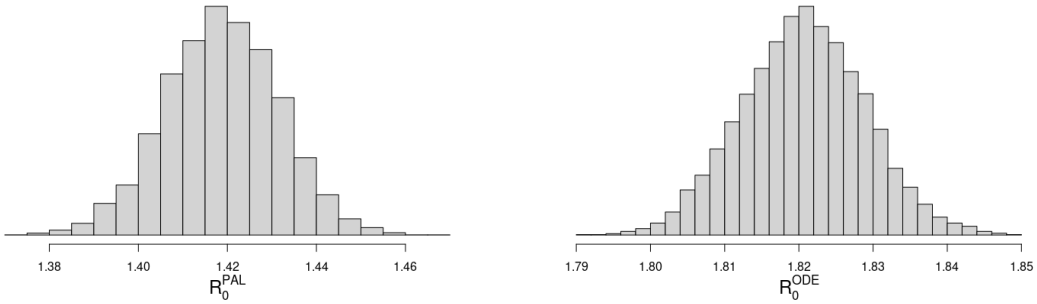

Figure 6.4: Age-structured 'flu example. Approximate posterior distributions for R_0 under the PAL and ODE procedures.

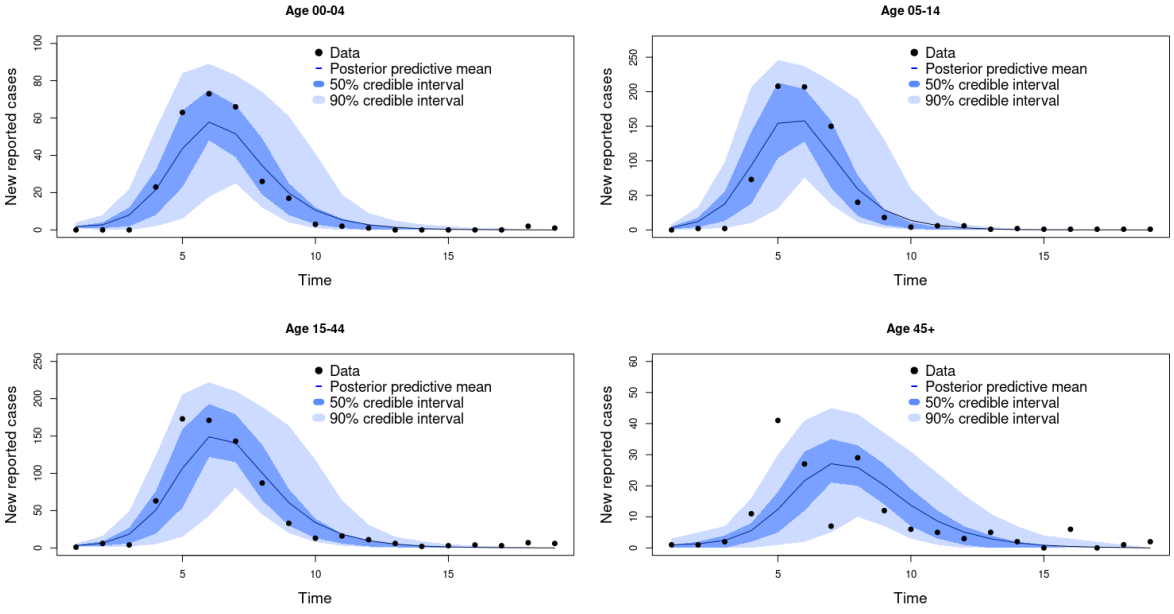

Figure 6.5: Age-structured 'flu example. Posterior predictive distributions obtained from inference under the stochastic model using the PAL within Stan.

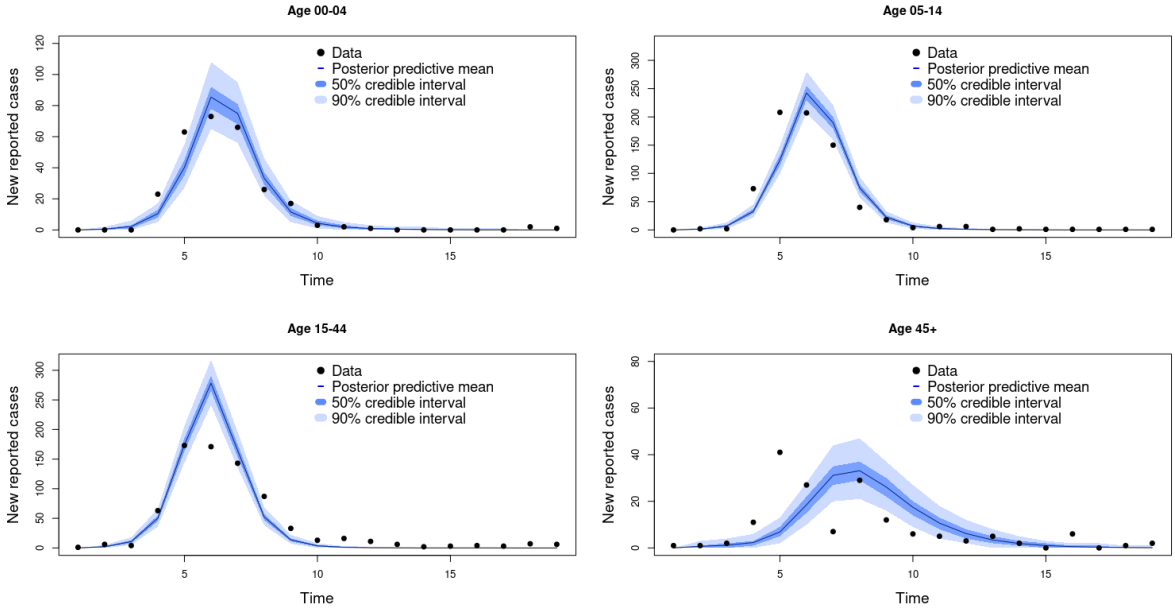

Figure 6.6: Age-structured 'flu example. Posterior predictive distributions obtained from inference under the ODE model using Stan.

6.3 Comparison of over-dispersion mechanisms in a model of rotavirus

In this section we explore a model selection task in which an equi-dispersed model is nested within a larger class of models including over-dispersion, using the approach of chapter 5. The rotavirus data and model we consider are inspired by Stocks et al. (2020), who assessed the fit of a family of continuous time, stochastic models with varying degrees of over-dispersion using the Akaike Information Criterion (AIC) (Akaike, 1974).

Models

The data considered consist of weekly incidence counts of rotavirus infections in Germany for 3 age groups over the 8 year period 2001-2008. We consider a discrete-time version of the model of Stocks et al. (2020) which compartmentalises a population of $n = 82,372,825$ into an age stratified SIR model $\{S_{k,t}, I_{k,t}, R_{k,t}\}_{k=1}^{3}$ comprising 3 age groups: $0-4$, $5-59$, and $60-99$. Given the number of susceptibles in age group k at time t after immigration, which we denote $\bar{S}_{k,t}$, and the number of infected individuals in each age group $\boldsymbol{I}_t = [I_{t,1}\, I_{t,2}\, I_{t,3}]^\top$, for $t \geq 1$ the number of new infected individuals in each age group $k = 1,2,3$ at time step t is conditionally distributed:

$$B_{k,t} \sim \text{Binom}\left(\bar{S}_{k,t}, 1 - \exp\left\{-\frac{\boldsymbol{\beta}_k^\top \boldsymbol{I}_t}{n}\chi_t\right\}\right), \tag{6.2}$$

with $\boldsymbol{\beta}_k = [\beta_k\ \beta_k\ \beta_k]^\top$ where $\beta_k > 0$ denotes the force of infection experienced by age group k, and $\chi_t = \big(1 + \rho\cos(2\pi t/w + \phi)\big)$ denotes a deterministic seasonality component with amplitude $\rho \in [0,1]$, phase $\phi \in [0, 2\pi]$, and period length $w > 0$, which we set to correspond to 1 year. Other details of the latent compartmental model are given in section B.3 of the appendix. We assume an aggregated transmission model, with weekly observations coming at times $\tau_r = 4r$ for $r = 1, \ldots, R$. For each age group observations are conditionally distributed $Y_{k,r} \sim \text{Binom}\left(\sum_{t=\tau_{r-1}}^{\tau_r} B_{k,t}, q_{k,r}\right)$.

We consider three variants of this model:

EqEq: a fully equi-dispersed model, in which $q_{k,r} = \mu_q \in [0,1]$, and μ_q is assumed known as in Stocks et al. (2020);

EqOv: an equi-dispersed latent compartmental model and an over-dispersed observation model, the same as EqEq except that $q_{k,r} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_q, \sigma_q^2)_{\geq 0, \leq 1}$ where $\sigma_q^2 > 0$ is to be estimated;

OvOv: over-dispersion in both the latent and observation models, the same as EqOv except that we augment χ_t in equation (6.2) to $\chi_t \xi_r$where for $r \geq 1$, $\xi_r \overset{\text{i.i.d}}{\sim} \text{Gamma}(\sigma_\xi, \sigma_\xi)$ are multiplicative disturbances with mean 1 and $\sigma_\xi > 0$ is to be estimated.

Inference

The parameters we estimate in each instance of the model are given by: EqEq: $\boldsymbol{\theta} = [\beta_1\ \beta_2\ \beta_3\ \phi\ \rho]$; EqOv: $\boldsymbol{\vartheta} = [\beta_1\ \beta_2\ \beta_3\ \phi\ \rho]$ with $\{\bar{\boldsymbol{\theta}}_r\}_{r\geq 1} = \{\mathbf{q}_r\}_{r\geq 1}$ and $\boldsymbol{\varphi} = \sigma_q$; OvOv: $\boldsymbol{\vartheta} = [\beta_1\ \beta_2\ \beta_3\ \phi\ \rho]$ with $\{\bar{\boldsymbol{\theta}}_r\}_{r\geq 1} =$

$\{[\mathbf{q}_r\ \xi_r]\}_{r\geq 1}$ and $\boldsymbol{\varphi} = [\sigma_q\ \sigma_\xi]$. The PALSMC algorithm for this model is given section B.3 in the appendix. For parameter estimation the approximate likelihoods of each of the models EqEq, EqOv, OvOv, obtained from PALSMC were maximised using a finite-difference coordinate ascent algorithm; we ran the optimisation 100 times, initialised randomly over a range of feasible values. Plots evidencing convergence are in section B.3 of the appendix. The algorithm was implemented using R and Rcpp on a node of the University of Bristol's BluePebble cluster, although we did exploit parallelisation.

We note that a PAL, e.g. the exponential of the r.h.s. of (3.10), is a valid likelihood function associated with a product of vector-Poisson distributions whose intensity parameters are defined through the corresponding filtering algorithm, e.g. algorithm 6. Similarly the output from PALSMC, e.g. (5.2) from algorithm 10, is a Monte Carlo approximation to a valid likelihood for a mixture of products of vector Poisson distributions. This validity justifies the use of AIC for model comparison but with the log-PAL, or the log-output from PALSMC, substituted in place of the usual log-likelihood.

Model	AIC	Ave. comp. time
EqEq	98866.65	30 sec
EqOv	15154.75	2 hr
OvOv	13778.08	3 hr
Stocks et al. (2020)	20134.38	11 hr

Table 6.4: Rotavirus example. Model assessment and computation time.

Parameter	EqEq	EqOv	OvOv
β_1	12.15	12.74	11.48
β_2	0.22	0.21	0.25
β_3	0.34	0.31	0.35
ϕ	0.017	0.14	0.14
ρ	0.022	0.19	0.16
σ_q^2	n/a	0.042	0.021
σ_ξ	n/a	n/a	66.89

Table 6.5: Rotavirus example. Parameter estimates.

As a benchmark comparison, we fitted an ARMA(2,0,1) model to the log-transformed data, which gives an AIC of 23043 (details are given in the appendix section B.3). Table 6.4 gives the AIC values for each of our models, along with the best AIC value reported by Stocks et al. (2020), which was for a model with over-dispersion in the transition model in the form of multiplicative gamma distributed noise, and over-dispersion in the observation model through negative-binomial reporting. This model of Stocks et al. (2020) is therefore qualitatively most similar to our model OvOv. The average computation times were calculated over 100 runs of the coordinate ascent procedure. We find that, whilst we can fit EqEq with high computational efficiency, our two over-dispersed models achieve a substantially better AIC score, indicating a much better fit with increasing over-dispersion. Both EqOv and OvOv outperform Stocks et al. (2020) AIC and computation time, although of course the latter is implementation-dependent. Figure 6.7 demonstrates the increase in goodness of fit that an over-dispersed model provides for the rotavirus data, we see that prediction intervals for OvOv drastically outperform those for EqEq in terms of coverage.

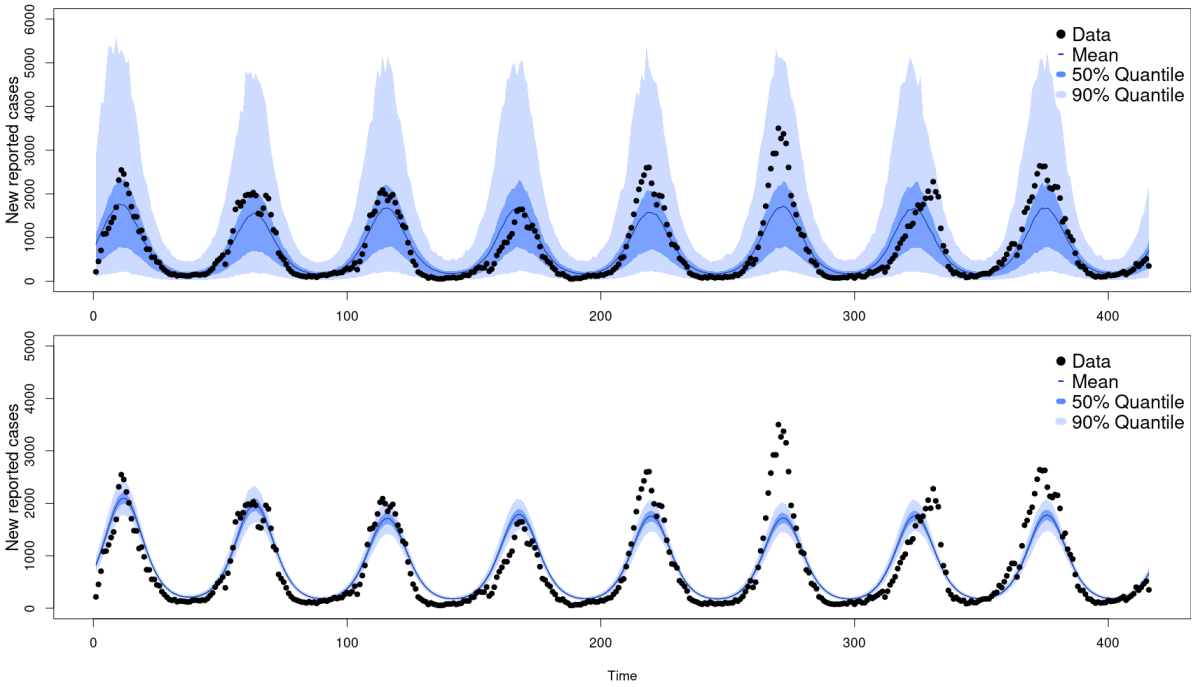

Figure 6.7: Rotavirus example. Prediction intervals for age group $0-4$ corresponding to 1000 realisations of OvOv (top panel) and EqEq (bottom panel), using maximum PALSMC parameter estimates.

The estimated values of β_1 and β_2 we find for all three models EqEq, EqOv and OvOv (table 6.5) are quite similar to those reported by Stocks et al. (2020), but we find a slightly lower value of β_3. For EqOv and OvOv we find a similar seasonal amplitude ρ but slightly larger phase ϕ than Stocks et al. (2020). The seasonal R_0 ranges for each model are: EqEq $(0.98, 1.027)$,EqOv $(0.83, 1.22)$, and OvOv $(0.82, 1.14)$ compared to $(0.855, 1.152)$ obtained by Stocks et al. (2020). The better fit of EqOv and OvOv compared to Stocks et al. (2020) may thus be attributed to some combination of quite subtle differences in estimates of parameters related to disease transmission, together with the difference between the negative binomial observation model in Stocks et al. (2020) and the way EqOv and OvOv treat the $q_{k,r}$ as latent variables.

6.4 Evaluating the role of unit-specific parameters in a meta-population model of measles

In this section we illustrate how the PAL framework can be used to calibrate a more complex, larger-scale model, and compare the fit of sub-models with different levels of unit-specific parameters.

Model

We consider a discrete time version of a measles model originally presented by Xia et al. (2004), subsequently extended into a spatio-temporal framework by He et al. (2010) and recently explored by Park and Ionides (2020) using guided intermediate resampling filter (GIRF) techniques.

The model describes the evolution of recurrent pre-vaccination measles epidemics in $J = 40$ cities across the UK over the 15 year period 1950 – 1965. The model has susceptible (S), exposed (E), infective (I), and removed (R) compartments for each of the $J = 40$ cities. For each city $k = 1,\dots,J$ the initial state of the epidemic is given by $[S_{k,0}\, E_{k,0}\, I_{k,0}\, R_{k,0}]^\top \sim \text{Mult}(n_{k,0}, \boldsymbol{\pi}_{k,0})$, where the probability vector $\boldsymbol{\pi}_{k,0}$ is a possibly city -specific initial distribution parameter, and $n_{k,t}$ for $t \geq 0$ denotes time varying population size. For each city $k = 1,\dots,J$ the population evolves twice per week with the following dynamic:

$$
\begin{aligned}
S_{k,t+1} &= S_{k,t} - B_{k,t} - F^{(S)}_{k,t} + A_{k,t}, \qquad & E_{k,t+1} &= E_{k,t} + B_{k,t} - C_{k,t} - F^{(E)}_{k,t}, \\
I_{k,t+1} &= I_{k,t} + C_{k,t} - D_{k,t} - F^{(I)}_{k,t}, \qquad & R_{k,t+1} &= R_{k,t} + D_{k,t} - F^{(R)}_{k,t},
\end{aligned}
$$

where $F^{(\cdot)}_{k,t}$ and $A_{k,t}$ model emigration (deaths) and immigration (births), respectively; and $C_{k,t}$ and $D_{k,t}$ are binomially distributed (details in the appendix). The term $B_{k,t}$ represents the number of new infections in the kth city and is distributed

$$
B_{k,t} \sim \text{Bin}\left(S_{k,t} - F^{(S)}_{k,t}, 1 - e^{-h b_{k,r}}\right),
$$

where:

$$
b_{k,r} = \beta_{k,r}\xi_{k,r} \cdot \left[\left(\frac{I_{k,\tau_r}}{n_{k,\tau_r}}\right) + \sum_{l \neq k} \frac{v_{k,l}}{n_{k,\tau_r}} \left\{\left(\frac{I_{l,\tau_r}}{n_{l,\tau_r}}\right) - \left(\frac{I_{k,\tau_r}}{n_{k,\tau_r}}\right)\right\}\right], \tag{6.3}
$$

for $r \geq 1$, $t = \tau_r,\dots,\tau_{r+1} - 1$. Here $\beta_{k,r}$ denotes a possibly city-specific seasonal transmission coefficient and $\xi_{k,r} \overset{\text{iid}}{\sim} \text{Gamma}(\sigma_\xi, \sigma_\xi)$, for $\sigma_\xi > 0$, is mean-1 multiplicative noise which achieves over-dispersion in the marginal distribution of $B_{k,t}$.

The summation term in (6.3) encodes the intercity interaction under a 'gravity model' – see Truscott and Ferguson (2012) for background on these kind of models in epidemiology. The strength of the interaction $v_{k,l}$ is computed as:

$$
v_{k,l} = g \frac{\bar{s}}{\bar{n}} \frac{n_{k,0} n_{l,0}}{s_{k,l}},
$$

where g is called the 'gravitational' constant parameter, $\bar{n}$ is the average of the initial populations, $\bar{s}$ is the average inter-city distance and $s_{k,l}$ denotes the distance between cities k and l. The interpretation of the gravity model is thus that the strength of the interaction between two cities is directly proportional to their populations and inversely proportional to their distance.

The observations are aggregated incidence data in the form of cumulative fortnightly transitions from infective to recovered for each of the 40 cities, at times $\tau_r = 4r$ for $r = 1,\dots,R$. Our observation model, which allows for over-dispersion, is described in section B.4 of the appendix,

along with the distributions of $C_{k,t}$, $D_{k,t}$, $F^{(\cdot)}_{k,t}$, and $A_{k,t}$, and an explanation of how we write the model as an instance of the Latent Compartmental model with $h = 3.5$ days, corresponding to bi-weekly transitions.

We consider three variants of this model all with over-dispersion in both the dynamics and observation mechanisms, but with increasing levels of city-specific parameters:

A: the initial distribution vectors $\boldsymbol{\pi}_{k,0}$ and force of infection parameters $\beta_{k,r}$ are shared across cities, i.e. constant in k;

B: $\boldsymbol{\pi}_{k,0}$ is city-specific and $\beta_{k,r}$ is shared across cities;

C: $\boldsymbol{\pi}_{k,0}$ and $\beta_{k,r}$ are city-specific.

Here we are inspired by an investigation conducted by Ionides et al. (2022), where sub-models with increasing numbers of city-specific parameters were fitted to a dataset on a smaller spatial scale, comprising 20 cities compared to the 40 we consider. Ionides et al. (2022) suggested that approximation techniques may be needed to analyse larger data sets, our application of the PAL framework is a step in that direction. However, we note that the 20-city dataset analysed by Ionides et al. (2022) is not a subset of the 40-city dataset we consider here, so direct comparisons of model fit may not be made. Never-the-less we shall compare our results to those obtained by (Park and Ionides, 2020) for a model in which parameters are shared across cities, fitted to the same 40-city dataset we consider.

Inference

In section B.4 of the appendix we give the details of a PALSMC algorithm in which the PAL is embedded within a block particle filter (Ionides et al., 2022; Rebeschini and Van Handel, 2015), to numerically approximate the log-likelihood. We used data-informed proposals and lookahead resampling to improve efficiency. For each of the models A,B,C, the approximate log-likelihood obtained from this PALSMC algorithm with 5000 particles was maximised with respect to the model parameters through Sequential Least Squares Programming. The procedures were implemented using Python and TensorFlow on a 32gb Tesla V100 GPU available on the HEC (High-End Computing) facility from Lancaster University.

Table 6.6 details PALSMC approximate log-likelihood and AIC values for each of the models A,B,C, along with an approximate log-likelihood reported by Park and Ionides (2020) for comparison. The GIRF used by Park and Ionides (2020) consists of a simulator for a continuous in time latent process combined with a particle filter which uses guide functions for intermediate propagation and resampling, parameters of the model are estimated via an iterated filtering scheme. Together with Monte Carlo adjusted profile methodology (Ionides et al., 2017) they are able to generate profile likelihood estimates for confidence interval estimation. Frequentist uncertainty interval calculation is out of the scope of the current work and would require results on the asymptotic distribution of the maximum PAL estimator, see chapter 7 for a discussion.

Our model A is similar to that of Park and Ionides (2020) in the sense that both these models have parameters shared across cities, but we find model A performs better in terms of log-likelihood and AIC. As we move from model A to models B and C, by making more parameters city-specific, we see an improvement in log-likelihood and AIC. We also note that the computation time for fitting model A is orders of magnitude smaller than that of Park and Ionides (2020). The computation time is of course implementation-dependent, but we note that we have not devised a bespoke optimisation algorithm to maximise the PALSMC approximation, but rather applied a standard 'black-box' optimiser. As prompted by an anonymous reviewer, we fitted an ARMA(2,0,1) model to the log-transformed data for a benchmark comparison; this gave a log-likelihood of -69168 (details are given in the appendix section B.4).

Estimates of the city-specific parameters $\beta_{k,r}$ in model C can be used to estimate city-specific R_0 values, calculated as in Ionides et al. (2022). We find that across the 40 cities these estimated R_0 values lie in the range $5.63 - 16.65$. The fitted mean latent and infective periods for model C were 8.49 and 9.53 respectively; these values are in line with previous inferences on the behaviour of measles epidemics (Guerra et al., 2017), (Delamater et al., 2019). Full details of our numerical results are in the appendix section B.4.

Model	No. parameters	Log-likelihood (sd)	AIC	Comp. time
A	11	$-63579\,(62)$	127180	45 min
B	128	$-61257\,(28)$	122770	10 hr
C	167	$-61169\,(34)$	122672	24 hr
Park and Ionides (2020)	12	$-70000^{\dagger}$	$140024^{\dagger}$	30 hr*

Table 6.6: Measles example. Mean log-likelihood values for models A, B, and C, with Monte Carlo standard deviation (sd) over 100 runs of PALSMC with 5000 particles. 'No. parameters' is the number of parameters estimated by maximising the log-likelihood for each model. †Approximate values read from figure 3 in (Park and Ionides, 2020). *We note that the 30hr reported by Park and Ionides (2020) includes confidence interval calculation via Monte Carlo adjusted profile methodology.

Figure 6.8 shows projected case numbers for the 4 fortnights following the end of the data record, obtained using model C with parameters fixed to the estimated values, full details are in the appendix section B.4. We see a general increase in forecast uncertainty as the time horizon increases, this reflecting the over-dispersed nature of model C. We also see that the forecasts generally exhibit higher certainty for cities with a larger population, as might be expected if a larger sub-population size allows latent variables and parameters which are specific to that sub-population to be estimated more accurately.

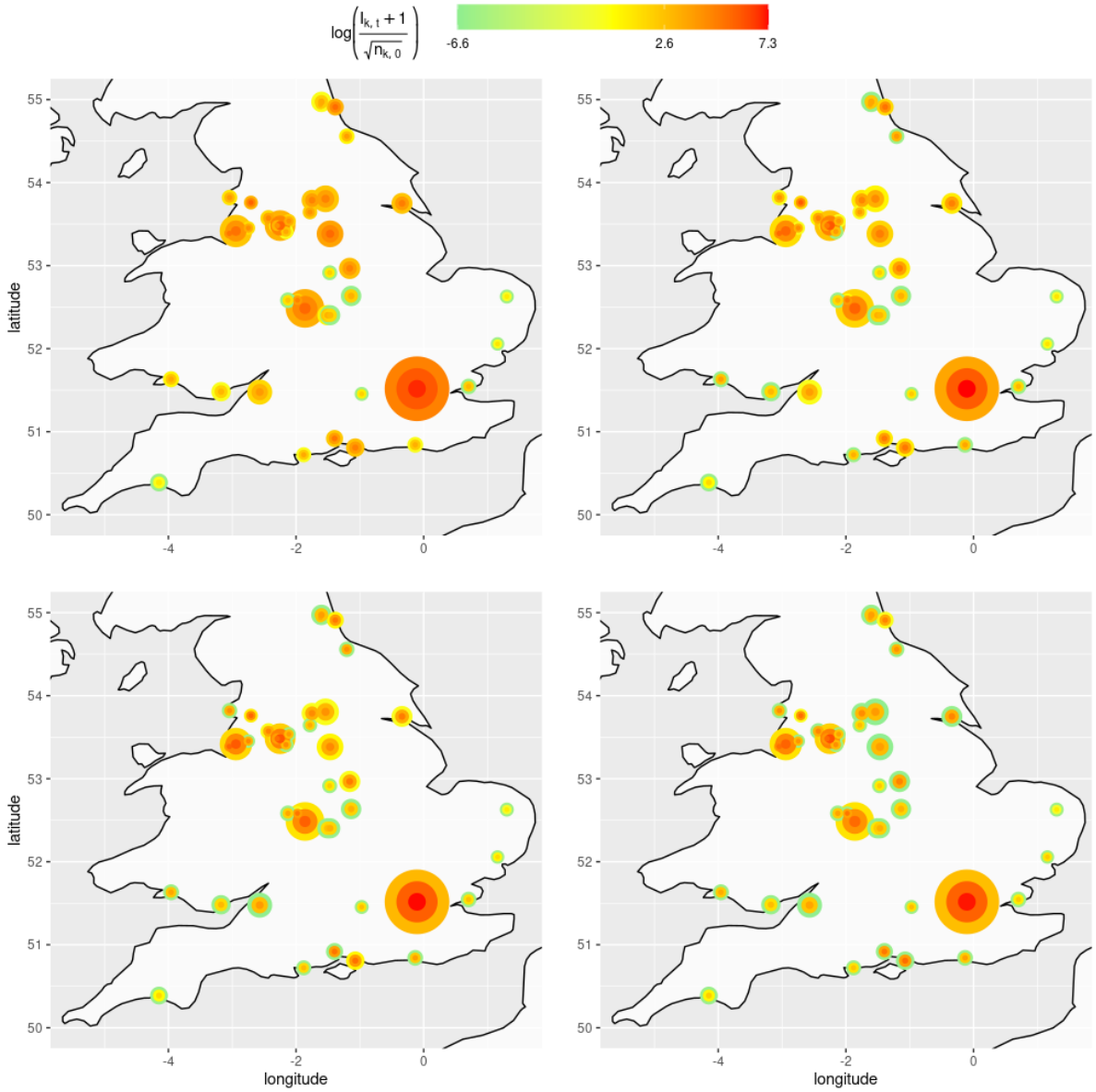

Figure 6.8: Measles example. Projected case numbers for the 4 fortnights (ordered top-left, top-right, bottom-left, bottom-right) following the end of the data record. For each town/city, the diameter of the outermost concentric ring represents log-population size. The shade of the outer concentric ring corresponds to the lower 5% quantile of the simulated case numbers, the shade of the middle concentric ring to the mean, and the inner concentric ring to the upper 95% quantile.

CHAPTER 7

CONCLUSIONS

Although this thesis presents PAL derivations for a fairly broad class of models, they seem tied to the various Poisson and binomial distribution assumptions appearing in the model definitions. The PAL approach is therefore clearly not as general as, say, inference using SMC or ABC, which in principle require little more than the ability to simulate from the model, but which in practice may involve various tuning parameters and incur a substantial computational cost. There is clearly a trade-off here between the generality of the model being accommodated and the resources needed to perform inference. It could be of interest to further broaden the class of models for which PAL-type approximations can be derived.

Recently Ju et al. (2021) devised sophisticated SMC algorithms to fit models in which individuals in the population each carry covariates influencing, for example, the probabilities that they come into contact, and hence the probabilities of disease spreading from one individual to the next. When these covariates are discrete or discretisable taking only finitely many distinct values, for example the subdivision of the population into age groups as in the age-structured example from section 6.3, then they can be handled in the latent compartmental modelling framework by simply introducing extra compartments and specifying an appropriate observation model. However, covariates taking infinitely many distinct values cannot be handled this way. Rimella et al. (2023) have suggested methods related to PALs to construct efficient proposal distributions for SMC in individual-based models. Further research may expand the applicability of PAL-like approximations in this direction.

Our consistency results provide rigorous assurances about the convergence of maximum PAL estimators, but it would be useful to obtain associated confidence intervals. A step towards this would be a central limit theorem for maximum PAL estimators.

Is there a statistical price to pay for using a PAL versus an exact likelihood? Whilst we

have proved that maximum PAL estimators can be consistent, we have not proved consistency of the estimators obtained by maximising the corresponding exact likelihoods. Indeed, for general classes of partially observed compartmental models there is a lack of such results in the literature. If central limit theorems for maximum PAL estimators and maximum exact likelihood estimators could be obtained, then that might help shed light on their relative statistical efficiency. This is, however, a somewhat academic question since the intractability of exact likelihoods for compartmental models, in general, seems to rule out the possibility of computing a maximum exact likelihood estimator in many practical situations.

In chapter 5 we introduced a procedure for PAL-based inference in over-dispersed models via sequential Monte Carlo. It would be of interest to develop a deterministic, and presumably faster, counterpart to this methodology. One way to do this could be similar in flavour to the integrated nested Laplace approximation (INLA) (Rue et al., 2009). Furthermore, whilst the large population theory is useful in choosing proposal distributions and filtering latent variables in the over-dispersed setting, further work is needed to assess the theoretical properties of maximum PALSMC estimators.

APPENDIX A

CONSISTENCY THEORY: SUPPORTING RESULTS

A.1 Laws of Large Numbers

Preliminaries

In this section we present some useful results for proving the main laws of large numbers.

Lemma 6. *Let $\lambda, \lambda_n \in \mathbb{R}_{\geq 0}$ and $X_n \sim Pois(\lambda_n)$ for $n = 1, 2, \ldots$. Assume that for all n, $|\frac{\lambda_n}{n} - \lambda| < cn^{-(\frac{1}{4}+\gamma)}$ for some $c > 0$ and $\gamma > 0$, then there exist constants b and $\bar{\gamma}$ such that:*

$$\mathbb{E}\left[\left|\frac{X_n}{n} - \lambda\right|^4\right]^{\frac{1}{4}} \leq bn^{-(\frac{1}{4}+\bar{\gamma})},$$

furthermore

$$\frac{X_n}{n} \xrightarrow{a.s.} \lambda.$$

Proof. By recurrence relations for the central moments of Poisson random variables, see e.g Kendall et al. (1946), we can write

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{X_n}{n} - \frac{\lambda_n}{n}\right|^4\right] &= n^{-4}\lambda_n \sum_{k=0}^{2} \binom{3}{k} \mathbb{E}\left[(X_n - \lambda_n)^k\right] \\
&= n^{-4}\lambda_n \left\{1 + 0 + \lambda_n \frac{3!}{2!1!}\right\} \\
&= n^{-2}\left\{n^{-2}\lambda_n + 3\left(\frac{\lambda_n}{n}\right)^2\right\} \leq an^{-2}, \qquad \text{(A.1)}
\end{aligned}$$

for some $a > 0$ since the curly bracketed term in (A.1) defines a convergent sequence. Hence, by the Minkowski inequality:

$$\begin{aligned}\mathbb{E}\left[\left|\frac{X_n}{n}-\lambda\right|^4\right]^{\frac{1}{4}} &\leq \mathbb{E}\left[\left|\frac{X_n}{n}-\frac{\lambda_n}{n}\right|^4\right]^{\frac{1}{4}}+\left|\frac{\lambda_n}{n}-\lambda\right| \\ &\leq a^{\frac{1}{4}}n^{-\frac{1}{2}}+cn^{-(\frac{1}{4}+\gamma)} \\ &\leq b\max\left(n^{-\frac{1}{2}},n^{-(\frac{1}{4}+\gamma)}\right) \\ &\leq bn^{-(\frac{1}{4}+\bar{\gamma})},\end{aligned}$$

where $\mathrm{b} = a^{\frac{1}{4}}+c$ and $\bar{\gamma} = \min\left(\gamma,\frac{1}{4}\right)$. Now let $\varepsilon > 0$, by Markov's inequality:

$$\mathbb{P}\left(\left|\frac{X_n}{n}-\lambda\right|>\varepsilon\right) \leq \varepsilon^{-4}\mathbb{E}\left[\left|\frac{X_n}{n}-\lambda\right|^4\right] \leq \epsilon^{-4}b^4n^{-(1+\bar{4}\gamma)},$$

So that:

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{X_n}{n}-\lambda\right|>\varepsilon\right) \leq \epsilon^{-4}b^4\sum_{n=1}^{\infty}n^{-(1+4\bar{\gamma})}<\infty.$$

Then $n^{-1}X_n \to \lambda$ almost surely by the Borel-Cantelli lemma. ■

Corollary 1. *If $\mathbf{x}_n \sim \mathrm{Pois}(\boldsymbol{\lambda}_n)$ for a sequence $(\boldsymbol{\lambda}_n)_{n\geq 1} \in \mathbb{R}^m$ such that there exists $c > 0$ and $\gamma > 0$ such that $\|n^{-1}\boldsymbol{\lambda}_n - \boldsymbol{\lambda}\|_\infty < cn^{-(\frac{1}{4}+\gamma)}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^m$, then for any vector $\boldsymbol{f} \in \mathbb{R}^m$ there exists constants $b > 0$ and $\bar{\gamma} > 0$ such that:*

$$\mathbb{E}\left[\left|\frac{1}{n}\mathbf{x}_n^\top\boldsymbol{f}-\boldsymbol{\lambda}^\top\boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq bn^{-(\frac{1}{4}+\bar{\gamma})}.$$

Proof. Apply lemma 6 in an element-wise fashion. ■

Lemma 7. *Let $\mathscr{F}$ be a filtration and $\Delta^{(i)}$ for $i = 1,2,\ldots$ be random variables which are conditionally independent given $\mathscr{F}$, are bounded by a constant $|\Delta^{(i)}| \leq M < \infty$ almost surely, and satisfy $\mathbb{E}\left[\Delta^{(i)} \mid \mathscr{F}\right] = 0$. Let a_n be a non-negative integer valued random variable such that $\sigma(a_n) \subseteq \mathscr{F}$ and assume there exist constants $a > 0$, $b > 0$, and $\gamma > 0$ such that for all $n \in \mathbb{N}$:*

$$\mathbb{E}\left[\left|\frac{a_n}{n}-a\right|^4\right]^{\frac{1}{4}} \leq bn^{-(\frac{1}{4}+\gamma)}.$$

Then there exists a constant $d > 0$ such that:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{a_n}\Delta^{(i)}\right|^4\right] \leq dn^{-2}.$$

Proof. Recalling that a sum over an empty set is equal to zero by convention, we have that:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^{a_n}\Delta^{(i)}\right)^4\Big|\mathscr{F}\right] &= \mathbb{E}\left[\sum_{\{k_1+\cdots+k_{a_n}=4;k_i\geq 0\}}\binom{4}{k_1,\ldots,k_{a_n}}\prod_{i=1}^{a_n}\left(\Delta^{(i)}\right)^{k_i}\Big|\mathscr{F}\right] \\
&= \sum_{\{k_1+\cdots+k_{a_n}=4;k_i\geq 0\}}\binom{4}{k_1,\ldots,k_{a_n}}\prod_{i=1}^{a_n}\mathbb{E}\left[\left(\Delta^{(i)}\right)^{k_i}\Big|\mathscr{F}\right] \\
&= \sum_{i=1}^{a_n}\mathbb{E}\left[\left(\Delta^{(i)}\right)^4\Big|\mathscr{F}\right] + 6\sum_{\{(i,j)\in[a_n]^2;i\neq j\}}\mathbb{E}\left[\left(\Delta^{(i)}\right)^2\Big|\mathscr{F}\right]\mathbb{E}\left[\left(\Delta^{(j)}\right)^2\Big|\mathscr{F}\right] \\
&\leq a_n M^4 + 3a_n(a_n-1)M^4 \\
&\leq c a_n^2,
\end{aligned}$$

for some constant $c>0$. The first equality holds by the multinomial theorem. The second equality holds through conditional independence of the $\Delta^{(i)}$. The third equality comes from the fact that all terms of the sum where $k_i=1$ for some i disappear since $\mathbb{E}\left[\left(\Delta^{(i)}\right)^1\Big|\mathscr{F}\right]=0$; hence, we need only count the terms with exclusively even k_i's. The first term after the inequality arises since there are a_n terms with a 4th power, each of which we can bound $\mathbb{E}\left[\left(\Delta^{(j)}\right)^4\Big|\mathscr{F}\right]<M^4$. The term $3a_n(a_n-1)$ comes from counting the number of terms with exactly 2 of the k_i's equal to 2 with the rest equalling 0; there are $\binom{a_n}{2}=a_n(a_n-1)/2$ such pairs, multiplying this by $\binom{4}{k_1,\ldots,k_{a_n}}=\binom{4}{2,2,0,\ldots}=6$ gives a total of $3a_n(a_n-1)$, then we bound each of the $\mathbb{E}\left[\left(\Delta^{(i)}\right)^2\Big|\mathscr{F}\right]\mathbb{E}\left[\left(\Delta^{(j)}\right)^2\Big|\mathscr{F}\right]\leq M^2M^2=M^4$ for all $(i,j)\in[a_n]^2$. So we have:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{a_n}\Delta^{(i)}\right|^4\Big|\mathscr{F}\right]\leq c\left(\frac{a_n}{n}\right)^2 n^{-2},$$

by the Lyapunov inequality:

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{a_n}{n}\right|^2\right]^{\frac{1}{2}} &\leq \mathbb{E}\left[\left|\frac{a_n}{n}-a+a\right|^4\right]^{\frac{1}{4}} \\
&\leq \mathbb{E}\left[\left|\frac{a_n}{n}-a\right|^4\right]^{\frac{1}{4}}+a \\
&\leq b n^{-(\frac{1}{4}+\gamma)}+a \\
&\leq b+a<\infty.
\end{aligned}$$

We now apply a tower law argument to the above to see that, for constant $d=c(b+a)^2$:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{a_n}\Delta^{(i)}\right|^4\right]\leq d n^{-2}.$$

■

Lemma 8. *Let* $\mathbf{x}\in\mathbb{R}^m_{\geq 0}$, $\boldsymbol{f}\in\mathbb{R}^m$, $c>0$, *and* $n\in\mathbb{N}$. *Then:*

$$\left|\boldsymbol{\eta}(\mathbf{x})^\top\boldsymbol{f}-\frac{n^{-1}\mathbf{x}^\top\boldsymbol{f}}{c}\right|\leq\left|\boldsymbol{\eta}(\mathbf{x})^\top\boldsymbol{f}\right|c^{-1}\left|n^{-1}\mathbf{1}_m^\top\mathbf{x}-c\right|$$

Proof. If $\mathbf{x} = \mathbf{0}$ then the result is trivial. Now, for $\mathbf{x} \neq \mathbf{0}$ we have:

$$\begin{aligned}
\left|\boldsymbol{\eta}(\mathbf{x})^\top \boldsymbol{f} - \frac{n^{-1}\mathbf{x}^\top \boldsymbol{f}}{c}\right| &= \left|\frac{\mathbf{x}^\top}{\mathbf{1}_m^\top \mathbf{x}}\boldsymbol{f} - \frac{n^{-1}\mathbf{x}^\top \boldsymbol{f}}{c}\right| \\
&\leq \left|\mathbf{x}^\top \boldsymbol{f}\right| \left|\frac{1}{\mathbf{1}_m^\top \mathbf{x}} - \frac{n^{-1}}{c}\right| \\
&= \left|\mathbf{x}^\top \boldsymbol{f}\right| \left|\frac{c - n^{-1}\mathbf{1}_m^\top \mathbf{x}}{c\mathbf{1}_m^\top \mathbf{x}}\right| \\
&\leq \left|\frac{\mathbf{x}^\top \boldsymbol{f}}{\mathbf{1}_m^\top \mathbf{x}}\right| c^{-1} \left|n^{-1}\mathbf{1}_m^\top \mathbf{x} - c\right| \\
&= \left|\boldsymbol{\eta}(\mathbf{x})^\top \boldsymbol{f}\right| c^{-1} \left|n^{-1}\mathbf{1}_m^\top \mathbf{x} - c\right|.
\end{aligned}$$

■

Case (I)

Define the sequence of vectors:

$$\begin{aligned}
\boldsymbol{\nu}_0(\boldsymbol{\theta}^*) &:= \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*), \\
\boldsymbol{\nu}_{t+1}(\boldsymbol{\theta}^*) &:= \left[\left(\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}^*)\right)^\top \mathbf{K}_{t+1,\boldsymbol{\eta}(\boldsymbol{\nu}_t(\boldsymbol{\theta}^*)\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}^*))}\right]^\top + \boldsymbol{\alpha}_{t+1,\infty}(\boldsymbol{\theta}^*).
\end{aligned}$$

Lemma 9. *Let assumptions 2-4 hold. For all $t \geq 0$ there exists $\gamma_t > 0$ and for all $\boldsymbol{f} \in \mathbb{R}^m$ there exists $c_t > 0$ such that:*

$$\mathbb{E}\left[\left|\frac{\mathbf{x}_t^\top}{n}\boldsymbol{f} - \boldsymbol{\nu}_t(\boldsymbol{\theta}^*)^\top \boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq c_t n^{-(\frac{1}{4}+\gamma_t)}. \tag{A.2}$$

Proof. Explicit dependence of some quantities on $\boldsymbol{\theta}^*$ and n is omitted throughout the proof to avoid over-cumbersome notation where the dependence is unambiguous. We proceed to prove the above by induction on t. At time 0 we have for some $c_0 > 0$ and $\gamma_0 > 0$:

$$\mathbb{E}\left[\left|\frac{\mathbf{x}_0^\top}{n}\boldsymbol{f} - \boldsymbol{\nu}_0^\top \boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq c_0 n^{-(\frac{1}{4}+\gamma_0)},$$

by assumption 4. Now for $t \geq 1$ assume (A.2) holds for $t-1$. Recall $\mathbf{x}_t = \tilde{\mathbf{x}}_t + \hat{\mathbf{x}}_t$, $\tilde{x}_t^{(j)} = \sum_{i=1}^{n_{t-1}} \mathbb{I}\{\phi_t^{(i)} = 1\}\mathbb{I}\{\xi_t^{(i)} = j\}$ and $\bar{x}_{t-1}^{(j)} = \sum_{i=1}^{n_{t-1}} \mathbb{I}[\xi_{t-1}^{(i)} = j]\mathbb{I}[\phi_t^{(i)} = 1]$. We make the following decomposition:

$$\begin{aligned}
\frac{\mathbf{x}_t^\top}{n}\boldsymbol{f} - \boldsymbol{\nu}_t^\top \boldsymbol{f} = \frac{\mathbf{x}_t^\top}{n}\boldsymbol{f} &- \left[\frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} + \boldsymbol{\alpha}_{t,\infty}^\top\right]\boldsymbol{f} && \text{(A.3)} \\
&+ \left[\frac{\bar{\mathbf{x}}_{t-1}}{n} - \boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t\right]^\top \left[\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\boldsymbol{f}\right] && \text{(A.4)} \\
&+ (\boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t)^\top \left[\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} - \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\right]\boldsymbol{f}. && \text{(A.5)}
\end{aligned}$$

Consider (A.3). We make the further decomposition:

$$
\begin{aligned}
&\frac{\mathbf{x}_t^\top}{n}\boldsymbol{f} - \left[\frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} + \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f} \\
&\quad= \frac{(\tilde{\mathbf{x}}_t + \hat{\mathbf{x}}_t)^\top}{n}\boldsymbol{f} - \left[\frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} + \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f} \\
&\quad= \frac{\tilde{\mathbf{x}}_t^\top}{n}\boldsymbol{f} - \left[\frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right]\boldsymbol{f} + \left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f} \\
&\quad= \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n_{t-1}} \mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\}\right) f^{(j)} - \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n_{t-1}} \mathbb{I}\{\xi_{t-1}^{(i)}=j\}\mathbb{I}\{\phi_t^{(i)}=1\}\right)\mathbf{K}^{j,\cdot}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\boldsymbol{f} \\
&\quad+ \left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f}. \\
&\quad= \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n_{t-1}} \mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\}\right) f^{(j)} - \frac{1}{n}\sum_{i=1}^{n_{t-1}} \mathbb{I}\{\phi_t^{(i)}=1\}\mathbf{K}^{\xi_{t-1}^{(i)},\cdot}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\boldsymbol{f} + \left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f}. \\
&\quad= \frac{1}{n}\sum_{i=1}^{n_{t-1}}\sum_{j=1}^{m}\left\{\mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\} - \mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right\} f^{(j)} \qquad \text{(A.6)} \\
&\quad+ \left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f}.
\end{aligned}
$$

The term $\left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f}$ converges to 0 in L^4 by assumption 2 and lemma 6, that is there exists an $\hat{c}_t > 0$ and $\hat{\gamma}_t > 0$ such that:

$$
\mathbb{E}\left[\left|\left[\frac{\hat{\mathbf{x}}_t^\top}{n} - \boldsymbol{\alpha}_{t,n}^\top\right]\boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq \hat{c}_t n^{-(\frac{1}{4}+\hat{\gamma}_t)}.
$$

Now, turning to (A.6), let $\mathscr{G}_t := \sigma(\{\xi_t^{(i)}\}_{i=1,\dots,n_t})$ and $\mathscr{F}_t := \sigma(\{\phi_t^{(i)}\}_{i=1,\dots,n_t})$. See that:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{m}\left\{\mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\} - \mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right\} f^{(j)}\middle|\mathscr{G}_{t-1}\vee\mathscr{F}_t\right] \\
&=\sum_{j=1}^{m}\left\{\mathbb{E}\left[\mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\}\middle|\mathscr{G}_{t-1}\vee\mathscr{F}_t\right] - \mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right\} f^{(j)} \\
&=\sum_{j=1}^{m}\left\{\mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} - \mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right\} f^{(j)} \\
&=0,
\end{aligned}
$$

since, given $\xi_{t-1}^{(i)}$, $\phi_t^{(i)} \sim \text{Bernoulli}\left(\delta_t^{(\xi_{t-1}^{(i)})}\right)$ and, conditional on $\phi_t^{(i)}=1$ and $\mathscr{G}_{t-1}$, $\xi_t^{(i)}$ is a draw from the $\xi_{t-1}^{(i)}$th row of $\mathbf{K}_{t,\boldsymbol{\eta}(\bar{x}_{t-1})}$; and if $\phi_t^{(i)}=0$ then $\xi_t^{(i)}=0$. Moreover:

$$
\left|\sum_{j=1}^{m}\left\{\mathbb{I}\{\phi_t^{(i)}=1\}\mathbb{I}\{\xi_t^{(i)}=j\} - \mathbb{I}\{\phi_t^{(i)}=1\}K^{(\xi_{t-1}^{(i)}),j}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\right\} f^{(j)}\right| \leq m\|\boldsymbol{f}\|_\infty =: M.
$$

Define:

$$\Delta_t^{(i)} := \sum_{j=1}^{m} \left\{ \mathbb{I}\{\phi_t^{(i)} = 1\}\mathbb{I}\{\xi_t^{(i)} = j\} - \mathbb{I}\{\phi_t^{(i)} = 1\} K_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}^{(\xi_{t-1}^{(i)}),j} \right\} f^{(j)}.$$

The $\Delta_t^{(i)}$ are conditionally independent and mean zero given $\mathscr{G}_{t-1} \vee \mathscr{F}_t$, and $\sigma(n_{t-1}) \subset \mathscr{G}_{t-1} \vee \mathscr{F}_t$. Also note that, since $\frac{n_{t-1}}{n}$ is equal to $\frac{\mathbf{x}_{t-1}^\top}{n}\mathbf{1}_m$, we can invoke the induction hypothesis with test vector $\mathbf{1}_m$ to see there exist constants c_{t-1} and γ_{t-1} such that:

$$\mathbb{E}\left[\left| \frac{n_{t-1}}{n} - \mathbf{1}_m^\top \boldsymbol{\nu}_{t-1} \right|^4 \right]^{\frac{1}{4}} \leq c_{t-1} n^{-(\frac{1}{4}+\gamma_{t-1})}$$

so that we satisfy the conditions of lemma 7. Hence there exists a constant $\tilde{c}_t > 0$:

$$\mathbb{E}\left[\left| \frac{1}{n} \sum_{i=1}^{n_{t-1}} \Delta_t^{(i)} \right|^4 \right]^{\frac{1}{4}} \leq \tilde{c}_t n^{-\frac{1}{2}}.$$

Before analysing (A.4) and (A.5) we will prove an intermediary result. Consider the decomposition:

$$\left| \frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \boldsymbol{f} - (\boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t)^\top \boldsymbol{f} \right| \leq \left| \frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \boldsymbol{f} - \left(\frac{\mathbf{x}_{t-1}}{n} \odot \boldsymbol{\delta}_t \right)^\top \boldsymbol{f} \right| \tag{A.7}$$

$$+ \left| \left(\frac{\mathbf{x}_{t-1}}{n} \odot \boldsymbol{\delta}_t \right)^\top \boldsymbol{f} - (\boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t)^\top \boldsymbol{f} \right|. \tag{A.8}$$

The term in (A.8) converges to 0 in L^4 at the required rate by the induction hypothesis with test vector $\boldsymbol{\delta}_t \odot \boldsymbol{f}$. Now for (A.7) see that:

$$\frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \boldsymbol{f} - \left(\frac{\mathbf{x}_{t-1}}{n} \odot \boldsymbol{\delta}_t \right)^\top \boldsymbol{f} = \frac{1}{n} \sum_{i=1}^{n_{t-1}} \underbrace{\sum_{j=1}^{m} \mathbb{I}\left\{ \xi_{t-1}^{(i)} = j \right\} \left(\mathbb{I}\left\{ \phi_t^{(i)} = 1 \right\} - \delta_t^{(\xi_{t-1}^{(i)})} \right) f^{(j)}}_{\bar{\Delta}_t^{(i)}}, \tag{A.9}$$

and note that the $\bar{\Delta}_t^{(i)}$ are mean 0, bounded, and independent given $\mathscr{G}_{t-1}$ so that by lemma 7 we have that (A.9) converges to 0 in L^4 at the required rate. Combining this with the Minkowski inequality, we have that for some positive constants $\bar{c}_t$ and $\bar{\gamma}_t$:

$$\mathbb{E}\left[\left| \frac{\bar{\mathbf{x}}_{t-1}}{n}^\top \boldsymbol{f} - (\boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t)^\top \boldsymbol{f} \right|^4 \right]^{\frac{1}{4}} \leq \bar{c}_t n^{-(\frac{1}{4}+\bar{\gamma})}. \tag{A.10}$$

Now we look at (A.4):

$$\begin{aligned}
\left| \left[\frac{\bar{\mathbf{x}}_{t-1}}{n} - \boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t \right]^\top \left[\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} \boldsymbol{f} \right] \right| &\leq \left\| \frac{\bar{\mathbf{x}}_{t-1}}{n} - \boldsymbol{\nu}_{t-1} \right\|_1 \left\| \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} \boldsymbol{f} \right\|_\infty \\
&\leq \|\boldsymbol{f}\|_\infty \sum_{i=1}^{m} \left| \frac{\bar{x}_{t-1}^{(i)}}{n} - \nu_{t-1}^{(i)} \delta_t^{(i)} \right| \\
&\leq \|\boldsymbol{f}\|_\infty \sum_{i=1}^{m} \left| \frac{\bar{\mathbf{x}}_{t-1}^\top}{n} \boldsymbol{e}_i - (\boldsymbol{\nu}_{t-1} \odot \boldsymbol{\delta}_t)^\top \boldsymbol{e}_i \right|.
\end{aligned}$$

The first inequality here uses Holder's inequality and the second uses the fact that the row sums of the matrix $\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}$ are equal to 1. By (A.10) with $\boldsymbol{f} = \boldsymbol{e}_i$ in conjunction with the Minkowski inequality there exists $\check{c}_t > 0$ and $\check{\gamma}_t > 0$ such that:

$$\mathbb{E}\left[\left|\left[\frac{\bar{\mathbf{x}}_{t-1}}{n} - \boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t\right]^\top\left[\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})}\boldsymbol{f}\right]\right|^4\right]^{\frac{1}{4}} \leq \check{c}_t n^{-(\frac{1}{4}+\check{\gamma}_t)}.$$

Now looking at (A.5) we see using assumption 3 that there exists a $c > 0$ such that:

$$\begin{aligned}\left|[\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t]^\top\left[\mathbf{K}_{t,\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1})} - \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\right]\boldsymbol{f}\right| &\leq c\|\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t\|_\infty\|\boldsymbol{f}\|_\infty\|\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1}) - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)\|_\infty \\ &\leq c\|\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t\|_\infty\|\boldsymbol{f}\|_\infty\sum_{i=1}^m\left|(\boldsymbol{\eta}(\bar{\mathbf{x}}_{t-1}) - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t))^\top\boldsymbol{e}_i\right|\end{aligned}$$

If $\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t) = 0$ then $\mathbf{1}_m^\top\bar{\mathbf{x}}_{t-1} = 0$ $\mathbb{P}^{\theta^*} - a.s.$ by lemmas 13 and 15, which we state and prove in section A.2, in which case the right hand side of the above is 0 and therefore satisfies all positive bounds. Henceforth, assume $\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t) > 0$. Consider:

$$\begin{aligned}&\left|\boldsymbol{\eta}(\bar{\mathbf{x}}_{\mathbf{t-1}})^\top\boldsymbol{f} - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}\right| \\ &\quad= \left|\boldsymbol{\eta}(\bar{\mathbf{x}}_{\mathbf{t-1}})^\top\boldsymbol{f} - \frac{(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\right| \\ &\quad= \left|\boldsymbol{\eta}(\bar{\mathbf{x}}_{\mathbf{t-1}})^\top\boldsymbol{f} + \frac{n^{-1}\bar{\mathbf{x}}_{t-1}^\top}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\boldsymbol{f} - \frac{n^{-1}\bar{\mathbf{x}}_{t-1}^\top}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\boldsymbol{f} - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}\right| \\ &\quad\leq \left|\boldsymbol{\eta}(\bar{\mathbf{x}}_{\mathbf{t-1}})^\top\boldsymbol{f} - \frac{n^{-1}\bar{\mathbf{x}}_{t-1}^\top}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\boldsymbol{f}\right| + \left|\frac{n^{-1}\bar{\mathbf{x}}_{t-1}^\top}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\boldsymbol{f} - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}\right| \\ &\quad\leq \left|\boldsymbol{\eta}(\bar{\mathbf{x}}_{\mathbf{t-1}})^\top\boldsymbol{f}\right|(\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t))^{-1}\left|\frac{\mathbf{1}_m^\top\bar{\mathbf{x}}_{t-1}}{n} - \mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)\right| \qquad \text{(A.11)} \\ &\quad+ (\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t))^{-1}\left|\frac{\bar{\mathbf{x}}_{t-1}^\top}{n}\boldsymbol{f} - (\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}\right| \\ &\quad\leq \frac{m\|\boldsymbol{f}\|_\infty}{\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)}\left|\frac{\mathbf{1}_m^\top\bar{\mathbf{x}}_{t-1}}{n} - \mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)\right| \qquad \text{(A.12)} \\ &\quad+ (\mathbf{1}_m^\top(\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t))^{-1}\left|\frac{\bar{\mathbf{x}}_{t-1}^\top}{n}\boldsymbol{f} - (\boldsymbol{\nu}_{t-1}\odot\boldsymbol{\delta}_t)^\top\boldsymbol{f}\right|. \qquad \text{(A.13)}\end{aligned}$$

Where we use lemma 8 in line (A.11). We can again invoke (A.10) to give L^4 convergence of (A.12) and (A.13) at the required rate. We can now combine all of the above, along with the Minkowski inequality to show that:

$$\mathbb{E}\left[\left|\frac{\mathbf{x}_t^\top}{n}\boldsymbol{f} - \boldsymbol{\nu}_t^\top\boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq c_t n^{-(\frac{1}{4}+\gamma_t)},$$

where $c_t = \hat{c}_t + \tilde{c}_t + \bar{c}_t + \check{c}_t$, and $\gamma_t = \min(\hat{\gamma}_t, \frac{1}{4}, \bar{\gamma}_t, \check{\gamma}_t)$. ■

Lemma 10. *Let assumptions 2 - 4 hold. Then there exists a constant $\rho_t > 0$ for each $\boldsymbol{f} \in \mathbb{R}^m$ and $t \geq 1$, and a constant $a_t > 0$ such that:*

$$\mathbb{E}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - \left([\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) \odot \mathbf{q}_t(\boldsymbol{\theta}^*)]^\top \mathbf{G}_t(\boldsymbol{\theta}^*) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}^*)^\top\right)\boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq a_t n^{-(\frac{1}{4}+\rho_t)}.$$

Proof. Explicit dependence of some quantities on $\boldsymbol{\theta}^*$ and n is omitted throughout the proof to avoid over-cumbersome notation where the dependence is unambiguous. First note that:

$$\frac{\mathbf{y}_t}{n} = \frac{\tilde{\mathbf{y}}_t}{n} + \frac{\hat{\mathbf{y}}_t}{n},$$

and

$$\mathbb{E}\left[\left|\frac{\hat{\mathbf{y}}_t^\top}{n}\boldsymbol{f} - \boldsymbol{\kappa}_{t,\infty}^\top \boldsymbol{f}\right|^4\right]^{\frac{1}{4}} < \hat{a}_t n^{-(\frac{1}{4}+\hat{\rho}_t)}, \tag{A.14}$$

for some $\hat{a}_t > 0$ and $\hat{\rho}_t > 0$ by corollary 1 and assumption 2. Write:

$$\begin{aligned}\frac{\tilde{\mathbf{y}}_t^\top}{n}\boldsymbol{f} - (\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t \boldsymbol{f} = \frac{\tilde{\mathbf{y}}_t^\top}{n}\boldsymbol{f} - \left(\frac{\mathbf{x}_t}{n} \odot \mathbf{q}_t\right)^\top \mathbf{G}_t \boldsymbol{f} \\ + \left(\frac{\mathbf{x}_t}{n} \odot \mathbf{q}_t\right)^\top \mathbf{G}_t \boldsymbol{f} - (\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t \boldsymbol{f}.\end{aligned} \tag{A.15}$$

We have that

$$\mathbb{E}\left[\left|\left(\frac{\mathbf{x}_t}{n} \odot \mathbf{q}_t\right)^\top \mathbf{G}_t \boldsymbol{f} - (\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t \boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq \bar{a}_t n^{-(\frac{1}{4}+\bar{\rho}_t)} \tag{A.16}$$

for some $\bar{a}_t > 0$ and $\bar{\rho}_t > 0$ by lemma 9 using test function $[(\mathbf{q}_t \otimes \mathbf{1}_m) \odot \mathbf{G}_t]\boldsymbol{f}$.

Furthermore, we have:

$$\begin{aligned}\frac{\tilde{\mathbf{y}}_t^\top}{n}\mathbf{f} &= \frac{1}{n}\sum_{j=1}^{m} \tilde{\mathbf{y}}^{(j)} \boldsymbol{f}^{(j)} \\ &= \frac{1}{n}\sum_{i=1}^{n_t}\sum_{j=1}^{m}\sum_{k=1}^{m} \mathbb{I}\{\xi_t^{(i)} = k\}\mathbb{I}\{\zeta_t^{(i)} = 1\}\mathbb{I}\{\varsigma_t^{(i)} = j\}\boldsymbol{f}^{(j)},\end{aligned}$$

where $\zeta_t^{(i)} \sim \text{Bernoulli}(\mathbf{q}_t^{(\xi_t^{(i)})})$ and $\varsigma_t^{(i)} \sim \text{Categorical}(\mathbf{G}^{(\xi_t^{(i)},\cdot)})$ indicates the compartment in which it is observed. Notice that for (A.15) :

$$\begin{aligned}&\frac{\tilde{\mathbf{y}}_t^\top}{n}\boldsymbol{f} - \left(\frac{\mathbf{x}_t}{n} \odot \mathbf{q}_t\right)^\top \mathbf{G}_t \boldsymbol{f} \\ &\quad = \frac{1}{n}\underbrace{\sum_{i=1}^{n_t}\sum_{j=1}^{m}\sum_{k=1}^{m}\left[\mathbb{I}\{\xi_t^{(i)} = k\}\mathbb{I}\{\zeta_t^{(i)} = 1\}\mathbb{I}\{\varsigma_t^{(i)} = j\} - \mathbb{I}\{\xi_t^{(i)} = k\}q_t^{(k)}G_t^{(k,j)}\right]\boldsymbol{f}^{(j)}}_{=:\Xi_t^{(i)}}.\end{aligned}$$

The $\Xi_t^{(i)}$ are mean zero and independent conditioned on $\mathscr{G}_t$. Furthermore:

$$|\Xi_t^{(i)}| \leq m^2 \|\boldsymbol{f}\|_\infty,$$

almost surely and $\sigma(n_t) \subseteq \mathscr{G}_t$ where $\mathscr{G}_t$ is defined as in lemma 9. An application of lemma 7, yields:

$$\mathbb{E}\left[\left|\frac{\tilde{\mathbf{y}}_t^\top}{n}\boldsymbol{f} - \left(\frac{\mathbf{x}_t}{n} \odot \mathbf{q}_t\right)^\top \mathbf{G}_t \boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq \tilde{a}_t n^{(\frac{1}{4}+\tilde{\rho}_t)}, \tag{A.17}$$

for some constants $\tilde{a}_t > 0$ and $\tilde{\rho}_t > 0$. Combining (A.14), (A.16), and (A.17) with the Minkowski inequality yields:

$$\mathbb{E}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - \left((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top\right)\boldsymbol{f}\right|^4\right]^{\frac{1}{4}} \leq a_t n^{-(\frac{1}{4}+\rho_t)},$$

where $a_t = \hat{a}_t + \bar{a}_t + \tilde{a}_t$ and $\rho_t = \min(\hat{\rho}_t, \bar{\rho}_t, \tilde{\rho}_t)$.

■

Proposition 1. *Let assumptions 2 - 4 hold. Then for all $t \geq 1$:*

$$\frac{\mathbf{y}_t^\top}{n} \xrightarrow[a.s.]{\boldsymbol{\theta}^*} [\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) \odot \mathbf{q}_t(\boldsymbol{\theta}^*)]^\top \mathbf{G}_t(\boldsymbol{\theta}^*) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}^*)^\top. \tag{A.18}$$

Proof. By lemma 10 there exists constants $a_t > 0$ and $\rho_t > 0$ such that:

$$\mathbb{E}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t)^\top \boldsymbol{f}\right|^4\right] \leq a_t^4 n^{-(1+4\rho_t)}.$$

By Markov's inequality:

$$\begin{aligned}\mathbb{P}^{\boldsymbol{\theta}^*}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}\right| > \varepsilon\right] &\leq \varepsilon^{-4}\mathbb{E}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}\right|^4\right] \\ &\leq \varepsilon^{-4} a_t^4 n^{-(1+4\rho_t)}.\end{aligned}$$

This implies that:

$$\sum_{n=1}^{\infty} \mathbb{P}^{\boldsymbol{\theta}^*}\left[\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}\right| > \varepsilon\right] < \infty. \tag{A.19}$$

We now appeal to the Borel-Cantelli lemma which tells us that (A.19) implies the event:

$$\left\{\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}\right| > \varepsilon\right\},$$

happens for infinitely many n with probability 0, and that:

$$\mathbb{P}^{\boldsymbol{\theta}^*}\left(\lim_{n\to\infty}\left|\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} - ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}\right| > \varepsilon\right) = 0,$$

for all $\varepsilon > 0$. Hence we have shown that:

$$\frac{\mathbf{y}_t^\top}{n}\boldsymbol{f} \xrightarrow[a.s.]{\boldsymbol{\theta}^*} ((\boldsymbol{\nu}_t \odot \mathbf{q}_t)^\top \mathbf{G}_t + \boldsymbol{\kappa}_t^\top)\boldsymbol{f}.$$

■

Case (II)

Define:

$$\begin{aligned}
&\boldsymbol{\nu}_0(\boldsymbol{\theta}^*) := \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*),\\
&\mathbf{N}_t(\boldsymbol{\theta}^*) := \left(\boldsymbol{\nu}_{t-1}(\boldsymbol{\theta}^*) \otimes \mathbf{1}_m\right) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1}(\boldsymbol{\theta}^*))}(\boldsymbol{\theta}^*),\\
&\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) := (\mathbf{1}_m^\top \mathbf{N}_t(\boldsymbol{\theta}^*))^\top.
\end{aligned}$$

Lemma 11. *Let assumptions 2 - 4 hold. For all $t \geq 1$ there exists a $\gamma_{t_z} > 0$, and for all vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^m$ a constant $b_t > 0$, such that:*

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top \mathbf{Z}_t \boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{N}_t(\boldsymbol{\theta}^*)\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq b_t n^{-(\frac{1}{4}+\gamma_{t_z})}, \text{ for all } t \geq 0.$$

Proof. Recall from section 3.2.2 that in case (II) there is no immigration or emigration, $n_t = n$ and hence also $\mathbf{x}_t = \bar{\mathbf{x}}_t$ with probability 1 for all $t \geq 0$.

Consider the decomposition:

$$\begin{aligned}
&\left|n^{-1}\boldsymbol{f}_1^\top \mathbf{Z}_t \boldsymbol{f}_2 - \boldsymbol{f}_1^\top [\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2\right|\\
&\qquad \leq \left|n^{-1}\boldsymbol{f}_1^\top \mathbf{Z}_t \boldsymbol{f}_2 - \boldsymbol{f}_1^\top \left[\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})} \boldsymbol{f}_2\right| \qquad \text{(A.20)}\\
&\qquad + \left|\boldsymbol{f}_1^\top \left[\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})} \boldsymbol{f}_2 - \boldsymbol{f}_1^\top \left[\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2\right| \qquad \text{(A.21)}\\
&\qquad + \left|\boldsymbol{f}_1^\top \left[\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2 - \boldsymbol{f}_1^\top [\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2\right|. \qquad \text{(A.22)}
\end{aligned}$$

Notice that by assumption 2 with vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, there exists a constant $c > 0$ such that the term (A.21) satisfies:

$$\begin{aligned}
&\left|\boldsymbol{f}_1^\top \left[\left(\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right) \odot \left(\mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})} - \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\right)\right] \boldsymbol{f}_2\right|\\
&= \left|\left(\boldsymbol{f}_1 \odot \frac{\mathbf{x}_{t-1}}{n}\right)^\top \left(\mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})} - \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\right) \boldsymbol{f}_2\right|\\
&\leq c\|\boldsymbol{f}_1\|_\infty \|\boldsymbol{f}_2\|_\infty \left\|\boldsymbol{\eta}(\mathbf{x}_{t-1}) - \boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})\right\|_\infty\\
&\leq c\|\boldsymbol{f}_1\|_\infty \|\boldsymbol{f}_2\|_\infty \left\|\frac{\mathbf{x}_{t-1}}{n} - \boldsymbol{\nu}_{t-1}\right\|_\infty\\
&\leq c\|\boldsymbol{f}_1\|_\infty \|\boldsymbol{f}_2\| \sum_{i=1}^m \left|\left(\frac{\mathbf{x}_{t-1}}{n} - \boldsymbol{\nu}_{t-1}\right)^\top \boldsymbol{e}_i\right|.
\end{aligned}$$

By the Minkowski inequality and lemma 9 there exist constants $\bar{b}_t > 0$ and $\bar{\gamma}_{t_z} > 0$ such that:

$$\mathbb{E}\left[\left|\boldsymbol{f}_1^\top \left[\left(\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right) \odot \left(\mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})} - \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\right)\right] \boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq \bar{b}_t n^{-(\frac{1}{4}+\bar{\gamma}_{t_z})}$$

Moreover, (A.22) is equal to:

$$\left|\frac{\mathbf{x}_{t-1}^\top}{n} [\boldsymbol{f}_1 \otimes \mathbf{1}_m] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2 - \boldsymbol{\nu}_{t-1}^\top [\boldsymbol{f}_1 \otimes \mathbf{1}_m] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})} \boldsymbol{f}_2\right|,$$

therefore we can invoke lemma 9 with test vector $\left[\boldsymbol{f}_1 \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\boldsymbol{f}_2$, this tells us there exists constants $\hat{b}_t > 0$ and $\hat{\gamma}_{t_z} > 0$ such that:

$$\mathbb{E}\left[\left|\frac{\mathbf{x}_{t-1}^\top}{n}\left[\boldsymbol{f}_1 \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\boldsymbol{f}_2 - \boldsymbol{\nu}_{t-1}^\top\left[\boldsymbol{f}_1 \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_{t-1})}\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq \hat{b}_t n^{-(\frac{1}{4}+\hat{\gamma}_{t_z})}.$$

We now recall that $\mathbf{Z}_t^{(j,k)} = \sum_{i=1}^n \mathbb{I}\{\xi_{t-1}^{(i)} = j, \xi_t^{(i)} = k\}$, so that the term (A.20) is equal to:

$$\frac{1}{n}\sum_{i=1}^n \underbrace{\left[\sum_{j=1}^m\sum_{k=1}^m \left(\mathbb{I}\{\xi_{t-1}^{(i)} = j, \xi_t^{(i)} = k\} - \mathbb{I}\{\xi_{t-1}^{(i)} = j\}K_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})}^{(j,k)}\right) f_1^{(j)} f_2^{(k)}\right]}_{=:\Delta_t^{(i)}}.$$

Since, conditioned on $\mathscr{G}_{t-1} := \sigma(\{\xi_{t-1}^{(i)}\}_{i=1,\dots,n_t})$, $\xi_t^{(i)}$ is a draw from the $\xi_{t-1}^{(i)}$th row of $\mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})}$, we have $\mathbb{E}\left[\Delta_t^{(i)} \mid \mathscr{G}_{t-1}\right] = 0$. Furthermore, given $\mathscr{G}_{t-1}$ the $\Delta_t^{(i)}$ are independent and $\left|\Delta_t^{(i)}\right| \leq m^2\|\boldsymbol{f}\|_\infty^2$. An application of lemma 7 yields that for some constants $\tilde{b}_t > 0$ and $\tilde{\gamma}_{t_z} > 0$:

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top\left[\frac{\mathbf{x}_{t-1}}{n} \otimes \mathbf{1}_m\right] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\mathbf{x}_{t-1})}\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq \tilde{b}_t n^{-(\frac{1}{4}+\tilde{\gamma}_{t_z})}.$$

Finally, use of the Minkowski inequality yields the result:

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top[\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m] \odot \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\nu}_t)}\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq b_t n^{-(\frac{1}{4}+\gamma_{t_z})},$$

where $b_t = \bar{b}_t + \hat{b}_t + \tilde{b}_t$ and $\gamma_{t_z} = \min(\bar{\gamma}_{t_z}, \hat{\gamma}_{t_z}, \tilde{\gamma}_{t_z})$.

■

Lemma 12. *Let assumptions 2 - 4 hold. For all $t \geq 1$ there exists a $\bar{\gamma}_Y > 0$, and for all vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^m$ a constant $c_Y > 0$, such that:*

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top\left[\mathbf{N}_t(\boldsymbol{\theta}^*) \odot \mathbf{Q}_t(\boldsymbol{\theta}^*)\right]\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq c_Y n^{-(\frac{1}{4}+\bar{\gamma}_Y)}.$$

Proof. Write

$$\begin{aligned}
&|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top\left[(\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\nu}_{t-1}} \odot \mathbf{Q}_t\right]\boldsymbol{f}_2| \\
&\qquad \leq \left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2 - n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t \odot \mathbf{Q}_t\boldsymbol{f}_2\right| \\
&\qquad + \left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t \odot \mathbf{Q}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top\left[(\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\nu}_{t-1}} \odot \mathbf{Q}_t\right]\boldsymbol{f}_2\right|. \qquad \text{(A.23)}
\end{aligned}$$

By lemma 11 there exists $a_Y > 0$ and $\gamma_{Y_1} > 0$ such that:

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t \odot \mathbf{Q}_t\boldsymbol{f}_2 - \boldsymbol{f}_1^\top\left[(\boldsymbol{\nu}_{t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\nu}_{t-1}} \odot \mathbf{Q}_t\right]\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}} \leq a_Y n^{-(\frac{1}{4}+\gamma_{Y_1})}$$

Now, we can write (A.23) as:

$$n^{-1}\sum_{i=1}^{n}\underbrace{\sum_{j=1}^{m}\sum_{k=1}^{m}\left[\mathbb{I}\{\xi_{t-1}^{(i)}=j,\xi_t^{(i)}=k\}\mathbb{I}\{\zeta^{(i)}=1\}-\mathbb{I}\{\xi_{t-1}^{(i)}=j,\xi_t^{(i)}=k\}Q^{(j,k)}\right]f_1^{(j)}f_2^{(k)}}_{=:\Xi_t^{(i)}}.$$

Where $\zeta^{(i)}$ given $\mathcal{G}_{t-1}\vee\mathcal{G}_t$ (where $\mathcal{G}_t$ is defined as in lemma 9) is distributed Bernoulli($Q^{(\xi_{t-1}^{(i)},\xi_t^{(i)})}$). Hence, $\mathbb{E}\left[\Xi_t^{(i)}\mid\mathcal{G}_{t-1}\vee\mathcal{G}_t\right]=0$. Furthermore, given $\mathcal{G}_{t-1}\vee\mathcal{G}_t$ the $\Xi_t^{(i)}$ are independent and $\left|\Xi_t^{(i)}\right|<m^2\|\boldsymbol{f}\|_\infty^2$. An application of lemma 7 yields:

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2-n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t\odot\mathbf{Q}_t\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}}\leq b_Y n^{-(\frac{1}{4}+\gamma_{Y_2})},$$

for some $b_y>0$ and $\gamma_{Y_2}>0$. Use of the Minkowski inequality yields the result:

$$\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2-\boldsymbol{f}_1^\top\left[(\boldsymbol{v}_{t-1}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{v}_{t-1}}\odot\mathbf{Q}_t\right]\boldsymbol{f}_2\right|^4\right]^{\frac{1}{4}}\leq(a_Y+b_Y)n^{-(\frac{1}{4}+\bar{\gamma}_Y)},$$

for $\bar{\gamma}_Y=\min(\gamma_{Y_1},\gamma_{Y_2})$.

■

Proposition 2. *Let assumptions 2 - 4 hold. Then for all $t\geq 1$:*

$$n^{-1}\mathbf{Y}_t\xrightarrow[a.s.]{\boldsymbol{\theta}^*}\mathbf{N}_t(\boldsymbol{\theta}^*)\odot\mathbf{Q}_t(\boldsymbol{\theta}^*),$$

and for all $r\geq 1$,

$$n^{-1}\bar{\mathbf{Y}}_r\xrightarrow[a.s.]{\boldsymbol{\theta}^*}\sum_{t=\tau_{r-1}+1}^{\tau_r}\mathbf{N}_t(\boldsymbol{\theta}^*)\odot\mathbf{Q}_t(\boldsymbol{\theta}^*).\tag{A.24}$$

Proof. We have that by lemma 12 for all t there exists $c>0$ and $\gamma>0$ such that:

$$\begin{aligned}\mathbb{P}^{\boldsymbol{\theta}^*}&\left(\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2-\boldsymbol{f}_1^\top\left[(\boldsymbol{v}_{t-1}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{v}_{t-1}}\odot\mathbf{Q}_t\right]\boldsymbol{f}_2\right|>\varepsilon\right)\\&\leq\varepsilon^{-4}\mathbb{E}\left[\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2-n^{-1}\boldsymbol{f}_1^\top\mathbf{Z}_t\odot\mathbf{Q}_t\boldsymbol{f}_2\right|^4\right]\\&\leq\varepsilon^{-4}cn^{-(1+\gamma)}.\end{aligned}$$

It follows that:

$$\sum_{n=1}^{\infty}\mathbb{P}^{\boldsymbol{\theta}^*}\left(\left|n^{-1}\boldsymbol{f}_1^\top\mathbf{Y}_t\boldsymbol{f}_2-\boldsymbol{f}_1^\top\left[(\boldsymbol{v}_{t-1}\otimes\mathbf{1}_m)\odot\mathbf{K}_{t,\boldsymbol{v}_{t-1}}\odot\mathbf{Q}_t\right]\boldsymbol{f}_2\right|>\varepsilon\right)<\infty.$$

This result along with a Borel-Cantelli argument, as in proposition 1, completes the proof of the first claim of the proposition. The second claim follows from the first since $\bar{\mathbf{Y}}_r=\sum_{t=\tau_{r-1}+1}^{\tau_r}\mathbf{Y}_t$. ■

A.2 Filtering intensity limits

Case (I)

Define the vectors, or $t\geq 1$:

$$\bar{\boldsymbol{\lambda}}_{0,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) := \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}), \tag{A.25}$$

$$\begin{aligned}
\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &:= \left[(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\odot\boldsymbol{\delta}_t(\boldsymbol{\theta}))}(\boldsymbol{\theta})\right]^\top + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta}), \\
\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &:= \left[\left(\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta})\right)^\top \mathbf{G}_t(\boldsymbol{\theta})\right]^\top + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}), \\
\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &:= \Big[\mathbf{1}_m - \mathbf{q}_t(\boldsymbol{\theta}) \\
&\quad + \left(\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \oslash \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right]^\top \left(\left[\mathbf{1}_m \otimes \mathbf{q}_t(\boldsymbol{\theta})\right] \odot \mathbf{G}_t(\boldsymbol{\theta})^\top\right)\right)^\top \Big] \odot \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}),
\end{aligned} \tag{A.26}$$

where by convention, if we encounter 0/0 in the element-wise division operation we replace that ratio by 0.

Our main objective in section A.2 is to show these vectors are the $\mathbb{P}^{\boldsymbol{\theta}^*}$-a.s. limits of the corresponding finite-n quantities evaluated at $\boldsymbol{\theta}$, computed using algorithm 6. This is the subject of proposition 3.

Proposition 3. *Let assumptions 2 - 4 hold. Then for all $\boldsymbol{\theta} \in \Theta$ and $t \geq 1$:*

$$\begin{aligned}
n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \\
n^{-1}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}) &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \\
n^{-1}\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}).
\end{aligned}$$

The proof is postponed until later in section A.2.

Remark 1. *By writing out the above definitions it can be checked that $\boldsymbol{\nu}_t(\boldsymbol{\theta}^*) = \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$, hence lemma 9 implies by a Borel-Cantelli argument $n^{-1}\mathbf{x}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$; and that $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ is equal to the right hand side of* (A.18) *in proposition 1, hence $n^{-1}\mathbf{y}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$. Therefore proposition 3 implies that if algorithm 6 is run with the model specified by the DGP $\boldsymbol{\theta}^*$, thus computing $\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}^*)$ and $\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)$, that when rescaled by n^{-1} these vectors converge as $n \to \infty$ to the same $\mathbb{P}^{\boldsymbol{\theta}^*}$-almost sure limits as $n^{-1}\mathbf{x}_t$ and $n^{-1}\mathbf{y}_t$. We provide empirical evidence for this remark in section 4.4.*

As preliminaries to the proof of proposition 3 we need to verify that certain quantities in algorithm 6 and the vectors defined at the start of section A.2 are $\mathbb{P}^{\boldsymbol{\theta}^*}$-a.s. well-defined and finite. This is the purpose of lemma 13 and lemma 14. In algorithm 6, if $\boldsymbol{\mu}_{t,n}^{(i)}(\boldsymbol{\theta}) = 0$ and $y_t^{(i)} > 0$, then line 3 would entail dividing a finite number by zero. Lemma 13 establishes that this happens with probability zero.

Lemma 13. *Let assumptions 2-4 hold. For any $\boldsymbol{\theta} \in \Theta$, $n \in \mathbb{N}$, $i \in [m]$, and $t \geq 1$,*

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0\right) > 0 \implies y_t^{(i)} = 0, \quad \mathbb{P}_n^{\boldsymbol{\theta}^*}\text{-}a.s.$$

Proof. Fix arbitrary $\boldsymbol{\theta} \in \Theta$ and $n \in \mathbb{N}$. All a.s. statements in the proof are with respect to $\mathbb{P}_n^{\boldsymbol{\theta}^*}$. We will show that for all $j \in [m]$ and $t \geq 1$, the following two implications hold:

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(j)}(\boldsymbol{\theta}) = 0\right) > 0 \implies x_t^{(j)} = 0, \quad a.s., \tag{A.27}$$

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0\right) > 0 \implies y_t^{(i)} = 0, \quad a.s. \tag{A.28}$$

The proof is inductive in t. To initialise the induction at $t = 1$, let $j \in [m]$ and suppose that $\mathbb{P}_n^{\boldsymbol{\theta}^*}(\lambda_{1,n}^{(j)}(\boldsymbol{\theta}) = 0) > 0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\sum_{k=1}^{m} \bar{\lambda}_{0,n}^{(k)}(\boldsymbol{\theta})\delta_1^{(k)}(\boldsymbol{\theta})K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{0,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_1(\boldsymbol{\theta}))}(\boldsymbol{\theta}) + \alpha_{1,n}^{(j)}(\boldsymbol{\theta}) = 0\right) > 0,$$

then $\alpha_{1,n}^{(j)}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $\alpha_{1,n}^{(j)}(\boldsymbol{\theta}^*) = 0$ and hence $\hat{x}_1^{(j)} = 0$, a.s. Furthermore, for all $k \in [m]$ we must have that either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{0,n}^{(k)}(\boldsymbol{\theta}) = 0\right) > 0$, which, since $\lambda_{0,n}^{(k)}(\boldsymbol{\theta})$ is a deterministic quantity, implies $\lambda_{0,n}^{(k)}(\boldsymbol{\theta}) = 0$, in turn by assumption 4 this implies $\lambda_{0,n}^{(k)}(\boldsymbol{\theta}^*) = 0$ so that $x_0^{(k)} = 0$ a.s. and $\bar{x}_0^{(k)} = 0$ a.s.; or
- $\delta_1^{(k)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $\delta_1^{(k)}(\boldsymbol{\theta}^*) = 0$ which means $\bar{x}_0^{(k)} = 0$ a.s.; or
- $K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{0,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_1(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$, which by assumptions 2, 3, and 4 implies $K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\mathbf{x}}_0)}(\boldsymbol{\theta}^*) = 0$ a.s.

Hence we have for all $k \in [m]$ either $\bar{x}_0^{(k)} = 0$ a.s. or $K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\mathbf{x}}_0)}(\boldsymbol{\theta}^*) = 0$ a.s. Since, given $\mathbf{x}_0$, $\tilde{x}_1^{(j)} \sim \sum_{k=1}^{m} \mathrm{Bin}\left(x_0^{(k)}, K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\mathbf{x}}_0)}(\boldsymbol{\theta}^*)\right)$ we must have that $\tilde{x}_1^{(j)} = 0$ a.s., therefore we have that $x_1^{(j)} = \tilde{x}_1^{(j)} + \hat{x}_1^{(j)} = 0$ a.s. We have thus proved (A.27) in the case $t = 1$.

Now let us prove (A.28) in the case $t = 1$. Suppose that for some $i \in [m]$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\mu_{1,n}^{(i)}(\boldsymbol{\theta}) = 0\right) > 0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\sum_{j=1}^{m} \lambda_{1,n}^{(j)}(\boldsymbol{\theta})q_1^{(j)}(\boldsymbol{\theta})G_1^{(j,i)}(\boldsymbol{\theta}) + \kappa_{1,n}^{(i)}(\boldsymbol{\theta}) = 0\right) > 0.$$

Then $\kappa_{1,n}^{(i)}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $\kappa_{1,n}^{(i)}(\boldsymbol{\theta}^*) = 0$ and hence $\hat{y}_1^{(i)} = 0$ a.s. Furthermore, for all $j \in [m]$ we must have that either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{1,n}^{(j)}(\boldsymbol{\theta}) = 0\right) > 0$, which implies $x_1^{(j)} = 0$ a.s. which implies $\bar{y}_1^{(j)} = 0$ a.s.; or
- $q_1^{(j)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies that $q_1^{(j)}(\boldsymbol{\theta}^*) = 0 \implies \bar{y}_1^{(j)} = 0$ a.s.; or
- $G_1^{(j,i)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $G_1^{(j,i)}(\boldsymbol{\theta}^*) = 0$.

Given, $\bar{\mathbf{y}}_1$, $\tilde{y}_1^{(i)} \sim \sum_{j=1}^{m} \mathrm{Bin}\left(\bar{y}_1^{(j)}, G_1^{(j,i)}(\boldsymbol{\theta}^*)\right)$. This means that $\tilde{y}_1^{(i)} = 0$ a.s., and furthermore that $y_1^{(i)} = \tilde{y}_1^{(i)} + \hat{y}_1^{(i)} = 0$ a.s. This completes the proof of (A.28) in the case $t = 1$.

As an induction hypothesis suppose that (A.27) and (A.28) hold at t. We shall show that $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta}) = 0\right) > 0 \implies x_{t+1}^{(j)} = 0$ a.s. Firstly we will show that, for all $k \in [m]$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta}) = 0\right) >$

$0 \implies \mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$ which, by the induction hypothesis, would imply $x_t^{(k)}=0$ a.s. Suppose that for some $k\in[m]$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left((1-q_t^{(k)}(\boldsymbol{\theta}))\lambda_{t,n}^{(k)}(\boldsymbol{\theta})+\sum_{j=1}^{m} y_t^{(j)}\frac{\lambda_{t,n}^{(k)}(\boldsymbol{\theta})q_t^{(k)}(\boldsymbol{\theta})G_t^{(k,j)}(\boldsymbol{\theta})}{\mu_{t,n}^{(j)}(\boldsymbol{\theta})}=0\right)>0.$$

Firstly, $\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta})$ is almost surely well defined by the induction hypothesis, since the event $\mu_{t,n}^{(j)}(\boldsymbol{\theta})=0$ and $y_t^{(j)}>0$ has probability 0 for each $j\in[m]$. Now if the above displayed inequality holds we must have that either:

- $q_t^{(k)}(\boldsymbol{\theta})<1$, in which case we must have $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$; or
- $q_t^{(k)}(\boldsymbol{\theta})=1$, in which case we must have $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(k)}(\boldsymbol{\theta})G_t^{(k,j)}(\boldsymbol{\theta})=0\right)>0$ for all j so that the sum is equal to 0 with positive probability, and since $\mathbf{G}_t$ is row-stochastic matrix, there must exist a $j\in[m]$ such that $G_t^{(k,j)}(\boldsymbol{\theta})>0$, hence $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$.

We have thus shown $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0 \implies \mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$ which by the induction hypothesis implies $x_t^{(k)}=0$ a.s. so that further $\bar{x}_t^{(k)}=0$ a.s. Now if for some $j\in[m]$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta})=0\right)>0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\sum_{k=1}^{m}\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta})\delta_{t+1}^{(k)}(\boldsymbol{\theta})K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta})+\alpha_{t+1,n}^{(j)}(\boldsymbol{\theta})=0\right)>0,$$

then $\alpha_{t+1,n}^{(j)}(\boldsymbol{\theta})=0$, which by assumption 2 implies $\alpha_{t+1,n}^{(j)}(\boldsymbol{\theta}^*)=0$, hence $\hat{x}_t^{(j)}=0$ a.s. Furthermore, for all $k\in[m]$ we must have that either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta})=0\right)>0$, which implies $x_t^{(k)}=0$ a.s. $\implies \bar{x}_t^{(k)}=0$ a.s.; or
- $\delta_{t+1}^{(k)}(\boldsymbol{\theta})=0$, which by assumption 2 implies $\delta_{t+1}^{(k)}(\boldsymbol{\theta}^*)=0 \implies \bar{x}_t^{(k)}=0$ a.s.; or
- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta})=0\right)>0$. We claim this implies that $K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t)}(\boldsymbol{\theta})=0$, a.s. Suppose, for contradiction, that $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta})=0\right)>0$ and $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t)}(\boldsymbol{\theta})>0\right)>0$. Then there exist $E,E'\subseteq\Omega_n$ with $\mathbb{P}_n^{\boldsymbol{\theta}^*}(E)>0$ and $\mathbb{P}_n^{\boldsymbol{\theta}^*}(E')>0$ such that for all $\omega\in E$ and all $\omega'\in E'$:

  $$K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta},\omega)\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta})=0 \text{ and } K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t(\omega'))}(\boldsymbol{\theta})>0,$$

  which implies:

  $$\operatorname{supp}\left(\mathbf{K}^{(k,\cdot)}_{t+1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta},\omega)\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta})\right)\not\subseteq\operatorname{supp}\left(\mathbf{K}^{(k,\cdot)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t(\omega'))}(\boldsymbol{\theta})\right).$$

  By assumption 3 this implies:

  $$\operatorname{supp}\left(\bar{\mathbf{x}}_t(\omega')\right)\not\subseteq\operatorname{supp}\left(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta},\omega)\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta})\right),$$

i.e. there exists l such that:

$$\left(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta},\omega)\odot\boldsymbol{\delta}_{t+1}(\boldsymbol{\theta})\right)^{(l)}=0 \text{ and } \left(\bar{\mathbf{x}}_t(\omega')\right)^{(l)}>0.$$

But since $\mathbb{P}_n^{\boldsymbol{\theta}^*}(E)>0$ and $\mathbb{P}_n^{\boldsymbol{\theta}^*}(E')>0$ this implies that:

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}(\bar{\lambda}_{t,n}^{(l)}(\boldsymbol{\theta})=0)>0 \text{ and } \mathbb{P}_n^{\boldsymbol{\theta}^*}(\bar{x}_t^{(l)}>0)>0.$$

This contradicts the observation in the first bullet point, hence $K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t)}(\boldsymbol{\theta})=0$ a.s. Then by assumption 2 we have $K^{(k,j)}_{t+1,\boldsymbol{\eta}(\bar{\mathbf{x}}_t)}(\boldsymbol{\theta}^*)=0$ a.s.

Hence, similarly to the argument used in the case $t=1$, we must have that $\tilde{x}_{t+1}^{(j)}=0$ a.s. so that $x_{t+1}^{(j)}=\tilde{x}_{t+1}^{(j)}+\hat{x}_{t+1}^{(j)}=0$ a.s. Thus (A.27) holds with t replace by $t+1$.

It remains to show that (A.28) holds with t replaced by $t+1$. So suppose that for some $i\in[m]$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\mu_{t+1,n}^{(i)}(\boldsymbol{\theta})=0\right)>0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\sum_{j=1}^{m}\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta})q_{t+1}^{(j)}(\boldsymbol{\theta})G_{t+1}^{(j,i)}(\boldsymbol{\theta})+\kappa_{t+1,n}^{(i)}(\boldsymbol{\theta})=0\right)>0,$$

then we must have $\kappa_{t+1,n}^{(i)}(\boldsymbol{\theta})=0$ which by assumption 2 implies $\kappa_{t+1,n}^{(i)}(\boldsymbol{\theta}^*)=0$ hence $\hat{y}_{t+1}^{(i)}=0$ a.s. Furthermore, for all $j\in[m]$ we must have either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta})=0\right)>0$, which implies that $x_{t+1}^{(j)}=0\implies\bar{y}_{t+1}^{(j)}=0$ a.s.; or
- $q_{t+1}^{(j)}(\boldsymbol{\theta})=0$, which by assumption 2 implies $q_{t+1}^{(j)}(\boldsymbol{\theta}^*)=0\implies\bar{y}_{t+1}^{(j)}=0$ a.s.; or
- $G_{t+1}^{(j,i)}(\boldsymbol{\theta})=0$, which by assumption 2 implies $G_{t+1}^{(j,i)}(\boldsymbol{\theta}^*)=0$.

Hence, using the same reasoning as in the $t=1$ case, we have $\tilde{y}_{t+1}^{(i)}=0$ a.s. and furthermore $y_{t+1}^{(i)}=\tilde{y}_{t+1}^{(i)}+\hat{y}_{t+1}^{(i)}=0$ a.s. This completes the proof of (A.28) with t replaced by $t+1$. The induction is therefore complete. ■

If $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})=0$ and $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)>0$ then $\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ would involve division of a finite number by zero. Lemma 14 establishes that this situation cannot arise.

Lemma 14. *Let assumptions 2 - 4 hold. For any $\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta$, $i\in[m]$ and $t\geq 1$,*

$$\bar{\lambda}_{t,\infty}^{(j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})=0\implies\bar{\lambda}_{t,\infty}^{(j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')=0,$$
$$\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})=0\implies\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')=0.$$

Proof. Fix arbitrary $\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta$. By symmetry we only need to prove the implication in one direction. We will show that the following two implications hold for all $i,j\in[m]$ and $t\geq 1$:

$$\lambda_{t,\infty}^{(j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})=0\implies\lambda_{t,\infty}^{(j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')=0, \tag{A.29}$$
$$\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})=0\implies\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')=0. \tag{A.30}$$

For the $t = 1$ case, if $\lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, i.e.,

$$\sum_{k=1}^{m} \lambda^{(k)}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \delta^{(k)}_1(\boldsymbol{\theta}) K^{(k,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) + \alpha^{(j)}_{1,\infty}(\boldsymbol{\theta}) = 0,$$

then $\alpha^{(j)}_{1,\infty}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $\alpha^{(j)}_{1,\infty}(\boldsymbol{\theta}') = 0$. Furthermore, for each $j \in [m]$ we must have either:

- $\lambda^{(k)}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \lambda^{(k)}_{0,\infty}(\boldsymbol{\theta}) = 0$, which by assumption 4 implies $\lambda^{(k)}_{0,\infty}(\boldsymbol{\theta}') = 0$; or
- $\delta^{(k)}_1(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $\delta^{(k)}_1(\boldsymbol{\theta}') = 0$; or
- $K^{(k,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$, which by assumptions 2, 3, and 4 implies $K^{(k,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \odot \boldsymbol{\delta}(\boldsymbol{\theta}'))}(\boldsymbol{\theta}') = 0$.

Hence we have:

$$\lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = \sum_{k=1}^{m} \lambda^{(k)}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \delta^{(k)}_1(\boldsymbol{\theta}') K^{(k,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \odot \boldsymbol{\delta}_1(\boldsymbol{\theta}'))}(\boldsymbol{\theta}') + \alpha^{(j)}_{1,\infty}(\boldsymbol{\theta}') = 0,$$

so (A.29) holds with $t = 1$. In order to establish (A.30) with $t = 1$, consider

$$\mu^{(i)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q^{(j)}_1(\boldsymbol{\theta}) G^{(i,j)}_1(\boldsymbol{\theta}) + \kappa^{(i)}_{1,\infty}(\boldsymbol{\theta}) = 0,$$

hence $\kappa^{(i)}_{1,\infty}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $\kappa^{(i)}_{1,\infty}(\boldsymbol{\theta}') = 0$. Furthermore, for each $j \in [m]$ we must have either:

- $\lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which by the above implies $\lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$; or
- $q^{(j)}_1(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $q^{(j)}_1(\boldsymbol{\theta}') = 0$; or
- $G^{(i,j)}_1(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $G^{(i,j)}_1(\boldsymbol{\theta}') = 0$.

Hence:

$$\mu^{(i)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = \sum_{j=1}^{m} \lambda^{(j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') q^{(j)}_1(\boldsymbol{\theta}') G^{(i,j)}_1(\boldsymbol{\theta}') + \kappa^{(j)}_{1,\infty}(\boldsymbol{\theta}') = 0.$$

Thus we have shown that (A.30) holds with $t = 1$.

For the induction hypothesis, assume that (A.29) and (A.30) with hold for some $t \geq 1$. Then for each $k \in [m]$ write:

$$\bar{\lambda}^{(k)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = (1 - q^{(k)}_t(\boldsymbol{\theta})) \lambda^{(k)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) + \sum_{j=1}^{m} \mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \frac{\lambda^{(k)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q^{(k)}_t(\boldsymbol{\theta}) G^{(k,j)}_t(\boldsymbol{\theta})}{\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})} = 0.$$

This is well defined by the induction hypothesis choosing $\boldsymbol{\theta}' = \boldsymbol{\theta}^*$. Furthermore, we must have either:

- $q^{(k)}_t(\boldsymbol{\theta}) < 1$, in which case we must have $\lambda^{(k)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda^{(k)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$; or

- $q_t^{(k)}(\boldsymbol{\theta}) = 1$, in which case we must have $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) G_t^{(k,j)}(\boldsymbol{\theta}) = 0$ a.s. for all j so that the sum is equal to 0. Since $\mathbf{G}_t$ is row-stochastic matrix, we know there must exist a $j \in [m]$ such that $G_t^{(k,j)}(\boldsymbol{\theta}) > 0$ and so we must have $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$.

So we have $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$, indeed the reverse implication is also true by definition of $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ so that $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \iff \lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$. Now consider

$$\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{k=1}^{m} \bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \delta_{t+1}^{(k)}(\boldsymbol{\theta}) K^{(k,j)}_{1, \boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) + \alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}) = 0,$$

then $\alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}) = \alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}') = 0$ and for all $k \in [m]$ we must have either:

- $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which implies by the above that $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0 \implies \bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}')$; or
- $\delta_{t+1}^{(k)}(\boldsymbol{\theta}) = 0$, which by assumption 2 that $\delta_{t+1}^{(k)}(\boldsymbol{\theta}') = 0$; or
- $K^{(k,j)}_{1, \boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$ which by assumptions 2, 3, and the induction hypothesis implies $K^{(k,j)}_{1, \boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}'))}(\boldsymbol{\theta}') = 0$.

Hence

$$\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = \sum_{k=1}^{m} \bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \delta_{t+1}^{(k)}(\boldsymbol{\theta}') K^{(k,j)}_{1, \boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}'))}(\boldsymbol{\theta}') + \alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}') = 0.$$

Now, if for some $i \in [m]$:

$$\mu_{t+1,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q_{t+1}^{(j)}(\boldsymbol{\theta}) G_{t+1}^{(j,i)}(\boldsymbol{\theta}) + \kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}) = 0,$$

then $\kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $\kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}') = 0$ and for all $j \in [m]$ we have either:

- $\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which by the above implies $\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$; or
- $q_{t+1}^{(j)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $q_{t+1}^{(j)}(\boldsymbol{\theta}') = 0$; or
- $G_{t+1}^{(j,i)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $G_{t+1}^{(j,i)}(\boldsymbol{\theta}') = 0$.

Hence

$$\mu_{t+1,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = \sum_{j=1}^{m} \lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') q_{t+1}^{(j)}(\boldsymbol{\theta}') G_{t+1}^{(i,j)}(\boldsymbol{\theta}') + \kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}') = 0,$$

and the inductive proof is complete. ■

The following lemma will be used in the proof of lemma 5.

Lemma 15. *Let assumptions 2- 4 hold. For all $\boldsymbol{\theta} \in \Theta$, $n \in \mathbb{N}$, and $i \in [m]$:*

$$\lambda_{t,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,n}^{(j)}(\boldsymbol{\theta}) = 0 \quad a.s.,$$
$$\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0 \quad a.s..$$

Proof. Fix arbitrary $\boldsymbol{\theta}, \in \Theta$ and $n \in \mathbb{N}$. All almost sure statements in the proof are made with respect to $\mathbb{P}_n^{\boldsymbol{\theta}^*}$. We will show by induction that the following two implications hold for all $t \geq 1$ and $i, j \in [m]$.

$$\lambda_{t,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,n}^{(j)}(\boldsymbol{\theta}) = 0 \quad a.s., \tag{A.31}$$

$$\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0 \quad a.s. \tag{A.32}$$

For $t = 1$ consider:

$$\lambda_{1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{k=1}^{m} \bar{\lambda}_{0,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \delta_1^{(k)}(\boldsymbol{\theta}) K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_1(\boldsymbol{\theta}))}^{(k,j)}(\boldsymbol{\theta}) + \alpha_{1,\infty}^{(j)}(\boldsymbol{\theta}) = 0,$$

then $\alpha_{1,\infty}^{(j)}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $\alpha_{1,n}^{(j)}(\boldsymbol{\theta}) = 0$, and for all $j \in [m]$ we must have either:

- $\bar{\lambda}_{0,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \lambda_{0,\infty}^{(k)}(\boldsymbol{\theta}) = 0$, which by assumption 4 implies $\lambda_{0,n}^{(k)}(\boldsymbol{\theta}) = 0$; or
- $\delta_1^{(k)}(\boldsymbol{\theta}) = 0$; or
- $K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}(\boldsymbol{\theta}))}^{(k,j)}(\boldsymbol{\theta}) = 0$ which by assumptions 3 and 4 implies $K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}(\boldsymbol{\theta}))}^{(k,j)}(\boldsymbol{\theta}) = 0$.

Hence we have:

$$\lambda_{1,n}^{(j)}(\boldsymbol{\theta}) = \sum_{k=1}^{m} \lambda_{0,n}^{(k)}(\boldsymbol{\theta}') \delta_1^{(k)}(\boldsymbol{\theta}') K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}') \odot \boldsymbol{\delta}_1(\boldsymbol{\theta}'))}^{(k,j)}(\boldsymbol{\theta}') + \alpha_{1,n}^{(j)}(\boldsymbol{\theta}') = 0, \quad a.s.$$

Now consider:

$$\mu_{1,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_{1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q_1^{(j)}(\boldsymbol{\theta}) G_1^{(i,j)}(\boldsymbol{\theta}) + \kappa_{1,\infty}^{(i)}(\boldsymbol{\theta}) = 0,$$

then $\kappa_{1,\infty}^{(i)}(\boldsymbol{\theta}) = 0$, which by assumption 2 implies $\kappa_{1,n}^{(i)}(\boldsymbol{\theta}) = 0$, furthermore for al $j \in [m]$ we must have either:

- $\lambda_{1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which by the above implies $\lambda_{1,n}^{(j)}(\boldsymbol{\theta}') = 0 \quad \mathbb{P}^{\boldsymbol{\theta}^*} a.s.$; or
- $q_1^{(j)}(\boldsymbol{\theta}) = 0$; or
- $G_1^{(i,j)}(\boldsymbol{\theta}) = 0$.

Hence:

$$\mu_{1,n}^{(i)}(\boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_{1,n}^{(j)}(\boldsymbol{\theta}) q_1^{(j)}(\boldsymbol{\theta}) G_1^{(k,j)}(\boldsymbol{\theta}) + \kappa_{1,n}^{(i)}(\boldsymbol{\theta}) = 0.$$

For the induction hypothesis, assume (A.31) and (A.32) hold. Then for each $k \in [m]$,

$$\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = (1 - q_t^{(k)}(\boldsymbol{\theta})) \lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) + \sum_{j=1}^{m} \mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \frac{\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q_t^{(k)}(\boldsymbol{\theta}) G_t^{(k,j)}(\boldsymbol{\theta})}{\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta})} = 0.$$

This is well defined by the induction hypothesis. Furthermore, in order for this equality with zero to hold we must have either:

- $q_t^{(k)}(\boldsymbol{\theta}) < 1$, in which case we must have $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$ which by the induction hypothesis implies $\lambda_{t,n}^{(k)}(\boldsymbol{\theta}) = 0 \quad a.s.$; or
- $q_t^{(k)}(\boldsymbol{\theta}) = 1$, in which case we must have $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) G_t^{(k,j)}(\boldsymbol{\theta}) = 0 \quad a.s.$ for all j so that the sum is equal to 0. Since $\mathbf{G}_t$ is row-stochastic matrix, we know there must exist a $j \in [m]$ such that $G_t^{(k,j)}(\boldsymbol{\theta}) > 0$ and so we must have $\lambda_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$ which by the induction hypothesis implies $\lambda_{t,n}^{(k)}(\boldsymbol{\theta}) = 0$ a.s.

So we have $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \lambda_{t,n}^{(k)}(\boldsymbol{\theta}) = 0$ a.s., furthermore $\lambda_{t,n}^{(k)}(\boldsymbol{\theta}) = 0$ a.s. $\implies \bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta}) = 0$ a.s.. Now if for some $j \in [m]$:

$$\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{k=1}^{m} \bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \delta_{t+1}^{(k)}(\boldsymbol{\theta}) K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) + \alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}) = 0,$$

then $\alpha_{t+1,\infty}^{(j)}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $\alpha_{t+1,n}^{(j)}(\boldsymbol{\theta}) = 0$ and for each $k \in [m]$ we must have either:

- $\bar{\lambda}_{t,\infty}^{(k)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which we have already shown implies $\lambda_{t,n}^{(k)}(\boldsymbol{\theta}) = 0 \quad a.s. \implies \bar{\lambda}_{t,n}^{(k)}(\boldsymbol{\theta}) = 0 \quad a.s.$; or
- $\delta_{t+1}^{(k)}(\boldsymbol{\theta}) = 0$; or
- $K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$, which together with assumption 3 implies $K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$.

Hence

$$\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta}) = \sum_{k=1}^{m} \bar{\lambda}_{t,n}^{(k)} \boldsymbol{\theta}) \delta_{t+1}^{(k)}(\boldsymbol{\theta}) K^{(k,j)}_{1,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) \odot \boldsymbol{\delta}_{t+1}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) + \alpha_{t+1,n}^{(j)}(\boldsymbol{\theta}) = 0 \quad a.s.$$

Now, if

$$\mu_{t+1,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q_{t+1}^{(j)}(\boldsymbol{\theta}) G_{t+1}^{(k,j)}(\boldsymbol{\theta}) + \kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}) = 0,$$

then $\kappa_{t+1,\infty}^{(i)}(\boldsymbol{\theta}) = 0$ which by assumption 4 implies $\kappa_{t+1,n}^{(i)}(\boldsymbol{\theta}) = 0$ and for all $j \in [m]$ we have either:

- $\lambda_{t+1,\infty}^{(j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$, which implies $\lambda_{t+1,n}^{(j)}(\boldsymbol{\theta}) = 0$ a.s.; or
- $q_{t+1}^{(j)}(\boldsymbol{\theta}) = 0$; or
- $G_{t+1}^{(k,j)}(\boldsymbol{\theta}) = 0$.

Hence

$$\mu_{t+1,n}^{(i)}(\boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_{t+1,n}^{(j)}(\boldsymbol{\theta}^) q_{t+1}^{(j)}(\boldsymbol{\theta}) G_{t+1}^{(k,j)}(\boldsymbol{\theta}) + \kappa_{t+1,\infty}^{(j)}(\boldsymbol{\theta}) = 0 \quad a.s.,$$

and the inductive proof is complete.

■

Proof of Proposition 3 Fix any $\boldsymbol{\theta} \in \Theta$. We proceed by induction to show that for all $t \geq 1$,

$$n^{-1}\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}),$$

with the other claims of the proposition proved along the way.

Using assumption 4 we have:

$$n^{-1}\bar{\boldsymbol{\lambda}}_{0,n}(\boldsymbol{\theta}) = n^{-1}\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}) = \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}).$$

Now, for $t \geq 1$ assume that $n^{-1}\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. We have:

$$\begin{aligned} n^{-1}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}) &= \left[(n^{-1}\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta})\odot\boldsymbol{\delta}_t(\boldsymbol{\theta}))}\right]^\top + n^{-1}\boldsymbol{\alpha}_{t,n}(\boldsymbol{\theta}) \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \left[(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\odot\boldsymbol{\delta}_t(\boldsymbol{\theta}))}\right]^\top + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta}) \\ &= \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \end{aligned}$$

by the continuous mapping theorem (CMT) and assumptions 2 and 3. A further application of the CMT and assumption 2 yields:

$$\begin{aligned} n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) &= \left[\left(n^{-1}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta})\right)^\top \mathbf{G}_t(\boldsymbol{\theta})\right]^\top + n^{-1}\boldsymbol{\kappa}_{t,n}(\boldsymbol{\theta}) \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \left[\left(\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta})\right)^\top \mathbf{G}_t(\boldsymbol{\theta})\right]^\top + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}) \\ &= \boldsymbol{\mu}_{t,\infty}\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}\right) \end{aligned}$$

Recalling from remark 1 that $n^{-1}\mathbf{y}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$ and applying the CMT, we have:

$$\begin{aligned} n^{-1}\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) &= \Big[\mathbf{1}_m - \mathbf{q}_t(\boldsymbol{\theta}) \\ &\quad + \left(\left[n^{-1}\mathbf{y}_t \oslash n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})\right]^\top \left(\left[\mathbf{1}_m \otimes \mathbf{q}_t(\boldsymbol{\theta})\right) \odot \mathbf{G}_t(\boldsymbol{\theta})\right]^\top\right)^\top \Big] \odot n^{-1}\boldsymbol{\lambda}_{t,n}(\boldsymbol{\theta}) \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \Big[\mathbf{1}_m - \mathbf{q}_t(\boldsymbol{\theta}) \\ &\quad + \left(\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \oslash \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]^\top \left(\left[\mathbf{1}_m \otimes \mathbf{q}_t(\boldsymbol{\theta})\right] \odot \mathbf{G}_t(\boldsymbol{\theta})^\top\right)\right)^\top \Big] \odot \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\ &= \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \end{aligned}$$

We note this limit is almost surely well defined since by lemma 14 for any $i \in [m]$,

$$\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0 \iff \mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$$

and by lemma 13 if $\mu^{(i)}_{t,n}(\boldsymbol{\theta}) = 0$ with positive probability then $y_t = 0$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-a.s. In both these cases we are working under the convention $\frac{0}{0} := 0$. ■

Lemma 16. *Let assumptions 2- 4 hold. For all $\boldsymbol{\theta}^* \in \Theta$ and $t \geq 1$ the function $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous on Θ.*

Proof. Fix an arbitrary $\boldsymbol{\theta}^* \in \Theta$. Note that $\bar{\boldsymbol{\lambda}}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \bar{\boldsymbol{\lambda}}_{0,\infty}(\boldsymbol{\theta})$ is continuous by assumption 4. We will now show that for any $t \geq 1$, continuity of $\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ implies continuity of $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, and $\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, from which the claim of the lemma follows.

Henceforth assume that $\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous and recall that by definition of $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$,

$$\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \left[(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t, \boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))}(\boldsymbol{\theta}) \right]^\top + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta}).$$

Continuity of $\boldsymbol{\delta}_t(\boldsymbol{\theta})$ and $\boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ holds directly by assumptions 2 and 4. By assumption 3 we know that $\mathbf{K}_{t,\boldsymbol{\eta}}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Hence, to show continuity of $\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ we shall show that $\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))$ is continuous in $\boldsymbol{\theta}$. The function $\boldsymbol{\eta} : \mathbb{R}^m_{\geq 0} \to \mathbb{R}^m_{\geq 0}$ is continuous everywhere except at $\mathbf{0}_m$, we now show that, by virtue of our assumptions, this discontinuity is immaterial. Consider the two following cases:

- There exists $\boldsymbol{\theta}' \in \Theta$ such that $\bar{\boldsymbol{\lambda}}_{t-1,\infty}\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}'\right) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}') = \mathbf{0}_m$. In this case, by assumption 2 and lemma 14 we have that $\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}) = \mathbf{0}_m$ for all $\boldsymbol{\theta} \in \Theta$, from which it follows that $\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta})) = \mathbf{0}_m$ for all $\boldsymbol{\theta} \in \Theta$, so that the continuity of $\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))$ in $\boldsymbol{\theta}$ on Θ holds trivially;
- For all $\boldsymbol{\theta} \in \Theta$, $\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}) \neq \mathbf{0}_m$. In this case the continuity of $\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))$ in $\boldsymbol{\theta}$ on Θ follows from the continuity of $\boldsymbol{\eta}$ on $\mathbb{R}^m_{\geq 0} \setminus \{\mathbf{0}_m\}$.

Hence, $\boldsymbol{\theta} \mapsto \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous. Recall that:

$$\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \left[\left(\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta}) \right)^\top \mathbf{G}_t(\boldsymbol{\theta}) \right]^\top + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta}).$$

Due to the continuity of $\lambda_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ and assumption 2, this is a composition of continuous functions and hence $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is itself continuous. Now consider

$$\begin{aligned} \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \Big[& \mathbf{1}_m - \mathbf{q}_t(\boldsymbol{\theta}) \\ & + \left(\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \oslash \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \right]^\top \left(\left[\mathbf{1}_m \otimes \mathbf{q}_t(\boldsymbol{\theta}) \right] \odot \mathbf{G}_t(\boldsymbol{\theta})^\top \right) \right)^\top \Big] \odot \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}). \end{aligned}$$

Each component of this function is trivially continuous on Θ except the $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \oslash \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ term, we will now prove its continuity. By lemma 14, for each $i \in [m]$ we need only consider the two cases:

- either $\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Theta$, in which case we have by convention $\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)/\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := 0$, which is continuous; or
- $\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \neq 0$ for all $\boldsymbol{\theta} \in \Theta$, in which case $\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)/\mu^{(i)}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous.

Hence we have elementwise continuity of $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \oslash \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ which gives us continuity of $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$.

We have shown that continuity of $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ on Θ implies continuity of $\boldsymbol{\theta} \mapsto \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ and $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ on Θ, which completes the proof. ■

Case (II)

Define:

$$\bar{\boldsymbol{\lambda}}_{0,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \coloneqq \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}), \tag{A.33}$$

and for $r = 1,\dots,R$ and $t = \tau_{r-1}+1,\dots,\tau_r - 1$,

$$\begin{aligned}
\boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq \left(\bar{\boldsymbol{\lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \otimes \mathbf{1}_m\right) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}(\boldsymbol{\theta}^*,\boldsymbol{\theta}),\\
\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq (\mathbf{1}_m^\top \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))^\top,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Lambda}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq \left(\bar{\boldsymbol{\lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \otimes \mathbf{1}_m\right) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}(\boldsymbol{\theta}),\\
\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq \sum_{s=\tau_{r-1}+1}^{\tau_r} \Lambda_{s,\infty}(\boldsymbol{\theta}) \odot \mathbf{Q}_s(\boldsymbol{\theta}),\\
\bar{\boldsymbol{\Lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq \left[\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right] \odot \boldsymbol{\Lambda}_{\tau_r}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\\
&\qquad + \left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \oslash \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right] \odot \left[\boldsymbol{\Lambda}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right],\\
\bar{\boldsymbol{\lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &\coloneqq (\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))^\top.
\end{aligned} \tag{A.34}$$

where if we encounter 0/0 in the element-wise division operation we set the entry to 0 by convention. The main result of section A.2 is proposition 4 concerning the convergence to the above of the associated finite-n quantities computed using algorithm 8.

Proposition 4. *Let assumptions 2 - 4 hold. For any $\boldsymbol{\theta} \in \Theta$ and $r \geq 1$ and $t \geq 1$:*

$$\begin{aligned}
n^{-1}\mathbf{M}_{r,n}(\boldsymbol{\theta}) &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}),\\
n^{-1}\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \Lambda_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}),
\end{aligned}$$

The proof is postponed until later in section A.2.

Remark 2. *Similarly to properties of case (I) pointed out in remark 1, by writing out the above definitions it can be checked that $\mathbf{N}_t(\boldsymbol{\theta}^*) = \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$, thus $n^{-1}\mathbf{Z}_t \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$; and that $\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ is equal to the right hand side of* (A.24)*, thus $n^{-1}\bar{\mathbf{Y}}_r \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$.*

Similarly to as in section A.2, in order to prove proposition 4 we need to check that certain quantities are almost surely well defined. For the update step of algorithm 8 to be $\mathbb{P}^{\boldsymbol{\theta}^*}$-a.s. well defined for all $\boldsymbol{\theta} \in \Theta$ we need that if $M_{r,n}^{(i,j)}(\boldsymbol{\theta}) = 0$ occurs with positive probability then $\bar{Y}_r^{(i,j)} = 0$ $\mathbb{P}^{\boldsymbol{\theta}^*}$- a.s. This is established in the following lemma.

Lemma 17. *Let assumptions 2 - 4 hold. For any $\boldsymbol{\theta} \in \Theta$, $n \in \mathbb{N}$, $(i,j) \in [m]^2$ and $r = 1,\dots,R$:*

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(M_{r,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0 \implies \bar{Y}_r^{(i,j)} = 0, \quad \mathbb{P}_n^{\boldsymbol{\theta}^*}\, a.s.$$

Proof. Fix arbitrary $\boldsymbol{\theta} \in \Theta$ and $n \in \mathbb{N}$. All almost sure statements made throughout the proof are with respect to $\mathbb{P}_n^{\boldsymbol{\theta}^*}$. We will prove by induction that for all $r = 1,\dots,R$ we have that for all $s \in \{\tau_{r-1}+1,\dots,\tau_r\}$ and $(i,j) \in [m]^2$, the following two implications hold.

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{s,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0 \implies Z_s^{(i,j)} = 0, \quad a.s., \tag{A.35}$$

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(M_{r,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0 \implies \bar{Y}_r^{(i,j)} = 0 \quad a.s. \tag{A.36}$$

Consider the case $r = 1$. We will first show that for all $s \in \{\tau_0+1,\dots,\tau_1\}$ if, for some $(i,j) \in [m]^2$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{s,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$ then $Z_s^{(i,j)} = 0$ a.s. by induction on s. Suppose that for some $(i,j) \in [m]^2$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{1,\infty}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, i.e,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{0,n}^{(i)}(\boldsymbol{\theta}) K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}))}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0.$$

This implies that either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\lambda_{0,n}^{(i)}(\boldsymbol{\theta}) = 0\right) > 0$, which since $\lambda_{0,n}^{(i)}(\boldsymbol{\theta})$ is deterministic implies that $\lambda_{0,n}^{(i)}(\boldsymbol{\theta}) = 0$ which by assumption 4 implies that $\lambda_{0,n}^{(i)}(\boldsymbol{\theta}^*) = 0 \implies x_0^{(i)} = 0$ a.s.; or
- $K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}^{(i)})}^{(i,j)}(\boldsymbol{\theta}) = 0$, which implies $K_{1,\boldsymbol{\eta}(\mathbf{x}_0)}^{(i,j)}(\boldsymbol{\theta}^*) = 0$ by assumptions 2, 3, and 4.

Together this implies imply $Z_1^{(i,j)} = 0$ a.s.. Now let $s \in \{\tau_0+1,\dots,\tau_1\}$ and assume that if, for some $(i,j) \in [m]^2$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{s-1,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$ then $Z_{s-1}^{(i,j)} = 0$ a.s.. Now suppose for some $(i,j) \in [m]^2$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{s,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,n}^{(\cdot,i)}(\boldsymbol{\theta})\right) K_{1,\boldsymbol{\eta}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,n}^{(\cdot,i)}(\boldsymbol{\theta})\right)}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0,$$

This implies that either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{0,n}^{(\cdot,i)}(\boldsymbol{\theta}) = 0\right) > 0$, which by the induction hypothesis implies $\mathbf{1}_m^\top \mathbf{Z}_{s-1}^{(\cdot,i)} = 0$ a.s., which in turn implies $x_{s-1}^{(i)} = 0$ a.s.; or
- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(K_{1,\boldsymbol{\eta}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,n}^{(\cdot,i)}(\boldsymbol{\theta})\right)}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$ which by assumptions 2 and 3 and the induction hypothesis implies $K_{1,\boldsymbol{\eta}(\mathbf{x}_{s-1})}^{(i,j)}(\boldsymbol{\theta}^*) = 0$ a.s.,

which together imply $Z_s^{(i,j)} = 0$ a.s.. Now suppose for some $(i,j) \in [m]^2$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(M_{1,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, i.e.,

$$\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\sum_{s=\tau_0+1}^{\tau_1} \Lambda_{s,n}^{(i,j)}(\boldsymbol{\theta}) \odot Q_s^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0,$$

then for all $s = \tau_0+1,\dots,\tau_1$ either:

- $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{s,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, which implies $Z_s^{(i,j)} = 0$ hence $Y_s^{(i,j)} = 0$ a.s.; or
- $Q_s^{(i,j)}(\boldsymbol{\theta}) = 0$ which by assumption 2 implies $Q_s^{(i,j)}(\boldsymbol{\theta}^*) = 0$ hence $Y_s^{(i,j)} = 0$ a.s.,

and hence $\bar{Y}_1^{(i,j)} = \sum_{s=\tau_0+1}^{\tau_1} Y_s^{(i,j)} = 0$ a.s., this completes the proof of (A.35) and (A.36) for $r = 1$.

For the induction hypothesis, suppose that (A.35) and (A.36) hold for some $r \geq 1$. Notice that:

$$\bar{\Lambda}_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) = \left[1 - Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta})\right] \Lambda_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) + \frac{\bar{Y}_r^{(i,j)}}{M_{r,n}^{(i,j)}(\boldsymbol{\theta})} \left[\Lambda_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) \odot Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta})\right] = 0$$

is almost surely well defined by the induction hypothesis since we divide positive $\bar{Y}_r^{(i,j)}$ by 0 with probability 0. Now suppose, for some $(i,j) \in [m]$, that $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\bar{\Lambda}_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, then either:

- $Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta}) < 1$, which implies $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, so that the first term of the sum is 0 with positive probability, which then implies $Z_{\tau_r}^{(i,j)} = 0 \quad a.s.$ by the induction hypothesis; or
- $Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta}) = 1$, which implies $\mathbb{P}_n^{\boldsymbol{\theta}^*}\left(\Lambda_{\tau_r,n}^{(i,j)}(\boldsymbol{\theta}) = 0\right) > 0$, so that the second term in the sum is 0 with positive probability, which then implies $Z_{\tau_r}^{(i,j)} = 0$ a.s. by the induction hypothesis.

Using this and identical reasoning to that in the $r = 1$ case completes the induction. ■

If, for some $i,j \in [m]$, $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$ and $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) > 0$, then $\bar{\boldsymbol{\Lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ would involve division of a finite number by zero. The following lemma implies this situation does not arise.

Lemma 18. *Let assumptions 2 - 4 hold. Then for all $\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta$, $(i,j) \in [m]^2$ and $r = 1,\ldots,R$:*

$$M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \iff M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0.$$

Proof. It is enough to establish the implication in one direction for arbitrary $\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta$. We will prove by induction that for all $r = 1,\ldots,R$, $s = \tau_{r-1}+1,\ldots,\tau_r$ and $i,j \in [m]$,

$$\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0,$$
$$M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0.$$

Consider the case $r = 1$. We will first show that for all $s \in \{\tau_0+1,\ldots,\tau_1\}$, $\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$ by induction on s. To this end suppose that for some $(i,j) \in [m]^2$:

$$\Lambda_{1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \lambda_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}^{(i,j)}(\boldsymbol{\theta}) = 0.$$

Then either:

- $\lambda_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \lambda_{0,n}^{(i)}(\boldsymbol{\theta}) = 0$, which by assumption 4 implies $\lambda_{0,n}(\boldsymbol{\theta}') = \lambda_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$; or
- $K_{1,\boldsymbol{\eta}(\lambda_{0,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}^{(i,j)}(\boldsymbol{\theta}) = 0$, which by assumptions 2, 3, and 4 implies $K_{1,\boldsymbol{\eta}(\lambda_{0,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}'))}^{(i,j)}(\boldsymbol{\theta}') = 0$.

Hence:

$$\Lambda_{1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = \lambda_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') K_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}'))}^{(i,j)}(\boldsymbol{\theta}') = 0.$$

Now assume that for $s = tau_0+1,\ldots,\tau_1$ that $\Lambda_{s-1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{s-1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$, then if:

$$\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right) K_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}^{(i,j)}(\boldsymbol{\theta}) = 0,$$

we must have either:

- $\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right) = 0$, which by the induction hypothesis implies $\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')\right) = 0$; or
- $K^{(i,j)}_{1,\boldsymbol{\eta}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right)}(\boldsymbol{\theta}) = 0$, which by the above and assumptions 2 and 3 implies $K^{(i,j)}_{1,\boldsymbol{\eta}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')\right)}(\boldsymbol{\theta}') = 0$.

We therefore find:

$$\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = \left(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}'))}(\boldsymbol{\theta}') = 0.$$

completing the intermediary induction on s. Now consider:

$$M_{1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \sum_{s=1}^{\tau_1} \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot Q_s^{(i,j)}(\boldsymbol{\theta}) = 0,$$

then for all $s = \tau_0 + 1, \ldots, \tau_1$ either:

- $\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$; or
- $Q_s^{(i,j)}(\boldsymbol{\theta}) = 0 \implies Q_s^{(i,j)}(\boldsymbol{\theta}') = 0$,

and hence:

$$M_{1,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = \sum_{s=1}^{\tau_1} \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') \odot Q_s^{(i,j)}(\boldsymbol{\theta}') = 0,$$

completing the $r = 1$ case.

Now assume that for all $s \in \{\tau_{r-1} + 1, \ldots, \tau_r\}$ that $\Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{s,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$ and that $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$. Then we have that if:

$$\begin{aligned}\bar{\Lambda}_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = &\left[1 - Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta})\right] \odot \Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \\ &+ \frac{M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})} \odot \left[\Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta})\right] = 0,\end{aligned}$$

which is well defined by the inductive hypothesis, then either:

- $Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta}) < 1$, and then $\Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') \implies \bar{\Lambda}_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$, hence the first term is 0, or
- $Q_{\tau_r}^{(i,j)}(\boldsymbol{\theta}) = 1$, and then $\Lambda_{\tau_r}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') \implies \bar{\Lambda}_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$, so that the right hand term is 0.

This along with using the same reasoning used in the $r = 1$ case gives:

$$\Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0,$$

implies

$$\Lambda_{\tau_r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = \left(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,\infty}^{(\cdot,i)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}'))}(\boldsymbol{\theta}') = 0.$$

Using this and further using identical inductive reasoning to the $r = 1$ case we see that for all $s = \tau_r + 1, \ldots, \tau_{r+1}$, $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$ and further that $M^{(i,j)}_{r+1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies M^{(i,j)}_{r+1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$. This completes the inductive proof. ■

The following lemma is used in the proof of lemma 6.

Lemma 19. *Let assumptions 2 - 4 hold. For all $\boldsymbol{\theta} \in \Theta$, $n \in \mathbb{N}$, $(i,j) \in [m]^2$, and $r \in \{1, \ldots, R\}$:*

$$M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies M^{(i,j)}_{r,n}(\boldsymbol{\theta}) = 0, \quad \mathbb{P}^{\boldsymbol{\theta}^*}_n\text{-}a.s.$$

Proof. Fix arbitrary $\boldsymbol{\theta} \in \Theta$ and $n \in \mathbb{N}$. We will prove that for all $r = 1, \ldots, R$, $s = \tau_{r-1} + 1, \ldots, \tau_r$ and $i, j \in [m]$ the following two implications hold:

$$\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{s,n}(\boldsymbol{\theta}) = 0, \quad a.s.$$
$$M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies M^{(i,j)}_{r,n}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0, \quad a.s.$$

The induction is on r and s. Consider $r = 1$. We will first show that for all $s \in \tau_0 + 1, \ldots, \tau_1$ that $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = 0$ by induction on s. We have for $t - 1$ case:

$$\Lambda^{(i,j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \lambda^{(i)}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) K^{(i,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0,$$

which implies that either:

- $\lambda^{(i)}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \lambda^{(i)}_{0,\infty}(\boldsymbol{\theta}) = 0$, in which case $\lambda^{(i)}_{0,n}(\boldsymbol{\theta}) = 0$ or
- $K^{(i,j)}_{1,\boldsymbol{\eta}(\lambda_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$, in which case $K^{(i,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$,

so that:

$$\Lambda^{(i,j)}_{1,n}(\boldsymbol{\theta}^*, \boldsymbol{\theta}') = \lambda^{(i)}_{0,n}(\boldsymbol{\theta})) K^{(i,j)}_{1,\boldsymbol{\eta}(\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0.$$

Now assume that given $\Lambda^{(i,j)}_{s-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{s-1,n}(\boldsymbol{\theta}) = 0 \quad a.s.$, then:

$$\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \boldsymbol{\Lambda}^{(\cdot,i)}_{s-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0,$$

which in turn implies either:

- $\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}^{(\cdot,i)}_{s-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right) = 0_m$, which implies $\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}^{(\cdot,i)}_{s-1,n}(\boldsymbol{\theta})\right) = 0$ a.s.; or
- $K^{(i,j)}_{1,\boldsymbol{\eta}\left(\mathbf{1}_m^\top \boldsymbol{\Lambda}^{(\cdot,i)}_{s-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right)}(\boldsymbol{\theta}) = 0$, which implies $K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,n}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0$ a.s..

Together we find:

$$\Lambda^{(i,j)}_{s,n}(\boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \boldsymbol{\Lambda}^{(\cdot,i)}_{s-1,n}(\boldsymbol{\theta})\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{s-1,n}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0,$$

completing the intermediary induction on s. Now consider:

$$M^{(i,j)}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \sum_{s=1}^{\tau_1} \Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot Q^{(i,j)}_s(\boldsymbol{\theta}) = 0,$$

then for all $s = \tau_0 + 1, \ldots, \tau_1$ either

- $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$, which implies $\Lambda^{(i,j)}_{s,n}(\boldsymbol{\theta}) = 0$; or
- $Q^{(i,j)}_{s}(\boldsymbol{\theta}) = 0$,

and hence:

$$M^{(i,j)}_{1,n}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = \sum_{s=1}^{\tau_1} \Lambda^{(i,j)}_{s,n}(\boldsymbol{\theta}) \odot Q^{(i,j)}_{s}(\boldsymbol{\theta}) = 0.$$

This completes the case $r = 1$.

Now assume that for all $s = \tau_{r-1}+1,\ldots,\tau_r$, $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$, which implies $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$ and that $M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$, which implies $M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') = 0$. Then

$$\begin{aligned}\bar{\Lambda}^{(i,j)}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = &\left[1 - Q^{(i,j)}_{\tau_r}(\boldsymbol{\theta})\right] \odot \Lambda^{(i,j)}_{\tau_r}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \\ &+ \frac{M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{M^{(i,j)}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})} \odot \left[\Lambda^{(i,j)}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot Q^{(i,j)}_{\tau_r}(\boldsymbol{\theta})\right] = 0,\end{aligned}$$

and either:

- $Q^{(i,j)}_{\tau_r}(\boldsymbol{\theta}) < 1$, which implies $\Lambda^{(i,j)}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{\tau_r,n}(\boldsymbol{\theta}) \implies \bar{\Lambda}^{(i,j)}_{\tau_r,n}(\boldsymbol{\theta}) = 0$ a.s., so that the first term is 0; or
- $Q^{(i,j)}_{\tau_r}(\boldsymbol{\theta}) = 1$ which implies $\Lambda^{(i,j)}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{\tau_r,n}(\boldsymbol{\theta}) \implies \bar{\Lambda}^{(i,j)}_{\tau_r,n}(\boldsymbol{\theta}) = 0$, a.s., so that the right hand term is 0.

This along with using the same reasoning used in the $r = 1$ case tells us that given:

$$\Lambda^{(i,j)}_{\tau_r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}^{(\cdot,i)}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0,$$

which implies

$$\Lambda^{(i,j)}_{\tau_r,n}(\boldsymbol{\theta}) = \left(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}^{(\cdot,i)}_{\tau_r-1,n}(\boldsymbol{\theta})\right) K^{(i,j)}_{1,\boldsymbol{\eta}(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r-1,n}(\boldsymbol{\theta}))}(\boldsymbol{\theta}) = 0 \quad a.s.$$

Using this and further using identical inductive reasoning to the $r = 1$ case we see that for all $s \in \{\tau_r+1,\ldots,\tau_{r+1}\}$ we have $\Lambda^{(i,j)}_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0 \implies \Lambda^{(i,j)}_{s,n}(\boldsymbol{\theta}) = 0$ a.s. and further that $M^{(i,j)}_{r+1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$, which implies $M^{(i,j)}_{r+1,n}(\boldsymbol{\theta}) = 0$ a.s. This completes the inductive proof. ■

Lemma 20. *Let assumptions 2- 4 hold. For all $\boldsymbol{\theta}^* \in \Theta$ and $r \geq 1$, the function $\boldsymbol{\theta} \mapsto \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ is continuous on Θ.*

Proof. The arguments are very similar to those in the proof of 16, but making use of lemma 18, so we omit them. ■

Proof of Proposition 4 The proof is by induction on r. Consider $r = 1$. Note that $n^{-1}\bar{\boldsymbol{\lambda}}_{0,n}(\boldsymbol{\theta}) := n^{-1}\boldsymbol{\lambda}_{0,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ by assumption 4. Now let $t = 1, \dots, \tau_1 - 1$ and assume that $n^{-1}\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. Then:

$$\begin{aligned} n^{-1}\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) &= n^{-1}(\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,n}(\boldsymbol{\theta}))} \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \left(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \otimes \mathbf{1}_m\right) \odot \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))} \\ &= \Lambda_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}). \end{aligned}$$

By the CMT, a further application yields:

$$n^{-1}\bar{\boldsymbol{\lambda}}_{t,n}(\boldsymbol{\theta}) = n^{-1}(\mathbf{1}_m^\top \boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}))^\top \xrightarrow[a.s.]{\boldsymbol{\theta}^*} (\mathbf{1}_m^\top \Lambda_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))^\top = \bar{\boldsymbol{\lambda}}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}).$$

Then by induction on t we have that:

$$n^{-1}\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \Lambda_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}),$$

for all $t = 1, \dots, \tau_1$, this means that:

$$\begin{aligned} n^{-1}\mathbf{M}_{1,n}(\boldsymbol{\theta}) &= n^{-1} \sum_{s=1}^{\tau_1} \boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) \odot \mathbf{Q}_s(\boldsymbol{\theta}) \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \sum_{s=1}^{\tau_1} \boldsymbol{\Lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{Q}_s(\boldsymbol{\theta}) = \mathbf{M}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}). \end{aligned}$$

Now for general $r = 1, \dots, R$ assume that $\bar{\boldsymbol{\lambda}}_{\tau_{r-1},n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \bar{\boldsymbol{\lambda}}_{\tau_{r-1},\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. Using identical reasoning to the $r = 1$ case, we find that for all $t = \tau_{r-1} + 1, \dots, \tau_r$:

$$n^{-1}\boldsymbol{\Lambda}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \Lambda_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}),$$

which in turn implies by the CMT that:

$$n^{-1}\mathbf{M}_{r,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}),$$

Writing out the definition of $M_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$, proposition 2 gives $n^{-1}\bar{\mathbf{Y}}_r \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$. Then by the CMT,

$$\begin{aligned} n^{-1}\bar{\boldsymbol{\Lambda}}_{\tau_r,n}(\boldsymbol{\theta}) &= \left(\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}^*_{\tau_r}\right) \odot n^{-1}\boldsymbol{\Lambda}_{\tau_r,n}(\boldsymbol{\theta}) \\ &\quad + \left[n^{-1}\bar{\mathbf{Y}}_r \oslash n^{-1}\mathbf{M}_{r,n}(\boldsymbol{\theta})\right] \odot \left(\left[n^{-1}\boldsymbol{\Lambda}_{\tau_r,n}(\boldsymbol{\theta}) \odot \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right]\right) \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \left(\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right) \odot \Lambda_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\ &\quad + \left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \oslash \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right] \odot \left[\Lambda_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right] \\ &= \bar{\Lambda}_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}). \end{aligned}$$

We note here that the left hand side of the above display is almost surely well defined since, by lemma 17, for all $n \in \mathbb{N}$ and $i, j \in [m]$ if there is positive probability that $M^{(i,j)}_{r,n}(\boldsymbol{\theta}) = 0$ then

$\bar{Y}_r^{(i,j)} = 0$ $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-a.s., in which case we invoke the convention $\frac{0}{0} \coloneqq 0$. The right hand side of the limit is well defined since for all $i, j \in [m]$ we have $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0 \iff M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$ by lemma 18, in which case we again invoke the convection $\frac{0}{0} \coloneqq 0$. A further application of the CMT gives:

$$n^{-1}\bar{\boldsymbol{\lambda}}_{\tau_r,n}(\boldsymbol{\theta}) = n^{-1}\left(\mathbf{1}_m^\top \bar{\boldsymbol{\Lambda}}_{\tau_r}(\boldsymbol{\theta})\right)^\top \xrightarrow[a.s.]{\boldsymbol{\theta}^*} (\mathbf{1}_m^\top \bar{\Lambda}_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}))^\top = \bar{\boldsymbol{\lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}).$$

This completes the proof. ■

A.3 Contrast functions

Definition 1. *Let $(\mathscr{H}_n)_{n\geq 1}$ be a sequence of random functions $\mathscr{H}_n : \theta \in \Theta \mapsto \mathscr{H}_n(\theta) \in \mathbb{R}$ where Θ is a metric space. We say that $(\mathscr{H}_n)_{n\geq 1}$ are stochastically equicontinuous if there exists an event M of probability 1, such that for all $\varepsilon > 0$ and $\omega \in M$, there exists $N(\omega)$ and $\delta > 0$ such that $n > N(\omega)$ implies:*

$$\sup_{|\theta_1 - \theta_2| < \delta} |\mathscr{H}_n(\omega, \theta_1) - \mathscr{H}_n(\omega, \theta_2)| < \epsilon.$$

Lemma 21. *Assume Θ is a compact metric space and let $(\mathscr{H}_n)_{n\geq 1}$ be a sequence of random functions $\mathscr{H}_n : \theta \in \Theta \to \mathscr{H}_n(\theta) \in \mathbb{R}$. If there exists a continuous function $\mathscr{H}$ such that for all $\theta \in \Theta$ we have $|\mathscr{H}_n(\theta) - \mathscr{H}(\theta)| \xrightarrow{a.s.} 0$, and $(\mathscr{H}_n)_{n\geq 1}$ are stochastically equicontinuous, then:*

$$\sup_{\theta \in \Theta} |\mathscr{H}_n(\theta) - \mathscr{H}(\theta)| \xrightarrow{a.s.} 0.$$

That is $\mathscr{H}_n(\theta)$ converges to $\mathscr{H}(\theta)$ almost surely as $n \to \infty$, uniformly in θ.

Proof. See Andrews (1992). ■

Case (I)

We have:

$$\begin{aligned} n^{-1}\ell_n(\boldsymbol{\theta}) - n^{-1}\ell_n(\boldsymbol{\theta}^*) = \sum_{t=1}^{T} \Bigg\{ & \frac{\mathbf{y}_t^\top}{n} \log\left(\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) \oslash \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right) \\ & - n^{-1}\mathbf{1}_m^\top \left[\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right] \Bigg\}. \end{aligned} \tag{A.37}$$

The following proposition details the limit of (A.37).

Proposition 5. *Let assumptions 1-4 hold. Then:*

$$n^{-1}\ell_n(\boldsymbol{\theta}) - n^{-1}\ell_n(\boldsymbol{\theta}^*) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\sum_{t=1}^{T} \mathrm{KL}\left(\mathrm{Pois}\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] \| \mathrm{Pois}\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\right), \tag{A.38}$$

uniformly in $\boldsymbol{\theta}$.

Proof. For $n \in \mathbb{N}$, we define a random function $\mathscr{C}_n : \Theta \to \mathbb{R}$, as $\mathscr{C}_n(\boldsymbol{\theta}) \coloneqq \sum_{t=1}^{T} \mathscr{C}_{t,n}(\boldsymbol{\theta})$ where:

$$\begin{aligned}\mathscr{C}_{t,n}(\boldsymbol{\theta}) &\coloneqq \frac{\mathbf{y}_t^\top}{n} \log\left(\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) \oslash \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right) - n^{-1}\mathbf{1}_m^\top\left[\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right] \\ &= \sum_{i=1}^{m} \frac{y_t^{(i)}}{n} \log \frac{\mu_{t,n}^{(i)}(\boldsymbol{\theta})}{\mu_{t,n}^{(i)}(\boldsymbol{\theta}^*)} - n^{-1}\left[\mu_{t,n}^{(i)}(\boldsymbol{\theta}) - \mu_{t,n}^{(i)}(\boldsymbol{\theta}^*)\right],\end{aligned}$$

with the convention $0 \log 0 \coloneqq 0$. To see that $\mathscr{C}_{t,n}(\boldsymbol{\theta})$ is almost surely well defined, consider the following cases for each $i \in [m]$. If both $\mu_{t,n}^{(i)}(\boldsymbol{\theta}) > 0$ and $\mu_{t,n}^{(i)}(\boldsymbol{\theta}^*) > 0$, $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-a.s., then the log of the ratio of these terms is almost surely well defined. If $\mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0$ or $\mu_{t,n}^{(i)}(\boldsymbol{\theta}^*) = 0$ with positive probability, then $y_t^{(i)} = 0$ $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-a.s. by lemma 13 and we invoke the convention $0 \log 0 \coloneqq 0$.

We shall show that for $t = 1, \ldots, T$,

$$\mathscr{C}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\mathrm{KL}(\mathrm{Pois}[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] \| \mathrm{Pois}[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})]), \text{ uniformly in } \boldsymbol{\theta}.$$

The proof consists of showing pointwise convergence and then stochastic equicontinuity of $\mathscr{C}_{t,n}(\boldsymbol{\theta})$. Uniform almost sure convergence then follows by lemma 21.

Fix $t \in \{1, \ldots, T\}$ and note that $\frac{\mathbf{y}_t}{n} \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$ by proposition 1 (see remark 1), and by proposition 3, $n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \boldsymbol{\mu}_t(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. We claim that by the CMT:

$$\begin{aligned}\mathscr{C}_{t,n}(\boldsymbol{\theta}) &= \frac{\mathbf{y}_t^\top}{n} \log\left(\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) \oslash \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right) - n^{-1}\left[\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}^*)\right]^\top \mathbf{1}_m \\ &= \sum_{i=1}^{m} \frac{y_t^{(i)}}{n} \log \frac{\mu_{t,n}^{(i)}(\boldsymbol{\theta})}{\mu_{t,n}^{(i)}(\boldsymbol{\theta}^*)} - n^{-1}\left[\mu_{t,n}^{(i)}(\boldsymbol{\theta}) - \mu_{t,n}^{(i)}(\boldsymbol{\theta}^*)\right] \qquad \text{(A.39)} \\ &\xrightarrow[a.s.]{\boldsymbol{\theta}^*} \sum_{i=1}^{m} \mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \log \frac{\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)} - \left[\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] \qquad \text{(A.40)} \\ &= -\mathrm{KL}\left(\mathrm{Pois}\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] \| \mathrm{Pois}\left[\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\right).\end{aligned}$$

To see that the limit is well defined consider the cases for each $i \in [m]$, either:

- $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) > 0$ and $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) > 0$. In this case all functions in the sequence $\{\mathscr{C}_{t,n}(\boldsymbol{\theta})\}_{n \geq 1}$ and its limit are well defined; or
- $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) > 0$ and $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0$, or $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 0$ and $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) > 0$. This case is prohibited by lemma 14; or
- $\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0$. In this case, by lemmas 13 and 15, we have that for all $n \in \mathbb{N}$ $\mu_{t,n}^{(i)}(\boldsymbol{\theta}^*) = 0$, $\mu_{t,n}^{(i)}(\boldsymbol{\theta}) = 0$, and $y_t^{(i)} = 0 \quad \mathbb{P}_n^{\boldsymbol{\theta}^*}\, a.s.$, so that the ith term disappears from (A.39) and (A.40) with probability 1 by the convention $0 \log 0 \coloneqq 0$.

Hence we have shown the convergence of:

$$\mathscr{C}_{t,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\mathrm{KL}(\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) \| \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})),$$

point-wise in $\boldsymbol{\theta} \in \Theta$.

Next we show that $(\mathscr{C}_{t,n})_{n\geq 1}$ are stochastically equicontinuous. Let $\boldsymbol{f} \in \mathbb{R}^m$ and $E \subset \Omega$ such that $\mathbb{P}^{\boldsymbol{\theta}^*}(E) = 1$. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $\omega \in E$, and $\varepsilon > 0$. Firstly we will show the stochastic equicontinuity of $n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})^\top \boldsymbol{f}$ for any $\boldsymbol{f} \in \mathbb{R}^m$. Let $\varepsilon_0 > 0$ and write by the triangle inequality:

$$\begin{aligned}\left|n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_1)^\top \boldsymbol{f} - n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_2)^\top \boldsymbol{f}\right| \leq{}& \left|n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_1)^\top \boldsymbol{f} - \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_1)^\top \boldsymbol{f}\right| \\ &+ \left|n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_2)^\top \boldsymbol{f} - \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_2)^\top \boldsymbol{f}\right| \\ &+ \left|\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_1)^\top \boldsymbol{f} - \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_2)^\top \boldsymbol{f}\right|.\end{aligned}$$

There exists $N(\omega) < \infty$ such that for $n > N(\omega)$ the first two terms are bounded by $\varepsilon_0/3$ by proposition 3. Furthermore, since $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous by lemma 16 there exists a $\delta_0 > 0$ such that:

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta_0 \implies \left|\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_1)^\top \boldsymbol{f} - \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_2)^\top \boldsymbol{f}\right| < \varepsilon_0/3.$$

Hence we have shown stochastic equicontinuity of $(n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta})^\top \boldsymbol{f})_{n\geq 1}$. Now, consider $\mathscr{C}_{t,n}$:

$$\begin{aligned}\left|\mathscr{C}_{t,n}(\boldsymbol{\theta}_1) - \mathscr{C}_{t,n}(\boldsymbol{\theta}_2)\right| \leq{}& \left|\sum_{i=1}^m n^{-1} y_t^{(i)} \log \frac{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1)}{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)}\right| \\ &+ \left|n^{-1}\left[\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_1)^\top - \boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_2)^\top\right]\mathbf{1}_m\right|. \qquad \text{(A.41)}\end{aligned}$$

By what has been proven already we can choose δ_1 and $N_1(\omega)$ to bound (A.41) by $\varepsilon/2$. Let $\varepsilon_2 > 0$, by proposition 1 there exists $N_2(\omega)$ such that for $n > N_2(\omega)$:

$$\left|\sum_{i=1}^m n^{-1} y_t^{(i)} \log \frac{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1)}{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)}\right| < \sum_{i=1}^m |\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_2| \left|\log \frac{n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1)}{n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)}\right| \qquad \text{(A.42)}$$

Furthermore, for each $i \in [m]$ either:

- there is positive probability that either $\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1) = 0$ or $\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2) = 0$, then the ith term of the sum on the l.h.s. of (A.42) disappears since $y_t^{(i)} = 0$ with probability 1 by lemma 13, and we invoke the convention $0 \log 0 := 0$; or

- $\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1) > 0$ and $\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2) > 0$ almost surely. Then by continuity of log on $\mathbb{R}_{>0}$ there exists a $\delta_3^{(i)} > 0$ such that for $|n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1) - n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)| < \delta_3^{(i)}$:

$$\begin{aligned}\left|\log \frac{n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1)}{n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)}\right| &= \left|\log n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1) - \log n^{-1}\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)\right| \\ &< \frac{\varepsilon}{2m|\mu_{t,\infty}^{(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_2|}.\end{aligned}$$

By stochastic equicontinuity of $(n^{-1}\boldsymbol{\mu}_{t,n}^{\top}\boldsymbol{f})_{n\geq 1}$ there exists $N_3(\omega)$ and δ_2 such that for $n > N_3(\omega)$ and $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta_2$ we have that $\|n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_1) - n^{-1}\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_2)\|_\infty < \min_i \delta_3^{(i)}$ so that:

$$\begin{aligned}\left|\sum_{i=1}^{m} n^{-1} y_t^{(i)} \log \frac{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_1)}{\mu_{t,n}^{(i)}(\boldsymbol{\theta}_2)}\right| &< \sum_{i=1}^{m} |\mu_{t,\infty}^{*(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_2| \frac{\varepsilon}{2m|\mu_{t,\infty}^{*(i)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_2|} \\ &= \varepsilon/2.\end{aligned}$$

Hence choosing $\delta = \min(\delta_1, \delta_2)$ and $N(\omega) = \max(N_1(\omega), N_2(\omega), N_3(\omega))$ we have that for $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta$ and $n > N(\omega)$:

$$\left|\mathscr{C}_{t,n}(\boldsymbol{\theta}_1) - \mathscr{C}_{t,n}(\boldsymbol{\theta}_2)\right| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Hence we have established the stochastic equicontinuity of $\mathscr{C}_{t,n}$. This along with the already proven pointwise convergence establishes uniform almost sure convergence by lemma 21 and completes the proof. ■

Case (II)

We have:

$$\begin{aligned}n^{-1}\mathscr{L}_n(\boldsymbol{\theta}) - n^{-1}\mathscr{L}_n(\boldsymbol{\theta}^*) = \sum_{r=1}^{R} \Big\{ &\mathbf{1}_m^{\top}\left[n^{-1}\bar{\mathbf{Y}}_r \odot \log\left(\mathbf{M}_{r,n}(\boldsymbol{\theta}) \oslash \mathbf{M}_{r,n}(\boldsymbol{\theta}^*)\right)\right]\mathbf{1}_m \\ &+ n^{-1}\mathbf{1}_m^{\top}\left[\mathbf{M}_{r,n}(\boldsymbol{\theta}) - \mathbf{M}_{r,n}(\boldsymbol{\theta}^*)\right]\mathbf{1}_m \Big\}.\end{aligned}$$

Proposition 6. *Let assumptions 1-4 hold. Then*

$$n^{-1}\mathscr{L}_n(\boldsymbol{\theta}) - n^{-1}\mathscr{L}_n(\boldsymbol{\theta}^*) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\sum_{r=1}^{R} \mathrm{KL}\left(\mathrm{Pois}\left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] \| \mathrm{Pois}\left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\right), \tag{A.43}$$

uniformly in $\boldsymbol{\theta}$.

Proof. The details are similar to lemma 5. For $n \in \mathbb{N}$ define the sequence of random functions $(\mathscr{D}_n(\boldsymbol{\theta}))_{n\geq 1}$, $\mathscr{D}_n(\boldsymbol{\theta}) := \sum_{r=1}^{R} \mathscr{D}_{r,n}(\boldsymbol{\theta})$, where:

$$\begin{aligned}\mathscr{D}_{r,n}(\boldsymbol{\theta}) &:= \mathbf{1}_m^{\top}\left[n^{-1}\bar{\mathbf{Y}}_r \odot \log\left(\mathbf{M}_{r,n}(\boldsymbol{\theta}) \oslash \mathbf{M}_{r,n}(\boldsymbol{\theta}^*)\right)\right]\mathbf{1}_m \\ &\quad + n^{-1}\mathbf{1}_m^{\top}\left[\mathbf{M}_{r,n}(\boldsymbol{\theta}) - \mathbf{M}_{r,n}(\boldsymbol{\theta}^*)\right]\mathbf{1}_m \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \bar{Y}_r^{(i,j)} \log \frac{M_{r,n}^{(i,j)}(\boldsymbol{\theta})}{M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*)} + \left[M_{r,n}^{(i,j)}(\boldsymbol{\theta}) - M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*)\right]\end{aligned}$$

With the convention $0\log 0 := 0$. To see that this mapping is almost surely well defined, consider the following cases for each $(i,j) \in [m]^2$. If both $M_{r,n}^{(i,j)}(\boldsymbol{\theta}) > 0$, or $M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*) > 0$ $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-$a.s.$, then the log of each of these terms is almost surely well defined. If either $M_{r,n}^{(i,j)}(\boldsymbol{\theta}) = 0$, or $M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*) = 0$ with positive probability, then $\bar{Y}_r^{(i,j)} = 0$ $\mathbb{P}_n^{\boldsymbol{\theta}^*}$-$a.s.$ by lemma 17 and we invoke the convention $0\log 0 := 0$.

It is enough to show that for each $r \in \{1, \ldots, R\}$

$$\mathscr{D}_{r,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\mathrm{KL}\left(\mathrm{Pois}\left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\right] \| \mathrm{Pois}\left[\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right]\right), \text{ uniformly in } \boldsymbol{\theta}.$$

We show pointwise almost sure convergence and then stochastic equicontinuity. Fix $r \in \{1,\dots,R\}$ and note that by proposition 4 $n^{-1}\mathbf{M}_{r,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ and $r = 1,\dots,R$. Furthermore by proposition 2 $n^{-1}\bar{\mathbf{Y}}_r \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ for all $r = 1,\dots,R$. We claim that by the CMT:

$$\begin{aligned}
\mathscr{D}_{r,n}(\boldsymbol{\theta}) =& \mathbf{1}_m^\top \left[n^{-1}\bar{\mathbf{Y}}_r \odot \log\left(\mathbf{M}_{n,r}(\boldsymbol{\theta}) \oslash \mathbf{M}_{n,r}(\boldsymbol{\theta}^*)\right)\right]\mathbf{1}_m \\
&+ n^{-1}\mathbf{1}_m^\top \left[\mathbf{M}_{n,r}(\boldsymbol{\theta}) - \mathbf{M}_{n,r}(\boldsymbol{\theta}^*)\right]\mathbf{1}_m \\
=& \sum_{i=1}^m \sum_{j=1}^m n^{-1}\bar{Y}_r^{(i,j)} \log \frac{M_{r,n}^{(i,j)}(\boldsymbol{\theta})}{M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*)} + n^{-1}\left[M_{r,n}^{(i,j)}(\boldsymbol{\theta}) - M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*)\right] \qquad \text{(A.44)} \\
\xrightarrow[a.s.]{\boldsymbol{\theta}^*}& \sum_{i=1}^m \sum_{j=1}^m M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \log \frac{M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)} + \left[M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\right] \qquad \text{(A.45)} \\
=& -\mathrm{KL}(\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \| \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})).
\end{aligned}$$

To see that this limit is indeed almost surely well defined consider the cases for each $i = 1,\dots,m$ and
$j \in [m]$, either:

- $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) > 0$ and $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) > 0$. In this case all functions in the sequence and its limit are well defined. Or
- $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) > 0$ and $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = 0$, or $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = 0$ and $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) > 0$. This case is prohibited by lemma 14. Or
- $M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = 0$. In this case, by lemmas 13 and 15, we have $M_{r,n}^{(i,j)}(\boldsymbol{\theta}^*) = 0$ and $\bar{Y}_r^{(i,j)} = 0 \quad \mathbb{P}_n^{\boldsymbol{\theta}^*} a.s.$, so that the (i,j)th term disappears from the sums in (A.44) and (A.45) by the convention $0\log 0 \coloneqq 0$.

Hence we have shown:

$$\mathscr{D}_{r,n}(\boldsymbol{\theta}) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} -\mathrm{KL}\left(\mathrm{Pois}\left(\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\right) \| \mathrm{Pois}\left(\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right)\right), \text{ pointwise in } \boldsymbol{\theta}.$$

We now prove stochastic equicontinuity of $(\mathscr{D}_{r,n})_{n\geq 1}$. Firstly we will show the stochastic equicontinuity of $(n^{-1}\boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta})\boldsymbol{f}_2)_{n\geq 1}$ for any vectors $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^m$. Let $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^m$ and $E \subseteq \Omega$ such that $\mathbb{P}^{\theta^*}(E) = 1$. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $\omega \in E$, and $\varepsilon > 0$. Let $\varepsilon_0 > 0$ and write by the triangle inequality:

$$\begin{aligned}
\left|n^{-1}\boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta}_1)\boldsymbol{f}_2 - n^{-1}\boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta}_2)\boldsymbol{f}_2\right| \leq& \left|n^{-1}\boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta}_1)\boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}_1)\boldsymbol{f}_2\right| \\
&+ \left|n^{-1}\boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta}_2)\boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}_2)\boldsymbol{f}_2\right| \\
&+ \left|\boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}_1)\boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}_2)\boldsymbol{f}_2\right|.
\end{aligned}$$

There exists $N(\omega)$ such that for $n > N(\omega)$ the first two terms are bounded by $\varepsilon_0/3$ by proposition 3. Furthermore, since $\boldsymbol{\theta} \mapsto \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is continuous by lemma 20 there exists a δ_0 such that:

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta_0 \implies \left| \boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_1) \boldsymbol{f}_2 - \boldsymbol{f}_1^\top \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_2) \boldsymbol{f}_2 \right| < \varepsilon_0/3.$$

Hence we have shown stochastic equicontinuity of $(n^{-1} \boldsymbol{f}_1^\top \mathbf{M}_{r,n}(\boldsymbol{\theta}) \boldsymbol{f}_2)_{n \geq 1}$. Now, consider $\mathscr{D}_{r,n}$:

$$\begin{aligned} \left| \mathscr{D}_{r,n}(\boldsymbol{\theta}_1) - \mathscr{D}_{r,n}(\boldsymbol{\theta}_2) \right| \leq & \left| n^{-1} \mathbf{1}_m^\top \left[\bar{\mathbf{Y}}_r \odot \log \left(\mathbf{M}_{r,n}(\boldsymbol{\theta}_1) \oslash \mathbf{M}_{r,n}(\boldsymbol{\theta}_2) \right) \right] \mathbf{1}_m \right| \\ & + \left| n^{-1} \mathbf{1}_m^\top \left[\mathbf{M}_{r,n}(\boldsymbol{\theta}_1) - \mathbf{M}_{r,n}(\boldsymbol{\theta}_2) \right] \mathbf{1}_m \right| \end{aligned} \tag{A.46}$$

By what has already been proven, for any $\varepsilon > 0$ we can choose δ_1 and $N_1(\omega)$ to bound (A.46) by $\varepsilon/2$. Let $\varepsilon_1 > 0$, by proposition 2 there exists $N_2(\omega)$ such that for $n > N_2(\omega)$,

$$\left| \sum_{i,j=1}^m n^{-1} \bar{Y}_r^{(i,j)} \log \frac{M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1)}{M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2)} \right| < \sum_{i,j=1}^m |M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_1| \left| \log \frac{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1)}{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2)} \right|.$$

Furthermore, for each $(i,j) \in [m]^2$ either:

- $M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1) = 0$ or $M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2) = 0$ with positive probability. In this case the (i,j)th terms disappear from the sum on the left hand side since $\bar{Y}_r^{(i,j)} = 0$ with probability 1 by lemma 17; or
- $M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1) > 0$ and $M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2) > 0$ almost surely, by continuity of log on $\mathbb{R}_{>0}$ there exists a $\delta_3^{(i,j)} > 0$ such that if $|M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1) - M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2)| < \delta_3^{(i,j)}$ then:

$$\begin{aligned} \left| \log \frac{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1)}{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2)} \right| &= \left| \log n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1) - \log n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2) \right| \\ &\leq \frac{\varepsilon}{2m^2 |M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_1|}. \end{aligned}$$

Then by stochastic equicontinuity of $(n^{-1} \mathbf{M}_{r,n})_{n \geq 1}$ there exists $N_3(\omega)$ and δ_2 such that for $n > \max(N_2(\omega), N_3(\omega))$ and $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta_2$ we have that
$\|\mathbf{M}_{r,n}(\boldsymbol{\theta}_1) - \mathbf{M}_{r,n}(\boldsymbol{\theta}_2)\|_\infty < \min_{(i,j)} \delta_3^{(i,j)}$ so that:

$$\begin{aligned} \sum_{i,j=1}^m n^{-1} \bar{Y}_r^{(i,j)} \left| \log \frac{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_1)}{n^{-1} M_{r,n}^{(i,j)}(\boldsymbol{\theta}_2)} \right| &< \sum_{i,j=1}^m |M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_1| \frac{\varepsilon}{2m^2 |M_{r,\infty}^{(i,j)}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) + \varepsilon_1|} \\ &= \varepsilon/2. \end{aligned}$$

Choosing $\delta = \min(\delta_1, \delta_2)$ and $N(\omega) = \max(N_1(\omega), N_2(\omega), N_3(\omega))$ we have that for $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty < \delta$ and $n > N(\omega)$:

$$\left| \mathscr{D}_{r,n}(\boldsymbol{\theta}_1) - \mathscr{D}_{r,n}(\boldsymbol{\theta}_2) \right| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Hence we have established the stochastic equicontinuity of $(\mathscr{D}_{r,n})_{n \geq 1}$. This along with the already proven pointwise convergence establishes uniform almost sure convergence by lemma 21.

■

A.4 Convergence of Maximum PAL estimators

Proof of Theorem 1 Let $\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ be defined to be the r.h.s. of (A.38) and let $\mathscr{C}_n$ be as in the proof of proposition 5. We have that $\mathscr{C}_n(\hat{\boldsymbol{\theta}}_n) \geq \mathscr{C}_n(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta^*_{(I)}$. Furthermore $\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$. We can combine these inequalities to obtain:

$$\begin{aligned} 0 &\leq \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathscr{C}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}_n) \\ &\leq \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathscr{C}_n(\boldsymbol{\theta}^*) + \mathscr{C}_n(\boldsymbol{\theta}^*) - \mathscr{C}_n(\hat{\boldsymbol{\theta}}_n) + \mathscr{C}_n(\hat{\boldsymbol{\theta}}_n) - \mathscr{C}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}_n) \\ &\leq 2\sup_{\boldsymbol{\theta}\in\Theta} \left|\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - \mathscr{C}_n(\boldsymbol{\theta})\right| \xrightarrow[a.s.]{\boldsymbol{\theta}^*} 0. \end{aligned} \tag{A.47}$$

Hence $\mathscr{C}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}_n) \xrightarrow[a.s.]{\boldsymbol{\theta}^*} \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$.

Now assume for purposes of contradiction that there is some positive probability that $\hat{\boldsymbol{\theta}}_n$ does not converge to the set $\Theta^*_{(I)}$, i.e. assume that there is an event $E \subset \Omega$ with $\mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$ such that for all $\omega \in E$ there exists a $\delta > 0$ such that for infinitely many $n \in \mathbb{N}$ we have $\hat{\boldsymbol{\theta}}_n(\omega)$ is not in the open neighbourhood $B_\delta(\Theta^*) = \{\boldsymbol{\theta} \in \Theta : \exists \boldsymbol{\theta}' \in \Theta^* : \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta\}$. Since Θ is compact, the set $B_\delta(\Theta^*_{(I)})^c = \Theta \setminus B_\delta(\Theta^*_{(I)})$ is closed, bounded, and therefore compact. Furthermore, $\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. By the extreme value theorem this means that there exists a $\boldsymbol{\theta}' \in B_\delta(\Theta^*_{(I)})^c$ such that for all $\boldsymbol{\theta} \in B_\delta(\Theta^*_{(I)})^c$:

$$\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \leq \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}')$$

Furthermore, since $\boldsymbol{\theta}' \notin \Theta^*_{(I)}$ there exists $\varepsilon > 0$ such that:

$$\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') < \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \varepsilon.$$

By our assumption we have for each $\omega \in E$ there are infinitely many $n \in \mathbb{N}$ such that $\hat{\boldsymbol{\theta}}_n(\omega) \in B_\delta(\Theta^*_{(I)})^c$. But this implies that for each $\omega \in E$ there are infinitely many $n \in \mathbb{N}$ such that:

$$\mathscr{C}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}_n(\omega)) \leq \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}') < \mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \varepsilon,$$

$$\implies |\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathscr{C}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}_n(\omega))| > \varepsilon,$$

which contradicts (A.47). Hence we must have that $\hat{\boldsymbol{\theta}}_n$ converges to the set $\Theta^*_{(I)}$ $\mathbb{P}^{\boldsymbol{\theta}^*}$-a.s. The proof for case (II) follows the same arguments but with $\mathscr{C}_n$ and $\mathscr{C}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ replaced by $\mathscr{D}_n$ as in the proof of proposition 6 and $\mathscr{D}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ defined to be the r.h.s. of (A.43). ■

A.5 Identifiability

Proposition 7. *For any $\boldsymbol{\theta} \in \Theta$,*

$$\begin{aligned} \boldsymbol{\theta} \in \Theta^*_{(I)} &\iff \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*), \quad \forall t = 1,\dots,T \\ \boldsymbol{\theta} \in \Theta^*_{(II)} &\iff \mathbf{M}_{r,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*), \quad \forall r = 1,\dots,R. \end{aligned}$$

Proof. For the first equivalence in the statement, in order to prove the implication in the forward direction, assume that $\boldsymbol{\theta} \in \Theta^*_{(I)}$, i.e., $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$, for all $t = 1, \ldots, T$. Recall from the definitions in (A.25)-(A.26) that $\boldsymbol{\lambda}_{0,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}^*$, hence neither does $\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, and so:

$$\begin{aligned}
\boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)^\top &= \boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})^\top \\
&= (\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{q}_1(\boldsymbol{\theta}))^\top \mathbf{G}_1(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{1,\infty}(\boldsymbol{\theta})^\top \\
&= (\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \odot \mathbf{q}_1(\boldsymbol{\theta}))^\top \mathbf{G}_1(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{1,\infty}(\boldsymbol{\theta})^\top \\
&= \boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})^\top.
\end{aligned}$$

Now, for $t > 1$ assume that $\boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) = \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ and $\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) = \boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$. Then we have that:

$$\begin{aligned}
\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \Big[& \mathbf{1}_m - \mathbf{q}_{t-1}(\boldsymbol{\theta}) \\
& + \Big(\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})^\top \Big\{ [(\mathbf{1}_m \otimes \mathbf{q}_{t-1}(\boldsymbol{\theta})) \odot \mathbf{G}_{\mathbf{t-1}}(\boldsymbol{\theta})^\top] \\
& \oslash \left[\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \otimes \mathbf{1}_m \right] \Big\} \Big)^\top \Big] \odot \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
= & \left[1_m - \mathbf{q}_{t-1}(\boldsymbol{\theta}) + \mathbf{q}_{t-1}(\boldsymbol{\theta}) \right] \odot \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
= & \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
= & \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}),
\end{aligned}$$

so that

$$\begin{aligned}
\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})^\top &= (\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^* \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))} + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})^\top \\
&= (\boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))} + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})^\top \\
&= \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})^\top,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)^\top &= \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})^\top \\
&= (\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta}))^\top \mathbf{G}_t(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})^\top \\
&= (\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta}))^\top \mathbf{G}_t(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})^\top \\
&= \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})^\top.
\end{aligned}$$

By induction we have thus shown that $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $t = 1, \ldots, T$ and have completed the proof for the forward direction of the first implication in the statement.

For the backwards direction we need to show that $\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}) \implies \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$, for all $t = 1, \ldots, T$. Similarly as for the forwards direction:

$$\begin{aligned}\boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})^\top &= (\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \mathbf{q}_1(\boldsymbol{\theta}))^\top \mathbf{G}_1(\boldsymbol{\theta}) + \boldsymbol{\kappa}_1(\boldsymbol{\theta})^\top \\ &= (\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \odot \mathbf{q}_1(\boldsymbol{\theta}))^\top \mathbf{G}_1(\boldsymbol{\theta}) + \boldsymbol{\kappa}_1(\boldsymbol{\theta})^\top \\ &= \boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta})^\top \\ &= \boldsymbol{\mu}_{1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)^\top.\end{aligned}$$

Now, for $t > 1$ assume that $\boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$ and $\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) = \boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})$. Then we have that:

$$\begin{aligned}\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &= \Big[\mathbf{1}_m - \mathbf{q}_{t-1}(\boldsymbol{\theta}) \\ &\quad + \Big(\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})^\top \Big\{[(\mathbf{1}_m \otimes \mathbf{q}_{t-1}(\boldsymbol{\theta})) \odot \mathbf{G}_{\mathbf{t-1}}(\boldsymbol{\theta})^\top] \\ &\quad \oslash \left[\boldsymbol{\mu}_{t-1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \otimes \mathbf{1}_m\right]\Big\}\Big)^\top\Big] \odot \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \\ &= \left[1_m - \mathbf{q}_{t-1}(\boldsymbol{\theta}) + \mathbf{q}_{t-1}(\boldsymbol{\theta})\right] \odot \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \\ &= \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \\ &= \boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}),\end{aligned}$$

so that

$$\begin{aligned}\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})^\top &= (\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1,\infty}(\boldsymbol{\theta}^*\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))} + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})^\top \\ &= (\boldsymbol{\lambda}_{t-1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))^\top \mathbf{K}_{t,\boldsymbol{\eta}(\boldsymbol{\lambda}_{1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \odot \boldsymbol{\delta}_t(\boldsymbol{\theta}))} + \boldsymbol{\alpha}_{t,\infty}(\boldsymbol{\theta})^\top \\ &= \boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta},\boldsymbol{\theta})^\top,\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta})^\top &= (\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta}))^\top \mathbf{G}_t(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})^\top \\ &= (\boldsymbol{\lambda}_{t,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \odot \mathbf{q}_t(\boldsymbol{\theta}))^\top \mathbf{G}_t(\boldsymbol{\theta}) + \boldsymbol{\kappa}_{t,\infty}(\boldsymbol{\theta})^\top \\ &= \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta},\boldsymbol{\theta})^\top \\ &= \boldsymbol{\mu}_{t,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)^\top\end{aligned}$$

This completes the proof of the first implication in the statement of the proposition.

For the second implication, we will first show $\mathbf{M}_{r\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \mathbf{M}_{r\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \implies \mathbf{M}_{r\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = \mathbf{M}_{r\infty}(\boldsymbol{\theta},\boldsymbol{\theta})$, for all $r = 1,\ldots,R$. Recalling the definitions in (A.33)-(A.34), we have that for all $s \in \{1,\ldots,\tau_1\}$, $\boldsymbol{\Lambda}_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) = \boldsymbol{\Lambda}_{s,\infty}(\boldsymbol{\theta},\boldsymbol{\theta})$ and hence:

$$\begin{aligned}\mathbf{M}_{1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = \mathbf{M}_{1,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) &= \sum_{s=1}^{\tau_1} \Lambda_{s,\infty}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \odot \mathbf{Q}_s(\boldsymbol{\theta}) \\ &= \sum_{s=1}^{\tau_1} \Lambda_{s,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}) \odot \mathbf{Q}_s(\boldsymbol{\theta}) \\ &= \mathbf{M}_{1,\infty}(\boldsymbol{\theta},\boldsymbol{\theta}).\end{aligned} \tag{A.48}$$

Now let $r \geq 1$ and assume that, for all $s \in \{\tau_{r-1}+1, \ldots, \tau_r\}$, $\mathbf{\Lambda}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \mathbf{\Lambda}_{s,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})$. Then:

$$\begin{aligned}
\bar{\mathbf{\Lambda}}_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) &= \left[\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right] \odot \mathbf{\Lambda}_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
&\quad + \frac{\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\mathbf{M}_{r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta})} \left[\Lambda_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \odot \mathbf{Q}_{\tau_r}(\boldsymbol{\theta})\right] \\
&= \Lambda_{\tau_r,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
&= \Lambda_{\tau_r,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta}).
\end{aligned}$$

This then implies that for all $s \in \{\tau_r + 1, \ldots, \tau_{r+1}\}$, $\mathbf{\Lambda}_{s,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \mathbf{\Lambda}_{s,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})$, which in turn implies, as in (A.48) that $\mathbf{M}_{r+1,\infty}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \mathbf{M}_{r+1,\infty}(\boldsymbol{\theta}, \boldsymbol{\theta})$. The reverse direction follows by similar reasoning, as in mirroring the proof of the backwards direction of the first implication in the statement of the proposition, so the details are omitted. ■

APPENDIX B

SUPPORTING MATERIAL FOR CHAPTER 6

B.1 Supplementary materials for section 6.1

In this section we present the supporting traceplots, autocorrelation plots, and histograms for section 6.1.

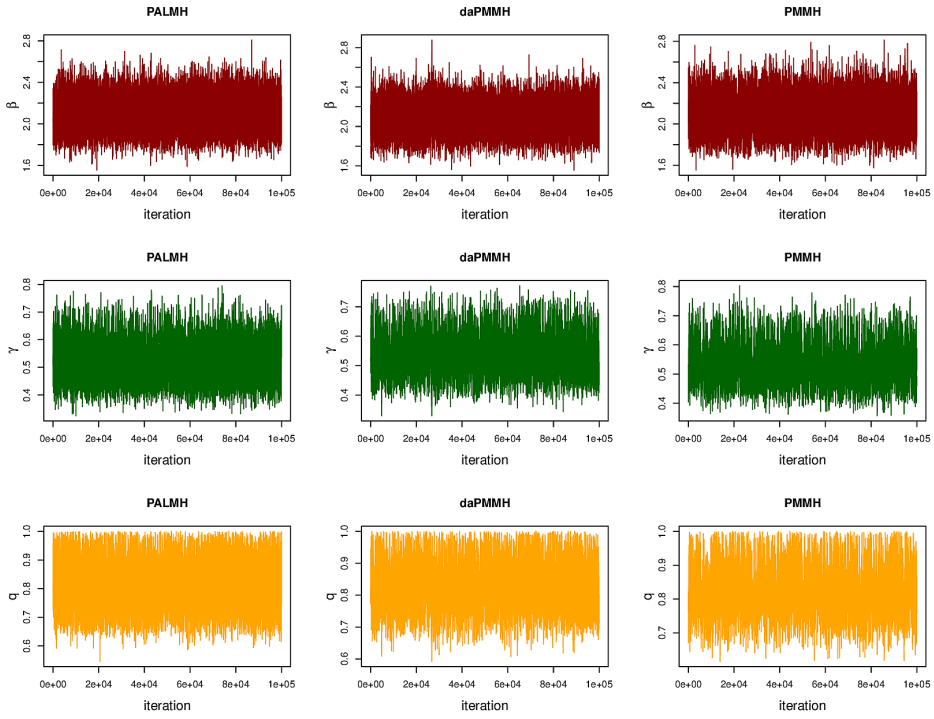

Figure B.1: Boarding school influenza example. Traceplots produced by the 3 procedures we have considered when run using synthetic data generated with parameters $\boldsymbol{\theta}^* = (\beta^*, \gamma^*, q^*) = (2, 0.5, 0.8)$. The plots display the first 10^5 iterations after the burn in period.

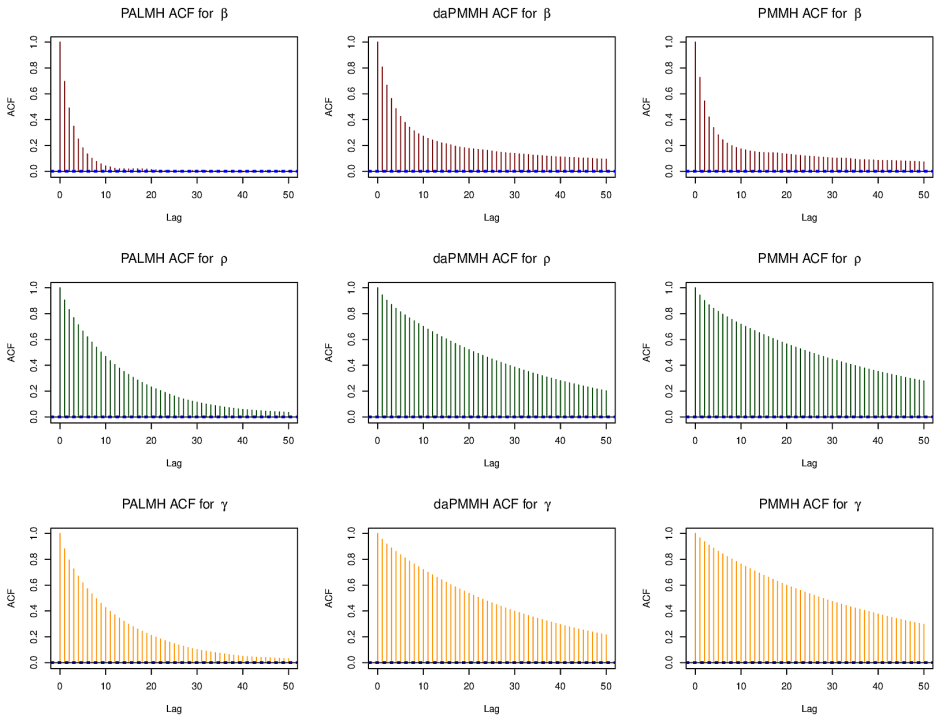

Figure B.2: Boarding school influenza example. ACF plots for each considered scheme when run using synthetic data generated with parameters $\boldsymbol{\theta}^* = (\beta^*, \gamma^*, q^*) = (2, 0.5, 0.8)$.

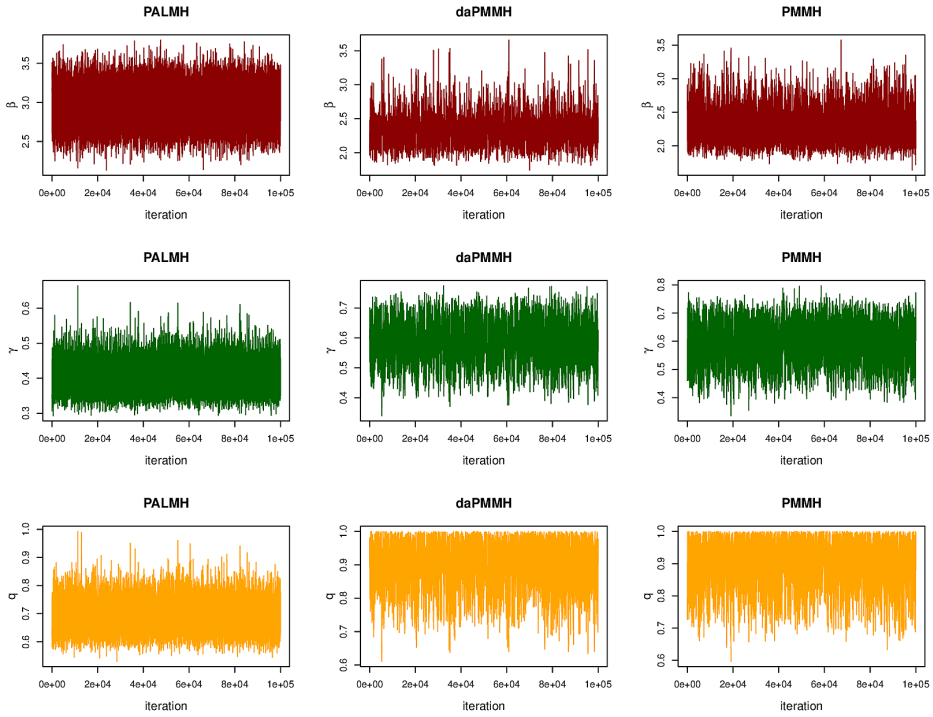

Figure B.3: Boarding school influenza example. Traceplots produced by the three considered schemes run using real data. The plots display the first 10^5 iterations after the burn in period.

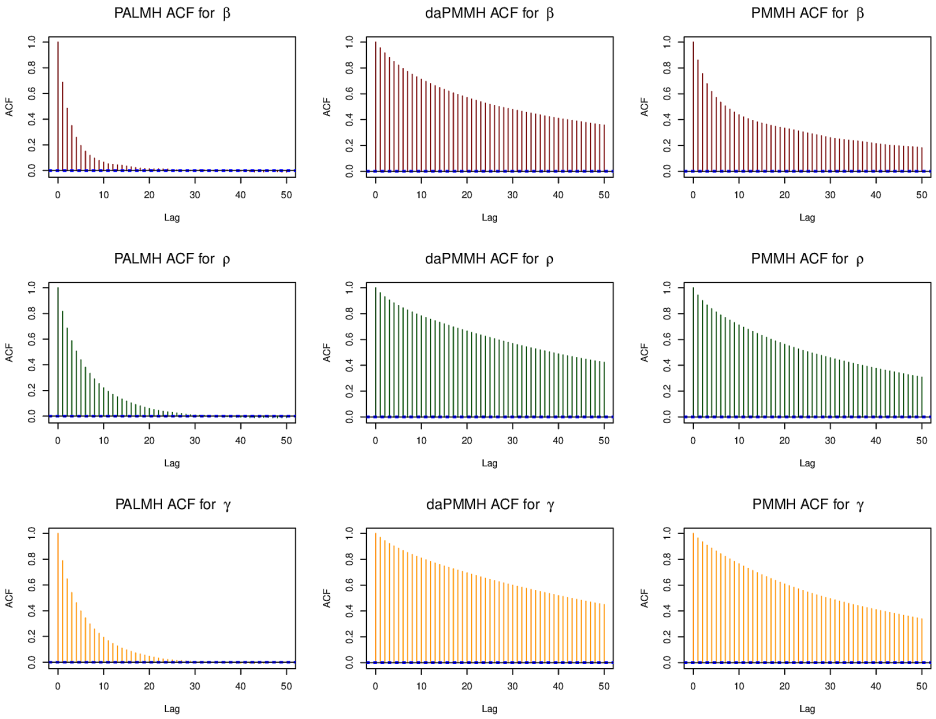

Figure B.4: Boarding school influenza example. ACF plots produced by the three schemes run using real data.

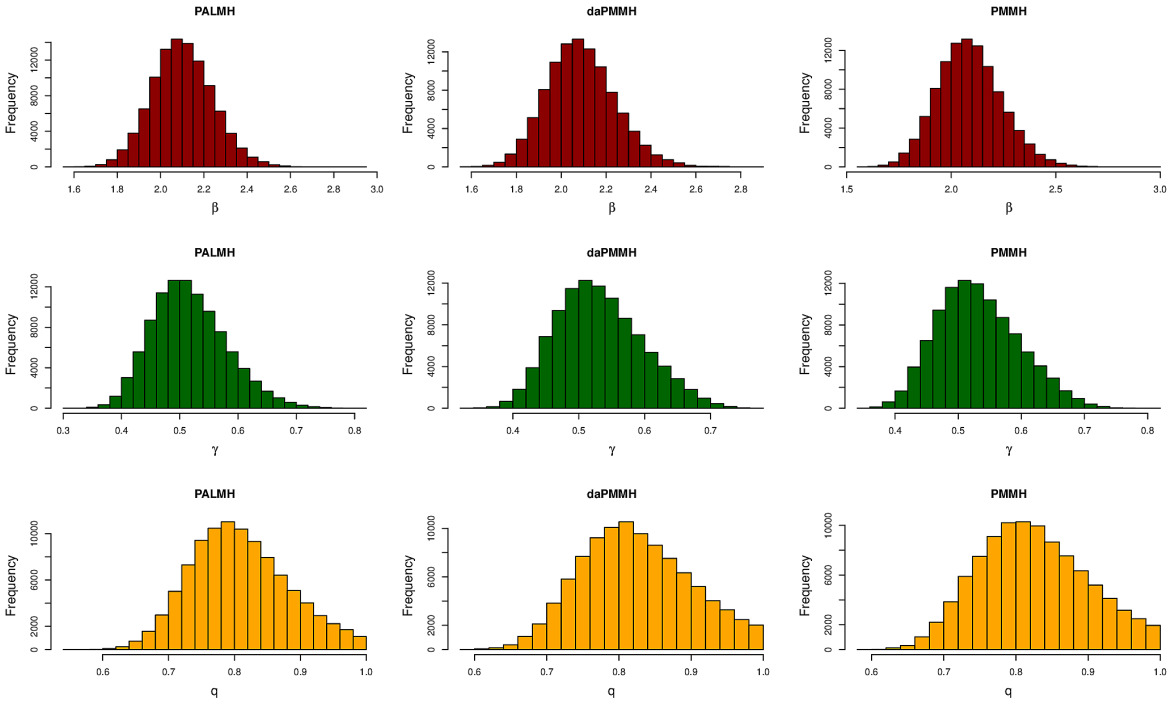

Figure B.5: Boarding school influenza example. Posterior marginals produced by the three algorithms when run using synthetic data generated with parameters $\boldsymbol{\theta}^* = [\beta^* \ \gamma^* \ q^*]^\top = [2 \, 0.5 \, 0.8]^\top$, the histograms are based on a thinned sample of 2.5×10^4.

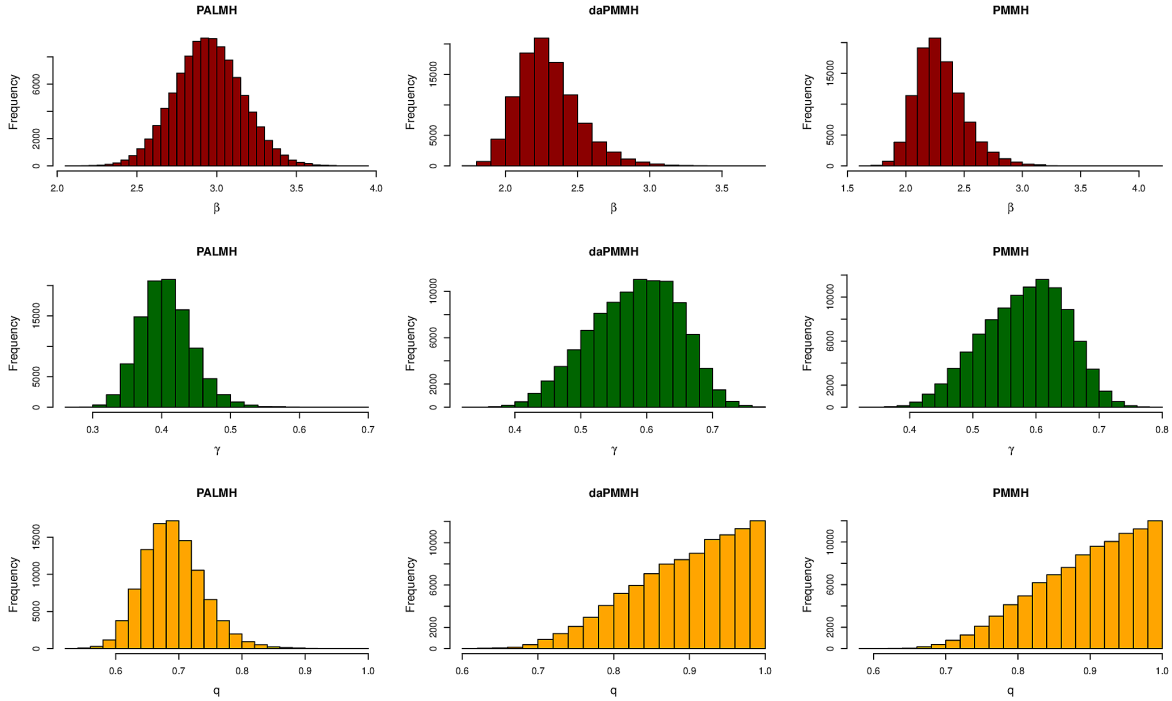

Figure B.6: Boarding school influenza example. Posterior samples produced by 3 considered schemes run using real data, the histograms are based on a thinned sample of 2.5×10^4.

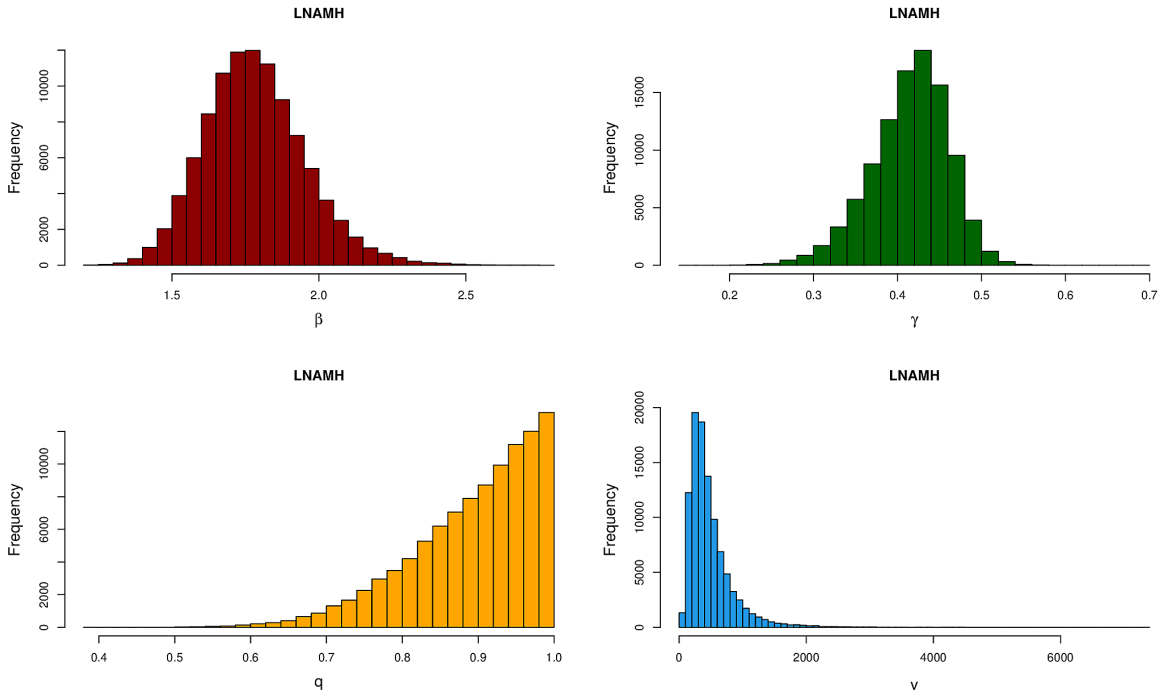

Figure B.7: Boarding school influenza example. Posterior samples produced by the LNA procedure, the histograms are based on a thinned sample of 2.5×10^4.

B.2 Supplementary material for section 6.2

This section presents supporting material for the age structured 'flu example.

To write the age structured model of section 6.2 as an instance of the Latent Compartmental Model we take $m = 16$ and identify vectors $\mathbf{x}_{k,t} := \left[S_{k,t}\, E_{k,t}\, I_{k,t}\, R_{k,t}\right]^\top$, $\mathbf{x}_t := \left[\mathbf{x}_{1,t}^\top \ldots \mathbf{x}_{4,t}^\top\right]^\top$ and matrices:

$$\mathbf{Z}_{k,t} := \begin{bmatrix} S_{k,t}-B_{k,t} & B_{k,t} & 0 & 0 \\ 0 & E_{k,t}-C_{k,t} & C_{k,t} & 0 \\ 0 & 0 & I_{k,t}-D_{k,t} & D_{k,t} \\ 0 & 0 & 0 & R_{k,t} \end{bmatrix}, \mathbf{Z}_t := \begin{bmatrix} \mathbf{Z}_{1,t} & & \cdots & 0 \\ & \mathbf{Z}_{2,t,} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{Z}_{4,t} \end{bmatrix}.$$

$$\mathbf{K}_{k,t,\boldsymbol{\eta}(\mathbf{x}_t)} := \begin{bmatrix} e^{-h\bar{\beta}_{k,t}} & 1-e^{-h\bar{\beta}_{k,t}} & 0 & 0 \\ 0 & e^{-h\rho} & 1-e^{-h\rho} & 0 \\ 0 & 0 & e^{-h\gamma} & 1-e^{-h\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{K}_{t,\boldsymbol{\eta}} := \begin{bmatrix} \mathbf{K}_{1,t,\boldsymbol{\eta}} & & \cdots & 0 \\ & \mathbf{K}_{2,t,\boldsymbol{\eta}} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{K}_{4,t,\boldsymbol{\eta}} \end{bmatrix},$$

where the $\bar{\beta}_{k,t}$ are the elements of the vector on the l.h.s. of (6.1). Due to the block-diagonal structure of the matrix $\mathbf{K}_{t,\boldsymbol{\eta}}$ for this example, algorithm 8 can be simplified to avoid performing various multiplications by zero. The resulting procedure is algorithm 13.

Parameter	ODE	PAL
q_1	0.93 (0.78, 0.99)	0.71 (0.53, 0.97)
q_2	0.96 (0.86, 0.99)	0.52 (0.49, 0.56)
q_3	0.28 (26, 0.30)	0.84 (0.61, 0.99)
q_4	0.28 (0.22, 0.34)	0.25 (0.19, 0.32)
β_{11}	4.34 (1.36, 8.83)	1.26 (0.44, 2.44)
β_{12}	2.91 (1.09, 5.45)	0.85 (0.56, 1.25)
β_{13}	3.51 (2.54, 4.59)	0.26 (0.09, 0.52)
β_{14}	1.33 (0.58, 2.29)	0.17 (0.05, 0.36)
β_{22}	2.55 (0.86, 5.11)	4.21 (3.98, 4.37)
β_{23}	6.89 (5.88, 8.18)	0.46 (0.35, 0.60)
β_{24}	0.72 (0.36, 1.12)	0.09 (0.04, 0.16)
β_{33}	18.08 (17.54, 18.50)	0.35 (0.15, 0.58)
β_{34}	0.14 (0.06, 0.25)	0.10 (0.01, 0.33)
β_{44}	21.34 (20.41, 22.26)	1.96 (1.59, 2.24)

Table B.1: Age-structured 'flu example. Posterior means and 95% credible intervals.

Algorithm 13 Filtering for the age-structured model

Initialise: $\bar{\boldsymbol{\lambda}}_0 \leftarrow \boldsymbol{\lambda}_0$.

$\bar{\boldsymbol{\lambda}}_0 \leftarrow \left[\bar{\boldsymbol{\lambda}}_{1,0}^\top \cdots \bar{\boldsymbol{\lambda}}_{4,0}^\top\right]^\top$
for $r \geq 1$:
  for $t = \tau_{r-1}+1, \ldots, \tau_r - 1$:
    for $k = 1, \ldots, 4$:
      $\Lambda_{k,t} \leftarrow (\bar{\boldsymbol{\lambda}}_{k,t-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{k,t,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1})}$
      $\bar{\boldsymbol{\lambda}}_{k,t} \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_{k,t})^\top$
    end for
    $\bar{\boldsymbol{\lambda}}_t \leftarrow \left[\bar{\boldsymbol{\lambda}}_{1,t}^\top \cdots \bar{\boldsymbol{\lambda}}_{4,t}^\top\right]^\top$
  end for
  for $k = 1, \ldots, 4$:
    $\Lambda_{k,\tau_r} \leftarrow (\bar{\boldsymbol{\lambda}}_{k,\tau_r-1} \otimes \mathbf{1}_m) \odot \mathbf{K}_{k,\tau_r,\boldsymbol{\eta}(\bar{\boldsymbol{\lambda}}_{t-1})}$
    $\mathbf{M}_{k,r} \leftarrow \sum_{t=\tau_{r-1}}^{\tau_r} \boldsymbol{\Lambda}_{k,t} \odot \mathbf{Q}_{k,t}$
    $\bar{\Lambda}_{k,\tau_r} \leftarrow \left(\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{k,\tau_r}\right) \odot \bar{\boldsymbol{\Lambda}}_{k,\tau_r} + \bar{\mathbf{Y}}_{k,r} \odot \boldsymbol{\Lambda}_{k,\tau_r} \odot \mathbf{Q}_{k,\tau_r} \oslash \mathbf{M}_{k,r}$
    $\bar{\boldsymbol{\lambda}}_{k,\tau_r} \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_{k,\tau_r})^\top$
  end for
  $\bar{\boldsymbol{\lambda}}_{\tau_r} \leftarrow \left[\bar{\boldsymbol{\lambda}}_{1,\tau_r}^\top \cdots \bar{\boldsymbol{\lambda}}_{4,\tau_r}^\top\right]^\top$
  $\mathscr{L}(\bar{\mathbf{Y}}_{1:4,r} | \bar{\mathbf{Y}}_{1:4,1:r-1}) \leftarrow \sum_{k=1}^4 -1_m^\top \mathbf{M}_{k,r} \mathbf{1}_m + \mathbf{1}_m^\top (\bar{\mathbf{Y}}_{k,r} \odot \mathbf{M}_{k,r}) \mathbf{1}_m - \mathbf{1}_m^\top \log(\bar{\mathbf{Y}}_{k,r}!) \mathbf{1}_m$
end for

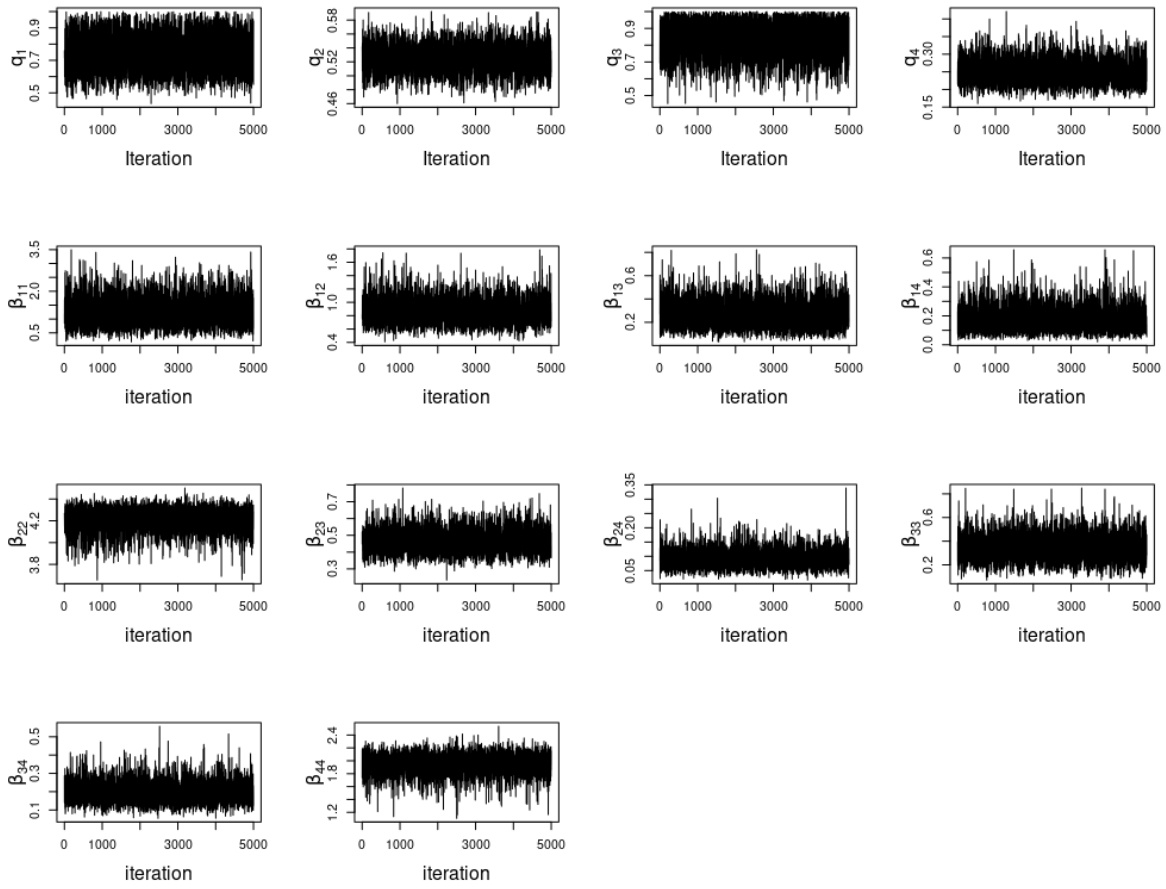

Figure B.8: Age-structured example. HMC posterior trace plots for the parameters of the stochastic model produced using Stan. The plots show the first 5^5 iterations after the burn in period.

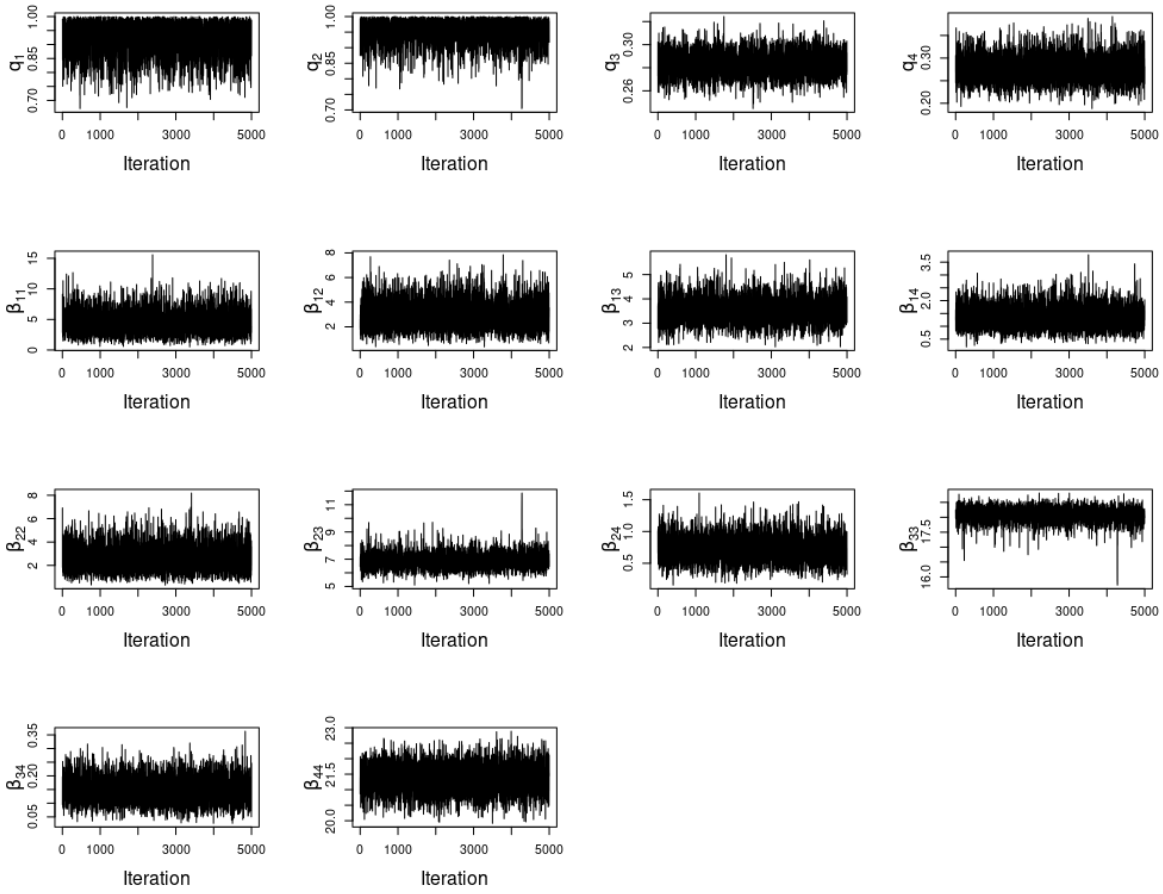

Figure B.9: Age-structured example. HMC posterior trace plots for the parameters of the ODE model produced using Stan. The plots show the first 5^5 iterations after the burn in period.

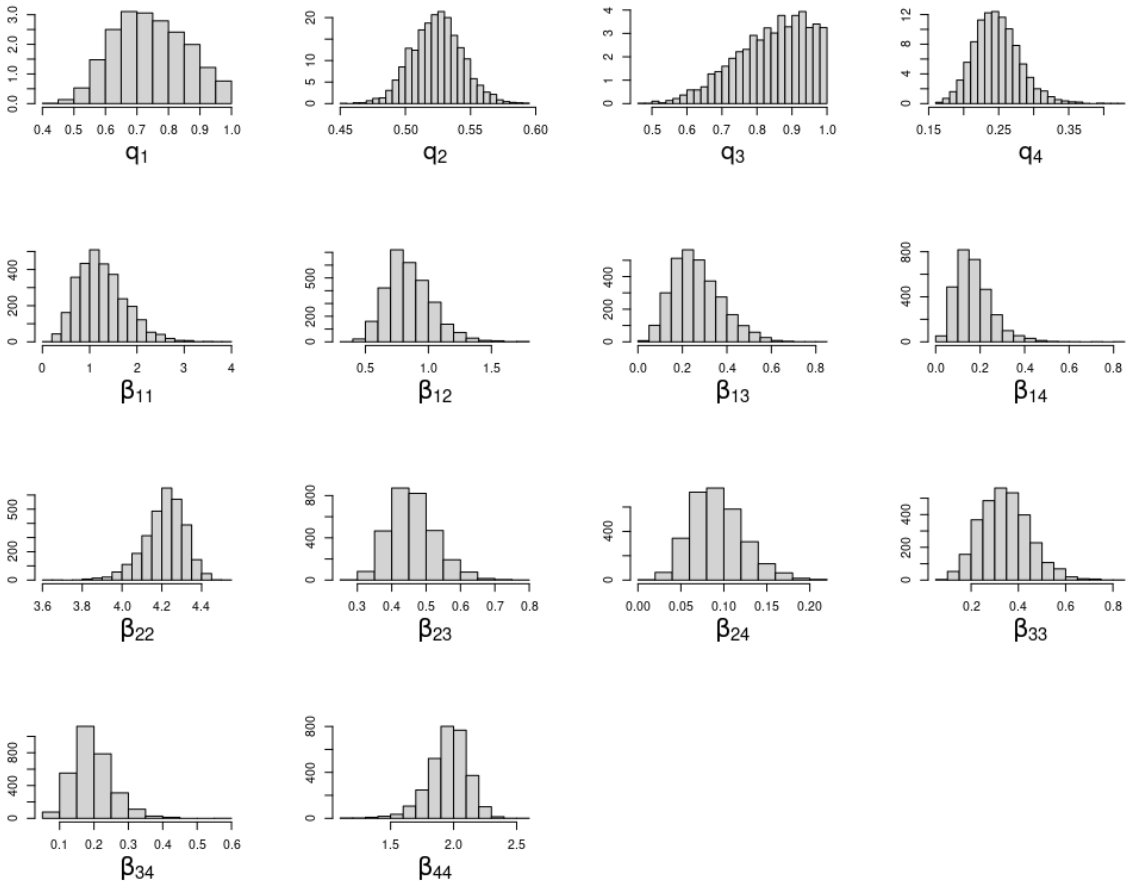

Figure B.10: Age-structured example. HMC posterior histograms for the parameters of the stochastic model produced using Stan.

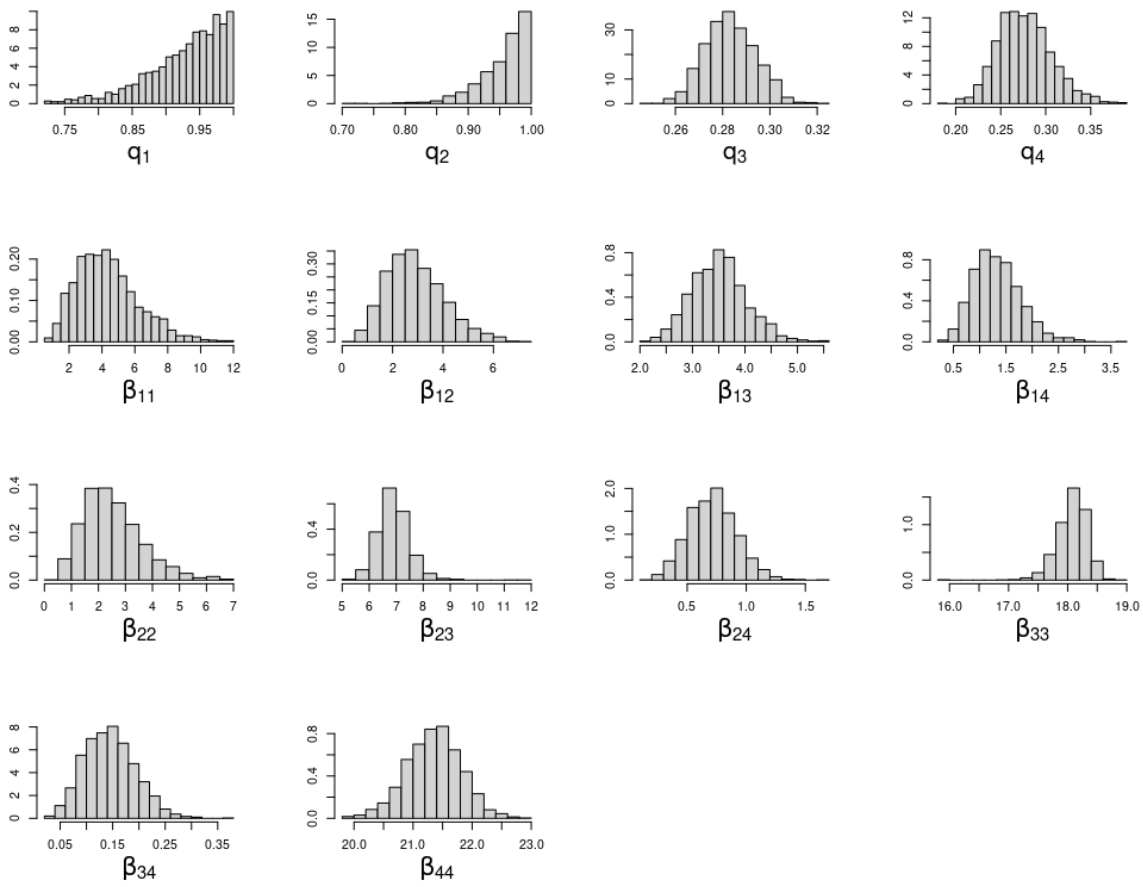

Figure B.11: Age-structured example. HMC posterior histograms for the parameters of the ODE model produced using Stan.

B.3 Supplementary material for section 6.3

This section presents supplementary materials for the rotavirus example, section 6.3.

Model

The evolution of the full age stratified rotavirus model at time t is given by:

$$
\begin{aligned}
S_{1,t+1} &= S_{1,t-1} + A_{1,t} + E_{1,t} - B_{1,t} - F^{(S)}_{1,t}, \\
I_{1,t+1} &= I_{1,t} + B_{1,t} - C_{1,t} - F^{(I)}_{1,t}, \\
R_{1,t+1} &= R_{1,t} + C_{1,t} - E_{1,t} - F^{(R)}_{1,t}, \\
S_{2,t+1} &= S_{2,t} + F^{(S)}_{1,t} + E_{2,t} - B_{2,t} - F^{(S)}_{2,t}, \\
I_{2,t+1} &= I_{2,t} + F^{(I)}_{1,t} + B_{2,t} - C_{2,t} - F^{(I)}_{2,t}, \\
R_{2,t+1} &= R_{2,t} + F^{(R)}_{1,t} + C_{2,t} - E_{2,t} - F^{(R)}_{2,t}, \\
S_{3,+1t} &= S_{3,t} + F^{(S)}_{2,t} + E_{3,t} - B_{3,t} - D^{(S)}_{t}, \\
I_{3,t+1} &= I_{3,t} + F^{(I)}_{2,t} + B_{3,t} - C_{3,t} - D^{(I)}_{t}, \\
R_{3,t+1} &= R_{3,t} + F^{(R)}_{2,t} + C_{3,t} - E_{3,t} - D^{(R)}_{t},
\end{aligned}
$$

where at time t: $A_{1,t} \sim \text{Pois}(\alpha_t)$, for some $\alpha_t \in \mathbb{R}$ represents new births, which is chosen according to historical birth record data; $B_{\cdot,t}$ represents new infectives; $C_{\cdot,t}$ represents recovering individuals; $D^{\cdot}_{t} \sim \text{Binom}(\cdot_{t-1}, 1-\delta)$ represents emigrating (dying) individuals; $E_{\cdot,t}$ represents individuals experiencing waning immunity; and $F_{\cdot,t}$ represents ageing individuals.

$$
\begin{bmatrix} B_{1,t} \\ F^{(S)}_{1,t} \\ S_{1,t} - B_{1,t} - F^{(S)}_{1,t} \end{bmatrix} \sim \text{Mult}\left(S_{1,t}, \begin{bmatrix} p_{1,t} \\ 1 - e^{(-hd_1)} \\ e^{(-hd_1)} - p_{k,t} \end{bmatrix} \right)
$$

$$
\begin{bmatrix} C_{1,t} \\ F^{(I)}_{1,t} \\ I_{1,t} - C_{1,t} - F^{(I)}_{1,t} \end{bmatrix} \sim \text{Mult}\left(I_{1,t}, \begin{bmatrix} 1 - e^{-h\gamma} \\ 1 - e^{-hd_1} \\ e^{-h\gamma} + e^{-hd_1} - 1 \end{bmatrix} \right)
$$

$$
\begin{bmatrix} E_{1,t} \\ F^{(R)}_{1,t} \\ R_{1,t} - E_{1,t} - F^{(R)}_{1,t} \end{bmatrix} \sim \text{Mult}\left(R_{1,t}, \begin{bmatrix} 1 - e^{-h\omega} \\ 1 - e^{-hd_1} \\ e^{-h\omega} + e^{-hd_1} - 1 \end{bmatrix} \right)
$$

$$
\begin{bmatrix} B_{2,t} \\ F^{(S)}_{2,t} \\ S_{2,t} - B_{2,t} - F^{(S)}_{2,t} \end{bmatrix} \sim \text{Mult}\left(S_{2,t}, \begin{bmatrix} p_{2,t} \\ 1 - e^{(-hd_2)} \\ e^{(-hd_2)} - p_{2,t} \end{bmatrix} \right)
$$

$$
\begin{bmatrix} C_{2,t} \\ F^{(I)}_{2,t} \\ I_{2,t} - C_{2,t} - F^{(I)}_{2,t} \end{bmatrix} \sim \text{Mult}\left(I_{2,t}, \begin{bmatrix} 1 - e^{-h\gamma} \\ 1 - e^{-hd_2} \\ e^{-h\gamma} + e^{-hd_2} - 1 \end{bmatrix} \right)
$$

$$
\begin{bmatrix} E_{2,t} \\ F^{(R)}_{2,t} \\ R_{2,t} - E_{2,t} - F^{(R)}_{2,t} \end{bmatrix} \sim \text{Mult}\left(R_{2,t}, \begin{bmatrix} 1 - e^{-h\omega} \\ 1 - e^{-hd_2} \\ e^{-h\omega} + e^{-hd_2} - 1 \end{bmatrix} \right)
$$

$$
\begin{bmatrix} B_{3,t} \\ S_{3,t} - D^{(S)}_t - B_{3,t} \end{bmatrix} \sim \text{Mult}\left(S_{3,t} - D^{(S)}_t, \begin{bmatrix} p_{3,t} \\ 1 - p_{3,t} \end{bmatrix} \right)
$$

$$
\begin{bmatrix} C_{3,t} \\ I_{3,t} - D^{(I)}_t - C_{2,t} \end{bmatrix} \sim \text{Mult}\left(I_{3,t} - D^{(I)}_t, \begin{bmatrix} 1 - e^{-h\gamma} \\ e^{-h\gamma} \end{bmatrix} \right)
$$

$$
\begin{bmatrix} E_{3,t} \\ R_{3,t} - D^{(R)}_t - E_{2,t} \end{bmatrix} \sim \text{Mult}\left(R_{3,t} - D^{(R)}_t, \begin{bmatrix} 1 - e^{-h\omega} \\ e^{-h\omega} \end{bmatrix} \right)
$$

To align notation with the model descriptions in section 3.1 collect observations at time r in the matrix $\bar{\mathbf{Y}}_r \in \mathbb{N}^{9\times 9}$ which has elements equal to zero except $\bar{Y}_r^{(3k-2,3k-1)} = Y_{r,k}$ for age groups $k = 1,2,3$, similarly collect reporting rates in $\mathbf{Q}_r \in \mathbb{N}^{9\times 9}$ which has elements equal to zero except $Q_r^{(3k-2,3k-1)} = q_{r,k}$ for $k = 1,2,3$. Define $\mathbf{x}_t = [S_{1,t}\, I_{1,t}\, R_{1,t}\, S_{2,t}\, I_{2,t}\, R_{2,t}\, S_{3,t}\, I_{3,t}\, R_{3,t}]$. Identify the matrix:

$$
\begin{aligned}
\mathbf{K}_{t,\boldsymbol{\eta}}^{(1,\cdot)} &= \begin{bmatrix} e^{-hd_1}-p_{1,t} & p_{1,t} & 0 & 1-e^{-hd_1} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(2,\cdot)} &= \begin{bmatrix} 0 & e^{-h\gamma}+e^{-hd_1}-1 & 1-e^{-h\gamma} & 0 & 1-e^{-hd_1} & 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(3,\cdot)} &= \begin{bmatrix} 1-e^{-h\omega} & 0 & 0 & e^{-h\omega}+e^{-hd_1}-1 & 0 & 1-e^{-hd_1} & 0 & 0 & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(4,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & e^{-hd_2}-p_{2,t} & p_{2,t} & 0 & 1-e^{-hd_2} & 0 & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(5,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & 0 & e^{-h\gamma}+e^{-hd_2}-1 & 1-e^{-h\gamma} & 0 & 1-e^{-hd_2} & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(6,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & 1-e^{-h\omega} & 0 & e^{-h\omega}+e^{-hd_1}-1 & 0 & 0 & 1-e^{-hd_1} \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(7,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{3,t} & p_{3,t} & 0 \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(8,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & e^{-h\gamma} & 1-e^{-h\gamma} \end{bmatrix}, \\
\mathbf{K}_{t,\boldsymbol{\eta}}^{(9,\cdot)} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1-e^{-h\omega} & 0 & e^{-h\omega} \end{bmatrix}.
\end{aligned}
$$

Where for models EqEq and EqOv we have $p_{k,t} = 1-\exp\left\{-\boldsymbol{\beta}_k^\top \frac{\boldsymbol{I}_t}{n}\chi_t\right\}$ for $k=1,2,3$, and for model OvOv we have $p_{k,t} = 1-\exp\left\{-\boldsymbol{\beta}_k^\top \frac{\boldsymbol{I}_t}{n}\chi_t\xi_r\right\}$ for $k=1,2,3$, in which case we will write $\mathbf{K}_{t,\boldsymbol{\eta}} = \mathbf{K}_{t,\boldsymbol{\eta},\xi}$.

For models EqOv and OvOv we have for $k=1,2,3$:

$$Q_r^{(3k-2,3k-1)} \sim \mathcal{N}(\mu_q,\sigma_q^2)_{\geq 0,\leq 1}$$

corresponding to the reporting rate of new infective individuals for each age group. Denote this prior density of $\mathbf{Q}_r$ as $f(\cdot \mid \mu_q,\sigma_q^2)$.

Inference

We assume that the values of $\alpha, d_1, d_2, \delta, \gamma, \omega$ and μ_q are known, we set them to the same values as assumed in Stocks et al. (2020), these are available on the GitHub page. All other parameters are to be estimated.

Laplace approximation proposals for the rotavirus example

Consider algorithm 14. We factorise the proposal of particles at time r, $[\xi_r^{(i)}, \mathbf{Q}_r^{(i)}]$, into sampling $\xi_r^{(i)}$ from its prior, then given this we seek a Laplace/PAL approximation to the distribution:

$$\hat{p}(\mathbf{Q}_r \mid \bar{\mathbf{Y}}_{1:r}, \mathbf{Q}_{1:r-1}, \boldsymbol{\xi}_{1:r}) := \frac{\exp \mathscr{L}(\bar{\mathbf{Y}}_r \mid \bar{\mathbf{Y}}_{1:r-1}, \mathbf{Q}_{1:r}, \boldsymbol{\xi}_{1:r}) f(\mathbf{Q}_r \mid \mu_q, \sigma_q^2)}{\int \exp \mathscr{L}(\bar{\mathbf{Y}}_r \mid \bar{\mathbf{Y}}_{1:r-1}, \mathbf{Q}_{1:r}, \boldsymbol{\xi}_{1:r}) f(\mathbf{Q}_r \mid \mu_q, \sigma_q^2) d\mathbf{Q}_r}.$$

Suppressing dependence on the particle, let $\mathbf{L}_r = \sum_{t=\tau_{r-1}+1}^{\tau_r} \boldsymbol{\Lambda}_t$ with $\boldsymbol{\Lambda}_t$ calculated as per line 5 of

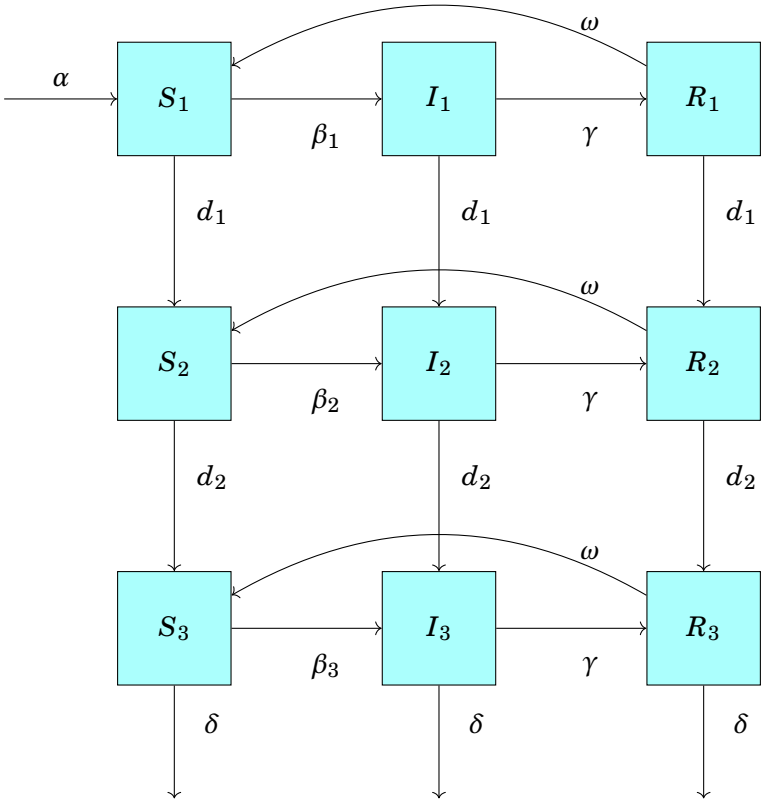

Figure B.12: Schema for the latent compartmental model of rotavirus transmission.

algorithm 14, we have for some constants C_1, C_2:

$$\begin{aligned}\log \hat{p}(\mathbf{Q}_r \mid \bar{\mathbf{Y}}_{1:r}, \mathbf{Q}_{1:r-1}, \boldsymbol{\xi}_{1:r}) &= \mathscr{L}(\bar{\mathbf{Y}}_r \mid \bar{\mathbf{Y}}_{1:r-1}, \mathbf{Q}_{1:r}, \boldsymbol{\xi}_{1:r}) + \log f(\mathbf{Q}_r \mid \mu_q, \sigma_q^2) + C_1 \\ &= \sum_{j=1}^{3} \Big\{ \bar{Y}_r^{(3j-2,3j-1)} \log(Q_r^{(3j-2,3j-1)} L_r^{(3j-2,3j-1)}) \\ &\quad - L_r^{(3j-2,3j-1)} Q_r^{(3j-2,3j-1)} - \bar{Y}_r^{(3j-2,3j-1)}! \\ &\quad - \frac{1}{2}\left(\frac{Q_r^{(3j-2,3j-1)} - \mu_q}{\sigma_q} \right)^2 \Big\} + C_2\end{aligned}$$

To get the mean of a Laplace approximation to the above we must find it's maximum w.r.t. $\mathbf{Q}_r$, hence for $j = 1, 2, 3$:

$$\begin{aligned}\frac{d \log \hat{p}(\mathbf{Q}_r \mid \mathbf{y}_r)}{d Q_r^{(3j-2,3j-1)}} &= \frac{\bar{Y}_r^{(3j-2,3j-1)}}{Q_r^{(3j-2,3j-1)}} - L_r^{(3j-2,3j-1)} - \frac{Q_r^{(3j-2,3j-1)} - \mu_q}{\sigma_q^2} = 0 \\ &\iff (Q_r^{(3j-2,3j-1)})^2 + (L_r^{(3j-2,3j-1)} \sigma_q^2 - \mu_q) Q_r^{(3j-2,3j-1)} - \bar{Y}_r^{(3j-2,3j-1)} \sigma_q^2 = 0 \\ &\implies Q_r^{(3j-2,3j-1)} = \frac{1}{2}\left(\mu_q - L_r^{(3j-2,3j-1)} \sigma_q^2 + \sqrt{(L_r^{(3j-2,3j-1)} \sigma_q^2 - \mu_q)^2 + 4 \bar{Y}_r^{(3j-2,3j-1)} \sigma_q^2} \right) \\ &=: \mu_r^{(j)}.\end{aligned}$$

For the variance we find the second derivative and evaluate it at $\mu_r^{(j)}$:

$$\frac{d^2 \log \hat{p}(\mathbf{q}_r \mid \mathbf{y}_r)}{d(Q_r^{(3j-2,3j-1)})^2} = -\frac{\bar{Y}_r^{(3j-2,3j-1)}}{(Q_r^{(3j-2,3j-1)})^2} - \frac{1}{\sigma_q^2}$$
$$\implies \left(\sigma_r^{(j)}\right)^2 = \left(\frac{\bar{Y}_r^{(3j-2,3j-1)}}{\left(\mu_r^{(j)}\right)^2} + \frac{1}{\sigma_q^2}\right)^{-1}.$$

Hence, having proposed ξ_r from its prior, we propose $\mathbf{Q}_r$ by setting all elements to be zero except:

$$Q_r^{(3j-2,3j-1)} \sim \mathcal{N}\left(\mu_r^{(j)}, (\sigma_r^{(j)})^2\right)_{\geq 0, \leq 1} \qquad \text{for } j = 1,2,3. \tag{B.1}$$

Let $\pi(\cdot \mid \bar{\mathbf{Y}}_{1:r}, \mathbf{Q}_{1:r-1}, \xi_{1:r})$ be the proposal density associated with (B.1). The resulting approximate

Algorithm 14 PAL within SMC for model of Rotavirus

initialise: $\bar{\boldsymbol{\lambda}}_0^{(i)} \leftarrow \boldsymbol{\lambda}_0$ for $i = 1$ to n_{part}.
1: **for** $r \geq 1$:
2: **for** $i = 1$ to n_{part}
3: $\xi_r^{(i)} \sim \text{Gamma}(\sigma_\xi, \sigma_\xi)$
4: **for** $t = \tau_{r-1}+1, \ldots, \tau_r - 1$:
5: $\boldsymbol{\Lambda}_t^{(i)} \leftarrow ((\bar{\boldsymbol{\lambda}}_{t-1}^{(i)} \odot \boldsymbol{\delta}_t) \otimes \mathbf{1}_m) \odot \mathbf{K}_{t, \boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{t-1}^{(i)} \odot \boldsymbol{\delta}_t\right), \xi_r^{(i)}} + \boldsymbol{\alpha}_t$
6: $\bar{\boldsymbol{\lambda}}_t^{(i)} \leftarrow (\mathbf{1}_m^\top \boldsymbol{\Lambda}_t^{(i)})^\top$
7: **end for**
8: $\boldsymbol{\Lambda}_{\tau_r}^{(i)} \leftarrow ((\boldsymbol{\lambda}_{\tau_r - 1}^{(i)} \odot \boldsymbol{\delta}_{\tau_r}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{\tau_r, \boldsymbol{\eta}(\boldsymbol{\lambda}_{\tau_r-1}^{(i)} \odot \boldsymbol{\delta}_{\tau_r}), \xi_r^{(i)}} + \boldsymbol{\alpha}_{\tau_r}$
9: $\mathbf{Q}_r^{(i)} \sim \pi(\cdot \mid \bar{\mathbf{Y}}_{1:r}, \mathbf{Q}_{1:r-1}^{(i)}, \xi_{1:r}^{(i)})$ calculated according to (B.1).
10: $\mathbf{M}_r^{(i)} \leftarrow \sum_{t=\tau_{r-1}+1}^{\tau_r} \boldsymbol{\Lambda}_t^{(i)} \odot \mathbf{Q}_r^{(i)}$
11: $\mathscr{L}(\bar{\mathbf{Y}}_r \mid \bar{\mathbf{Y}}_{1:r-1}, \mathbf{Q}_{1:r}^{(i)}, \xi_{1:r}^{(i)}) \leftarrow \mathbf{1}_m^\top \mathbf{M}_r \mathbf{1}_m + \mathbf{1}_m^\top (\bar{\mathbf{Y}}_r \odot \log \mathbf{M}_r) \mathbf{1}_m - \mathbf{1}_m^\top \log(\bar{\mathbf{Y}}_r!) \mathbf{1}_m$
12: $\log w_r^{(i)} \leftarrow \mathscr{L}(\bar{\mathbf{Y}}_r \mid \bar{\mathbf{Y}}_{1:r-1}, \mathbf{Q}_{1:r}^{(i)}, \xi_{1:r}^{(i)}) + f(\mathbf{Q}_r^{(i)} \mid \mu_q, \sigma_q) - \pi(\mathbf{Q}_r^{(i)} \mid \bar{\mathbf{Y}}_{1:r}, \mathbf{Q}_{1:r-1}^{(i)}, \xi_{1:r}^{(i)})$
13: $\bar{\boldsymbol{\Lambda}}_{\tau_r}^{(i)} \leftarrow (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_r^{(i)}) \odot \boldsymbol{\Lambda}_{\tau_r}^{(i)} + \bar{\mathbf{Y}}_r \odot \boldsymbol{\Lambda}_{\tau_r}^{(i)} \odot \mathbf{Q}_r^{(i)} \oslash \mathbf{M}_r^{(i)}$
14: $\bar{\boldsymbol{\lambda}}_{\tau_r}^{(i)} \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_{\tau_r}^{(i)})^\top$
15: **end for**
16: $\hat{\mathscr{L}}(\bar{\mathbf{Y}}_r | \bar{\mathbf{Y}}_{1:r-1}) \leftarrow \log\left(n_{part}^{-1} \sum_{j=1}^{n_{part}} w_r^{(j)}\right)$
17: $\bar{w}_r^{(i)} \leftarrow w_r^{(i)} / \sum_j w_r^{(j)}$ for $i = 1$ to n_{part}
18: **resample** $\left\{\bar{\boldsymbol{\lambda}}_{\tau_r}^{(i)}\right\}_{i=1}^{n_{part}}$ according to a systematic resampling scheme with weights $\{w_r^{(i)}\}_{i=1}^{n_{part}}$
19: **end for**

likelihood estimate for algorithm 14 is:

$$p(\bar{\mathbf{Y}}_{1:R}) \approx \sum_{r=1}^{R} \hat{\mathscr{L}}(\bar{\mathbf{Y}}_r | \bar{\mathbf{Y}}_{1:r-1}).$$

Convergence plots for coordinate ascent algorithm

For each of EqEq EqOv, and OvOv, we performed a finite differencing coordinate ascent optimisation. That is, for each parameter: fix all others to their current value and approximate the sign of

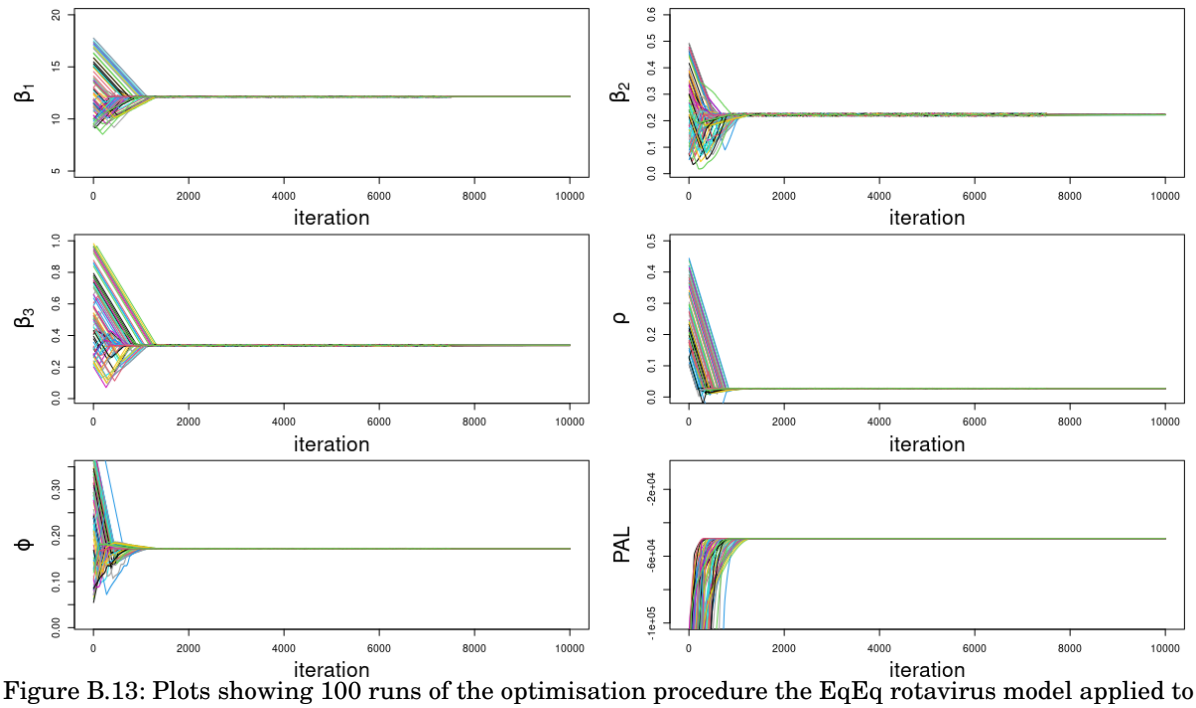

Figure B.13: Plots showing 100 runs of the optimisation procedure the EqEq rotavirus model applied to real data.

the gradient with finite differencing and take a step in positive gradient direction - cycle through parameters until convergence. Figures B.13, B.14, and B.15 demonstrate the convergence of this procedure for each model EqEq, EqOv, and OvOv respectively.

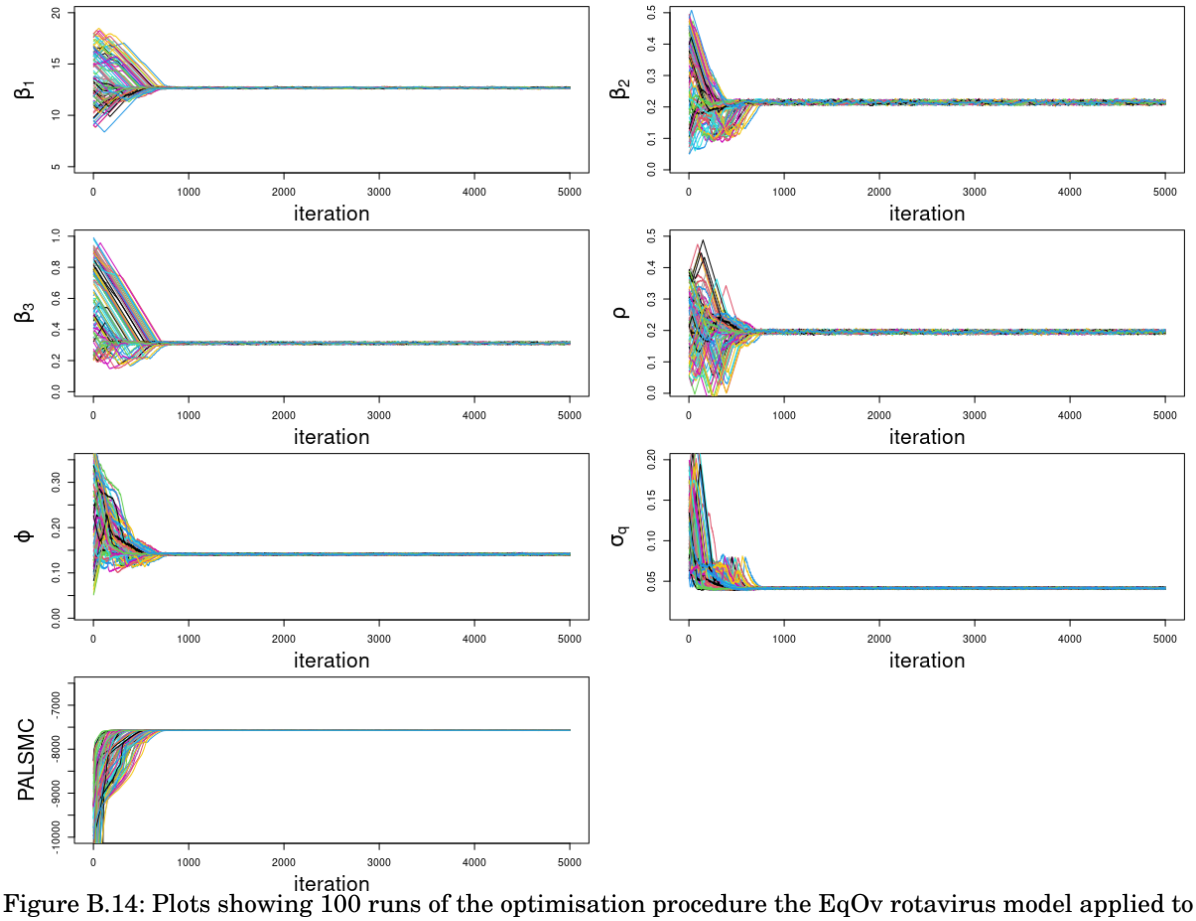

Figure B.14: Plots showing 100 runs of the optimisation procedure the EqOv rotavirus model applied to real data.

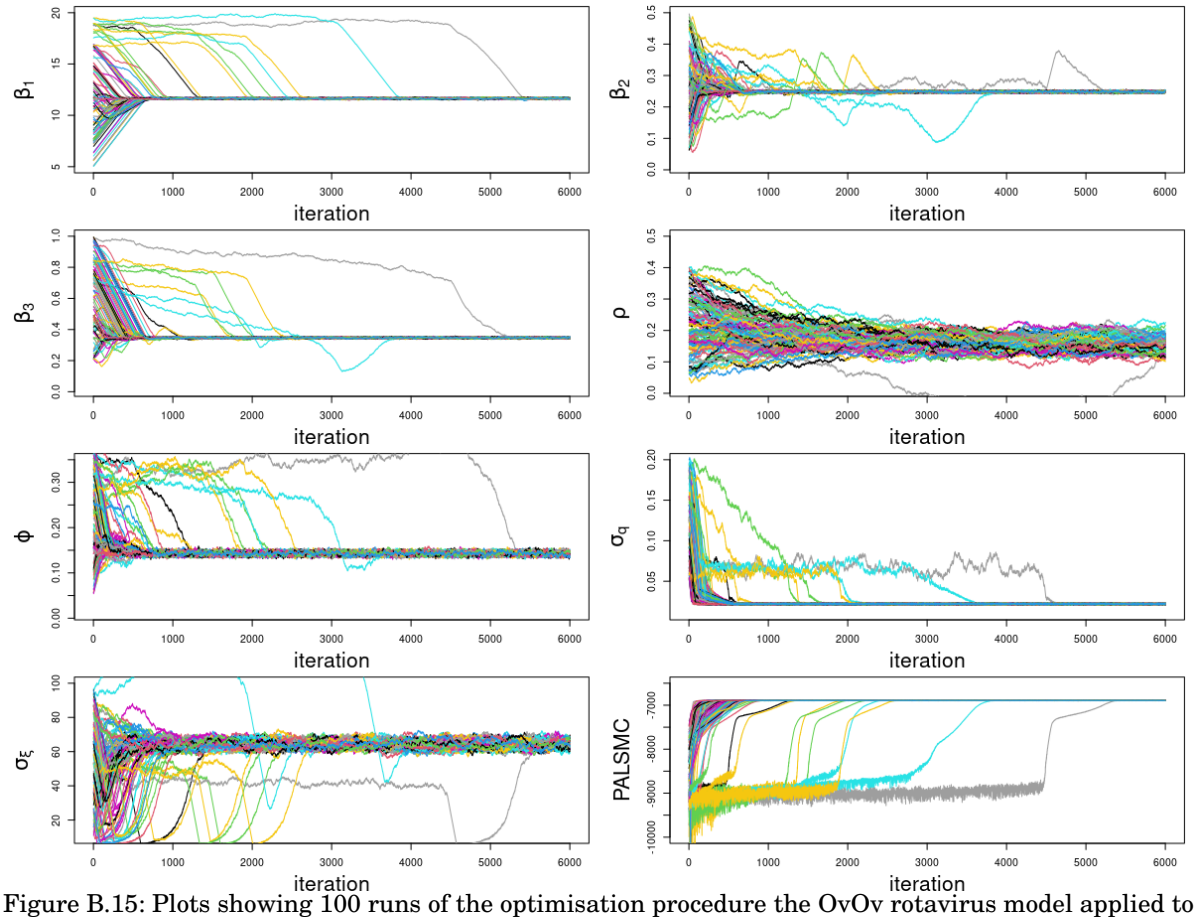

Figure B.15: Plots showing 100 runs of the optimisation procedure the OvOv rotavirus model applied to real data.

B.4 Supplementary material for section 6.4

This section presents supplementary material for the measles example of section 6.4

Model

As in Xia et al. (2004) and Park and Ionides (2020), we assume that $\beta_{k,r}$ follows the school year:

$$\beta_{k,r} = \begin{cases} (1+2(1-p)a)\bar{\beta}_k, & \text{during school term,} \\ (1-2pa)\bar{\beta}_k, & \text{during school holidays,} \end{cases}$$

where $p = 0.759$ is the proportion of the year taken up by school terms, $\bar{\beta}_k > 0$ is the mean transition rate for city k, and a is the relative effect of holidays on transmission. Finally, the new infected and new removed are:

$$C_{k,t} \sim \text{Bin}(E_{k,t} - F^{(E)}_{k,t}, 1-e^{-h\rho}), \qquad D_{k,t} \sim \text{Bin}(I_{k,t} - F^{(I)}_{k,t}, 1-e^{-h\gamma}),$$

with h/ρ mean time spent in the exposed compartment and h/γ mean recovery time. Given the vectors $\boldsymbol{\delta}_{k,t} = \left[\delta^{(S)}_{k,t}\, \delta^{(E)}_{k,t}\, \delta^{(I)}_{k,t}\, \delta^{(R)}_{k,t}\right]^\top \in \mathbb{R}^4_{\geq 0}$ and $\boldsymbol{\alpha}_{k,t} = \left[\alpha^{(1)}_{k,t}\, 0\, 0\, 0\right]^\top \in \mathbb{R}^4_{\geq 0}$, we have:

$$F^{(\cdot)}_{k,t} \sim \text{Bin}\left(\cdot_{k,t}, 1-\delta^{(\cdot)}_{k,t}\right), \qquad A_{k,t} \sim \text{Pois}\left(\alpha^{(1)}_{k,t}\right),$$

modelling the new births (immigration) into the susceptible population and the deaths (emigration) across compartments. Since there is no reinfection mechanism in the model (a realistic assumption for measles modelling), it is important to have new individuals enter the population to model the recurrent epidemic peaks present in the data. As already mentioned, for the model to capture recurrent peaks, it must accommodate recruitment into the susceptible compartments. Birthrate data for each city of the model is used to do this — as in Xia et al. (2004) — it is assumed that newborns enter the susceptible class after a delay of 4 years, corresponding to the age an individual enters the high-risk school-age demographic. There is a further 'cohort' effect aspect to the model: it is assumed that at the start of the school year, a fraction $c \in (0,1)$ of the lagged births enter the susceptible compartment, the remaining $1-c$ proportion enter at a constant rate throughout the year. This informs the assumed rate parameters $\boldsymbol{\alpha}_{k,t} = [\alpha^{(1)}_{k,t}\, 0\, 0\, 0\,]^\top$ and, similarly, death rate records inform choice of $\boldsymbol{\delta}_{k,t}$. The values used for $\boldsymbol{\alpha}_{k,t}$ and $\boldsymbol{\delta}_{k,t}$ are reported in the data available on the GitHub page.

The observations are aggregated incidence data in the form of cumulative fortnightly transitions from infective to recovered for each of the 40 cities subject to binomial under-reporting, at times $\tau_r = 4r$ for $r = 1,\ldots,R$. Observations are modelled as transitions from infective to recovered compartments because, on discovery, cases are treated with bed rest and hence removed from the population Park and Ionides (2020). Denoting observations as $\bar{\mathbf{Y}}_{k,r} = \sum_{t=\tau_{r-1}+1}^{\tau_r} \mathbf{Y}_{k,t}$ where each $\mathbf{Y}_{k,t} \in \mathbb{N}^{4\times 4}$ has each element equal to zero except for the $(3,4)$th element which, conditional on

$D_{k,t}$, is distributed:

$$Y_{k,t}^{(3,4)} \sim \text{Bin}\left(D_{k,t}, Q_{k,r}^{(3,4)}\right), \text{ for } t = \tau_{r-1}, \dots \tau_r, r \geq 1, \quad k = 1, \dots, J,$$

where $\mathbf{Q}_{k,r} \in [0,1]^{4\times 4}$ consists of all zeros apart from the $(3,4)$th entry, which is the reporting rate of transitions from infective to recovered. We assume that this rate follows $Q_{k,r}^{(3,4)} \sim \mathcal{N}(\mu_{q,k}, \sigma_q^2)_{\geq 0, \leq 1}$ for $k = 1, \dots, K$, denote this density with $f(\cdot \mid \cdot)$ for the purposes of algorithm 15. The mean under-reporting rate parameters, $\mu_{q,k} \in [0,1], k = 1, \dots, J$, are assumed known for each city and are set to the same values as Park and Ionides (2020), which are available in the data on the GitHub page, $\sigma_q^2 > 0$ is to be estimated.

Inference

To employ the algorithms described in the methodology section we need to specify the transition matrix $\mathbf{K}_{r,\boldsymbol{\eta},\xi}$, which in the case of this model is of size $4J \times 4J$. To be more succinct, we can write out a matrix $\mathbf{K}_{r,\tilde{\boldsymbol{\eta}}\xi,k}$ for each city $k = 1, \dots, 40$. We define our matrices $\mathbf{K}_{r,\tilde{\boldsymbol{\eta}},\xi,k}$:

$$\mathbf{K}_{r,\boldsymbol{\eta},\xi,k} = \begin{bmatrix} e^{-hg_k(\beta_{k,r},\boldsymbol{\eta},\xi)} & 1 - e^{-hg_k(\beta_{k,r},\boldsymbol{\eta},\xi)} & 0 & 0 \\ 0 & e^{-h\rho} & 1 - e^{-h\rho} & 0 \\ 0 & 0 & e^{-h\gamma} & 1 - e^{-h\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where

$$g_k(\beta, \boldsymbol{\eta}, \xi) = \beta\xi \cdot \left[\eta^{(k)} + \sum_{l \neq k} \frac{v_{kl}}{n_k} \left\{\eta^{(l)} - \eta^{(k)}\right\}\right],$$

One can identify matrices:

$$\mathbf{Z}_{k,t} := \begin{bmatrix} S_{k,t} - F_{k,t}^{(S)} - B_{k,t} & B_{k,t} & 0 & 0 \\ 0 & E_{k,t} - F_{k,t}^{(E)} - C_{k,t} & C_{k,t} & 0 \\ 0 & 0 & I_{k,t} - F_{k,t}^{(I)} - D_{k,t} & D_{k,t} \\ 0 & 0 & 0 & R_{k,t} - F_{k,t}^{(R)} \end{bmatrix},$$

and let $\mathbf{Z}_t$ be block-diagonal with blocks $\mathbf{Z}_{k,t}$, $k = 1, \dots, J$. One can take advantage of the block-diagonal structure of $\mathbf{K}_{t,\boldsymbol{\eta}}$ to implement an efficient block particle filter, see Rebeschini and Van Handel (2015) and Ning and Ionides (2021), with lookahead resampling scheme (Lin et al., 2013).

Proposals for algorithm 15

The details for the derivation of the proposals used in algorithm 15 lines 10 and 28 are similar to those of section B.3, so they are omitted. Suppressing dependence on the particle, let $\mathbf{L}_{r,k} = \sum_{t=\tau_{r-1}+1}^{\tau_r} \boldsymbol{\Lambda}_{t,k}$ with $\boldsymbol{\Lambda}_{t,k}$ calculated as per line 6 (resp. 24) of algorithm 14 and define:

$$\hat{\mu}_{r,k} = \frac{1}{2}\left(\mu_{q,k} - L_{r,k}^{(3,4)}\sigma_q^2 + \sqrt{(L_{r,k}^{(3,4)}\sigma_q^2 - \mu_{q,k})^2 + 4\bar{Y}_r^{(3,4)}\sigma_q^2}\right),$$
$$\left(\hat{\sigma}_{r,k}\right)^2 = \left(\frac{\bar{Y}_r^{(3,4)}}{\left(\hat{\mu}_{r,k}\right)^2} + \frac{1}{\sigma_q^2}\right)^{-1}.$$

Then in line 10 (resp. 28) we make the proposals:

$$Q_{k,r}^{(3,4)} \sim \mathcal{N}\left(\hat{\mu}_{r,k}, \hat{\sigma}_{r,k}^2\right)_{\geq 0, \leq 1}. \tag{B.2}$$

Inference

In each model instance, A,B, and C, described in 6.4, we can define $\boldsymbol{\vartheta}$, $\bar{\boldsymbol{\theta}}_{1:T}$, and $\boldsymbol{\varphi}$:

- A: $\boldsymbol{\vartheta} = [\boldsymbol{\pi}_0\ \bar{\beta}\ \rho\ \gamma\ g\ a\ c]$, $\{\bar{\boldsymbol{\theta}_r}\}_{r\geq 0} = \left\{\left[\xi_{1,r}\ \ldots\ \xi_{40,r}\ Q_{1,r}^{(3,4)}\ \ldots\ Q_{40,r}^{(3,4)}\right]\right\}_{r\geq 0}$, and $\boldsymbol{\varphi} = [\sigma_q^2, \sigma_\xi]$.
- B: $\boldsymbol{\vartheta} = [\boldsymbol{\pi}_{1,0}\ \ldots\ \boldsymbol{\pi}_{40,0}\ \bar{\beta}\ \rho\ \gamma\ g\ a\ c]$, $\{\bar{\boldsymbol{\theta}_r}\}_{r\geq 0} = \left\{\left[\xi_{1,r},\ldots,\xi_{40,r},Q_{1,r}^{(3,4)},\ldots,Q_{40,r}^{(3,4)}\right]\right\}_{r\geq 0}$, and $\boldsymbol{\varphi} = [\sigma_q^2, \sigma_\xi]$.
- C: $\boldsymbol{\vartheta} = [\boldsymbol{\pi}_{1,0}\ \ldots\ \boldsymbol{\pi}_{40,0}\ \bar{\beta}_1\ \ldots\ \bar{\beta}_{40}\ \rho\ \gamma\ g\ a\ c]$, $\{\bar{\boldsymbol{\theta}_r}\}_{r\geq 0} = \left\{\left[\xi_{1,r},\ldots,\xi_{40,r},Q_{1,r}^{(3,4)},\ldots,Q_{40,r}^{(3,4)}\right]\right\}_{r\geq 0}$, and $\boldsymbol{\varphi} = [\sigma_q^2, \sigma_\xi]$.

Each block, labelled $k = 1,\ldots,J$, corresponds to a specific city. This block structure allows one to perform proposals and weighting locally to each block, avoiding explicit high-dimensional filtering. At time r, the lookahead scheme consists of: performing a 'regular' particle propagation and reweighting step (the usual SMC iteration), then we propagate again each particle and run a PAL iteration for time $r+1$, with 'dummy' particles (used purely for weighting purposes, denoted with tildes in algorithm 15), we then weight the original particles proportionally to the joint likelihood of the regular and dummy particles at times r and $r+1$ - taking care to apply the appropriate correction in the likelihood calculation, dummy particles are then discarded. In practise, this scheme greatly reduced Monte Carlo error. See algorithm 15 for our implementation. The resulting approximate log-likelihood estimate associated with algorithm 15 is:

$$\log p(\bar{\mathbf{Y}}_{1:J,1:R}) \approx \sum_{r=1}^{R}\sum_{k=1}^{J} \hat{\mathscr{L}}(\bar{\mathbf{Y}}_{r,k}|\bar{\mathbf{Y}}_{1:r-1,k}).$$

The optimisation scheme that was used is described in figure B.16. We report the inferences for model C in table B.2.

Measles projection details.

The sample, size 300, of projected case numbers used to produce figure 6.8 in the main article were generated by the following workflow:

Algorithm 15 PAL within a lookahead block particle filter

initialise: $\bar{\boldsymbol{\lambda}}_{0,k}^{(i)} \leftarrow n_{k,0}\boldsymbol{\pi}_{k,0}$ set $\log\zeta_{0,k}^{(i)} \leftarrow 0$ and set $\log W_{0,k}^{(i)} \leftarrow 0$ for $i = 1$ to n_{part} and $k = 1,\ldots,K$.

for $r \geq 1$:
  for $k = 1,\ldots,J$:
    for $i = 1,\ldots,n_{part}$
      $\xi_{k,r}^{(i)} \sim \text{Gamma}(\sigma_\xi, \sigma_\xi)$
      for $t = \tau_{r-1}+1,\ldots,\tau_r - 1$:
        $\boldsymbol{\Lambda}_{t,k}^{(i)} \leftarrow ((\bar{\boldsymbol{\lambda}}_{t-1,k}^{(i)} \odot \boldsymbol{\delta}_{t,k}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{r,\boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{\tau_{r-1},1:J}^{(i)}\right),\xi_{k,r}^{(i)},k}$
        $\bar{\boldsymbol{\lambda}}_{t,k}^{(i)} \leftarrow (\mathbf{1}_m^\top \boldsymbol{\Lambda}_{t,k}^{(i)})^\top + \boldsymbol{\alpha}_{t,k}$
      end for
      $\boldsymbol{\Lambda}_{\tau_r,k}^{(i)} \leftarrow ((\boldsymbol{\lambda}_{\tau_r-1,k}^{(i)} \odot \boldsymbol{\delta}_{\tau_r,k}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{r,\boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{\tau_{r-1},1:J}^{(i)}\right),\xi_{r,k}^{(i)},k}$
      $\mathbf{Q}_{k,r}^{(i)} \sim \pi\left(\cdot \mid \left\{\boldsymbol{\Lambda}_{t,k}^{(i)}\right\}_{t=\tau_{r-1}+1}^{\tau_r}, \bar{\mathbf{Y}}_{r,k}, \boldsymbol{\varphi}\right)$ as per (B.2)
      $\mathbf{M}_{r,k}^{(i)} \leftarrow \sum_{t=\tau_{r-1}+1}^{\tau_r} \boldsymbol{\Lambda}_{t,k}^{(i)} \odot \mathbf{Q}_{k,r}^{(i)}$
      $\mathscr{L}(\bar{\mathbf{Y}}_{r,k} | \bar{\mathbf{Y}}_{1:r-1,k}) \leftarrow -\mathbf{1}_m^\top \mathbf{M}_{r,k}^{(i)} \mathbf{1}_m + \mathbf{1}_m^\top \left(\bar{\mathbf{Y}}_{r,k} \odot \log \mathbf{M}_{r,k}^{(i)}\right) \mathbf{1}_m - \mathbf{1}_m^\top \left(\log \bar{\mathbf{Y}}_{r,k}!\right) \mathbf{1}_m$
      $\log w_{r,k}^{(i)} \leftarrow \mathscr{L}(\bar{\mathbf{Y}}_{r,k} | \bar{\mathbf{Y}}_{1:r-1,k}) + \log\left(f(\mathbf{Q}_{k,r}^{(i)} \mid \mathbf{Q}_{k,1:r-1}^{(i)})\right) - \log\left(\pi\left(\mathbf{Q}_{k,r}^{(i)} \mid \left\{\boldsymbol{\Lambda}_{t,k}^{(i)}\right\}_{t=\tau_{r-1}+1}^{\tau_r}, \bar{\mathbf{Y}}_{r,k}, \boldsymbol{\varphi}\right)\right)$
      $\bar{W}_{r-1,k}^{(i)} \leftarrow W_{r-1,k}^{(i)} / \sum_j W_{r-1,k}^{(j)}$ for $i = 1$ to n_{part}
      $\log W_{r,k}^{(i)} \leftarrow \log \bar{W}_{r-1,k}^{(i)} + \log w_{r,k}^{(i)}$
      $\bar{\boldsymbol{\Lambda}}_{\tau_r,k}^{(i)} \leftarrow (\mathbf{1}_m \otimes \mathbf{1}_m - \mathbf{Q}_{k,r}^{(i)}) \odot \boldsymbol{\Lambda}_{\tau_r,k} + \bar{\mathbf{Y}}_{r,k} \odot \boldsymbol{\Lambda}_{\tau_r,k}^{(i)} \odot \mathbf{Q}_{k,r}^{(i)} \oslash \mathbf{M}_{r,k}^{(i)}$
      $\bar{\boldsymbol{\lambda}}_{\tau_r,k}^{(i)} \leftarrow (\mathbf{1}_m^\top \bar{\Lambda}_{\tau_r,k}^{(i)})^\top + \boldsymbol{\alpha}_{\tau_r,k}$
    end for
  end for
  for $k = 1,\ldots,J$:
    for $i = 1,\ldots,n_{part}$
      $\tilde{\xi}_{k,r+1}^{(i)} \sim \text{Gamma}(\sigma_\xi, \sigma_\xi)$
      for $t = \tau_r+1,\ldots,\tau_{r+1}-1$:
        $\boldsymbol{\Lambda}_{t,k}^{(i)} \leftarrow ((\bar{\boldsymbol{\lambda}}_{t-1,k}^{(i)} \odot \boldsymbol{\delta}_{t,k}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{r+1,\boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{\tau_r,1:J}^{(i)}\right),\tilde{\xi}_{r+1,k}^{(i)},k}$
        $\bar{\boldsymbol{\lambda}}_{t,k}^{(i)} \leftarrow (\mathbf{1}_m^\top \boldsymbol{\Lambda}_{t,k}^{(i)})^\top + \boldsymbol{\alpha}_{t,k}$
      end for
      $\boldsymbol{\Lambda}_{\tau_{r+1},k}^{(i)} \leftarrow ((\boldsymbol{\lambda}_{\tau_{r+1}-1,k}^{(i)} \odot \boldsymbol{\delta}_{\tau_{r+1},k}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{r+1,\boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{\tau_{r-1},1:J}^{(i)}\right),\tilde{\xi}_{r+1,k}^{(i)},k}$
      $\tilde{\mathbf{Q}}_{k,r+1}^{(i)} \sim \pi\left(\cdot \mid \left\{\boldsymbol{\Lambda}_{t,k}^{(i)}\right\}_{t=\tau_r+1}^{\tau_{r+1}}, \bar{\mathbf{Y}}_{r+1,k}, \boldsymbol{\varphi}\right)$ as per (B.2)
      $\mathbf{M}_{r+1,k}^{(i)} \leftarrow \sum_{t=\tau_r+1}^{\tau_{r+1}} \boldsymbol{\Lambda}_{t,k}^{(i)} \odot \tilde{\mathbf{Q}}_{r+1,k}^{(i)}$
      $\mathscr{L}(\bar{\mathbf{Y}}_{r+1,k} | \bar{\mathbf{Y}}_{1:r,k}) \leftarrow -\mathbf{1}_m^\top \mathbf{M}_{r+1,k}^{(i)} \mathbf{1}_m + \mathbf{1}_m^\top \left(\bar{\mathbf{Y}}_{r+1,k} \odot \log \mathbf{M}_{r+1,k}^{(i)}\right) \mathbf{1}_m - \mathbf{1}_m^\top \left(\log \bar{\mathbf{Y}}_{r+1,k}!\right) \mathbf{1}_m$
      $\log w_{r+1,k}^{(i)} \leftarrow \mathscr{L}(\bar{\mathbf{Y}}_{r+1,k} | \bar{\mathbf{Y}}_{1:r,k}) + \log\left(f(\mathbf{Q}_{k,r+1}^{(i)} \mid \mathbf{Q}_{k,1:r}^{(i)})\right) - \log\left(\pi\left(\mathbf{Q}_{k,r+1}^{(i)} \mid \left\{\boldsymbol{\Lambda}_{t,k}^{(i)}\right\}_{t=\tau_r+1}^{\tau_{r+1}}, \bar{\mathbf{Y}}_{r+1,k}, \boldsymbol{\varphi}\right)\right)$
      $\log \zeta_{r,k}^{(i)} \leftarrow \log W_{r,k}^{(i)} + \log w_{r+1,k}^{(i)}$
    end for
    $\hat{\mathscr{L}}(\bar{\mathbf{Y}}_{r,k} | \bar{\mathbf{Y}}_{1:r-1,k}) \leftarrow \log\left(\sum_j W_{r,k}^{(j)}\right)$
    $\bar{\zeta}_{r,k}^{(i)} \leftarrow \zeta_{r,k}^{(i)} / \sum_j \zeta_{r,k}^{(j)}$ for $i = 1$ to n_{part}
    resample $\left\{\bar{\boldsymbol{\lambda}}_{\tau_r,k}^{(i)}, W_{r,k}^{(i)}, \zeta_{r,k}^{(i)}\right\}_{i=1}^{n_{part}}$ with weights $\{\bar{\zeta}_{r,k}^{(i)}\}_{i=1}^{n_{part}}$
    $\log W_{r,k}^{(i)} \leftarrow \log W_{r,k}^{(i)} - \log \zeta_{r,k}^{(i)}$
end for

1. Presampling $\xi_{k,r}^{(i)} \sim \text{Gamma}(\sigma_\xi, \sigma_\xi)$ with σ_ξ set to our point estimate, for $k = 1,...,40$, $r = 1,...,4$, and $i = 1,\ldots,300$.

2. Running our PALSMC scheme on the original dataset with 300 particles and parameters set to our point estimates, taking as output a sample of final time-point population state intensity vectors $\bar{\boldsymbol{\lambda}}_{T,k}^{(i)}$.

3. For $i = 1,\ldots,300$ and $k = 1,\ldots,40$, propagate the intensity vectors through the transition kernel using the iteration for $t = 1,\ldots,16$ (corresponding to 8 weeks):

$$\boldsymbol{\Lambda}_{t,k}^{(i)} = ((\bar{\boldsymbol{\lambda}}_{t-1,k}^{(i)} \odot \boldsymbol{\delta}_{t,k}) \otimes \mathbf{1}_m) \odot \mathbf{K}_{r,\boldsymbol{\eta}\left(\bar{\boldsymbol{\lambda}}_{\tau_{r-1},1:J}^{(i)}\right),\xi_{r,k}^{(i)},k}$$
$$\bar{\boldsymbol{\lambda}}_{t,k}^{(i)} = (\mathbf{1}_m^\top \boldsymbol{\Lambda}_{t,k}^{(i)})^\top + \boldsymbol{\alpha}_{t,k}$$

   Where $\boldsymbol{\alpha}_{t,k}$ and $\boldsymbol{\delta}_{t,k}$ are chosen according to the assumption that birth rates and death rates remain constant.

4. Simulate $I_{k,t}^{(i)} \sim \text{Pois}(\bar{\boldsymbol{\lambda}}_{t,k}^{(i)})$ for t corresponding to weeks 2,4,6, and 8 for each sample $i = 1,\ldots,300$.

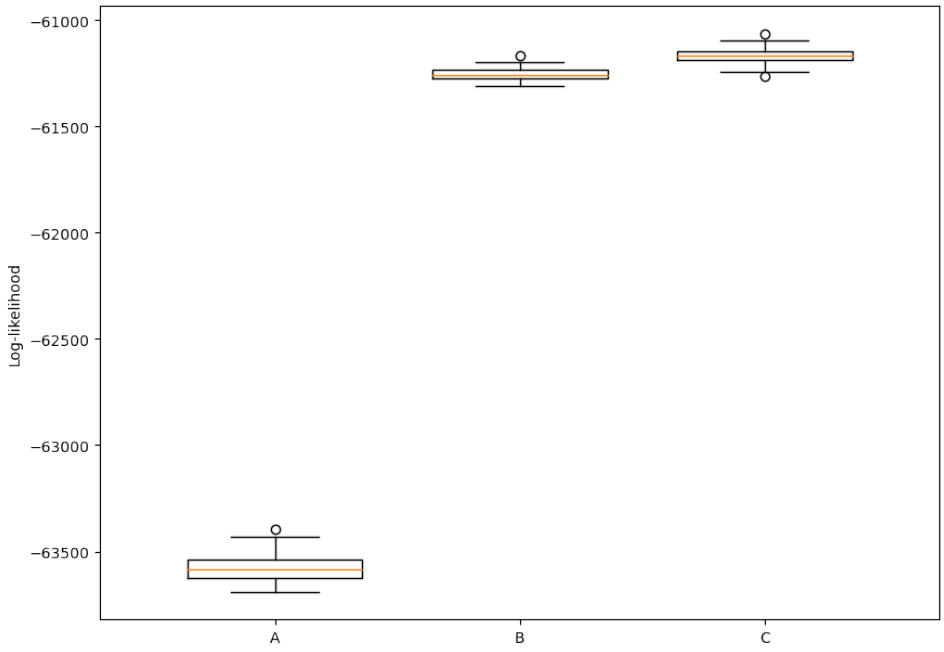

Figure B.16: Approximate log-likelihood values for the measles data under scenarios A, B, and C. For each scenario, the optimal combination of parameters was obtained through Sequential Least Squares Programming (SLSQP) with target function given by algorithm 15 with 5000 particles and lookahead resampling, this scheme was initialised randomly at 100 points over feasible values, the best attained values are presented. After the optimisation, algorithm 15 with 5000 particles and lookahead resampling is run 100 times on the optimised parameters to build the boxplots and estimate the variance of the approximate log-likelihood.

City	$\frac{n_{0,k}}{1000}$	$\pi_{0,k}^{(1)}$	$\pi_{0,k}^{(2)}$	$\pi_{0,k}^{(3)}$	$\pi_{0,k}^{(4)}$	R_0	$1/\rho$	$1/\gamma$
BIRKENHEAD	143	0.07594	0.00007	0.00013	0.92387	8.47	8.49	9.53
BIRMINGHAM	1118	0.04575	0.00005	0.00013	0.95408	5.63	8.49	9.53
BLACKPOOL	150	0.07210	0.00005	0.00259	0.92525	12.93	8.49	9.53
BOLTON	169	0.09337	0.00007	0.00120	0.90537	9.44	8.49	9.53
BOURNEMOUTH	140	0.12166	0.00006	0.00005	0.87822	10.62	8.49	9.53
BRADFORD	294	0.08243	0.00004	0.00044	0.91708	10.22	8.49	9.53
BRIGHTON	158	0.07625	0.00008	0.00035	0.92332	14.66	8.49	9.53
BRISTOL	443	0.07355	0.00009	0.00206	0.92430	8.63	8.49	9.53
CARDIFF	245	0.09190	0.00005	0.00058	0.90747	7.81	8.49	9.53
COVENTRY	257	0.11602	0.00004	0.00018	0.88376	8.16	8.49	9.53
DERBY	143	0.11061	0.00006	0.00008	0.88925	10.46	8.49	9.53
GATESHEAD	115	0.08601	0.00007	0.00006	0.91386	8.28	8.49	9.53
HUDDERSFIELD	130	0.09003	0.00007	0.00022	0.90968	10.78	8.49	9.53
HULL	302	0.06856	0.00009	0.00083	0.93051	9.28	8.49	9.53
IPSWICH	104	0.08528	0.00009	0.00000	0.91463	9.03	8.49	9.53
LEEDS	510	0.09935	0.00006	0.00168	0.89891	5.92	8.49	9.53
LEICESTER	288	0.07103	0.00005	0.00133	0.92759	9.00	8.49	9.53
LIVERPOOL	802	0.05754	0.00004	0.00025	0.94217	5.63	8.49	9.53
LONDON	3389	0.04575	0.00006	0.00021	0.95399	5.63	8.49	9.53
MANCHESTER	704	0.05658	0.00003	0.00145	0.94193	7.29	8.49	9.53
MIDDLESBOROUGH	146	0.06662	0.00007	0.00067	0.93264	11.32	8.49	9.53
NEWCASTLE	295	0.07129	0.00005	0.00024	0.92843	9.19	8.49	9.53
NORWICH	120	0.10958	0.00005	0.00000	0.89037	12.78	8.49	9.53
NOTTINGHAM	307	0.05794	0.00004	0.00068	0.94133	11.54	8.49	9.53
OLDHAM	119	0.09814	0.00007	0.00092	0.90087	11.57	8.49	9.53
PLYMOUTH	209	0.08388	0.00006	0.00077	0.91529	13.37	8.49	9.53
PORTSMOUTH	240	0.07339	0.00007	0.00295	0.92359	9.39	8.49	9.53
PRESTON	120	0.06501	0.00007	0.00242	0.93251	7.52	8.49	9.53
READING	116	0.07686	0.00005	0.00137	0.92172	12.93	8.49	9.53
SALFORD	178	0.08982	0.00007	0.00109	0.90903	8.82	8.49	9.53
SHEFFIELD	515	0.07818	0.00006	0.00308	0.91869	8.10	8.49	9.53
SOUTHAMPTON	181	0.11018	0.00006	0.00391	0.88585	11.14	8.49	9.53
SOUTHEND	152	0.11816	0.00008	0.00043	0.88132	16.65	8.49	9.53
ST.HELENS	112	0.09871	0.00008	0.00256	0.89864	9.89	8.49	9.53
STOCKPORT	142	0.09721	0.00004	0.00231	0.90044	9.03	8.49	9.53
STOKE	276	0.07614	0.00006	0.00071	0.92310	8.62	8.49	9.53
SUNDERLAND	178	0.06698	0.00006	0.00088	0.93208	14.90	8.49	9.53
SWANSEA	162	0.08195	0.00006	0.00052	0.91748	12.59	8.49	9.53
WALSALL	115	0.08378	0.00009	0.00080	0.91533	13.75	8.49	9.53
WOLVERHAMPTON	162	0.06376	0.00006	0.00021	0.93597	8.57	8.49	9.53

Table B.2: Measles example. Inferred quantities for model C.

BIBLIOGRAPHY

Akaike, H. (1974).
A new look at the statistical model identification.
IEEE transactions on automatic control 19(6), 716–723.

Allen, L. J. (2017).
A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis.
Infectious Disease Modelling 2(2), 128–142.

Andersson, H. and T. Britton (2012).
Stochastic epidemic models and their statistical analysis, Volume 151.
Springer Science & Business Media.

Andrade, J. and J. Duggan (2020).
An evaluation of Hamiltonian Monte Carlo performance to calibrate age-structured compartmental SEIR models to incidence data.
Epidemics 33, 100415.

Andrews, D. W. (1992).
Generic uniform convergence.
Econometric theory 8(2), 241–257.

Andrieu, C., A. Doucet, and R. Holenstein (2010).
Particle Markov chain Monte Carlo methods.
Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(3), 269–342.

Anon (1978).
Influenza in a boarding school.
The British Medical Journal, 587.

Ball, F., T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. S. Tomba (2015).
Seven challenges for metapopulation models of epidemics, including households models.
Epidemics 10, 63–67.

Bartlett, M. (1949).
Some evolutionary stochastic processes.

Journal of the Royal Statistical Society. Series B (Methodological) 11(2), 211–229.

Bartlett, M. S. (1966).
An introduction to stochastic processes.
University Press Cambridge.

Beaumont, M. A. (2010).
Approximate bayesian computation in evolution and ecology.
Annual review of ecology, evolution, and systematics 41, 379–406.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002).
Approximate bayesian computation in population genetics.
Genetics 162(4), 2025–2035.

Becker, N. (1993).
Martingale methods for the analysis of epidemic data.
Statistical Methods in Medical Research 2(1), 93–112.

Bretó, C. and E. L. Ionides (2011).
Compound markov counting processes and their applications to modeling infinitesimally over-dispersed systems.
Stochastic Processes and their Applications 121(11), 2571–2591.

Britton, T. (2010).
Stochastic epidemic models: a survey.
Mathematical biosciences 225(1), 24–35.

Caron, F., P. Del Moral, A. Doucet, and M. Pace (2011).
On the conditional distributions of spatial point processes.
Advances in Applied Probability 43(2), 301–307.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017).
Stan: A probabilistic programming language.
Journal of statistical software 76(1).

Cauchemez, S. and N. M. Ferguson (2008).
Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london.
Journal of the Royal Society Interface 5(25), 885–897.

Chopin, N., O. Papaspiliopoulos, et al. (2020).
An introduction to sequential Monte Carlo, Volume 4.
Springer.

Davies, J., A. Smith, E. Grilli, and T. Hoskins (1982).
Christ's hospital 1978–79: An account of two outbreaks of influenza a h1n1.
Journal of Infection 5(2), 151–156.

Delamater, P. L., E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen (2019).
Complexity of the basic reproduction number (r0).
Emerging infectious diseases 25(1), 1.

Demiris, N. and P. D. O'Neill (2005).
Bayesian inference for epidemics with two levels of mixing.
Scandinavian journal of statistics 32(2), 265–280.

Diekmann, O., J. A. P. Heesterbeek, and J. A. Metz (1990).
On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations.
Journal of mathematical biology 28(4), 365–382.

Doucet, A., S. Godsill, and C. Andrieu (2000).
On sequential monte carlo sampling methods for bayesian filtering.
Statistics and computing 10, 197–208.

Fasiolo, M., N. Pya, and S. N. Wood (2016).
A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology.
Statistical Science, 96–118.

Fearnhead, P., V. Giagos, and C. Sherlock (2014).
Inference for reaction networks using the linear noise approximation.
Biometrics 70(2), 457–466.

Fearnhead, P. and D. Prangle (2012).
Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation.
Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74(3), 419–474.

Fintzi, J., X. Cui, J. Wakefield, and V. N. Minin (2017).
Efficient data augmentation for fitting stochastic epidemic models to prevalence data.
Journal of Computational and Graphical Statistics 26(4), 918–929.

Funk, S., S. Bansal, C. T. Bauch, K. T. Eames, W. J. Edmunds, A. P. Galvani, and P. Klepac (2015).
Nine challenges in incorporating the dynamics of behaviour in infectious diseases models.
Epidemics 10, 21–25.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995).
Bayesian data analysis.
Chapman and Hall/CRC.

Gillespie, D. T. (1976).
A general method for numerically simulating the stochastic time evolution of coupled chemical reactions.
Journal of computational physics 22(4), 403–434.

Golightly, A., D. A. Henderson, and C. Sherlock (2015).
Delayed acceptance particle MCMC for exact inference in stochastic kinetic models.
Statistics and Computing 25(5), 1039–1055.

Gourieroux, C. and J. Jasiak (2021).
Temporally local maximum likelihood with application to sis model.
arXiv:2107.06971.

Guerra, F. M., S. Bolotin, G. Lim, J. Heffernan, S. L. Deeks, Y. Li, and N. S. Crowcroft (2017).
The basic reproduction number (r0) of measles: a systematic review.
The Lancet Infectious Diseases 17(12), e420–e428.

He, D., E. L. Ionides, and A. A. King (2010).
Plug-and-play inference for disease dynamics: measles in large and small populations as a case study.
Journal of the Royal Society Interface 7(43), 271–283.

Ionides, E. L., A. Bhadra, Y. Atchadé, and A. King (2011).
Iterated filtering.
The Annals of Statistics 39(3), 1776–1802.

Ionides, E. L., C. Breto, J. Park, R. Smith, and A. A. King (2017).
Monte carlo profile confidence intervals for dynamic systems.
Journal of The Royal Society Interface 14(132), 20170126.

Ionides, E. L., N. Ning, and J. Wheeler (2022).
An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters.
arXiv preprint arXiv:2206.03837.

Isham, V. (2005).
Stochastic models for epidemics.
Oxford statistical science series 33, 27.

Jasra, A. and P. Del Moral (2011).
Sequential monte carlo methods for option pricing.
Stochastic analysis and applications 29(2), 292–316.

Jasra, A., D. A. Stephens, A. Doucet, and T. Tsagaris (2011).
Inference for lévy-driven stochastic volatility models via adaptive sequential monte carlo.
Scandinavian Journal of Statistics 38(1), 1–22.

Ju, N., J. Heng, and P. E. Jacob (2021).
Sequential Monte Carlo algorithms for agent-based models of disease transmission.
arXiv preprint arXiv:2101.12156.

Kendall, D. G. (1956).
Deterministic and stochastic epidemics in closed populations.
In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health*, pp. 149–165. University of California Press.

Kendall, M. G. et al. (1946).
The advanced theory of statistics.
The advanced theory of statistics. 1(2nd Ed).

Kermack, W. O. and A. G. McKendrick (1927).
A contribution to the mathematical theory of epidemics.
Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115(772), 700–721.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani (2015).
Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola.
Proceedings of the Royal Society B: Biological Sciences 282(1806), 20150347.

King, A. A., D. Nguyen, and E. L. Ionides (2016).
Statistical Inference for Partially Observed Markov Processes via the R Package pomp.
Journal of Statistical Software 69(12), 1–43.

Kingman, J. F. C. (1992).
Poisson processes, Volume 3.
Clarendon Press.

Komorowski, M., B. Finkenstädt, C. V. Harper, and D. A. Rand (2009).
Bayesian inference of biochemical kinetic parameters using the linear noise approximation.
BMC bioinformatics 10(1), 1–10.

Kurtz, T. G. (1970).
Solutions of ordinary differential equations as limits of pure jump Markov processes.
Journal of Applied Probability 7(1), 49–58.

Kurtz, T. G. (1971).
Limit theorems for sequences of jump Markov processes approximating ordinary differential processes.
Journal of Applied Probability 8(2), 344–356.

Lin, M., R. Chen, and J. S. Liu (2013).
Lookahead strategies for sequential monte carlo.
Statistical Science 28(1), 69–94.

Lindenstrand, D. and Å. Svensson (2013).
Estimation of the malthusian parameter in an stochastic epidemic model using martingale methods.
Mathematical biosciences 246(2), 272–279.

Mahler, R. P. (2003).
Multitarget Bayes filtering via first-order multitarget moments.
IEEE Transactions on Aerospace and Electronic systems 39(4), 1152–1178.

Mahler, R. P., B.-T. Vo, and B.-N. Vo (2011).
Cphd filtering with unknown clutter rate and detection profile.
IEEE Transactions on Signal Processing 59(8), 3497–3513.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003).
Markov chain monte carlo without likelihoods.
Proceedings of the National Academy of Sciences 100(26), 15324–15328.

McKendrick, A. (1925).
Applications of mathematics to medical problems.
Proceedings of the Edinburgh Mathematical Society 44, 98–130.

McKinley, T., A. R. Cook, and R. Deardon (2009).
Inference in epidemic models without likelihoods.
The International Journal of Biostatistics 5(1).

Morsomme, R. and J. Xu (2022).
Uniformly ergodic data-augmented mcmc for fitting the general stochastic epidemic model to incidence data.
arXiv preprint arXiv:2201.09722.

Ning, N. and E. L. Ionides (2021).
Iterated block particle filter for high-dimensional parameter learning: Beating the curse of dimensionality.
arXiv preprint arXiv:2110.10745.

Park, J. and E. L. Ionides (2020).
Inference on high-dimensional implicit dynamic models using a guided intermediate resampling filter.
Statistics and Computing 30(5), 1497–1522.

Prangle, D., P. Fearnhead, M. P. Cox, P. J. Biggs, and N. P. French (2014).
Semi-automatic selection of summary statistics for abc model choice.
Statistical applications in genetics and molecular biology 13(1), 67–82.

Rebeschini, P. and R. Van Handel (2015).
Can local particle filters beat the curse of dimensionality?
The Annals of Applied Probability 25(5), 2809–2866.

Riley, S., K. Eames, V. Isham, D. Mollison, and P. Trapman (2015).
Five challenges for spatial epidemic models.
Epidemics 10, 68–71.

Rimella, L., C. Jewell, and P. Fearnhead (2022).
Approximating optimal smc proposal distributions in individual-based epidemic models.
arXiv preprint arXiv:2206.05161.

Rimella, L., C. Jewell, and P. Fearnhead (2023).
Approximating optimal smc proposal distributions in individual-based epidemic models.
To appear in Statistics Sinica.

Roberts, G. O. and O. Stramer (2001).
On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm.
Biometrika 88(3), 603–621.

Roberts, M., V. Andreasen, A. Lloyd, and L. Pellis (2015).
Nine challenges for deterministic epidemic models.
Epidemics 10, 49–53.

Rubin, D. B. (1984).
Bayesianly justifiable and relevant frequency calculations for the applied statistician.
The Annals of Statistics, 1151–1172.

Rue, H., S. Martino, and N. Chopin (2009).
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the royal statistical society: Series b (statistical methodology) 71(2), 319–392.

Saulnier, E., O. Gascuel, and S. Alizon (2017).
Inferring epidemiological parameters from phylogenies using regression-abc: A comparative study.
PLoS computational biology 13(3), e1005416.

Singh, S. S., B.-N. Vo, A. Baddeley, and S. Zuyev (2009).
Filters for spatial point processes.
SIAM Journal on Control and Optimization 48(4), 2275–2295.

Singh, S. S., N. Whiteley, and S. Godsill (2011).
Approximate likelihood estimation of static parameters in multi-target models.
In D. Barber, A. Cemgil, and S. Chiappa (Eds.), *Bayesian Time Series Models*, Chapter 11, pp. 225–244. Cambridge University Press.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson (2008).
Obstacles to high-dimensional particle filtering.
Monthly Weather Review 136(12), 4629–4640.

Stocks, T., T. Britton, and M. Höhle (2020).
Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in germany.
Biostatistics 21(3), 400–416.

Sun, L., C. Lee, and J. A. Hoeting (2015).
Parameter inference and model selection in deterministic and stochastic dynamical models via approximate bayesian computation: modeling a wildlife epidemic.
Environmetrics 26(7), 451–462.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf (2009).
Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems.
Journal of the Royal Society Interface 6(31), 187–202.

Truscott, J. and N. M. Ferguson (2012).
Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling.

Van den Driessche, P. and J. Watmough (2002).
Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission.
Mathematical biosciences 180(1-2), 29–48.

Vynnycky, E. and W. Edmunds (2008).
Analyses of the 1957 (Asian) influenza pandemic in the United Kingdom and the impact of school closures.
Epidemiology & Infection 136(2), 166–179.

Walker, J. N., J. V. Ross, and A. J. Black (2017).
Inference of epidemiological parameters from household stratified data.
Plos one 12(10), e0185910.

Whitehouse, M., N. Whiteley, and L. Rimella (2023).
Consistent and fast inference in compartmental models of epidemics using poisson approximate likelihoods.
Journal of the Royal Statistical Society Series B: Statistical Methodology.

Whiteley, N. and L. Rimella (2021).
Inference in stochastic epidemic models via multinomial approximations.
In *International Conference on Artificial Intelligence and Statistics*, pp. 1297–1305. PMLR.

Wikramaratna, P. S., A. Kucharski, S. Gupta, V. Andreasen, A. R. McLean, and J. R. Gog (2015).
Five challenges in modelling interacting strain dynamics.
Epidemics 10, 31–34.

Worden, L. and T. C. Porco (2017).
Products of compartmental models in epidemiology.
Computational and mathematical methods in medicine 2017.

Xia, Y., O. N. Bjørnstad, and B. T. Grenfell (2004).
Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics.
The American Naturalist 164(2), 267–281.