Statistical Methods for Complex Population Dynamics

Fasiolo, Matteo

Award date:
2016

Awarding institution:
University of Bath

Statistical Methods for Complex Population Dynamics

submitted by

Matteo Fasiolo

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

March 2016

SUMMARY

In recent years, the development of simulation-based, often approximate, statistical methods has been prompted by the challenges posed by complex models used in fields such ecology, epidemiology and system biology. A common issue with such models is that the likelihood function, so central to both Bayesian and classical approaches to statistical inference, is often unavailable or intractable. While intractable models could be dealt with using other methodologies, in this work we focus mainly on Synthetic Likelihood (SL). This is a simulation-based method based on summary statistics, rather than on the full data, and it is closely related to Approximate Bayesian Computation (ABC) methods.

The purpose of this thesis is twofold. First, we compare SL, ABC and other, less approximate, methods in the context of highly non-linear ecological and epidemiological models. We do this using a wide range of models, with both simulated and real data. The second part of the thesis is dedicated to improving SL. In particular, we address the computational cost of SL by proposing an efficient Maximum Synthetic Likelihood (MSL) algorithm, which exploits the Gaussian assumption used by SL. Finally, we relax this distributional assumption by proposing an original density estimator which, while being more flexible than a Gaussian estimator, scales well with the number of statistics used by SL.

ACKNOWLEDGEMENTS

I am very thankful to my supervisor Simon N. Wood for his support, guidance and advise during my PhD. Beside benefiting from Simon's technical ideas and insights, while collaborating with him I had to opportunity to appreciate and admire his way of thinking about problems and his scientific attitude, which certainly made me a better researcher.

I thank NCSE, EPSRC and the Department of Mathematical Sciences for financial support of the research that led to this thesis.

I am indebted to my girlfriend Valentina Noacco for supporting me in so many ways during these years. In particular, she has been incredibly patient with me when debugging-related stress was at its highest.

I would like to thank Chris Jennison for providing useful comments on an earlier version of this thesis. Similarly, I am thankful to Edward Ionides and Aaron King for their suggestions during, and after, the BIRS meeting in Banff and to Natalya Pya, Florian Hartig and Mark Bravington, who collaborated with me on different parts of the work appearing in this thesis.

I was lucky to have great office mates while in Bath. In particular, I would like to thank Yi Liu, Kuntalee Chaisee, Alex Griffiths, Marion Hesse, Christoph Hoeggerl, Elisabeth Ullmann and Eike Mueller for creating a great working atmosphere.

I thank Simon Maskell for his kindness and support during my postdoc at the University of Liverpool. I also would like to thank Flávio de Melo, Chinmay Mishra, Chongyang Liu, Richard Sloane, Lykourgos Kekempanos, Yifan Zhou, Joanna Hajne and Chloe Barrett-Pink for useful discussions and for making me feel at home in Liverpool.

Infine vorrei ringraziare la mia famiglia, il cui sostegno mi è stato indispensabile per arrivare fino a qui.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

In this chapter we describe some of the challenges faced by statistical ecologists and we point out that some of these can be addressed by approximate methods, which use summary statistics rather than the full data. In addition, we describe how many ecological models are characterized by unavailable or intractable likelihood functions. We define one class of such models, namely State Space Models, which will appear several times in the rest of this thesis. Finally, we give an outline of the thesis.

1.1 Intractable ecological models

Ecology aims to understand the abundance and distribution of organisms. This essentially quantitative task is made difficult by the complex web of interactions that exist between living things. In the face of such daunting ecological complexity, dynamic models play an important role in separating fundamental mechanisms from matters of detail. In particular, they allow theoretical ideas to be sharpened into well defined quantitative hypotheses, and this in turn opens up the possibility of testing these hypotheses using data.

But there is a catch. To be useful, ecological dynamic models must often resort to 'cartooning' of some ecological processes. Simplification is essential if the model is not to become a 'model-of-everything', hence a reasonably parsimonious model may not be intended to reproduce the full data, $\boldsymbol{y}^0$, in all its features. For example, while the full data might be characterized by a spatial structure, it is often convenient to use a lumped model that ignores this dimension. Similarly, when the data contains several classes of organisms, computational considerations might lead to a model that aggregates key statistics, such population counts, over different classes. Under these circumstances, reducing the full data to a set of summary statistics, $\boldsymbol{s}^0 = S(\boldsymbol{y}^0)$, might not lead to any loss of information during parameter estimation or model selection (Hartig et al., 2011).

Basing statistical inference on aggregate summary statistics might be necessary also when working with Individual Based Models, which are often used to understand ecological outcomes that depend intricately on the interactions of individuals within a population. Forest stand growth models are an example. In these models individual trees of many species may be grown to maturity, all competing continuously for light

and nutrients as they do so. Here the mismatch between data and model is of a different kind. For example, in a real forest we would obtain data consisting of measurements on individual trees. The same measurements can often be made on the model trees, but a particular model individual does not correspond to any real individual. We are left with no choice but to base inference on summary statistics, as done by Hartig et al. (2014) who uses Synthetic Likelihood (SL), an approximate method proposed by Wood (2010), to fit the Formind individual-based forest model to Ecuadorian tropical forest field data. While the model deals with individual trees, its output is summarized using 112 statistics such as biomass, growth rate and tree counts, obtained by aggregating trees over several diameter classes.

Other reasons for considering the use of summary statistics relate to highly non-linear dynamics, of the sort that are often found in populations of small animals, with high rates of fecundity and mortality. Indeed, even if our models are perfect descriptions of the driving ecological mechanisms, dynamic irregularity can make reliable inference very difficult to achieve by conventional means. If our models are less perfect, the interaction of such irregularity with small infelicities in the model's ability to match the data can lead to substantial inferential errors. Wood (2010) shows that these problems can arise in ecological systems as simple as the Ricker map (May, 1976), and illustrates how the extreme sensitivity of near chaotic systems to small changes in dynamically important parameters can cause minuscule moves in the parameter space to result in massive changes in likelihood values. In this circumstance, it is obviously appealing to base inference on summary statistics of the data that the model should be able to reproduce, rather than on the full data. Indeed, Wood (2010) argues that approximate methods can offer an appealing robustness here, provided that they are used in conjuction with appropriately robust statistics.

Even in the absence of the difficuties just discussed, ecological models can have tractability problems. Most of the conventional statistical tools used to find the parameter values or models that are most consistent with the data (and possibly with prior knowledge), rely on the likelihood function, $p(\boldsymbol{y}^0|\boldsymbol{\theta})$. Unfortunately, for many models of ecological interest, $p(\boldsymbol{y}^0|\boldsymbol{\theta})$ is not available directly or it is otherwise problematic, thus posing an obstacle to the whole inferential process. This difficulty can occur for several possible reasons, but one common problem is the presence of hidden or latent states. Specifically, we often know that the dynamics of an observed process $\boldsymbol{y}^0$ are related to those of other processes $\boldsymbol{n}$, which are hidden from us. In such cases the likelihood could ideally be obtained by integrating the latent states out of the joint probability density of data and hidden states

$$p(\boldsymbol{y}^0|\boldsymbol{\theta}) = \int p(\boldsymbol{y}^0, \boldsymbol{n}|\boldsymbol{\theta})\, d\boldsymbol{n}. \tag{1.1}$$

In practice this integration problem is usually analytically intractable, while the efficient implementation of numerical or Monte Carlo integration schemes often require additional assumptions, such as those detailed in Section 1.2.

Classical examples of partially observed systems of ecological interest are predator-prey systems, where the abundance of one of the two components is often completely unknown. For instance, in Chapter 3 we consider the prey-predator model proposed by Turchin and Ellner (2000), which has been used to describe the population dynamics of Fennoscandian voles. In that example trap data provides noisy estimates of voles abundance, but no such proxy is available for predatory weasels. A similar example is

provided by Kendall et al. (2005), who evaluate alternative explanations for the regular oscillations in population density of insect pest pine looper moths. They consider, among others, a parasitoid and a food quality model and they fit them using only data on moth population density. Given that ecological systems are observed with noise in most cases, the issue of hidden states is widespread and it appears in studies concerned with animal movement (Langrock et al., 2012; Morales et al., 2004), population abundance estimation (Farnsworth et al., 2007), and essentially whenever remote tracking data is available (Jonsen et al., 2005).

The rapid growth in computational resources has supported the development of several approaches meant to tackle the issue of intractable likelihoods. Some of these approaches exploit the fact that faster computation makes forward model simulation, that is simulation of data $\boldsymbol{y}$ from $p(\boldsymbol{y}|\boldsymbol{\theta})$, cheap enough that it can be repeated many thousands of times. In particular, it is possible to use forward simulations to find the set of parameter values or models that are able to closely reproduce the full data, $\boldsymbol{y}^0$, or more often some of its most informative features, $\boldsymbol{s}^0$. Approximate Bayesian Computation (ABC) (Beaumont, 2010) represents one class of such methods which, being based on a Bayesian framework, generally try to address questions regarding parameter estimation or model selection by approximately sampling the corresponding posteriors $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ and $p(M|\boldsymbol{s}^0)$.

In this thesis we will deal mainly with a particular family of intractable models: State Space Models. This very popular class of partially observed models will be described in Section 1.2.

1.2 State Space Models

State Space Models (SSMs) represent a special class of models with hidden or partially observed states. In these models the hidden states follow Markov processes, whose conditional pdf has the following property

$$p(\boldsymbol{n}_t|\boldsymbol{n}_1,\dots,\boldsymbol{n}_{t-1},\boldsymbol{\theta}) = p(\boldsymbol{n}_t|\boldsymbol{n}_{t-1},\boldsymbol{\theta}), \tag{1.2}$$

where $t \in \{1,\dots,T\}$ and $\boldsymbol{\theta}$ is a vector of static parameters. Property (1.2) implies that the future states are statistically independent of the past, upon conditioning on the present. Generally, the hidden ecological processes are coupled with an observation process according to which observed data points are conditionally independent, given the underlying states (King, 2014)

$$p(\boldsymbol{y}_t^0|\boldsymbol{n}_t,\boldsymbol{y}_1^0,\dots,\boldsymbol{y}_{t-1}^0,\boldsymbol{\theta}) = p(\boldsymbol{y}_t^0|\boldsymbol{n}_t,\boldsymbol{\theta}). \tag{1.3}$$

Typically the term SSMs is used to indicate partially observed Markov processes with continuous state spaces, while models with discretely valued states are called Hidden Markov Models (HMMs). In this thesis we focus mainly on SSMs, but most considerations apply also to HMMs.

As for most partially observed systems, the likelihood of SSMs is generally not available directly. Indeed, for such models $p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta})$, where $\boldsymbol{y}_{1:T}^0 = \{\boldsymbol{y}_1^0,\dots,\boldsymbol{y}_T^0\}$, is available analytically only if both $p(\boldsymbol{n}_t|\boldsymbol{n}_{t-1},\boldsymbol{\theta})$ and $p(\boldsymbol{y}_t^0|\boldsymbol{n}_t,\boldsymbol{\theta})$ are linear and Gaussian (Kalman, 1960). Fortunately, the Markov property (1.2) mitigates the intractability of these models, because it allows estimation of the likelihood by performing the required

T-dimensional integration efficiently. In particular, the Markov property is exploited by particle filters (Doucet and Johansen, 2009) to break down the integration problem into T sequential integration steps. These computational tools will be described in Chapter 2.

1.3 Objectives of thesis and outline

This first objective of this thesis is to compare approaches based on summary statistics, such as ABC and SL, with less approximate methods, such as particle filters. In particular, in Chapter 2 we consider simple SSMs with highly non-linear dynamics and we compare ABC and SL with two methods based on particle filters. We show that the full likelihood of these models can be highly multimodal in certain areas of the parameter space, and that approximate methods might be more robust than full likelihood methods under such circumstances. However, we illustrate that, when the likelihood is sufficiently smooth, particle filters generally outperform ABC and SL in terms of accuracy in parameter estimation.

The comparison work continues in Chapter 3, where we restrict our attention to SL and to Particle Marginal Metropolis Hastings (PMMH), a particle-filtering-based method described by Andrieu et al. (2010). We compare SL and PMMH on three complex models, using both simulated and real data. First, we consider the blowfly model of Gurney et al. (1980), modified by Wood (2010), and we fit it to Nicholson's experimental datasets (Nicholson, 1954, 1957). Second, we use the prey-predator model of Turchin and Ellner (2000) to analyse the dynamics of Fennoscandian voles. Third, we modify the compartmental model of King et al. (2008), and we fit three versions of it to a dataset concerning cholera-related deaths in the Bay of Bengal. The first two examples demonstrate that, when simulated data is used, PMMH is generally more accurate than SL for the purpose of parameter estimation. However, we show that SL can be more robust than PMMH in the presence of outliers or under model mis-specification.

While Chapters 2 and 3 focus on comparing existing methods, subsequent chapters are dedicated to improving the Synthetic Likelihood method. In Chapter 4 we propose a Maximum Synthetic Likelihood (MSL) procedure, aimed at reducing the computational cost of SL. More specifically, we describe how the synthetic likelihood can be maximized efficiently, by approximating its first and second derivatives using local regressions. We also discuss the relation between SL and the Continuous Updating Generalized Method of Moments (CUGMM), and we describe how CUGMM can make use of local regressions analogous to those used by MSL. We then compare MSL, CUGMM and a Metropolis Hastings implementation of SL (SLMH) on a sequence of simple examples. In all examples MSL and CUGMM outperfom SLMH, in terms of Mean Squared Error reduction of the parameter point estimates, as a function of the number of simulations.

In Chapter 5 we relax the distributional assumption made by SL. In particular, rather than modelling the summary statistics using a Multivariate Gaussian distribution, as suggested by Wood (2010), we propose to use a new, more flexible, density estimator. More specifically, we develop a modified version of the Empirical Saddlepoint approximation of Davison and Hinkley (1988), which we refer to as the Extended Empirical Saddlepoint approximation (ESA). We discuss how ESA's tuning parameter can be estimated by cross-validation and, within the context of SL, we prove the

consistency of the resulting parameter estimates. We then use a toy example and the Individual Based forest model of Dislich et al. (2009) to demonstrate that the new estimator clearly outperforms a Gaussian approximation, when the summary statistics are far from normal.

Chapters 2, 3 and 5 of this thesis are based on the following papers:

- Fasiolo M., Wood S. N., Approximate methods for dynamic ecological models. To appear in the *Handbook of Approximate Bayesian Computation*.
- Fasiolo M., Pya N., Wood S. N., Statistical inference for highly non-linear dynamical models in ecology and epidemiology. To appear in *Statistical Science*.
- Fasiolo M., Wood S. N., Hartig F., Bravington M. V., An Extended Empirical Saddlepoint Approximation for Intractable Likelihoods. Submitted to *Biometrics*.

The approach proposed in Chapter 4 has reached its current form only recently, hence the results discussed therein will require some additional work, before they can lead to a publication.

The work behind this thesis was performed by Matteo Fasiolo, who has benefited from the comments and ideas of the co-authors of the above papers. Hence, all the content of this thesis, and the code behind it, has been produced by Matteo Fasiolo, with the following exceptions:

- The code that calculates the exact likelihood of the discretize Ricker map (Chapter 2) was written by Simon N. Wood.
- The code that produces Figure 2-8 was written by Natalya Pya.
- The compiled object implementing the Formind forest model was provided by Florian Hartig.

CHAPTER 2

METHODS PRESENTATION AND A FIRST COMPARISON ON SIMPLE CHAOTIC MAPS

Highly non-linear, chaotic or near chaotic, dynamic models are important in fields such as ecology and epidemiology: for example, pest species and diseases often display highly non-linear dynamics. However, such models are problematic from the point of view of statistical inference. The defining feature of chaotic and near chaotic systems is extreme sensitivity to small changes in system states and parameters, and this can interfere with inference. There are two main classes of methods for circumventing these difficulties: information reduction approaches, such as Approximate Bayesian Computation or Synthetic Likelihood, and state space methods, such as Particle Markov chain Monte Carlo, Iterated Filtering or Parameter Cascading. The purpose of this Chapter is to present and compare the methods, in order to reach conclusions about how to approach inference with such models in practice. We show that state space methods can suffer multimodality problems in settings with low process noise. Information reduction methods avoid this problem but, under the correct model and with sufficient process noise. state space methods lead to substantially sharper inference than information reduction methods.

2.1 Introduction

Non-linear or near-chaotic dynamical systems represent a challenging setting for statistical inference. The chaotic nature of such systems implies that small variations in model parameters can lead to very different observed dynamics. This characteristic alone is enough to invalidate many conventional statistical methods, but in most cases additional complications are present. Firstly, the process under study is generally observed with errors. In addition, many models include a further layer of uncertainty, which we call process stochasticity. In ecology this is often environmental noise, driving the system dynamics. Process stochasticity increases the complexity of the model in a non-trivial way: apart from being unobservable, its presence makes every realized trajectory of the system essentially unique. This is particularly true for chaotic models where any amount of process noise will cause rapid divergence of two paths generated using identical parameters and initial conditions, in sharp contrast to the situation in

which dynamics lie on a stable attractor.

Developing statistical methods that can deal effectively with highly non-linear systems is not simply a matter of theoretical interest, since examples of non-linear or near-chaotic behaviour in ecological systems abound: lemmings (Kausrud et al., 2008), voles (Turchin and Ellner, 2000), mosquitos (Yang et al., 2008), moths (Kendall et al., 2005) and fish (Anderson et al., 2008). Similar degrees of non-linearity have been observed in experimental settings, for example: blowflies (Nicholson, 1957) and flour beetles (Desharnais et al., 2001).

The focus of epidemiologists often differs from that of ecologists. Both groups are concerned with explaining the persistence of the species under study, but epidemiologists and ecologists are often aiming respectively at causing and avoiding its extinction (Earn et al., 1998). Despite this divergence in objectives, the mathematical structures used to study population dynamics are often very similar. Hence, the role of non-linearities in the population dynamics of infectious diseases has attracted much attention in epidemiology as well. In the context of measles, Grenfell (1992) and Grenfell et al. (1995) describe how the interaction between seasonal forcing and observed heterogeneities, such as age structure or spatial coupling, can result in chaotic or stable dynamics, while Grenfell et al. (2002) address the issue of predictability under a time-series Susceptible Infected Recovered model. More recently King et al. (2008), Lavine et al. (2013) and Bhadra et al. (2011) use non-linear stochastic models with multiple compartments to analyse cholera, pertussis and malaria epidemics, respectively.

The relation between chaos, statistics and probability theory has been discussed by Berliner (1992) and Chan and Tong (2001), among others. We have a quite different focus, which is to review and compare the state of the art statistical methods for highly non-linear dynamic models in ecology and epidemiology, investigating the difficulties involved in their use, and attempting to establish the best approach to take in practical applications.

The chapter is organized as follows: in Section 2.2 we show that the likelihood function of simple dynamic models can be intractable in certain areas of the parameter space, while in Section 2.3 we briefly review the set of statistical methods most useful in the context of non-linear dynamic systems. Section 2.5 illustrates how these methods deal with the issue discussed in Section 2.2, while Section 2.6 discusses the relative merits and drawbacks of Synthetic Likelihood and Approximate Bayesian Computation methods. In Section 2.7 we compare the relative performance of these methodologies on an array of simple ecological maps that can show chaotic dynamics. Section 2.8 draws some initial conclusions, which we will expand upon in Chapter 3, in light of the results obtained therein.

2.2 Chaos and the likelihood function

To provide a simple example illustrating how the dynamics of an ecological model can challenge conventional statistical approaches, let us consider the noisily observed Ricker map

$$y_t \sim \text{Pois}(\phi n_t), \tag{2.1}$$

$$n_{t+1} = r n_t e^{-n_t + z_{t+1}}, \quad z_t \sim N(0, \sigma^2), \tag{2.2}$$

which can be used to describe the evolution in time t of a population n_t. Parameter r is the intrinsic growth rate of the population, controlling the dynamics of the system;

Model Name	Process Equation
Generalized Ricker	$n_{t+1} = rn_t e^{-n_t^\theta + z_t}$
Pennycuick	$n_{t+1} = \frac{rn_t}{1+e^{-a(1-n_t)}} e^{z_t}$
Maynard-Smith	$n_{t+1} = \frac{rn_t}{(1+n_t^b)} e^{z_t}$
Varley	$n_{t+1} = \begin{cases} rn_t e^{z_t} & \text{if } n_t \leq c; \\ rn_t^{1-b} e^{z_t} & \text{if } n_t > c. \end{cases}$

Table 2.1: *Five simple maps that can show chaotic dynamics. In each case* $y_t \sim Pois(\phi n_t)$ *and* $z_t \sim N(0, \sigma^2)$.

ϕ is a scale parameter. The process noise z_t can be interpreted as environmental noise.

Denote with $\boldsymbol{y}^0_{1:T} = \{\boldsymbol{y}^0_1, \boldsymbol{y}^0_2, \ldots, \boldsymbol{y}^0_T\}$ and $\boldsymbol{n}_{1:T} = \{\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_T\}$ the observations and hidden state sequence up to time T, where $\boldsymbol{y}^0_t \in \mathbb{R}^{d_y}$ and $\boldsymbol{n}_t \in \mathbb{R}^{d_n}$ for $t \in \{1, \ldots, T\}$. Equations (2.1) and (2.2) define a simple State Space Model (SSM), for which parameter inference is non-trivial: defining $\boldsymbol{\theta} = \{r, \phi, \sigma\}^T$, the likelihood $p(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta})$ is intractable in certain areas of the parameter space. For example, when $\sigma = 0$, the likelihood is analytically available, but extremely irregular for high values of r. The plot on the top left of Figure 2-1 shows a transect of the log-likelihood w.r.t. $\log(r)$, obtained using 50 observations, y_t, simulated using parameters $\log(r) = 3.8$, $\sigma = 0$ and $\phi = 10$. Given the ragged shape of the log-likelihood, estimating the parameters by maximum likelihood would be very challenging computationally, while having only limited theoretical motivation. Similarly, any standard MCMC algorithm targeting the parameter posterior distributions would hardly mix at all. This behaviour is generic to highly non-linear dynamic systems: Figure 2-1 shows likelihood transects for three more dynamic models, defined in Table 2.1, any of which could be used to make the same points made using the Ricker map.

Figure 2-1 reflects the extreme sensitivity of the likelihood of chaotic models to minuscule changes in parameters or process noise. The bifurcation diagram of the Ricker map (grey) shows the possible long term values n_t of the map, as a function of $\log(r)$. While the trajectories oscillate between two values for $\log(r) \approx 2$, increasing $\log(r)$ above 2.5 leads to a sequence of closely spaced bifurcations, each doubling the periodicity of the map. This period-doubling cascade has a direct effect on the likelihood. Notice that this function is smooth again for values of $\log(r)$ where stable periodic oscillations are recovered. Further increasing $\log(r)$ leads to more period-doubling phases and eventually to chaos.

Figure 2-2 illustrates the origin of this extreme multimodality. We generated two state paths, $\boldsymbol{n}_{1:50}$, using $\sigma = 0$ and the same initial value $n_1 = 7$, but different values of $\log(r)$: 3.8 (black) and 3.799 (red). The two paths are close to each other for the first steps, but the mismatch between them increases with time, and by $t = 15$ the peaks and troughs of the paths do not coincide any more. This sort of divergence of neighbouring

trajectories is the defining feature of chaotic dynamics (measured formally in terms of Lyapunov exponents).

The choice $\sigma = 0$ is quite peculiar. What does the likelihood look like when the process dynamics are stochastic?

Algorithm 1 Sequential Importance Re-Sampling (SIR) for likelihood estimation

This algorithm, originally proposed by Gordon et al. (1993), exploits the Markov property to approximate integral (2.3) in T sequential steps. Let $\boldsymbol{n}_0^{1:M}$ be a sample of particles from the prior distribution $p(\boldsymbol{n}_0)$. Then $p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta})$ is estimated as follows.
For $t = 1$ to T:

1: For $i = 1, \ldots, M$:
propagate the i-th particle forward

$$\boldsymbol{n}_t^i \sim p(\boldsymbol{n}_t^i|\boldsymbol{n}_{t-1}^i, \boldsymbol{\theta}),$$

and weight it using the t-th observation

$$w^i = p(\boldsymbol{y}_t^0|\boldsymbol{n}_t^i, \boldsymbol{\theta}).$$

2: Estimate the t-th likelihood component

$$\hat{p}(\boldsymbol{y}_t^0|\boldsymbol{y}_{1:t-1}^0, \boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} w^i.$$

3: Re-sample $\boldsymbol{n}_t^{1:M}$ with replacement, using probabilities proportional to $\boldsymbol{w}^{1:M}$.
4: Estimate the likelihood by using

$$\hat{p}(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}) = \hat{p}(\boldsymbol{y}_1^0|\boldsymbol{\theta}) \prod_{t=2}^{T} \hat{p}(\boldsymbol{y}_t^0|\boldsymbol{y}_{1:t-1}^0, \boldsymbol{\theta}).$$

As explained in Chapter 1, the likelihood of SSMs, $p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})$, must be obtained by integration

$$\begin{aligned} p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}) &= \int p(\boldsymbol{y}_{1:T}^0, \boldsymbol{z}_{1:T}|\boldsymbol{\theta})\, d\boldsymbol{z}_{1:T} \\ &= \int p(\boldsymbol{y}_{1:T}^0, \boldsymbol{n}_{1:T}|\boldsymbol{\theta})\, d\boldsymbol{n}_{1:T}, \end{aligned} \tag{2.3}$$

where the second integral is generally the more computationally tractable version. The plot on the right of Figure 2-2 shows a transect of the estimated log-likelihood of the Ricker map w.r.t. parameter $\log(r)$, obtained using the Sequential Importance Re-sampling (SIR) particle filter with 5×10^5 particles. Algorithm 1 details the main steps of this procedure, but we refer to Doucet and Johansen (2009) for a more detailed introduction to particle filters. The observed path $\boldsymbol{y}_{1:50}^0$ has been simulated using $\log(r) = 3.8$, $\sigma = 0.3$ and $\phi = 10$. In sharp contrast with the deterministic case (Figure 2-1), it appears that the injection of process noise ($\sigma > 0$) into the system has made the likelihood smooth and unimodal. At this point several questions arise: is the likelihood really smooth, as Figure 2-2 suggests, or is it possible that the particle filter is hiding

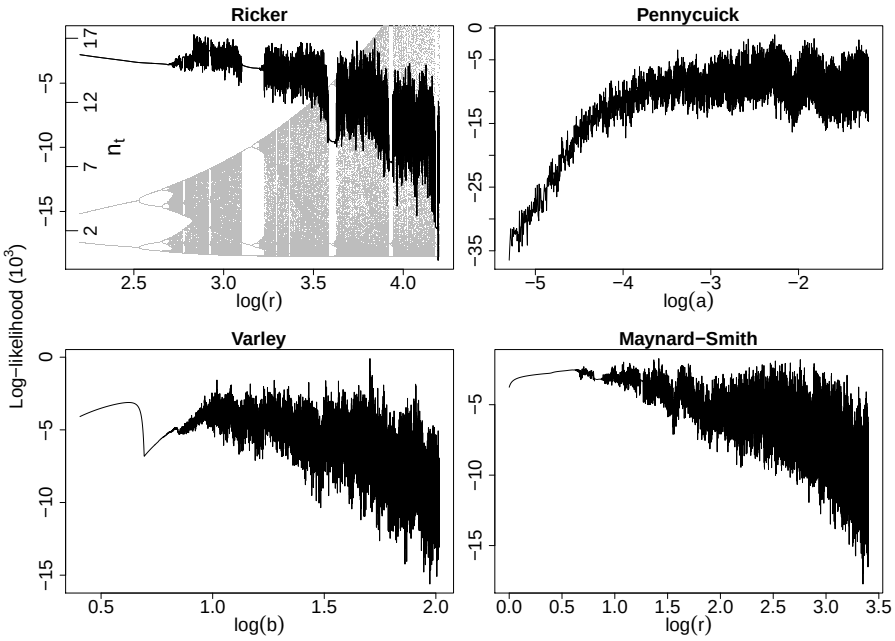

Figure 2-1: *Slices of the log-likelihoods of four simple models w.r.t different parameters (black). In each case $\sigma = 0$, hence the likelihoods are analytically available. For the Ricker map a bifurcation diagram is included (gray).*

the extreme multimodality of Figure 2-1, so that what we observe in Figure 2-2 is an artefact of Monte Carlo integration? If the likelihood is indeed smooth, how did the transition from Figure 2-1 to Figure 2-2 occur? How much noise σ should be present in order to obtain a smooth likelihood?

Checking the reliability of the estimates provided by a particle filter is difficult because, for non-linear and/or non-Gaussian models, Monte Carlo or numerical integration are the only ways to get an approximation to (2.3). To obtain a benchmark against which to compare the estimates of the likelihood provided by the filter, we have therefore discretized the state space of the Ricker map in 500 intervals. In this way we can calculate the likelihood exactly, since the integrations are replaced by efficiently computable summations over all the possible values of the states, as detailed in Appendix A.1. Obviously, we do not propose discretization as a viable alternative to particle filters, but we want to use a discretized SSM to compare the performance of a particle filter with the true likelihood. It is interesting to check whether the injection of any amount of noise is sufficient to smooth the likelihood, or whether there is a slow transition from the intractable likelihood shown in Figure 2-1 to the unimodal case of Figure 2-2. Perhaps unsurprisingly, Figure 2-3 shows that the latter is the case, since as we reduce the process noise the likelihood becomes firstly multimodal and then (for any practical purpose) non-differentiable for very low σ. SIR estimate of the likelihood deteriorates as multi-modality sets in: we will investigate this more fully in Section 2.5. Notice that the likelihood estimates shown here were obtained by running SIR on the discretized version of the Ricker map.

This suggests that there is an area of the parameter space, corresponding to high $\log(r)$ and low σ, where the likelihood is essentially intractable. For practical purposes

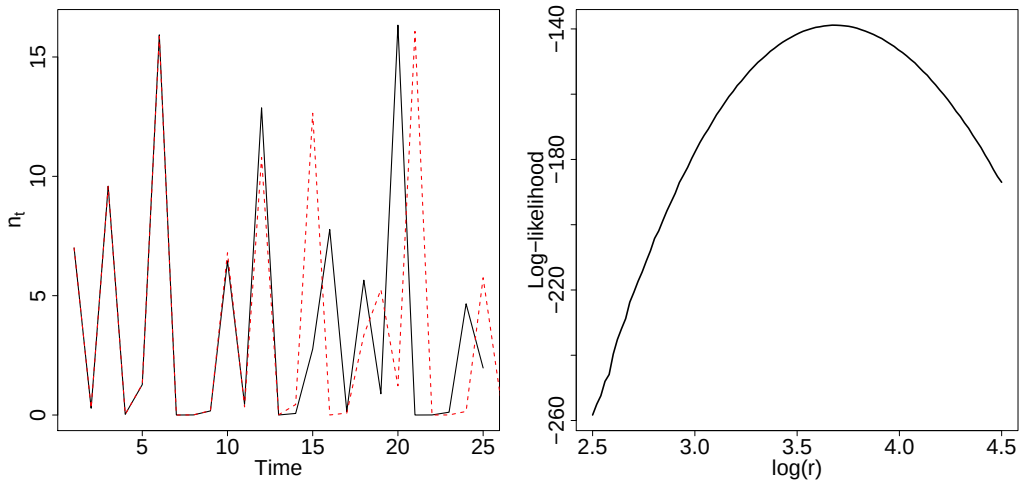

Figure 2-2: *Left: two trajectories* $\boldsymbol{n}_{1:T}$ *of the hidden state, generated using the same initialization, but slightly different values of* $\log(r)$*. Right: transect w.r.t.* $\log(r)$ *of the log-likelihood of the Ricker map with* $\sigma = 0.3$*, estimated using the SIR particle filter. The irregularities at* $\log(r) \approx 2.6$ *are due to Monte Carlo noise.*

it is therefore important to compare the robustness of alternative statistical methods across the parameter space, and to understand how alternative methods behave in the face of this difficulty. In particular, we need to avoid the possibility of concluding that a system's dynamics are relatively stable and noisy, not because they really are, but because that is the only case in which the likelihood is numerically tractable.

2.3 Available statistical methods

The literature contains two main classes of statistical methods for non-linear dynamical systems:

1. Information reduction: methods that discard the information in the data that is most sensitive to extreme divergence of trajectories, so that fitting objectives become more regular. Two methodologies belonging to this group will be described in Section 2.3.1.

2. State space: these work on the hidden states, $\mathbf{n}_{1:T}$, in order to estimate model parameters and/or the hidden states themselves. Some of these approaches work without modifying the model or the data in any way, by using advanced computational techniques based on particle filtering. We describe two members of this family in Section 2.3.2.

Given that the main purpose of this chapter and of Chapter 3 is to consider the applicability and relative performance of these methods in the context of near-chaotic dynamic systems, we will skip over the technical detail whenever they are not essential for the discussion. Obviously our analysis is by no means exhaustive, as we do not examine all the approaches that could be applied in this context. In Section 2.3.3 we briefly describe some of the alternatives to the methods considered here.

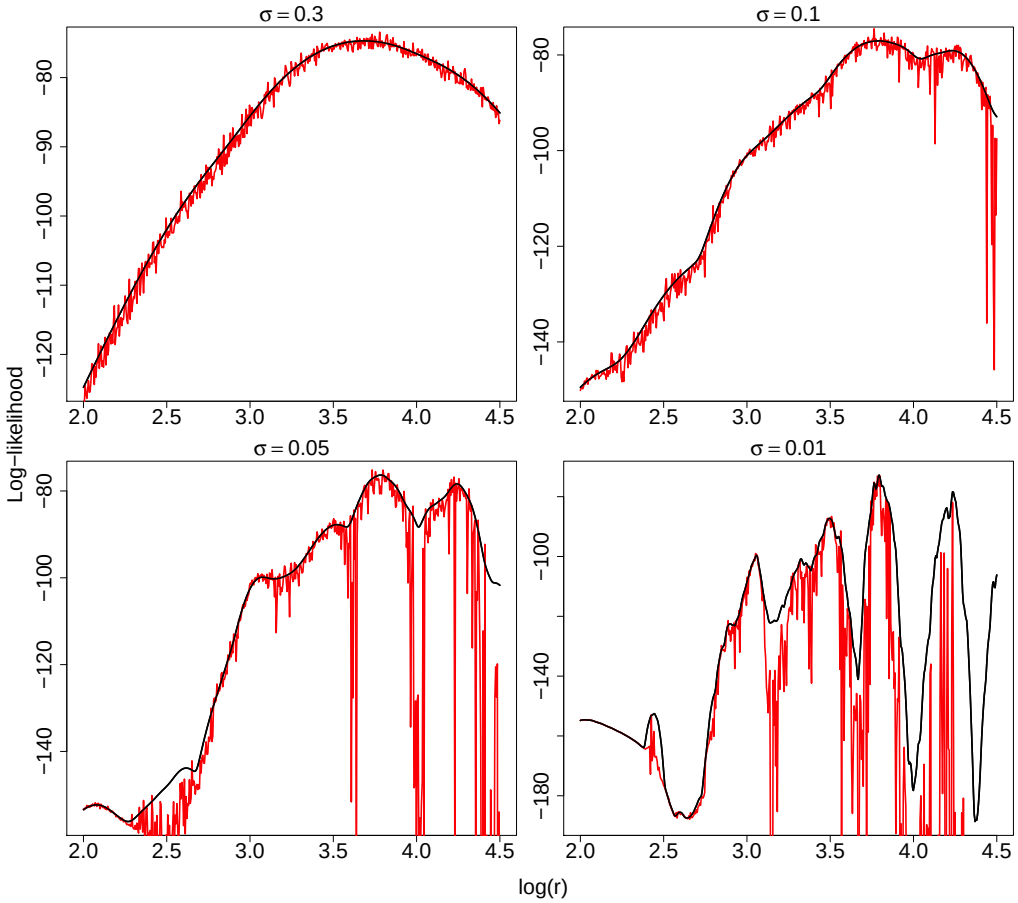

Figure 2-3: *Transects of the true log-likelihood (black) of the discrete Ricker map w.r.t.* $\log(r)$ *for decreasing values of* σ*. The red lines are SIR's estimates, using 1000 particles.*

2.3.1 Approaches based on information reduction

Since the trajectories of near chaotic systems are extremely sensitive to perturbations of parameters or system state, statistical methods that rely on recovering the true system state face a difficult task. At the same time it is often the case that the true state itself is only a nuisance for parameter estimation, and discarding some information regarding the particular observed trajectory might ease the inferential process.

To make this point clearer consider again the Ricker paths in Figure 2-2. Even though the two trajectories, which we indicate with $\boldsymbol{y}_{1:T}$ and $\boldsymbol{x}_{1:T}$, are very different in terms of Euclidean distance $||\boldsymbol{y}_{1:T} - \boldsymbol{x}_{1:T}||$, it is clear that they share some common features. A way around the impossibility of replicating the observed path, even when the simulations use the true or "best-fitting" parameters and initial value, is focusing on the relationship between some characteristic features of the data and the unknown parameters. One way of doing this is to transform the observed and simulated data into a set of summary statistics and to base subsequent inferences on these.

In the following we denote with $\boldsymbol{y}^0_{1:T}$ the observed path, and with $\boldsymbol{s}^0 = S(\boldsymbol{y}^0_{1:T})$ the vector of observed summary statistic. Often methods based on summary statistics involve two main approximations of the likelihood function. The first is implied by the use of $p(\boldsymbol{s}^0|\boldsymbol{\theta})$ as a proxy for $p(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta})$. The second approximation arises from the fact that $p(\boldsymbol{s}^0|\boldsymbol{\theta})$ itself is generally not available analytically and hence it has to be approximated or estimated by simulation.

We will focus on two approaches based on information reduction: Approximate Bayesian Computation (ABC) (Beaumont et al., 2002) (Fearnhead and Prangle, 2012) and Synthetic Likelihood (SL) (Wood, 2010).

Approximate Bayesian Computation

The main purpose of ABC algorithms is approximating the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y}^0_{1:T}) \propto p(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is the prior distribution of the model parameters, when the likelihood $p(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta})$ is unavailable or intractable. Given that the data is often transformed into a vector of summary statistics, these methods are generally aiming at sampling from $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ rather than $p(\boldsymbol{\theta}|\boldsymbol{y}^0_{1:T})$.

An elementary ABC algorithm iterates the following rejection procedure (Toni et al., 2009):

1. Sample a vector of parameters $\boldsymbol{\theta}^i$ from $p(\boldsymbol{\theta})$.
2. Simulate a path $\boldsymbol{y}^i_{1:T}$ from the model $p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}^i)$.
3. Transform $\boldsymbol{y}^i_{1:T}$ to a vector of summary statistics $\boldsymbol{s}^i = S(\boldsymbol{y}^i_{1:T})$.
4. Compare $\boldsymbol{s}^i$ to the observed statistics $\boldsymbol{s}^0$ using a pre-specified distance measure $d(\cdot,\cdot)$. If $d(\boldsymbol{s}^i, \boldsymbol{s}^0) \leq \epsilon$, where $\epsilon \geq 0$, accept $\boldsymbol{\theta}^*$ otherwise reject it.

The output of this algorithm will be distributed according to

$$p(\boldsymbol{\theta})p\{d(\boldsymbol{s}, \boldsymbol{s}^0) < \epsilon|\boldsymbol{\theta}\} \propto p\big\{\boldsymbol{\theta}|d(\boldsymbol{s}, \boldsymbol{s}^0) < \epsilon\big\},$$

which approximates the posterior density, $p(\boldsymbol{\theta}|\boldsymbol{s}_0)$, for sufficiently small ϵ.

In order to obtain higher computational efficiency it is possible to implement ABC through Monte Carlo Markov Chain (MCMC) or Sequential Monte Carlo (SMC) algorithms, such as those described by Beaumont (2010). In Section 2.7 we use the MCMC

implementation of ABC which was firstly proposed by Marjoram et al. (2003), and that uses the following iteration:

1. Let $\boldsymbol{\theta}_{i-1}$ be current position in the parameter space and propose a new parameter vector $\boldsymbol{\theta}^*$, according to the transition kernel $K(\boldsymbol{\theta}^*|\boldsymbol{\theta})$.

2. Simulate a single dataset $\boldsymbol{y}^*_{1:T} \sim p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}^*)$ and transform it to a vector of summary statistics $\boldsymbol{s}^* = S(\boldsymbol{y}^*_{1:T})$.

3. If $d(\boldsymbol{s}^0, \boldsymbol{s}^*) \leq \epsilon$ proceed to step 4, otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$ and return to step 1.

4. Calculate the acceptance probability

$$\alpha = min\left\{1, \frac{p(\boldsymbol{\theta}^*)K(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})K(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}$$

5. Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*$ with probability α, otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$. Return to step 1.

This algorithm, which we refer to as ABC-MCMC, generally leads to an higher acceptance rate than the rejection sampler described above, but it produces a dependent sample.

Synthetic Likelihood

Similarly to ABC, Synthetic Likelihood (SL) (Wood, 2010) can be used for problems where the likelihood is intractable, but it is still possible to simulate from the model. The main difference between ABC and SL is how $p(\boldsymbol{s}^0|\boldsymbol{\theta})$ is approximated. While ABC does not rely on any distributional assumption on $\boldsymbol{s}$, SL assumes that

$$S(\boldsymbol{y}) \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta), \tag{2.4}$$

at least approximately. In addition, while ABC methods explicitly aim at sampling from the approximate posterior $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$, SL uses assumption (2.4) to provide a parametric approximation, $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, to $p(\boldsymbol{s}^0|\boldsymbol{\theta})$. This synthetic likelihood can then be used within a Bayesian or a classical context.

A point-wise estimate of the synthetic likelihood at $\boldsymbol{\theta}$ can be obtained as follows:

1. Simulate N datasets $\boldsymbol{y}^1_{1:T}, \ldots, \boldsymbol{y}^N_{1:T}$ from the model $p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})$.

2. Transform each dataset $\boldsymbol{y}^i_{1:T}$ into a d-dimensional vector of summary statistics $\boldsymbol{S}^i = S(\boldsymbol{y}^i_{1:T})$.

3. Estimate the sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, using standard estimators

$$\hat{\boldsymbol{\mu}}_\theta = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{S}^i,$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \frac{1}{N-1}\sum_{i=1}^{N} \left\{\boldsymbol{S}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right\}\left\{\boldsymbol{S}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right\}^T,$$

or possibly more robust alternatives.

4. Estimate the synthetic likelihood

$$\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}}|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})\right\}.$$

Hence, SL explicitly provides point estimates of $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, which can, for instance, be used within a Metropolis-Hastings algorithm, targeting an approximation to $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$. We refer to this procedure as the Synthetic Likelihood Metropolis Hastings (SLMH) algorithm. Alternatively, the point estimates can be used within an optimizer aiming at maximizing the synthetic likelihood (see Chapter 4).

Notice that there exists a strong relationship between SL and the simulation-based approach of Diggle and Gratton (1984), who proposed to estimate the full likelihood $p(\boldsymbol{y}^0|\boldsymbol{\theta})$ point-wise, by simulating data from the model and approximating its distribution using a non-parametric density estimator.

2.3.2 State space methods

If discarding information through the use of summary statistics is not desirable, then it is necessary to deal with the hidden states explicitly. As previously stated, calculating the likelihood of SSMs involves integrating the hidden states $\boldsymbol{n}_{1:T}$ out of the joint density $p(\boldsymbol{y}^0_{1:T}, \boldsymbol{n}_{1:T}|\boldsymbol{\theta})$. The SIR particle filter can be used to obtain a Monte Carlo estimate of the likelihood, by employing a sequential integration scheme. The use of a sequential approach allows filters to direct the simulated trajectories of the hidden states toward values that are consistent with the observations. This feature is particularly attractive in the context of near-chaotic models, where simulated paths diverge rapidly (recall Figure 2-2). In this work we mainly focus on algorithms based on the SIR scheme, but many other approaches are available. For example, it is possible to use algorithms that sample directly from the joint posterior density of parameters and hidden states, thus circumventing the estimation of the likelihood. For overviews, see Andrieu et al. (2010) and Doucet et al. (2000).

Here we consider three state space approaches, two of which are based on particle filtering. In particular, we describe in turn: a sampler belonging to the family of Particle Markov chain Monte Carlo (PMCMC) methods (Andrieu et al., 2010), the Iterated Filtering (IF) algorithm (Ionides et al., 2011) and the Parameter Cascading approach, proposed by Ramsay et al. (2007).

Particle Marginal Metropolis-Hastings sampler

Filters such as the SIR algorithm can provide point estimates, $\hat{p}(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta})$, of the likelihood, which ideally converge to the true likelihood as the number of simulations increases. Andrieu et al. (2010) proposed to use these estimates of the likelihood to set up a Particle Marginal Metropolis-Hastings (PMMH) algorithm, which can be used to sample from the posterior distribution of the parameters. The algorithm is formed by the following steps:

- Step 1: Initialization $i = 0$.
  Given an estimate or a guess of the parameters $\boldsymbol{\theta}_0$, estimate the likelihood $p(\boldsymbol{y}^0_{1:T}|\boldsymbol{\theta}_0)$ using a particle filter.
- Iteration $i \geq 1$:

1. Sample a new vector of parameters $\boldsymbol{\theta}^*$ from a transition kernel $K(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1})$.
2. Using a particle filter estimate the likelihood $\hat{p}(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}^*)$.
3. With probability

$$min\left\{1, \frac{\hat{p}(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\hat{p}(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}_{i-1})p(\boldsymbol{\theta}_{i-1})} \frac{K(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}^*)}{K(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1})}\right\},$$

set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$.

This algorithm is exact in the sense that, despite the use of noisy estimates of $p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta})$ in the acceptance step, it will generate a dependent sample from $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T}^0)$. The conditions under which this occurs are discussed in Section 2.4.

Iterated filtering

The IF algorithm uses particle filters to provide approximate Maximum Likelihood estimates of the unknown parameters. As shown by Ionides et al. (2006), by including the unknown parameters in the state space and running a filtering operation, it is possible to estimate the gradient of the likelihood function, which can then be used within an optimization routine. In more detail, Ionides et al. (2006) treat the parameters as if they were following a multivariate random walk

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\psi}_t \text{ with } \boldsymbol{\psi}_t \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}). \tag{2.5}$$

With this choice we have that

$$E(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \boldsymbol{\theta}_{t-1},\ Var(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) = \sigma^2\boldsymbol{\Sigma},$$

$$E(\boldsymbol{\theta}_0) = \hat{\boldsymbol{\theta}} \text{ and } Var(\boldsymbol{\theta}_0) = c^2\sigma^2\boldsymbol{\Sigma},$$

where σ and c^2 are two variance multipliers, $\hat{\boldsymbol{\theta}}$ is an initial estimate, while $\boldsymbol{\Sigma}$ is typically a diagonal matrix, giving the respective scale of the parameters.

The main result underlying the IF algorithm is

$$\lim_{\sigma^2\to 0} \sum_{t=1}^{T} \boldsymbol{V}_t^{-1}(\hat{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1}) = \nabla \log p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta}), \tag{2.6}$$

where

$$\hat{\boldsymbol{\theta}}_t = E(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}^0) \text{ and } \boldsymbol{V}_t = Var(\boldsymbol{\theta}_t|\boldsymbol{y}_{1:t}^0),$$

can be estimated using the SIR particle filter. The IF algorithm is composed of the following steps:

- Choose initial value $\hat{\boldsymbol{\theta}}_0^{(0)}$, parameters σ^2, c^2, $\boldsymbol{\Sigma}$, $\alpha \in (0,1)$ and number of iterations M.
- Iterate for j in $1, \ldots, M$:
  1. Set $\sigma_j = \alpha^{j-1}$. Estimate $\hat{\boldsymbol{\theta}}_t^{(j)}$ and $\boldsymbol{V}_t^{(j)}$, for $t = 1, \ldots, T$, using a particle filter.

2. Update the parameter estimate

$$\hat{\boldsymbol{\theta}}_0^{(j+1)} = \hat{\boldsymbol{\theta}}_0^{(j)} + \boldsymbol{V}_1^{(j)} \sum_{t=1}^{T} (\boldsymbol{V}_t^{(j)})^{-1} (\hat{\boldsymbol{\theta}}_t^{(j)} - \hat{\boldsymbol{\theta}}_{t-1}^{(j)}).$$

- Then $\hat{\boldsymbol{\theta}}_0^{(M+1)}$ is an approximate Maximum Likelihood estimate of the parameters.

Notice that, as long as $\sigma > 0$, IF will not be fitting the original model, which will be recovered as $\sigma \to 0$. Ionides et al. (2011) give results concerning the theoretical foundation of IF and describe how slowly σ has to decrease to assure convergence.

Parameter Cascading

In the context of Ordinary Differential Equations (ODEs), Ramsay et al. (2007) proposed an approach to parameter estimation which can be adapted to the discrete-time models, such as the Ricker map. The estimation procedure is a nested optimization problem with three levels. Given λ and $\boldsymbol{\theta}$, the hidden states are estimated by minimizing an inner criterion

$$\begin{aligned} \boldsymbol{n}_{1:T}^{\boldsymbol{\theta}} &= \underset{\boldsymbol{n}_{1:T}}{\operatorname{argmin}}\, J(\boldsymbol{n}_{1:t}|\boldsymbol{\theta}, \lambda) \\ &= \underset{\boldsymbol{n}_{1:T}}{\operatorname{argmin}} \Big\{ -\textstyle\sum_{t=1}^{T} \log p(\boldsymbol{y}_t^0|\boldsymbol{n}_t, \boldsymbol{\theta}) + \lambda \psi(\boldsymbol{n}_{1:T}|\boldsymbol{\theta}) \Big\}, \end{aligned}$$

where

$$\psi(\boldsymbol{n}_{1:T}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \big\{ \boldsymbol{n}_t - E(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{\theta}) \big\}^2,$$

quantifies deviations of the estimated state from the model, while λ determines the trade-off between data fitting and model compliance. The parameters are estimated using the higher level criterion

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, H(\boldsymbol{\theta}|\boldsymbol{n}_{1:T}^{\boldsymbol{\theta}_0}, \lambda) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \big\{ -\textstyle\sum_{t=1}^{T} \log p(\boldsymbol{y}_t^0|\boldsymbol{n}_t^{\boldsymbol{\theta}_0}, \boldsymbol{\theta}) \big\}. \end{aligned}$$

A further level can be added in which an outer grid search is used to select λ. This method is especially useful for exploring multimodality problems in Section 2.5.

2.3.3 Alternative approaches

The methods described in the preceding sections represent a subset of those that could be used in the context of parameter estimation for non-linear state space models. Here we discuss some of the alternatives, describe their relation with the methods described above and detail our reasons for not including them in this work.

There exist a large variety of particle-filtering-based methods that can be used to obtain approximate Maximum Likelihood (ML) estimates of the static parameters, such as Andrieu et al. (2005), Andrieu and Doucet (2003), Malik and Pitt (2011), Poyiadjis et al. (2011) and Nemeth et al. (2013). IF belongs to this class of methods, and we

chose to include it, rather than some of the alternatives, in this work because (i) it is theoretically justified, as detailed in Ionides et al. (2011), (ii) it is has been tested on a variety of complex models, such as those described in King et al. (2008), He et al. (2010) and Bhadra et al. (2011), which are of direct interest to applied researchers in ecology and epidemiology, and (iii) the computational cost of a score function estimate is $O(M)$ in the number of particles, which, to our best knowledge, is the state of the art. Hence we argue that, by including IF, this work should adequately cover this class of methods.

Notably, this work does not include MCMC methods for parameter identification, such as those proposed by Carlin et al. (1992), Geweke and Tanizaki (2001), Polson et al. (2008) and Niemi and West (2010). One reason for this is that highly non-linear models, such as those considered here, are often characterized by strong dependencies between states and static parameters. Under such circumstances, implementing an efficient MCMC sampler requires the design of adequate conditional proposal densities, which is not trivial for non-linear non-Gaussian models (Andrieu et al., 2010; Kantas et al., 2014). In addition, the cholera model presented in Chapter 3 is a discretized version of a continuous time model, where the discretization error was limited by using a large number of intermediate states between each pair of observations. Sampling this enlarged state space using standard MCMC methods would be challenging, because the convergence rate of such schemes can be arbitrarily slow if the amount of augmentation is large (Roberts and Stramer, 2001). With the exception of Parameter Cascading, all the methods described in our work are less affected by this problem, because the intermediate states are simply simulated forward using $p(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \boldsymbol{\theta})$. This "plug-and-play" property is one of the reasons behind popularity of these methods (Ionides et al., 2011).

Apart from PMCMC and MCMC algorithms, the methods proposed by Kitagawa (1993) and Liu and West (2001) could also be used to sample the posterior distribution of $\boldsymbol{\theta}$. Analogously to IF, these filters include the parameters in the state space, and perturb them using an artificial noise process. Even though Liu and West (2001) counteract the resulting over-dispersion of the posterior by shrinking the perturbed parameters toward their mean, this does not entirely eliminate the information loss, if the posterior is far from Gaussian. Hence, in this work we preferred to target $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T})$ using PMMH, because of the convergence guarantees detailed in Andrieu and Roberts (2009). However, the computational cost of PMMH is fairly high, and the filter of Liu and West (2001) might be able to sample a close approximation to $p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T})$, using far fewer filtering operations.

Finally, the versions of IF and PMMH used here are based on the SIR algorithm, as described in Gordon et al. (1993) and Doucet et al. (2000). More sophisticated filters, such as those proposed by Pitt and Shephard (1999) and Klaas et al. (2012), might provide more accurate estimates of the likelihood, or of $\nabla p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})$ in the context of IF. Similarly, it might be possible to improve upon the MCMC implementation of ABC and SL used in Section 2.7, by using more sophisticated SMC samplers (Toni et al., 2009) or Gaussian Processes (Meeds and Welling, 2014), respectively. We do not explore these possibilities here, because doing so would increase the complexity of this work, without adding much to its main results.

2.4 Synthetic Likelihood and exact-approximate methods

As discussed in Sections 2.3.1 and 2.3.2, SLMH and PMMH use estimates of, respectively, $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ and $p(\boldsymbol{y^0}|\boldsymbol{\theta})$ at each Metropolis Hastings (MH) step. In this section we discuss the conditions under which these samplers will target the corresponding posterior densities exactly. In particular, in Section 2.4.1 we describe exact-approximate samplers, which provide samples from the correct posterior, despite being based on noisy likelihood estimates. This class of algorithms includes PMMH but, for reasons explained in Section 2.4.2, not SLMH. In Section 2.4.3 we propose a simulation-based approach meant to alleviate this problem.

2.4.1 PMMH as an exact-approximate sampler

The use of noisy likelihood estimates within MH algorithms appeared firstly in Lin et al. (2000) and Beaumont (2003), while the theoretical conditions under which the resulting noisy MH sampler targets the correct posterior distribution were later described in Andrieu and Roberts (2009). To explain their results, we consider PMMH and assume that $\hat{p}(\boldsymbol{y}^0|\boldsymbol{\theta}) = p(\boldsymbol{y}^0|\boldsymbol{\theta})z$, where z is a strictly positive random variable, with density $p(z|\boldsymbol{\theta})$. If this noisy estimate of the target density is used within MH, the acceptance ratio becomes

$$\alpha = \frac{\hat{p}(\boldsymbol{y}^0|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\hat{p}(\boldsymbol{y}^0|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{K(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{K(\boldsymbol{\theta}^*|\boldsymbol{\theta})} = \frac{z^* p(\boldsymbol{y}^0|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{z\, p(\boldsymbol{y}^0|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{K(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{K(\boldsymbol{\theta}^*|\boldsymbol{\theta})},$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ are the current and the proposed parameter vectors, while $K(\cdot|\cdot)$ is transition kernel. Notice that the acceptance ratio can be written as

$$\alpha = \frac{z^* p(\boldsymbol{y}^0|\boldsymbol{\theta}^*)p(z^*|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{z\, p(\boldsymbol{y}^0|\boldsymbol{\theta})p(z|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{p(z\,|\boldsymbol{\theta})K(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(z^*|\boldsymbol{\theta}^*)K(\boldsymbol{\theta}^*|\boldsymbol{\theta})},$$

which shows that the sampler is targeting a density proportional to $z\, p(\boldsymbol{y}^0|\boldsymbol{\theta})p(z|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Now, if we assume that $E(z|\boldsymbol{\theta}) = E(z) = c = \text{const}$, which implies that the bias of $\hat{p}(\boldsymbol{y}^0|\boldsymbol{\theta})$ is constant, we have that the marginal density of $\boldsymbol{\theta}$ is proportional to

$$\int z\, p(\boldsymbol{y}^0|\boldsymbol{\theta})p(\boldsymbol{\theta})p(z|\boldsymbol{\theta})\, dz = c\, p(\boldsymbol{y}^0|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

which implies that the MH algorithm is sampling the correct posterior, even though noisy estimates of the target density are being used at each step. Hence, this sampler, us can be said to be "exact-approximate". It is important to highlight that the sampler will target the correct density, as long as the the expected value of the noise z does not vary with $\boldsymbol{\theta}$. Hence, other features of the distribution of z, such as higher order moments, might depend on $\boldsymbol{\theta}$ without compromising the exactness of the sampler.

Andrieu and Roberts (2009) consider unbiased likelihood estimators, which are often obtained through importance sampling or particle filters. They call the resulting MH samplers "pseudo-marginal", to indicate that the likelihood estimates are often obtained by marginalizing a joint density using Monte Carlo methods. This is precisely what happens under PMMH, where the hidden states, $\boldsymbol{n}_{1:T}$, are approximately integrated out of the joint density, $p(\boldsymbol{y}^0_{1:T}, \boldsymbol{n}_{1:T}|\boldsymbol{\theta})$, to provide unbiased likelihood estimates. This demonstrates that PMMH is an exact approximate algorithm targeting $p(\boldsymbol{\theta}|\boldsymbol{y}^0)$.

2.4.2 Obtaining unbiased synthetic likelihood estimates

In order to simplify the notation, in the remaining part of Section 2.4 we temporarily suppress the dependencies on $\boldsymbol{\theta}$ and the use of the subscript SL to indicate the synthetic likelihood. Hence, we indicate the synthetic likelihood estimates with $p(\boldsymbol{s}^0)$ rather than with $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. As we discuss here, the pointwise synthetic likelihood estimates, $\hat{p}(\boldsymbol{s}^0)$, obtained as described in Section 2.3.1, are generally biased, hence an SLMH algorithm based on them cannot be considered an exact-approximate sampler.

We start by assuming that the summary statistics are normally distributed, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. From Braun and McAuliffe (2010) we have that

$$E\big\{\log|\hat{\boldsymbol{\Sigma}}|\big\} = d\log 2 + \log|\boldsymbol{\Sigma}| + \sum_{i=1}^{d}\psi\left(\frac{N-i}{2}\right) - d\log(N-1).$$

where ψ is the digamma function and, as before, N is the number of simulations used to estimate mean and covariance matrix. In addition, we have that (Hartlap et al., 2007)

$$\begin{aligned} E\big[(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}})\big] &= E\big[(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}})^T\big]E\big(\hat{\boldsymbol{\Sigma}}^{-1}\big)E\big[(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}})\big] \\ &= \tfrac{N-1}{N-d-2}(\boldsymbol{s}^0-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}^0-\boldsymbol{\mu}). \end{aligned}$$

where the first equality holds due to Basu's Theorem (Basu, 1955). Hence, we can derive an unbiased estimator of $\log p(\boldsymbol{s}^0)$, which is

$$\begin{aligned} \log\tilde{p}(\boldsymbol{s}^0) = &-\tfrac{d}{2}\log(2\pi) - \tfrac{1}{2}\left\{\log|\hat{\boldsymbol{\Sigma}}| + d\log(N-1) - d\log 2 - \textstyle\sum_{i=1}^{d}\psi\left(\frac{N-i}{2}\right)\right\} \\ &- \tfrac{1}{2}\tfrac{N-d-2}{N-1}(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{s}^0-\hat{\boldsymbol{\mu}}). \end{aligned} \tag{2.7}$$

Unfortunately, this does not direclty lead to an unbiased estimator of $p(\boldsymbol{s}^0)$, due to the non-linearity of the transformation. In particular, let us make the assumption, to be justified later, that $\log\tilde{p}(\boldsymbol{s}^0)$ is approximately normally distributed, with mean $\log p(\boldsymbol{s}^0)$ and variance σ^2. This implies that

$$E\{\tilde{p}(\boldsymbol{s}^0)\} = E\{e^{\log p(\boldsymbol{s}^0)+v}\} = p(\boldsymbol{s}^0)e^{\frac{\sigma^2}{2}},$$

where $v \sim N(0,\sigma^2)$. Hence, if σ^2 varies with $\boldsymbol{\theta}$, the bias of the synthetic likelihood estimates will also depend on the parameters, and the corresponding SLMH algorithm will not target the correct posterior density. We mitigate this issue in Section 2.4.3, where we correct $\tilde{p}(\boldsymbol{s}^0)$ using a simulation-based approach.

2.4.3 Bootstrapped synthetic likelihood

Under the assumptions detailed in the previous section, an unbiased estimator of $p(\boldsymbol{s}^0)$ is $\bar{p}(\boldsymbol{s}^0) = \tilde{p}(\boldsymbol{s}^0)e^{-\frac{\sigma^2}{2}}$. The problem with this estimator is that $\sigma^2 = var\{\log\tilde{p}(\boldsymbol{s}^0)|\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which are unknown. Following a parametric boostrap approach (Efron and Tibshirani, 1994), we propose to approximate $var\{\log\tilde{p}(\boldsymbol{s}^0)|\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ with an estimate of $var\{\log\tilde{p}(\boldsymbol{s}^0)|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}\}$. To reduce the computational effort, we exploit the fact that, if the summary statistics are normally distributed, then $\hat{\boldsymbol{\mu}} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}/N)$ and $(N-1)\hat{\boldsymbol{\Sigma}} \sim \text{Wish}(\boldsymbol{\Sigma},N-1)$.

Given a sample $\boldsymbol{S}_1, \dots, \boldsymbol{S}_N$ of simulated summary statistics, with empirical mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$, we obtain an approximately unbiased estimate of the synthetic likelihood as follows:

1. simulate M mean vectors, $\tilde{\boldsymbol{\mu}}_1, \dots, \tilde{\boldsymbol{\mu}}_M$, and covariance matrices, $\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_M$, from $N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}/N)$ and $\text{Wish}(\hat{\boldsymbol{\Sigma}}/(N-1), N-1)$, respectively.

2. Given the M mean vectors and covariance matrices, calculate the corresponding densities $\tilde{p}(\boldsymbol{s}^0)_1, \dots \tilde{p}(\boldsymbol{s}^0)_M$.

3. Estimate $p(\boldsymbol{s}^0)$ using $\bar{p}(\boldsymbol{s}^0) = \tilde{p}(\boldsymbol{s}^0) e^{-\frac{\hat{\sigma}^2}{2}}$, where $\hat{\sigma}^2$ is the empirical variance of $\tilde{p}(\boldsymbol{s}^0)_1, \dots \tilde{p}(\boldsymbol{s}^0)_M$.

In order to implement this bootstrapping procedure efficiently, it is critical to notice that evaluating $\tilde{p}(\boldsymbol{s}^0)$ requires only the Cholesky decompositions of $\tilde{\boldsymbol{\Sigma}}$. More precisely, indicate with $\boldsymbol{L}$ the unique lower triangular matrix with positive diagonal elements such that $\boldsymbol{L}\boldsymbol{L}^T = \tilde{\boldsymbol{\Sigma}}$. Then $|\tilde{\boldsymbol{\Sigma}}| = |\boldsymbol{L}|^2 = \prod_{j=1}^{d} \boldsymbol{L}_{jj}^2$ and $(\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}}) = ||\boldsymbol{L}^{-1}(\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}})||^2$, where $\boldsymbol{L}^{-1}(\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}})$ can be evaluated in $O(d^2)$ by forward-substitution. Hence, it would be advantageous to simulate $\boldsymbol{L}$ directly, rather than simulating $\tilde{\boldsymbol{\Sigma}}$ and then calculating its Cholesky decompositions, which is an $O(d^3)$ operation. This is easily achieved by noticing that, if $\boldsymbol{A}$ is a lower triangular matrix such that $\boldsymbol{A}_{jk} \sim N(0,1)$, for $1 \leq j < k < d$, and $\boldsymbol{A}_{jj} \sim \chi^2(N-j)$, for $j = 1, \dots, d$, then $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{D}^T \sim \text{Wish}(\hat{\boldsymbol{\Sigma}}/(N-1), N-1)$, where $\boldsymbol{D}$ is the lower triangular Cholesky factor of $\hat{\boldsymbol{\Sigma}}/(N-1)$ (Smith and Hocking, 1972). Hence, $\boldsymbol{L} = \boldsymbol{D}\boldsymbol{A}$ is the lower Cholesky factor of $\tilde{\boldsymbol{\Sigma}}$, whose computation requires generating $d(d-1)/2$ standard normal, d chi-squared random variables and evaluating the product of two lower triangular matrices, which is $O(d^3/3)$ (Golub and Van Loan, 2012). Fortunately, the matrix multiplication does not need to be performed, because to evaluate $||\boldsymbol{L}_i^{-1}(\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}}_i)||^2 = ||\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}(\boldsymbol{s}^0 - \tilde{\boldsymbol{\mu}}_i)||^2$ it is sufficient to use two forward sustitutions, while $|\boldsymbol{L}|^2 = |\boldsymbol{D}\boldsymbol{A}|^2 = |\boldsymbol{D}|^2|\boldsymbol{A}|^2 = \prod_{j=1}^{d} \boldsymbol{D}_{jj}^2 \boldsymbol{A}_{jj}^2$. Hence, simulating $\tilde{p}(\boldsymbol{s}^0)$ is an $O(d^2)$ operation, if we follow this numerical recipe.

The derivation of the corrected estimator, $\bar{p}(\boldsymbol{s}^0)$, is based on a normality assumption on $\log \tilde{p}(\boldsymbol{s}^0)$, which requires some justification. As $N \to \infty$, $\log|\hat{\boldsymbol{\Sigma}}|$ is asymptotically normally distributed, as proven by Cai et al. (2015) under three different regimes for $d = d(N)$: $\lim_{N\to\infty} d(N)/N = 0$, $\lim_{N\to\infty} d(N)/N = c$ and, remarkably, $d(N) = N$. Unfortunately, as $N \to \infty$, the distribution of $(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}})$ is not asymptotically normal. Indeed, under normality of $\boldsymbol{S}$, we have that

$$u = \frac{N-d}{d(N-1)} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}) \sim F(d, N-d),$$

where $F(\nu_1, \nu_2)$ is a non-central F distribution. As $N \to \infty$, we have that $d \times u$ converges to a non-central χ^2 distribution, with d degrees of freedom. This should guarantee the asymptotic normality of $\log \tilde{p}(\boldsymbol{s}^0)$ in the $\lim_{N\to\infty} d(N)/N = 0$ regime, as long as $\lim_{N\to\infty} d(N) = \infty$. If we expect the number of statistics used to grow proportionally to N, the $\lim_{N\to\infty} d(N)/N = c$ regime might be more interesting. Under this regime, Pan and Zhou (2011) proved the asymptotic normality of u, under the special case $E(\boldsymbol{S}) = E(\hat{\boldsymbol{\mu}}) = \boldsymbol{s}^0$. Analogous results might be expected to hold also in the non-centred case (i.e. $E(\boldsymbol{S}) \neq \boldsymbol{s}^0$), but we have not found them in the literature.

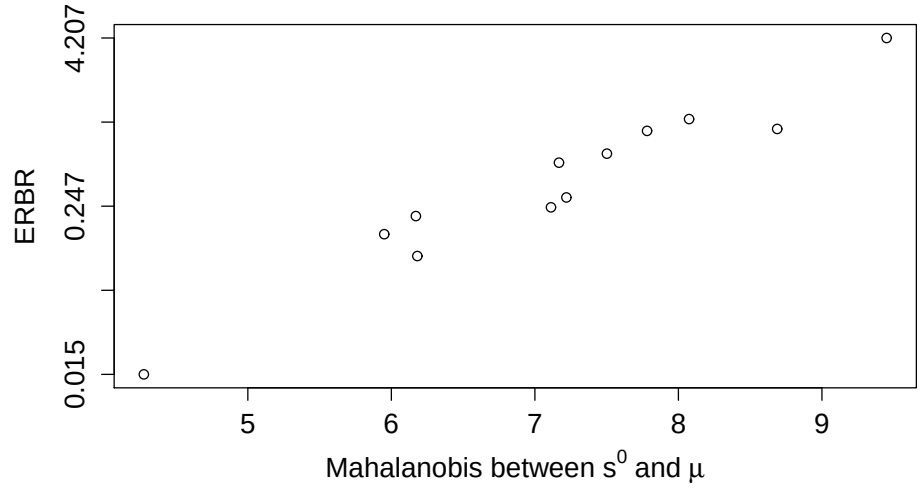

Figure 2-4: *ERBR (see Section 2.4.4 for a definition) as a function of the Mahalanobis distance between $\boldsymbol{s}^0$ and $\boldsymbol{\mu}$. The vertical axis uses a logarithmic scale.*

2.4.4 Toy example

To test whether the bias of $\bar{p}(\boldsymbol{s}^0)$ is lower than that of $\hat{p}(\boldsymbol{s}^0)$, we have considered the following scenario. The summary statistics are normally distributed with mean vector $\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{0}$ is a d-dimensional vector of zeros, and covariance matrix $\boldsymbol{\Sigma}$, which is a symmetric Toeplitz matrix with first row $\boldsymbol{\Sigma}_{1i} = 1 - (i-1)/d$, for $i = 1, \ldots, d$. We chose $d = 10$. We simulated $K = 12$ different locations, $\boldsymbol{s}_1^0, \ldots \boldsymbol{s}_K^0$, from a Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We then re-estimated $p(\boldsymbol{s}_1^0), \ldots, p(\boldsymbol{s}_K^0)$ $L = 100$ times, using both $\hat{p}(\boldsymbol{s}^0)$ and $\bar{p}(\boldsymbol{s}^0)$, with $N = 1000$ and $M = 1000$.

We quantified the improvement brought about by using $\bar{p}(\boldsymbol{s}^0)$, rather than $\hat{p}(\boldsymbol{s}^0)$, in terms of Empirical Relative Bias Reduction (ERBR), which we define as follows

$$\text{ERBR}(\boldsymbol{s}^0) = \frac{|\hat{\text{Bias}}\{\hat{p}(\boldsymbol{s}^0)\}| - |\hat{\text{Bias}}\{\bar{p}(\boldsymbol{s}^0)\}|}{p(\boldsymbol{s}^0)},$$

where

$$\hat{\text{Bias}}\{\hat{p}(\boldsymbol{s}^0)\} = \frac{1}{L}\sum_{i=1}^{L}\{\hat{p}(\boldsymbol{s}^0) - p(\boldsymbol{s}^0)\},$$

with $\hat{\text{Bias}}\{\bar{p}(\boldsymbol{s}^0)\}$ being defined analogously.

Figure 2-4 show how ERBR varies with the Mahalanobis distance between $\boldsymbol{s}^0$ and $\boldsymbol{\mu}$. The plot uses a logarithmic scale for the y-axis, and it shows an almost linear relation between ERBR and the Mahalanobis distance from the mean vector. This suggest that using $\bar{p}(\boldsymbol{s}^0)$ should lead to substantial gains, in terms of bias reduction, when $p(\boldsymbol{s}^0)$ is estimated deep in the tails.

2.5 Multimodality and state space methods

If the presence of process noise smooths the likelihood sufficiently, then methods that discard information should be outperformed by those that retain it. However, we can not generally prove that the likelihood for any particular model is smoothed and, as shown in Section 2.2, there exist models for which smoothing is only partial, and may be inadequate, when process noise is low. In this section we further investigate the impact of multimodality on state space methods, and show that information reduction methods can reduce the associated problems.

In order to evaluate the accuracy of the likelihood estimates given by the SIR algorithm for different levels of noise, we used the discretized SSM described in Section 2.2 and in Appendix A.1. We chose ten levels of process noise in the interval $\sigma \in [0.01, 0.3]$. For each level we simulated 1000 paths using the Ricker map, with $\log(r) = 3.8$, $\phi = 0.5$, and evaluated the likelihood of each of them at the true parameters. Figure 2-5 shows the results.

The plot on the top shows that, as the process noise decreases, the average bias of the likelihood estimated by the filter (solid) increases in absolute value. Indeed, while the true log-likelihood (not shown) is roughly constant (≈ -70) for different levels of σ, the mean filter's estimates drop from -65 for $\sigma = 0.3$ to -140 for $\sigma = 0.01$. The strong dependence between likelihood bias and σ suggests that a sampler using these likelihood estimates will never explore areas of the parameter space where σ is low. In addition, any model comparison criterion based on the biased likelihood estimates is unreliable.

On the bottom of Figure 2-5 we plotted the ratios between sample variance of the likelihood estimated by the filter and the sample variance of the true likelihood for each value of σ, that is

$$\frac{\hat{\mathrm{Var}}\big\{ \log \hat{p}(\boldsymbol{y}_{1:50}|\boldsymbol{\theta}) \big\}}{\hat{\mathrm{Var}}\big\{ \log p(\boldsymbol{y}_{1:50}|\boldsymbol{\theta}) \big\}}.$$

From the plot we see that the variance of the estimated log-likelihood increases exponentially as σ decreases, suggesting that Monte Carlo variability of the integration procedure dwarfs sampling variation for low σ. This has implications for algorithms based on particle filters: with such noisy likelihood estimates the PMMH algorithm will have an extremely low acceptance rate (Doucet et al., 2012), while the IF procedure will become quite unstable, due to the high variability of the estimated gradients.

The broken lines in Figure 2-5, show corresponding quantities for the synthetic likelihood, obtained using the set of 13 summary statistics proposed by Wood (2010) and reported in Appendix A.2. Interestingly, both the average and the variance of the synthetic likelihood estimates remain roughly constant for different degrees of process noise. This suggests that the SL approach is quite robust to the level of process noise in the system, as it gives stable estimates also when the process dynamics are near-deterministic. On the other hand, the variance of the synthetic likelihood is lower than that of the true likelihood for any σ, which might be a consequence of the information loss.

Note that to use synthetic likelihood when the system is (close to) deterministic, the initial values of the simulated paths have to be randomized ($N_1 \sim \mathrm{Unif}(0.1, 5)$), otherwise the variances of the summary statistics can be close to zero for very low process noise. Random initial values are consistent with the information reduction

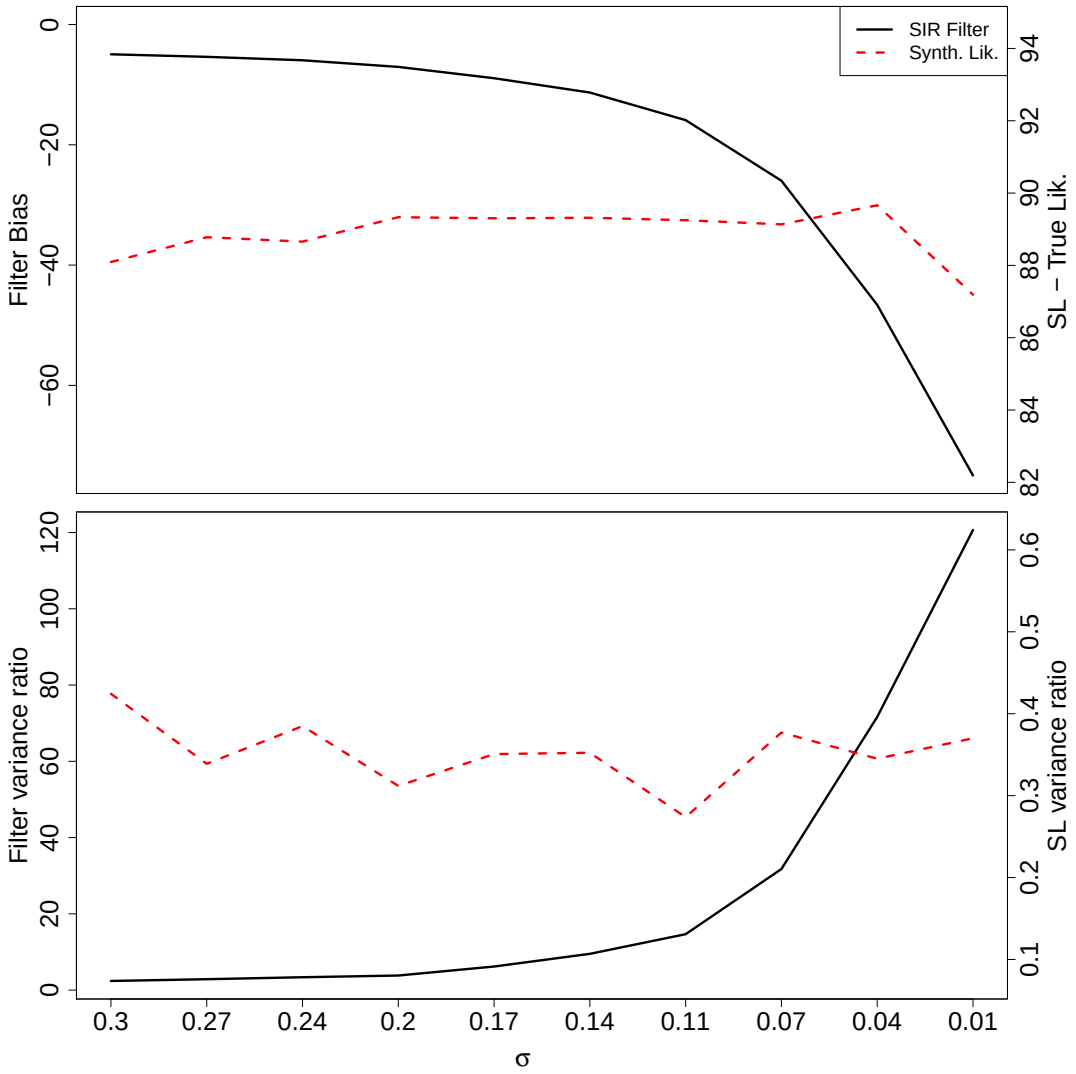

Figure 2-5: *Top: average difference between the full likelihood and the estimated full (solid) or synthetic likelihood (dashed) as a function of σ, obtained using respectively the SIR filter and SL. Bottom: ratio between the sample variance of estimated full (black line) or synthetic (broken red line) likelihoods and the true likelihood for several values of σ.*

philosophy: inference should be robust to the particular values of the hidden states. In this context we are confident that ABC, being based on summary statistics, would perform similarly to SL.

Figure 2-6 shows why the SIR algorithm is struggling to estimate the log-likelihood when σ is very low. Each of the 20 columns in the top image represents the true filtering density $p(n_t|\boldsymbol{y}_{1:t}, \boldsymbol{\theta})$ at each time step, when $\sigma = 0.3$. Areas of high density are represented in yellow, while area of lower density are coloured in red. With this level of process noise the filtering densities are smooth and unimodal, so the filter places the particles around each mode, thus providing a reliable estimate of the likelihood. In contrast, the image on the bottom of Figure 2-6 shows that for very low process noise the filtering densities are unimodal in the first couple of time step, but then they break into narrow multiple modes. Because of the irregularity of the filtering densities, the quality of the particle approximation is poor in this case (see time 19 in particular). The filter struggles to explore all the important modes of the filtering distributions, and hence the resulting estimates of the log-likelihood are very variable.

So Figure 2-6 helps to explain the variability in performance of the particle filter approach seen in Figures 2-3 and 2-5 as the process noise level changes. For models capable of showing chaotic or near-chaotic dynamics, there will be areas of the parameter space where the likelihood is highly multimodal. In these areas particle filtering methods will struggle to estimate the likelihood. In such situations most of the likelihood-based asymptotic theory will not be applicable, and even if it was possible to sample the corresponding parameter posterior exactly, it would not be obvious how the results should be interpreted. Hence, we argue that in such situations the use of approaches based on information reduction, which can provide a smooth proxy to likelihood, might be preferable from both a methodological and practical point of view.

To emphasise that the issue of multimodality is generic to the state space approach, rather than being specific to filtering, or a particular filtering implementation, or our discretized state space example, we illustrate how Parameter Cascading can encounter similar problems on the unmodified Ricker model. Figure 2-7 shows transects of the parameter fitting objective function, $H(\boldsymbol{\theta}|\boldsymbol{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$, (see Section 2.3.2) with respect to $\log(r)$ for four values of λ, and show that this function becomes more irregular as λ increases. For large λ, which is appropriate when σ is low, this hinders the optimization and makes estimating $\boldsymbol{\theta}$ problematic. In the following we illustrate that jumps in the objective function correspond to transitions between modes of the objective function for the state, $J(\boldsymbol{n}_{1:T}|\boldsymbol{\theta}, \lambda)$.

The upper plot of Figure 2-8 shows other transects of $H(\boldsymbol{\theta}|\boldsymbol{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$, for $\lambda = 65$. The solid line was obtained using the same initial value $\boldsymbol{n}^{\boldsymbol{\theta}}_{1:T} = \boldsymbol{y}_{1:T}/\phi$ for each value of $\log(r)$. The dashed lines show the $H(\boldsymbol{\theta}|\boldsymbol{n}^{\boldsymbol{\theta}}_{1:T}, \lambda)$ curves corresponding to two different modes of $J(\boldsymbol{n}_{1:t}|\boldsymbol{\theta}, \lambda)$ and have been obtained by carefully tracking of the modes. We refer to these modes as A and B. The plots on the bottom of Figure 2-8 represent the estimated hidden states $\boldsymbol{n}^{\theta}_{1:T}$ corresponding to two values of $\log(r)$ and to each mode. This shows that the same value of $\log(r)$ leads to two different modes in the state space, depending on the initialization. The similarity between the pairs A1-A2 and B1-B2 shows that these initialization-dependent modes are persistent along $\log(r)$.

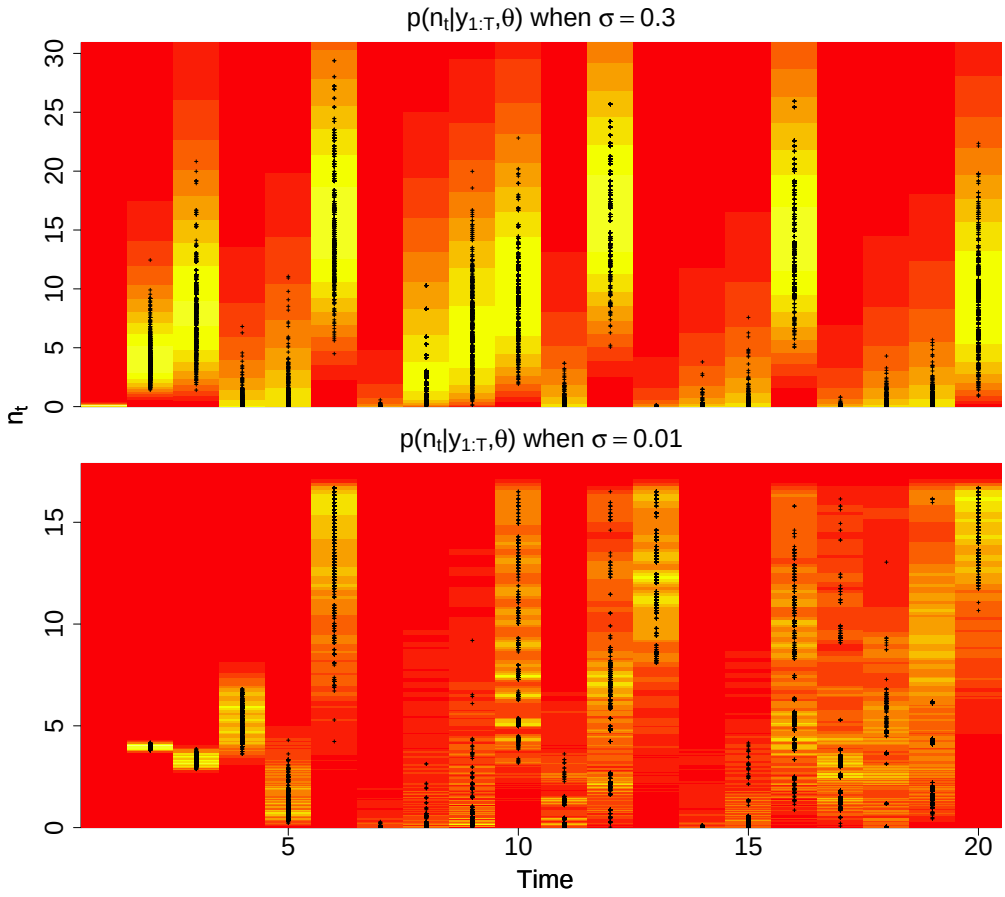

Figure 2-6: *Filtering densities* $p(n_t|\boldsymbol{y}_{1:t}, \boldsymbol{\theta})$ *for a single Ricker path generated using* $log(r) = 3.8$, $\phi = 10$ *and* $\sigma = 0.3$ *(top) or* $\sigma = 0.01$ *(bottom).*

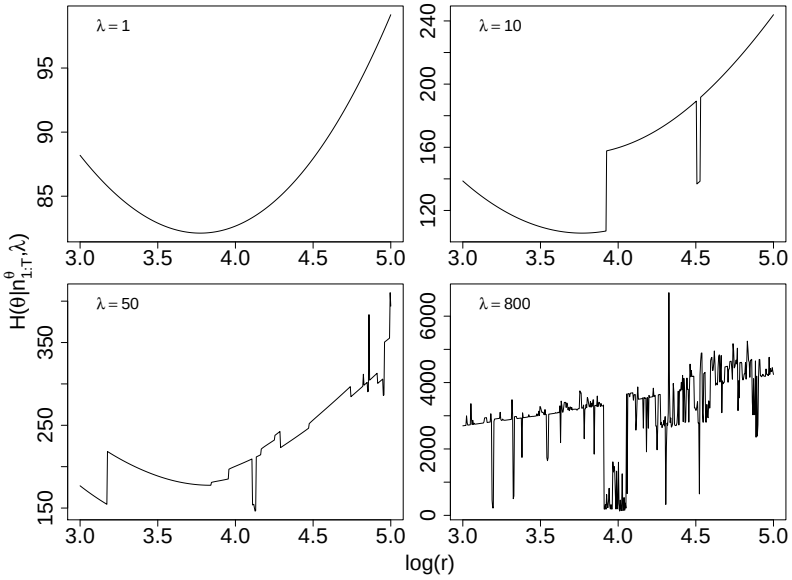

Figure 2-7: *Transects of* $H(\boldsymbol{\theta}|\boldsymbol{n}_{1:T}, \lambda)$ *w.r.t.* $\log(r)$, *as* λ *increases.*

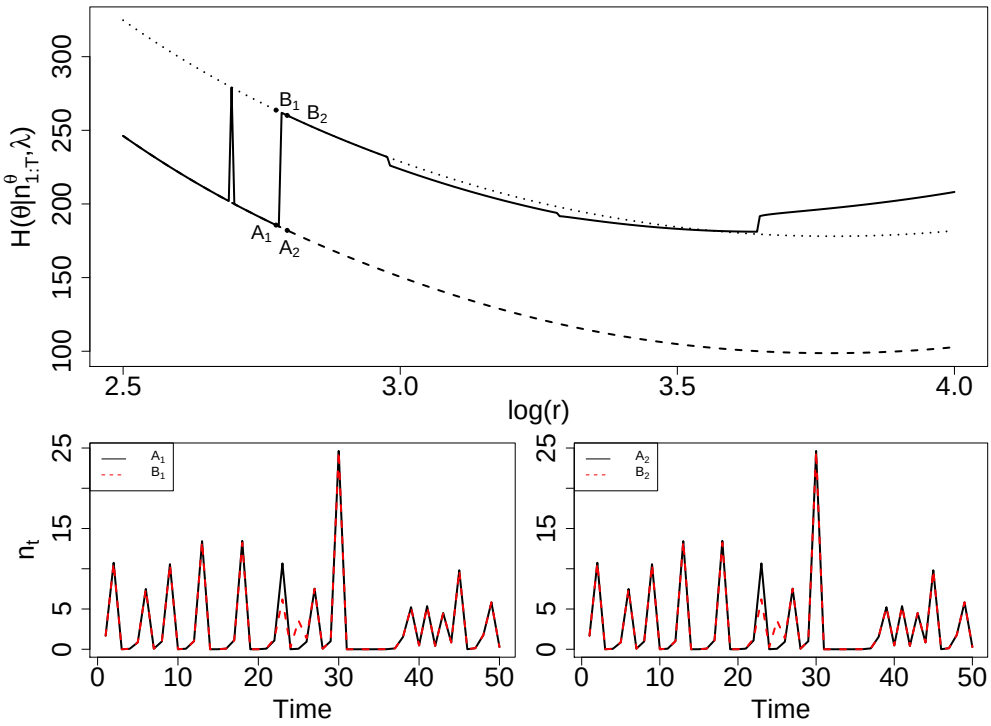

Figure 2-8: *Top: transects of* $H(\boldsymbol{\theta}|\lambda, \mathbf{n}_t)$ *with respect to* $\log(r)$*. Bottom: paths corresponding to two points 1 or 2 along the* $\log(r)$ *axis and to modes A or B in the state space.*

2.6 SL versus tolerance-based ABC

The choice of summary statistics is crucial for the performance of information reduction methods, hence the topic has been the subject of much research. For instance, see Blum et al. (2013) for a comprehensive review of methods for dimension reduction or statistics selection, in the context of ABC. SL and ABC methods share some requirements regarding the choice of summary statistics. More specifically, in parameter estimation problems the summary statistics should contain as much information as possible about the parameters, so that $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ will be approximately proportional to $p(\boldsymbol{\theta}|\boldsymbol{y}^0)$.

Beside this common ground, SL differs from ABC methods in several ways, and this entails some diverging requirements on the summary statistics. In particular, reducing the number summary statistics is more critical to ABC methods than to SL. In fact, the non-parametric approach followed by most ABC methods, implies that the convergence rate of the resulting posterior distributions slows down rapidly as the dimension of the statistics vector increases (Blum, 2010). On the other hand, the parametric likelihood estimator used by SL, ensures that this method is much less sensitive to the number of summary statistics used. This difference in scalability has important practical implications. In particular, SL allows practitioners to focus on the challenging task of identifying informative summary statistics, without having to worry too much about keeping their number low. Obviously SL's scalability in the number of statistics does not come without a cost, but it has to be paid for in parametric assumptions, whose effect might be hard to quantify.

Another potential issue with ABC algorithms is that they often measure the distance between the observed and simulated statistics using a quadratic form, that is

$$d(\boldsymbol{s}^0, \boldsymbol{S}) = ||\boldsymbol{s}^0, \boldsymbol{S}||^2_{\boldsymbol{A}} = (\boldsymbol{s}^0 - \boldsymbol{S})^T \boldsymbol{A} (\boldsymbol{s}^0 - \boldsymbol{S}),$$

where $\boldsymbol{A}$ is a positive-definite scaling matrix. This distance function was adopted by, for instance, Fearnhead and Prangle (2012). The choice of $\boldsymbol{A}$ is fundamental when the summary statistics have very different scales or when there are subsets of highly correlated statistics. A possible solution is to simulate N vectors of summary statistic at some location $\boldsymbol{\theta}_p$ in the parameters space and use the inverse of the empirical covariance matrix of the simulated summary statistics as scaling matrix $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{\theta}_p}$. Alternatively, it is possible to use a diagonal scaling matrix, such as $\boldsymbol{A} = \text{diag}\{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_p}\}^{-1}$. Because of the strong correlations between the summary statistics used in this section, we scale them using the full precision matrix. This simple choice works well in many cases, but it can lead to unsatisfactory results when the covariance of the summary statistics varies strongly with model parameters.

As an illustration of this problem, we consider the stochastic version of the Ricker map (2.1, 2.2) and the set of 13 summary statistic proposed by Wood (2010). To quantify the importance of the scaling matrix $\boldsymbol{A}$ in this setting, we performed the following simulation experiment:

- Define a sequence of equally space values v_k, for $k = 1, 2, \ldots, 50$, ranging from 2.8 to 3.8.
- For each value v_k:
    1. Simulate a path $\boldsymbol{y}_{1:T}$ from the Ricker map, using $T = 50$ and parameter values $\log r = 3.8$, $\sigma^2 = 0.3$ and $\phi = 10$. Define $\boldsymbol{s}^0 = S(\boldsymbol{y}_{1:T})$.

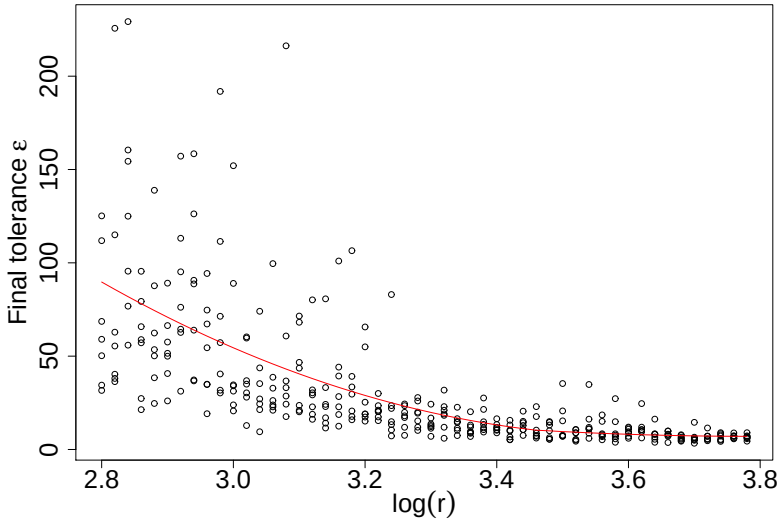

Figure 2-9: *Lowest achievable tolerance* ϵ *versus value of* $\log r$ *at which the scaling matrix is estimated. The red line is a quadratic regression fit.*

2. Set the initial parameter vector $\boldsymbol{\theta}_p$ to $\log r = v_k$, $\sigma^2 = 0.3$ and $\phi = 10$.
3. Simulate 10^4 paths from the model using parameters $\boldsymbol{\theta}_p$, transform each of them into a vector summary statistics and calculate their empirical covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_p}$.
4. Approximately sample $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ using the ABC-SMC routine proposed by Toni et al. (2009), where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_p}^{-1}$ is used as scaling matrix. We refer the reader to Toni et al. (2009) for details about this algorithm, but we point out that this is a sequential scheme where the tolerance ϵ is reduced at each step and that we terminated the algorithm when the acceptance ratio of the most recent iteration was below 1%.

We repeated the whole experiment 7 times and the results are illustrated in Figure 2-9. Here the x-axis represents the value of $\log(r)$ at which the scaling matrix was estimated, while the y-axis represents the lowest tolerance ϵ achieved before the termination of the ABC-SMC algorithm. This plot shows how crucial is the choice of scaling matrix in situations where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ varies widely with $\boldsymbol{\theta}$: if the scaling matrix is not adequate the tolerance cannot be reduced enough. In an applied ecological setting, where the true parameters are unknown and the model of interest is more complex than the one used here, this means that a practitioner might struggle to find either a reasonable guess for the scaling matrix or a set of summary statistics whose covariance is not strongly dependent on $\boldsymbol{\theta}$.

Another choice that has to be made, in order to use tolerance-based ABC procedures, is the selection of ϵ. The tolerance can be a small scalar constant as in the ABC-MCMC algorithm of Marjoram et al. (2003), or it can be a vector of decreasing tolerances as in the ABC-SMC algorithm of Toni et al. (2009). In order to obtain a better approximation to $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$, ϵ should be chosen to be as small as possible, but the acceptance probability will decrease with the tolerance. A common choice is to select

a tolerance that allows a predetermined acceptance ratio to be achieved, but in some cases this strategy can lead to invalid results, as detailed in Silk et al. (2013).

The regression adjustment of Beaumont et al. (2002) can be used to mitigate the discrepancy between the observed and the simulated statistics, which is proportional to the tolerance ϵ. However, the result of this correction is generally still dependent on ϵ, which controls the bias-variance trade-off of the regression (Beaumont et al., 2002). Hence, using this procedure does not necessarily lead to higher accuracy in parameter estimation. For example, Fearnhead and Prangle (2012) obtained worse results with the regression correction than from the raw ABC output, using the Ricker model and the same summary statistics considered here.

In this section we used $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_p}^{-1}$ to scale the summary statistic. This was done for the purpose of illustrating that the scaling the summary statistics correctly is critical to the performance of the ABC methods. A more commonly used approach is to simulate N parameters vectors, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$, from $p(\boldsymbol{\theta})$ and the corresponding statistics vectors, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$, from $p(\boldsymbol{s}|\boldsymbol{\theta})$, and to use the empirical covariance matrix of the simulated statistics as scaling matrix. In the context of ABC-MCMC, it is then possible to calculate the distances $d(\boldsymbol{s}_i, \boldsymbol{s}^0)$, for $i = 1, \ldots, L$, and to chose ϵ so that only 0.1% of the distances fall below this threshold. This approach worked relatively well with the simple models used in Section 2.7. However, the tuning tends to be much more laborious under more complex models, such as described those presented in Chapter 3. In particular, when the number of unknown parameters is high, training ϵ and $\boldsymbol{A}$ using simulations from the prior can be very inefficient, especially if the prior contains little information. Hence, for complex models, tuning ϵ and $\boldsymbol{A}$ might require a more sophisticated approach, possibly involving some degree of manual intervention.

SL is not afflicted by the difficulties just described, because it is tolerance-free and the summary statistics are scaled automatically and dynamically by the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Obviously this robustness comes at a cost: a single point-wise synthetic likelihood estimate requires a number of simulations sufficient to estimate the covariance matrix. In addition, even though for many commonly used statistics the Central Limit Theorem (CLT) assures asymptotic normality, in small samples the normal approximation might be crude, while in some contexts it might be difficult to devise asymptotically normal statistics.

As a simple example of the former problem, let us consider a sample of size N from an exponential distribution with rate α. Here the Maximum Likelihood (ML) estimator of α is given by the reciprocal of the sample average

$$s = \frac{1}{\bar{x}} = \left(\frac{\sum_{i=1}^{N} x_i}{N} \right)^{-1}.$$

Given that s is a sufficient statistic for α, the likelihood function can be factorized as follows

$$p(x|\alpha) = h(x) f(s, \alpha) \propto f(s, \alpha),$$

hence the likelihood is proportional to a function of only s and α. By the CLT, the distribution of s is asymptotically normal, but we want to verify how well we can approximate the likelihood using SL when $N = 10$. Figure 2-10 shows the log-likelihood (dashed) and the estimated synthetic log-likelihood (black) for $\alpha \in [0.5, 2]$. The true value of α is 1. With such a small sample size the distribution of the simulated statistic is far from normal, and in fact the synthetic log-likelihood is quite off target. In cases

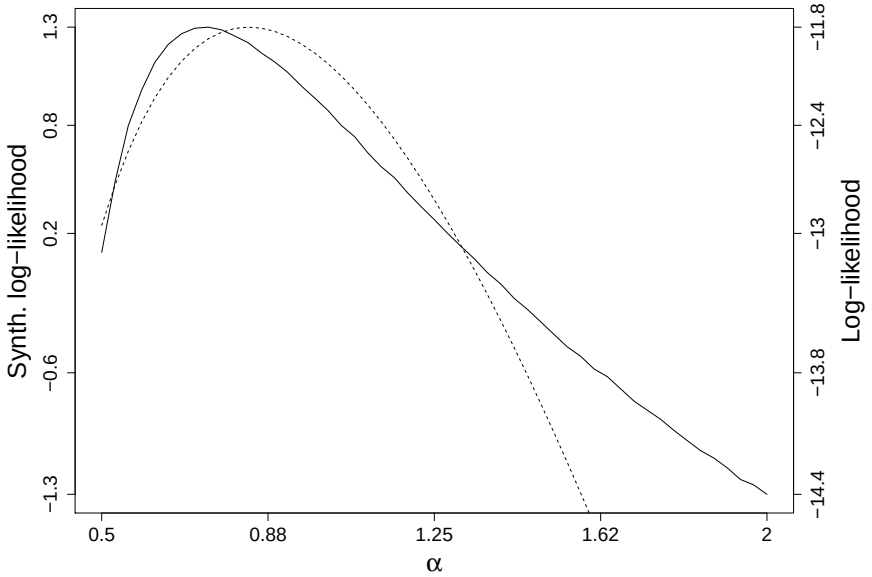

Figure 2-10: *Synthetic log-likelihood function (black line) vs true log-likelihood function (broken line) for a Exp*($\alpha = 1$) *distribution.*

such as this, where the number of summary statistics is low, it is straightforward to use transformations to improve to normality assumption, as proposed by Wood (2010). However, in an higher dimensional setting approximate multivariate normality might be difficult to assess or improve. More importantly, achieving multivariate normality for a certain set of parameter values does not assure that this approximation will hold elsewhere in the parameter space.

2.7 Comparison on simple chaotic maps

Here we consider the models summarized in Table 2.1, in addition to the Ricker map. The parameter values of each model, reported in the Appendix A.2, have been chosen so that the simulated paths show similar chaotic dynamics (Figure 2-11).

The data consist of 50 simulated paths $\boldsymbol{y}_{1:T}$, where $T = 50$, from each model. All paths were used to estimate the parameters using each method. For SLMH and for the ABC-MCMC algorithm of Marjoram et al. (2003) we have used 3×10^4 iterations to sample the posterior of each path. The PMMH algorithm had an extremely low acceptance rate unless the likelihood of the latest accepted position was re-estimated at each MCMC step. This doubled the computational effort, and hence we used only 1.5×10^4 iterations for this method. To check if recomputing the likelihood was biasing the results in favour of PMMH, we have implemented a version of SLMH (labelled SLMH-R) that uses the same approach. For SLMH and ABC-MCMC we have discarded 5000 iterations as burn-in, while for PMMH and SLMH-R 2500 iterations were discarded. For IF we have used 3000 optimization steps.

At each MH step, SLMH and PMMH estimated the (synthetic) likelihood by using 500 simulations from the model, while IF used 5000 simulations at each step of optimization step. ABC simulates only one sample at each step, but we stored an iteration every 500. Notice that, with this set-up, SLMH, SLMH-R, PMMH and ABC used the

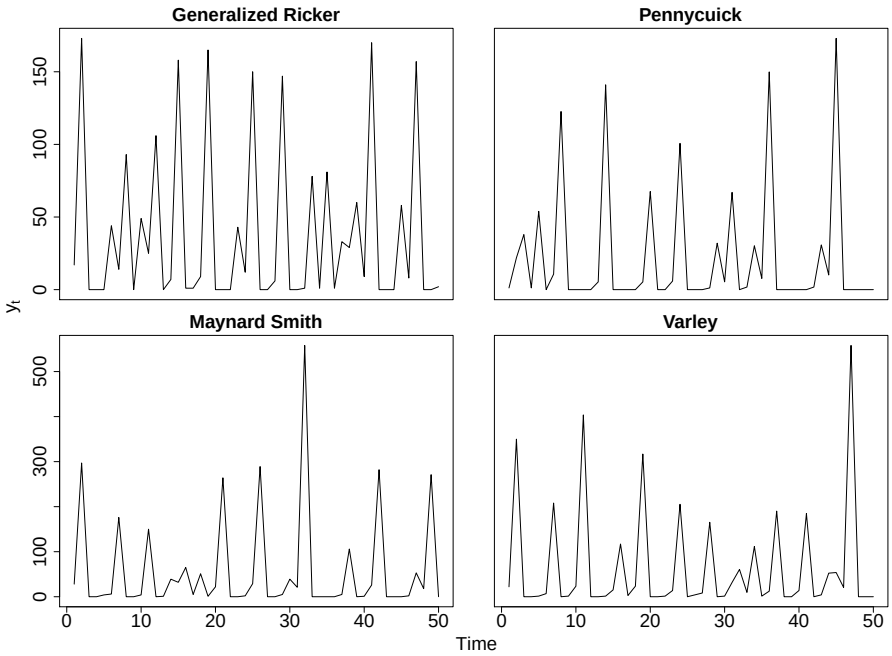

Figure 2-11: *Trajectories simulated using the four models described in Table 2.1.*

same number of simulations (1.5×10^7) from the model in order to fit each of the 250 simulated datasets. Given that the methods have very different implementation, basing the comparison on the number of simulations from the model, rather than CPU time, ensures fairness.

We used proper uniform priors for all parameters. IF does not support the use of priors, so we interpreted the priors as box constraints for the optimization. All methods were initialized at the same starting values which, together with the priors and other details, are included in Appendix A.2.

We evaluated the accuracy of different approaches in term of squared errors between point estimates and the true parameters. While IF provided point estimates directly, ABC-MCMC, SLMH and PMMH give dependent samples from the (approximate) parameter posteriors. Hence, for the latter group of methods, we have used the posterior means as point estimates.

Appendix A.2 reports the median squared errors for each model-method-parameter combination. Here we have summarized the results in Figure 2-12 which represents, for each model and method, the median and Inter-Quartile Range of the squared errors, averaged geometrically across the parameters. More precisely, let m, k, j and i be the indexes of model, method, dataset and parameter respectively, the average squared errors are then given by

$$\bar{e}_j^{m,k} = \left\{ \prod_{i=1}^{p_m} \left(\hat{\theta}_{j,i}^{m,k} - \theta_i^m\right)^2 \right\}^{\frac{1}{p_m}},$$

where p_m is the parameter count for model m.

Figure 2-12 shows that, on this set of simple models, methods based on particle filtering consistently outperform methods based on information reduction. The performance of IF and PMMH is quite similar, and the differences in average squared

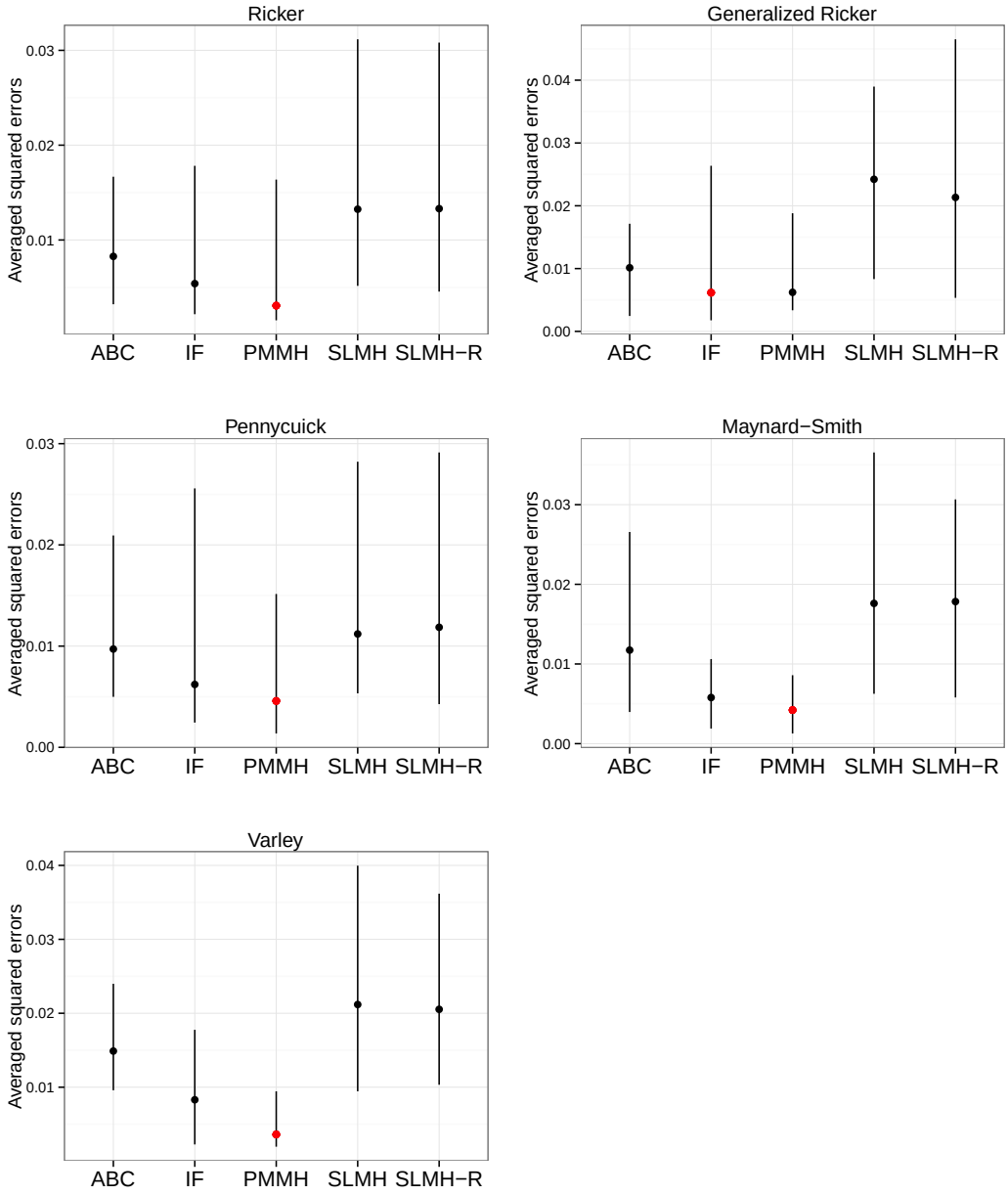

Figure 2-12: *Medians and Inter-Quartile Ranges of the averaged squared errors for each model and method.*

errors between these two methods might be due to the different type of point estimates used. ABC-MCMC seems to perform better that either SLMH or SLMH-R for all models. This performance gap might be attributable to the normal approximation used by SLMH, to the bias entailed by estimating $p(\boldsymbol{s}_0|\boldsymbol{\theta})$ using a finite sample or simply to particular set-up we have used for the experiment. The clear result here is that, given sufficient noise, the information reduction methods have noticeably worse performance than the state space methods for these simple toy models.

In this section we compared the accuracy of the point estimates produced by SLMH, ABC-MCMC, IF and PMMH, while giving the same computational budget to each method. However, given that PMMH is an exact-approximate method, as explained in Section 2.4, it is likely that the performance gap between PMMH and the remaining methods, all of which are approximate in nature, would increase with the computational budget.

2.8 Conclusions

In this chapter we described some of the difficulties that can be encountered when working with highly non-linear dynamical models, and we have shown how these issues influence the performance of some popular inferential approaches. In particular, in Section 2.5 we provided strong experimental evidence suggesting that, when the dynamics of the system are chaotic or near-chaotic, the likelihood function becomes increasingly multimodal as the process noise is reduced. While this directly undermines the performance of state space methods aiming at estimating the full likelihood, as in PMMH, or its derivatives, as in IF, approaches based on information reduction are less affected. This has practical implications because, in an applied setting, it is generally not known whether the best fitting parameters lay in an area of the parameter space where the stochasticity is too low for state space methods to work adequately. Beside the practical problem of obtaining reliable likelihood estimates, the existence of highly multimodal likelihood functions has deeper implications. Indeed, even if the likelihood were analytically available, as it is for the models in Table 2.1 in the absence of process noise, it is not clear whether it could be used to perform any meaningful statistical inference. Leaving aside the practical issue of the maximizing this multimodal likelihood, most of the likelihood-based asymptotic theory would not be applicable under such circumstance, and even if it were possible to sample the corresponding parameter posterior efficiently, it would not be obvious how the results should be interpreted. It is difficult to answer to these concerns in general, but it seems desirable to obtain a smoother, more interpretable, proxy to the likelihood. As we demostrated in Section 2.5, this can be achieved using information reduction methods, which focus on features of the data that are phase-independent.

So Figure 2-6 helps to explain the variability in performance of the particle filter approach seen in Figures 2-3 and 2-5 as the process noise level changes. Figure 2-3 demonstrates that the likelihood of a chaotic model might be highly multimodal in certain areas of the parameters space. In these areas particle filtering methods will struggle to estimate the likelihood. In such situations most of the likelihood-based asymptotic theory will not be applicable, and even if it was possible to sample the corresponding parameter posterior exactly, it would not be obvious how the results should be interpreted. Hence, we argue that in such situations the use of approaches

based on information reduction, which can provide a smooth proxy to likelihood, might be preferable from both a methodological and practical point of view.

The results reported in Section 2.7 suggest that reducing the data to a set of summary statistics generally entails a loss of accuracy in parameter estimation. Indeed, SLMH and ABC-MCMC are consistently outperformed by PMMH and IF in terms of MSEs. In Chapter 3 we will provide futher examples demonstrating that, when simulated data is used, this is generally the case. However, we will we show that when real data is used, and hence the true data generating process is unknown, the use of approximate methods might be appealing, due to their robustness properties.

All the methods described in this chapter, with the exception of Parameter Cascading, are computationally intensive. In particular, obtaining pointwise estimates of $p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta})$ or $\nabla p(\boldsymbol{y}_{1:T}^0|\boldsymbol{\theta})$ requires MT simulations from $p(\boldsymbol{n}_t|\boldsymbol{n}_{t-1}, \theta)$, where M is the number of particles, under SIR and IF respectively. Similarly, SL uses N simulations from $p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})$ to estimate $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. Within PMMH and SLMH, this price has to be paid at each iteration and the efficiency of the sampler will depend on the trade-off between the variance of likelihood estimates and the number of simulations used to obtain them (Sherlock et al., 2014). Similar considerations hold for IF, but the optimizer generally needs much fewer iterations to reach convergence. On the other hand IF does not directly provide parameter uncertainty estimates, which have to be obtained through an expensive likelihood profiling procedure (see Ionides et al. (2006)). On first sight ABC samplers seem more efficient than the above approaches, because they target $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ directly, by simulating a single statistics vector at the time. However, ABC samplers generally have a very low acceptance rate, because the latter increases with the tolerance ϵ, while their accuracy is inversely proportional to it.

CHAPTER 3

REAL DATA EXAMPLES

In Chapter 2 we compared SL, ABC, PMMH and IF using simple chaotic models and simulated data. Here we consider more realistic ecological and epidemiological models, using both simulated and real data. Hence, the content of this chapter is the natural continuation of the comparison work started in Chapter 2.

3.1 Introduction

In order to limit the computational and programming effort we will restrict our attention to PMMH and SLMH: that is, one method from each of the two inferential philosophies described in Chapter 2. We chose SL rather than ABC, because the former method does not require tuning of the tolerance and scaling matrix. We selected PMMH over IF, because PMMH and SLMH have very similar Metropolis-Hastings implementations, which should limit the influence of other implementational confounders on the results of the comparison.

In each of the first three sections of this chapter we compare SLMH and PMMH on a different model. In particular, in Section 3.2 we consider Wood's (2010) discretized version of the blowfly model of Gurney et al. (1980), and we fit it to simulated and experimental datasets. In Section 3.3 we propose a modified version of the Susceptible-Infected-Recovered-Susceptible (SIRS) model of King et al. (2008) and we fit it to real epidemiological data, concerning cholera-related deaths in the Bay of Bengal. The last example is the prey-predator model of Turchin and Ellner (2000), which we use to compare SLMH and PMMH on simulated and on Fennoscandian voles' trapping data. Section 3.5 interprets and discusses the results obtained in earlier sections, and expands upon the conclusions drawn at the end of Chapter 2.

3.2 Nicholson's blowflies

In this section we consider the results, reported by Nicholson (1954) and Nicholson (1957), of a series of laboratory experiments meant to elucidate the population dynamics of sheep blowfly *Lucilia cuprina* under resource limitation. Blowflies develop in four successive stages: eggs, larvae, pupae and adults. Feeding occurs only in the

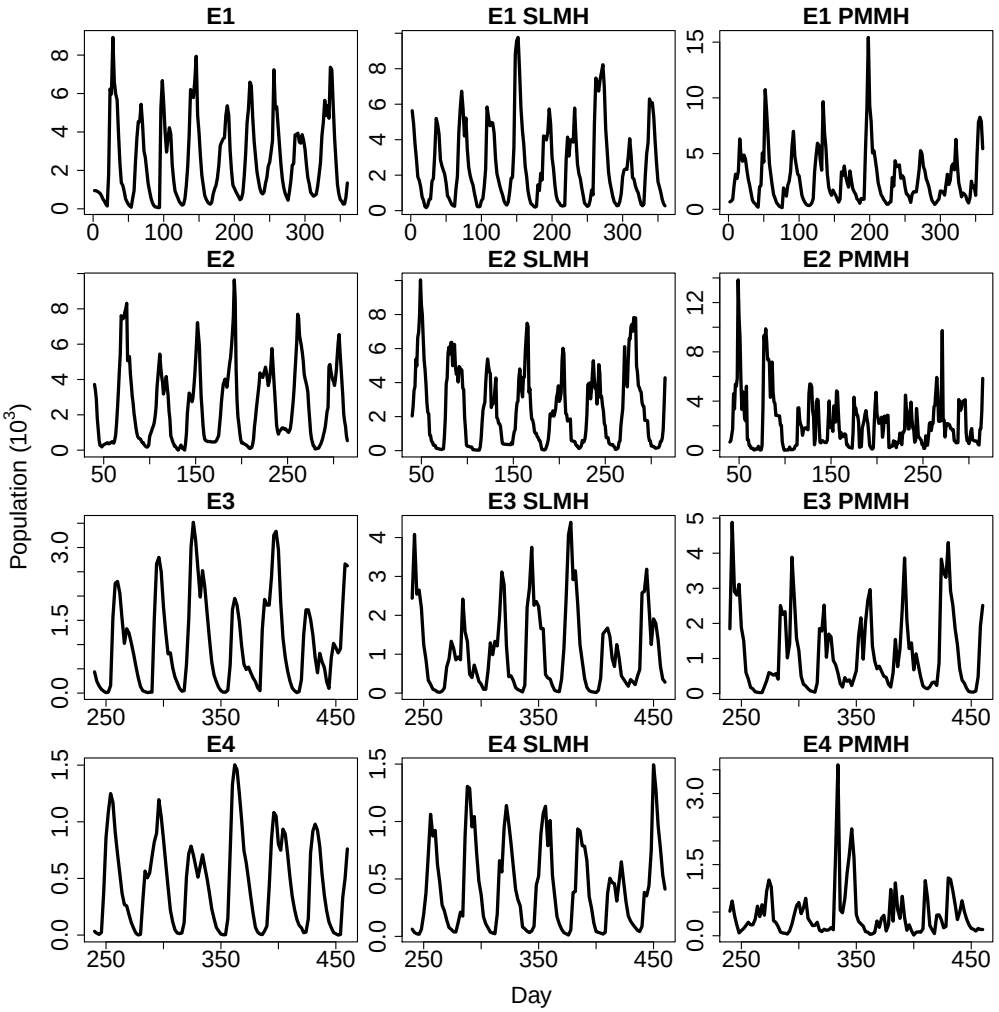

Figure 3-1: *Left column: the datasets reported by Nicholson (1954) and Nicholson (1957). Central and right columns: paths simulated from model 3.1 using parameters equal to the posterior means, obtained by fitting the four datasets using SLMH and PMMH.*

larval and adult stages. In two of the experiments (E1 and E2) the larvae had unlimited resources, while the adults had unlimited access to sugar and water, but were provided with a limited amount of protein, which is required for egg production. In another two experiments (E3 and E4) the larvae were supplied respectively with a moderately and severely restricted amount of food, while adults had unlimited resources. The resulting population dynamics are shown in the left column of Figure 3-1.

3.2.1 The model

A model potentially capable of explaining the observed dynamics of this population was proposed by Gurney et al. (1980), and it is represented by the following delayed differential equation

$$\frac{dn(t)}{dt} = Pn(t-\tau)e^{-\frac{n(t-\tau)}{n_0}} - \delta n(t), \tag{3.1}$$

where n represents the adult population, while P, τ, n_0 and δ are parameters. In order to fit the model to the available datasets, Wood (2010) proposed a discretized version of equation (3.1) and added a stochastic component to its deterministic structure. More precisely, he proposed the following model

$$n_t = r_t + s_t, \tag{3.2}$$

where

$$r_t \sim \text{Pois}(P n_{t-\tau} e^{-\frac{n_{t-\tau}}{n_0}} e_t),$$

represents delayed recruitment process, while

$$s_t \sim \text{binom}(e^{-\delta \epsilon_t}, n_{t-1}),$$

denotes the adult survival process. Finally, e_t and ϵ_t are independent gamma distributed random variables, with unit means and variances equal to σ_p^2 and σ_d^2 respectively.

3.2.2 Comparison using simulated data

In order to verify the accuracy of SLMH and PMMH for the blowfly model, we have tested them on simulated data. Firsly, notice that model (3.2) does not include any measurement noise: the number of blowflies, n_t, is assumed to be perfectly observed. This means that the model is not a SSM, hence it cannot be fitted using methods based on particle filtering directly. Our solution has been to introduce an artificial measurement process, when fitting the model using PMMH. More precisely, we use the following log-normal observational process

$$\log y_t \sim \text{N}(\log n_t, \sigma_o^2),$$

where the value of σ_o was predetermined, not estimated. Notice that, because of this modification, PMMH is fitting the wrong model and this procedure can be seen as an importance sampling ABC procedure, where σ_o plays the role of the tolerance. See Dean et al. (2011) for more details about the use ABC procedures in the context of SSMs with intractable observational processes. Despite having introduced an artificial measurement process, we decided to avoid estimating the initial values $n_1, \ldots, n_\tau$ when using PMMH, but we have fixed their values to that of the first τ observations.

Before moving to the results, it is important to point out that the likelihood of model (3.2) could be estimated, without modifying the model in any way, using the Monte Carlo integration approach described in Appendix B.1. In its current formulation this approach is fairly expensive to compute, but it might lead to more accurate results than the PMMH approach described here. However, an MCMC algorithm using the resulting likelihood estimates could not be considered to be a PMMH sampler, and given that the purpose of this chapter is comparing SLMH and PMMH, we have decided not to include it in the comparison.

For the comparison we simulated 24 datasets of length $T = 200$, using parameter values $\delta = 0.16$, $P = 6.5$, $n_0 = 400$, $\sigma_p^2 = 0.1$, $\tau = 14$, $\sigma_d^2 = 0.1$. We then estimated the parameters with both methods, using 2×10^4 MCMC iteration and 1000 simulation from the model at each step. The choice of σ_o was critical for the performance of PMMH. Obviously, we would like σ_o to be as small as possible, but lowering it increases

	δ	P	n_0	σ_p^2	τ	σ_d^2
SLMH0	0.08(0.83)	0.13(0.83)	0.102(0.79)	0.24(1)	0.035(0.92)	0.43(0.96)
PMMH0	0.06(0.67)	0.11(0.88)	0.071(0.88)	0.55(0.58)	0.02(0.92)	1.32(0.17)
p-value	0.414	0.197	0.01	0.359	0.03	< 0.001
Best	PMMH0	PMMH0	PMMH0	SLMH0	PMMH0	SLMH0
SLMH1	0.054(0.83)	0.14(0.75)	0.09(0.88)	0.25(1)	0.03(0.96)	0.43(1)
PMMH1	0.04(0.88)	0.06(0.92)	0.03(0.92)	0.18(1)	0.003(1)	0.17(0.96)
p-value	0.123	0.006	< 0.001	0.058	0.006	< 0.001
Best	PMMH1	PMMH1	PMMH1	PMMH1	PMMH1	PMMH1

Table 3.1: *Root MSEs(coverage) of the log-parameters for SLMH and PMMH for the blowflies model for realistic (0) and optimistic (1) starting values. The p-values for the differences in log-absolute errors have been calculated using t-tests.*

the variance of the importance weights and, in turn, of the estimated likelihood. In particular, if PMMH was initialized far from the true parameters, σ_o had to be increased in order to avoid particle depletion. Hence, we decided to include the results (PMMH0 and SL0) obtained using a realistic initialization ($\delta = 0.1$, $P = 4$, $n_0 = 200$, $\sigma_p^2 = 0.2$, $\tau = 10$, $\sigma_d^2 = 0.2$) and the results obtained by initializing the chains at the true parameters. In the first case σ_o was fixed to 0.05, while in the second to 0.01. For all parameters we used flat priors and for SLMH we used the set of 16 summary statistics proposed by Wood (2010) for this model. We report these details in Appendix B.1.

The resulting Mean Squared Errors (MSEs) of the log-parameters are reported in Table 3.1. The table includes the p-values for differences in MSEs, which clearly show that PMMH is more accurate when the lower value of σ_o is used. On the other hand, in the more realistic setting the performance of the two procedure is more comparable, as PMMH underestimates both σ_p^2 and σ_d^2, while SLMH performs slightly worse than PMMH on the remaining parameters.

3.2.3 Results using Nicholson's datasets

Fitting Nicholson's datasets was relatively straightforward with SLMH, and we used the same initial values ($\delta = 0.16$, $P = 6.5$, $n_0 = 400$, $\sigma_p^2 = 0.1$, $\tau = 14$, $\sigma_d^2 = 0.1$) for each dataset. Using this initialization was not possible for PMMH, as we would be forced to use values of σ_o as high as 0.2, in order to avoid failures in the Monte Carlo integration step. In particular, when this initialization was used, the dynamics simulated from the model were very different from the observed ones. This is due to two facts: this initialization is quite distant from the MLE and, as we will discuss later,

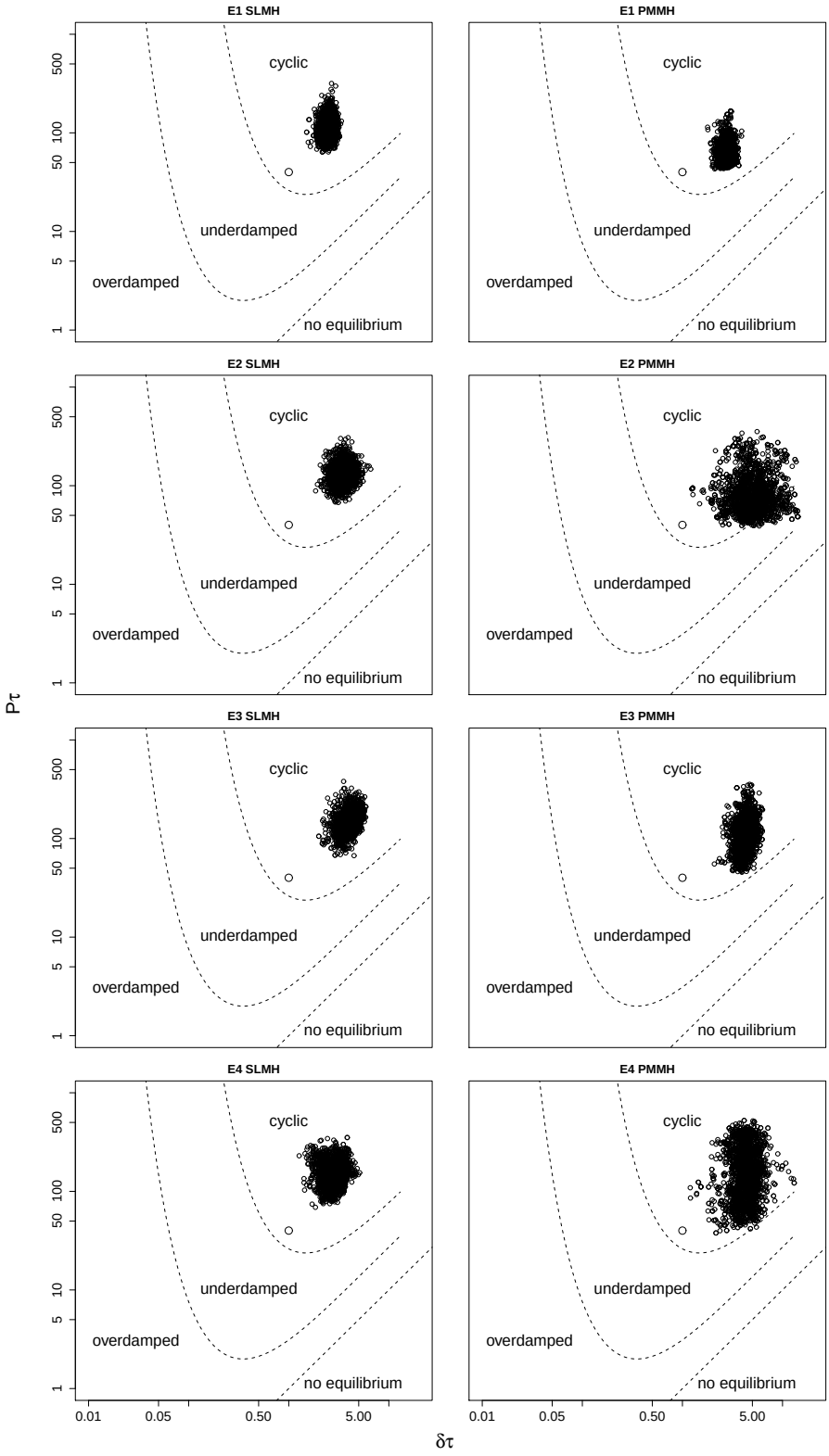

Figure 3-2: *Stability plots for the blowfly model, obtained by fitting Nicholson's datasets using SLMH and PMMH. The black dots are 2000 values of the $P\tau$ and $\delta\tau$ randomly sampled from each MCMC chain. The white circle represents the initial value used for SLMH. Notice that $P\tau$ and $\delta\tau$ indicate simply the product between these parameters.*

model (3.2) is probably misspecified. Hence, most of the particles simulated by SIR were very far from the corresponding observations and, when σ_o was set to a small value, they ended up being attributed almost zero weight. This lead to extremely noisy likelihood estimates. To avoid this problem, we initialized PMMH using values obtained through preliminary runs of SLMH on the four datasets. Still, we were forced to use values of σ_o equal to 0.1 for the second dataset and 0.05 for the others. For each dataset we used 3×10^4 MCMC iterations, of which the first 5000 were discarded as burn-in. The (synthetic) likelihood was estimated using 1000 particles or simulated paths at each step.

Figure 3-2 shows the stability diagrams for model (3.2), for each combination of dataset and fitting procedure. These plots show how the stability properties of the system depend on the parameter combinations $P\tau$ and $\delta\tau$. All posterior samples obtained through SLMH lay strictly in the cyclic region of the parameter space, indicating that observed oscillation of blowfly population are due to intrinsic blowfly biology, rather than stochastic perturbation of the system (Wood, 2010). On the other hand, the posteriors samples given by PMMH, in particular those corresponding to datasets E2 and E4, are closer to the under-damped region, where the oscillations are driven by the stochasticity rather than intrinsic effects. With the exception of E1, the PMMH posteriors are more dispersed, which is attributable to the high estimates of noise parameters σ_d^2 and σ_p^2, as shown in Table 3.2.

	δ	P	n_0	σ_p^2	τ	σ_d^2
E1 SLMH	0.17	7.57	395.30	0.70	14.44	0.47
E1 PMMH	0.19	4.45	653.93	1.54	14.82	0.30
E2 SLMH	0.22	8.70	407.61	0.21	15.95	1.77
E2 PMMH	0.37	6.26	576.30	2.35	15.02	3.47
E3 SLMH	0.29	10.48	184.38	0.64	14.62	0.55
E3 PMMH	0.28	7.71	229.32	1.56	15.18	0.53
E4 SLMH	0.22	12.81	59.16	0.71	12.91	0.55
E4 PMMH	0.30	12.10	88.33	2.42	14.46	1.23

Table 3.2: *Posterior means for model (3.2), obtained by fitting each of Nicholson's dataset using either SLMH or PMMH.*

Figure 3-1 compares the observed trajectories with those simulated from the model, using parameter values equal to the posterior means estimated by SLMH and PMMH. While using parameter values estimated through SLMH gives trajectories that are qualitatively similar to the observed ones in all cases, using the parameters estimated

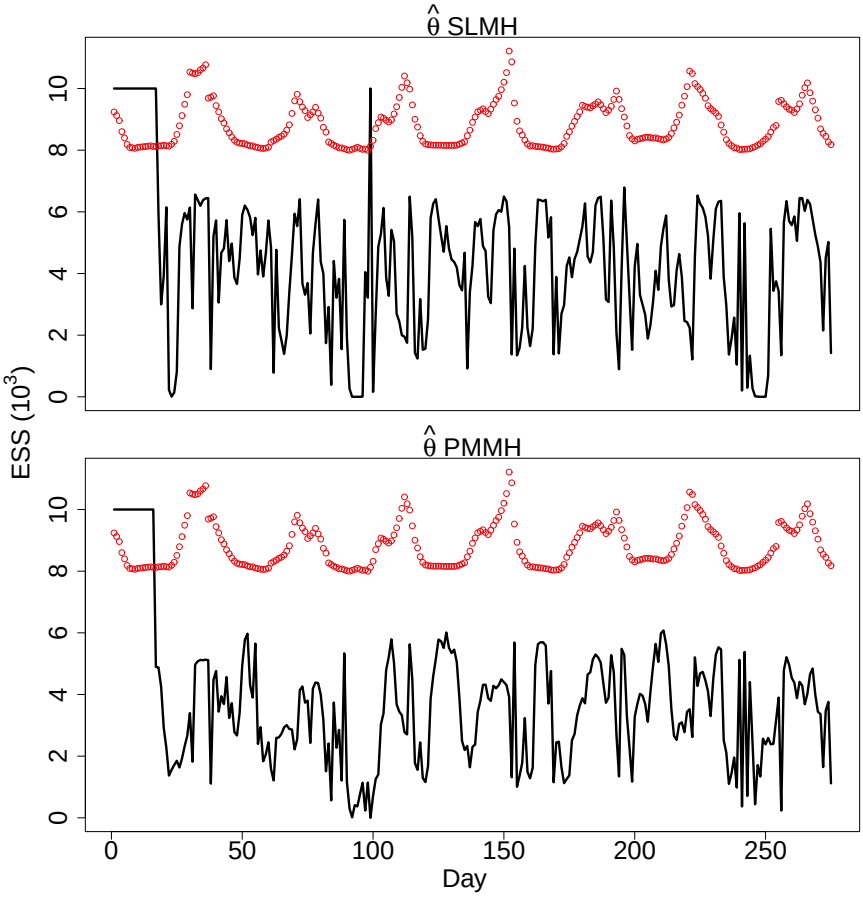

Figure 3-3: *Dynamics of the ESS (black line) for the E2 dataset (red points), using parameters equal to the posterior means given by SLMH (top) and PMMH (bottom). For the first τ steps the ESS is equal to the number of particles, because we have set $n_i = y_i$, for $i = 1, \ldots, \tau$, as stated in the main text.*

through PMMH gives a poor match for datasets E2 and E4.

To understand what happened, we ran a filtering operation using dataset E2, 10^4 particles and parameters equal to the posterior mean given by SLMH and PMMH. Figure 3-3 shows the dynamics of the Effective Sample Size (ESS) using either parameter set. From the top plot we see that the ESS drops to practically zero around the 25th, 95th and 250th observation, if SLMH estimates are used. On the other hand, PMMH gives much higher estimates of σ_p and σ_d and this keeps the ESS from dropping to zero in those occasions. This suggests that few idiosyncrasies or outliers in datasets E2 and E4 might be pushing PMMH toward the underdamped region. This is supported by the fact that, if PMMH is run using a log Student's t-distribution for the observational process

$$\frac{\log y_t - \log n_t}{\sigma_o} \sim \text{Student}(\nu = 2),$$

the resulting posterior estimates for E2 and E4 lay strictly inside the cyclic region, as shown in Figure 3-4. We comment on these results in Section 3.5.

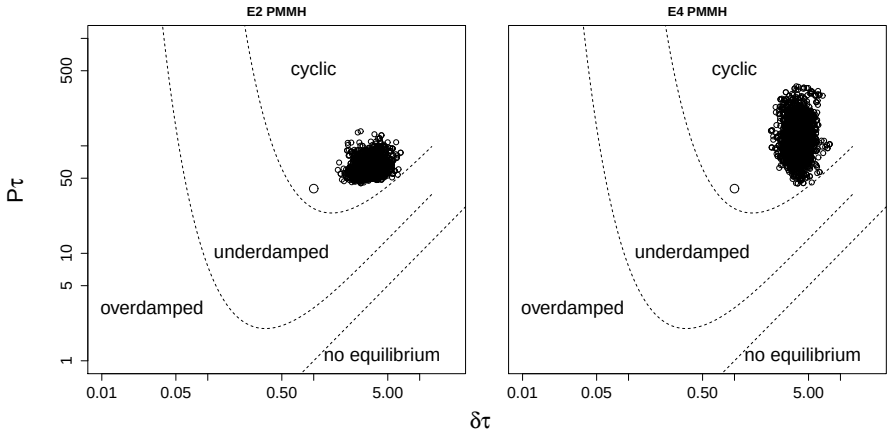

Figure 3-4: *Stability plots for datasets E2 and E4 using PMMH with log Student's t observational error.*

3.3 Cholera epidemics in the Bay of Bengal

In this section we consider a modified version of the Susceptible-Infected-Recovered-Susceptible (SIRS) model used by King et al. (2008) to explain cholera epidemics in the regions north of the Bay of Bengal. The dataset considered here corresponds to cholera-related mortality records in the former Dacca district of British East Indian province of Bengal, which is available within the *pomp* R-package (King et al., 2014). The data, depicted in Figure 3-5, consists of monthly deaths counts occurring between 1891 and 1941. See King et al. (2008) for additional details regarding the data.

3.3.1 The model

The model proposed by King et al. (2008) is composed of several classes, all of which are completely unobserved apart from the infected class, which is observed indirectly through the deaths count. In King et al. (2008) the model was represented by a system of differential equations, which was solved numerically using a Euler-Maruyama scheme. The main issue with their formulation is that the positivity of the states is not guaranteed. To address this problem, we propose an alternative model formulation, to be justified later, which results in the following system of difference equations

$$
\begin{aligned}
s_{t+1} &= s_t - s_t^o + \frac{r_{kt}^o k\epsilon}{k\epsilon+\delta} + \frac{y_t^o \rho}{\rho+\delta} + b_{t+1},\\
i_{t+1} &= i_t - i_t^o + c\frac{s_t^o \lambda_t}{\lambda_t+\delta},\\
y_{t+1} &= y_t - y_t^o + (1-c)\frac{s_t^o \lambda_t}{\lambda_t+\delta},\\
r_{1t+1} &= r_{1t} - r_{1t}^o + \frac{i_t^o \gamma}{m+\gamma+\delta},\\
r_{it+1} &= r_{it} - r_{it}^o + \frac{r_{i-1t}^o k\epsilon}{k\epsilon+\delta}, \quad \text{for } i = 2,\ldots,k,
\end{aligned}
$$

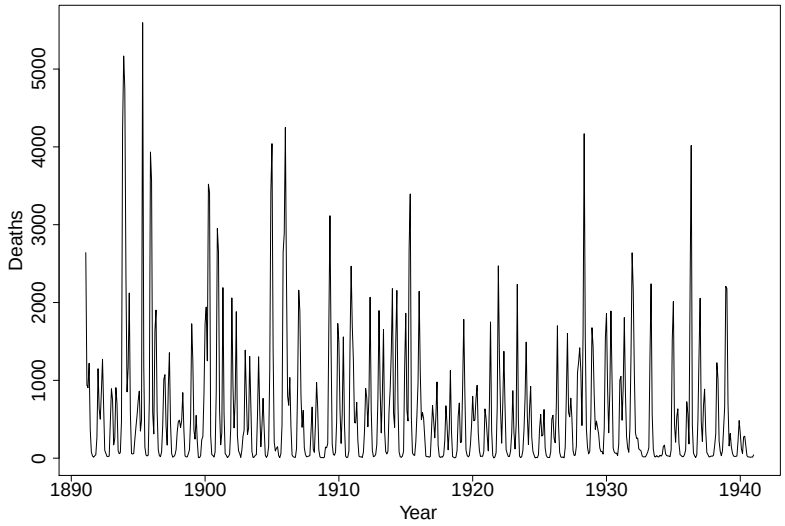

Figure 3-5: *Cholera-related monthly death count in the Dacca district between 1891 and 1941.*

where

$$
\begin{aligned}
b_{t+1} &= p_{t+1} - p_t + \frac{s_t^o \delta}{\lambda_t + \delta} + \frac{i_t^o \delta}{m + \gamma + \delta} + \frac{y_t^o \delta}{\rho + \delta} + \sum_{i=1}^{k} \frac{r_{it}^o \delta}{k\epsilon + \delta}, \\
s_t^o &= s_t(1 - e^{-(\lambda_t + \delta)\Delta t}), \\
i_t^o &= i_t(1 - e^{-(m+\gamma+\delta)\Delta t}), \\
y_t^o &= y_t(1 - e^{-(\rho+\delta)\Delta t}), \\
r_{it}^o &= r_{it}(1 - e^{-(k\epsilon+\delta)\Delta t}), \quad \text{for} \quad i = 1, \ldots, k.
\end{aligned}
\tag{3.3}
$$

Here b_{t+1} represents the number of births between time t and $t+1$, while p_t is the total population of the Dacca district at time t, characterized by constant birth-death rate δ. Susceptible individuals s are infected by cholera at time-varying rate λ_t, which will be explained in detail later. Parameter c determines the fraction of infected individuals that will undergo a full blown infection, represented by class i, rather than an asymptomatic infection, represented by class y. Individuals in i suffer from an excess death rate m and transition to the first Recovered class r_1 with rate γ. On the other hand, individuals in y have the same death rate as susceptible individuals and do not acquire any long term immunity, as they rejoin the s class directly at rate ρ. The duration of immunity is gamma distributed, with mean $1/\epsilon$ and variance k/ϵ^2.

The rationale behind our discretized model needs to be clarified. Consider, for instance, y_t. To obtain y_{t+1} we model inputs and outputs involving y in turn, rather than simultaneously. Firstly, we obtain the number of individuals, y_t^o, leaving the asymptomatic infected class by solving

$$dy_s = -(\rho + \delta) y_s ds,$$

between t and $t+1$. The resulting solution is an exponential decay, which ensures the positivity of y_{t+1}. Then y_t^o is divided between b_{t+1} and s_{t+1}, with proportions

determined by the output rates δ and ρ. This solution preserves the positivity of all classes and mass-balance, both of which are essential for a realistic model. In addition, our formulation becomes equivalent to the Euler-Maruyama scheme of King et al. (2008), as $\Delta t \to 0$.

The force of infection λ_t is given by

$$\lambda_t = \omega_t + e^{\beta t}\beta_t \frac{i_t}{p_t}\frac{\Delta w}{\Delta t}, \tag{3.4}$$

where $\Delta w \sim \Gamma(\Delta t/\sigma^2, 1/\sigma^2)$, so that $\Delta w/\Delta t$ represents multiplicative gamma noise with unit mean and variance equal to σ^2. We preferred this choice to the additive Gaussian noise originally used by King et al. (2008), because the multiplicative version assures the positivity of λ_t.

In (3.4), ω_t and β_t represent respectively the environmental and human feedback components of the force of infection

$$\omega_t = \exp\left(\sum_{i=1}^{6} \omega_i g_i(t)\right),$$

$$\beta_t = \exp\left(\sum_{i=1}^{6} \beta_i g_i(t)\right),$$

where $g_i(t)$, for $i = 1, \ldots, 6$, are a periodic B-spline basis. Parameter β is the long term trend in human-to-human transmission.

The observed number of deaths registered during the n-th month, is assumed to follow a negative binomial distribution

$$e_n \sim \mathrm{NB}\left(q_n, \frac{1}{\tau^2}\right),$$

with mean q_n and variance $q_n + q_n^2/\tau^2$, where q_n is the accumulated number of cholera-related deaths between the previous and the current month

$$q_n = \sum_{s=t_{n-1}}^{t_n} m i_s.$$

In the original model e_n was normally distributed around q_n, but that choice often produces negative death counts when the model is simulated. See King et al. (2008) for further model details.

In the Section 3.3.2 we fit three versions of model 3.3 to the Dacca dataset. Our results roughly agree with those of King et al. (2008), with some notable exceptions, which will be illustrated in Section 3.3.2. Given that the version of the model presented here has the desirable property of enforcing the positivity of all the states and of the force of infection, we argue that the results presented here, in particular those obtained under PMMH, should be preferred to those of King et al. (2008).

3.3.2 Set-up and results using the Dacca dataset

Similarly to King et al. (2008), we do not fit the full model, but we consider:

- a seasonal model where the y class is not included ($c = 1$);
- a two-path model were the environmental force of infection is constant ($\omega_s(t) = \omega_s$);
- a basic SIRS model where $c = 1$, $\omega_s(t) = \omega_s$ and $\beta_s(t) = \beta_s$.

We fitted each model to the Dacca dataset using SLMH and PMMH. For both methods we used 1.4×10^6 MCMC iterations, the first half of which was discarded as burn-in period, and 2000 simulations to estimate the (synthetic) likelihood at each step. We used uniform or diffuse priors for all parameters. We report them, together with the 26 summary statistics used by SLMH, in Appendix B.2.

Table 3.3 reports the estimated Akaike Information Criterion (AIC) for each model and method. SLMH and PMMH agree in selecting the seasonal reservoir model, while the two paths mechanisms does not improve the fit enough, relatively to the SIRS model, to justify the additional complexity. This is in contrast with the results of King et al. (2008), whose second-order AIC estimate was lower for the two paths than for the SIRS model.

Almost all the marginal posterior variances were higher when SLMH was used, with a median increase equal to 7.2, 2.6 and 2.2 for the seasonal, two paths and SIRS model, respectively. The variance increases were highest for the seasonal coefficients, $\omega_{1:6}$, of the force of infection, which suggest that the amount of information lost through the use of summary statistics is sizeable.

Method	Seasonal	Two Paths	SIRS
AIC_{SLMH}	-38.4	-31.6	-34.6
AIC_{PMMH}	7458	7532.6	7528.2

Table 3.3: *Estimated AICs for each model, using SLMH and PMMH.*

One important hypothesis examined by King et al. (2008) was that the mean duration of immunity, $d_L := 1/\epsilon$, might be much shorter than previously thought. Our analysis partially supports this conclusion, as shown by Figure 3-6. The plots in the top row show the marginal densities of d_L under each model. Under the seasonal model, most of the posterior mass lies close to the lower prior boundary, corresponding to unrealistically low periods of immunity (shorter than one week). The posterior given by SLMH under the SIRS model is slightly less extreme, but it still suggests period of immunity of one to three months, which is much shorter than the 3 to 10 years timescale suggested by several sources (Cash et al., 1974; Glass et al., 1982; Koelle et al., 2005). One surprising result is that, under the two paths model, d_L is still estimated to be lower than one month. This is in contrast with the results of King et al. (2008), who estimates d_L to be around 1.4 years, under the same model and dataset. The mean duration of immunity after mild infections $d_S = 1/\rho$ is estimated to be shorter than three weeks under PMMH, while SLMH seems to have lost information regarding d_S, as the corresponding marginal posterior is bimodal and highly dispersed.

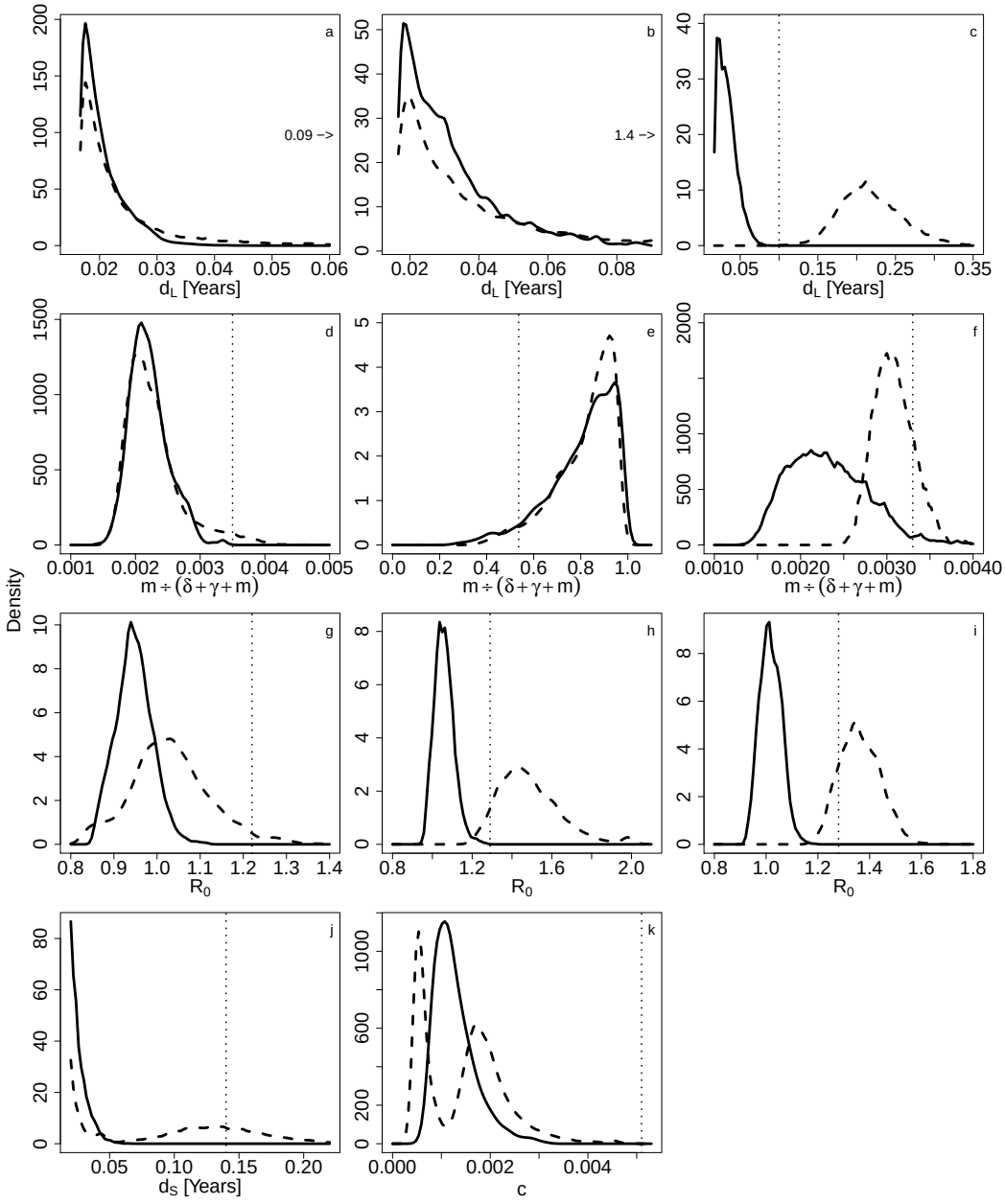

Figure 3-6: *Posterior marginal distributions from PMMH (solid) and SLMH (dashed). The estimates of King et al. (2008) correspond to the vertical dotted lines, substituted by annotations when out of range. The first three rows contain the marginals of immunity duration after full-blow infections, fatality and basic reproductive number for the seasonal (a, d, g), two paths (b, e, h) and SIRS (c, f, i) model. The last row shows the marginals of immunity duration after mild infections (j) and of the fraction of severe infections (k) for the two paths model.*

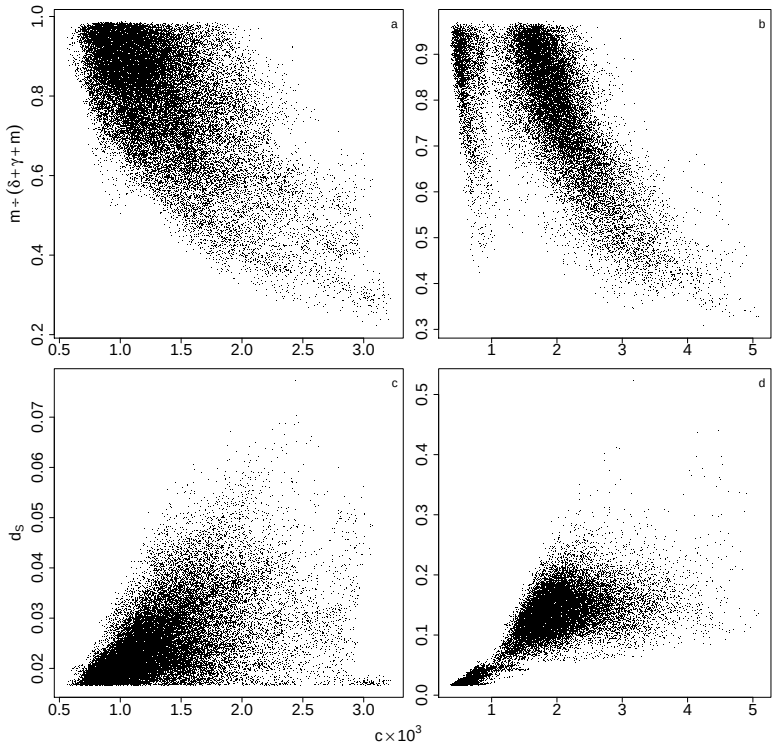

Figure 3-7: *Joint posterior samples for fraction of symptomatic infections vs fatality and duration of short term immunity under PMMH (a, c) and SLMH (b, d).*

Figure 3-6 shows also the marginal distributions of the cholera-related death probability $f = m/(\delta + \gamma + m)$. Under the seasonal and the SIRS models our estimates roughly agree with those of King et al. (2008), but our fatality estimate is much higher than theirs when asymptomatic infections are included in the model. Similarly to King et al. (2008), we estimate the fraction of infection that are symptomatic to be very low under the two path model.

Our results suggest that including asymptomatic infections does not improve the fit and does not provide more realistic estimates of immunity duration, following full-blown infections. In addition, this model is difficult to identify, because there is a trade-off between parameters c, d_S and m, which is captured by Figure 3-7. The correlations observed in the PMMH joint posterior sample are explained by the fact that an increase in the fraction of individuals with full infection can be compensated by decreasing their mortality rate or by increasing the duration of long short term immunity (thus delaying individuals with mild infection from rejoining the susceptible). Under SLMH this identifiability issue is more severe, and the corresponding posteriors are bimodal and more dispersed.

Another question addressed by King et al. (2008) is the relative importance of the environmental reservoir and of the human habitat for *V.Cholerae* persistence. They found that the basic reproductive number, R_0, which quantifies the strength of human-to-human transmission, was consistently low (around 1.5) across model and geographic area. Figure 3-6 shows that our estimates of R_0 are very low under all models and methods, thus supporting the hypothesis that humans might be only a marginal habitat for *V.Cholerae*.

3.4 Fennoscandian Voles

Here we consider a modified version of the prey-predator model proposed by Turchin and Ellner (2000), which has been used to describe the dynamics of Fennoscandian voles (*Microtus* and *Clethrionomys*). More specifically, the model was an attempt at explaining the shift in voles abundance dynamics from low-amplitude oscillations in central Europe and southern Fennoscandia to high-amplitude fluctuations in the north. One of the possible drivers of this shift is the absence of generalist predators in the north, where voles are hunted primarily by weasels (*Mustela nivalis*) (Turchin and Ellner, 2000). According to this hypothesis, the lack of the stabilizing effect of generalist predators is the main factor determining the observed instability of voles abundances in the north.

The predator-prey dynamics are given by the following system of differential equations (Turchin and Ellner, 2000)

$$\frac{dN}{dt} = r(1 - e\sin 2\pi t)N - \frac{r}{K}N^2 - \frac{GN^2}{N^2 + H^2} - \frac{CNP}{N + D} + \frac{N}{K}\frac{dw}{dt},$$

$$\frac{dP}{dt} = s(1 - e\sin 2\pi t)P - sQ\frac{P^2}{N}, \tag{3.5}$$

where $dw(t_2) - dw(t_1) \sim N[0, \sigma^2(t_2 - t_1)]$, with $t_2 > t_1$, is a Brownian motion process with constant volatility σ. The model is formulated in continuous time, because voles do not reproduce in discrete generations (Turchin and Hanski, 1997). Here N and P indicate voles and weasels abundances, respectively. In the absence of predators, voles abundance grows at a seasonal logistic rate. Parameters r and s represents the intrinsic population growth rates of voles and weasels, while K is the carrying capacity of the former. These parameters are averaged over the seasonal component, which is modelled through a sine function with amplitude e and period equal to one year, with peak growth achieved in the summer. Generalist predation is modelled through a type III functional response, under which generalists progressively switch from alternative prey to hunting voles, as voles density increases. The maximal rate of mortality inflicted by generalists is G, while H is the half saturation parameter.

Predation by weasels follows a type II response, where C is the maximal predation rate of individual weasels and D is the half saturation prey density. No prey-switching behaviour occurs under this functional response, which is consistent with weasels being specialist predators. Weasel abundance grows at a seasonal logistic rate, where the carrying capacity depends on prey density. Parameter Q specifies the number of voles needed to support and replace an individual weasel and it determines the ratio of prey to predator densities at equilibrium.

Differently from Turchin and Ellner (2000), who include environmental stochasticity in the system by randomly perturbing all model parameters using Gaussian noise with pre-specified volatility, we choose to explicitly perturb the prey equation using a Brownian motion process and to include its volatility σ in the vector of unknown parameters.

Vole abundance is not observed directly, but a proxy is provided by trapping data. We assume that the number of trapped voles is Poisson distributed

$$Y_t \sim \text{Pois}(\Phi N_t),$$

where $t \in \{1, \dots, T\}$ is the set of discrete times when trapping took place. No such proxy is available for weasels density, hence predator abundance represents a completely hidden state.

Following Turchin and Ellner (2000) model (3.5) is not fitted directly to data, but it is rescaled to a dimensionless form first. In particular, if we define

$$n = \frac{N}{K}, \quad p = \frac{QP}{K}, \quad d = \frac{D}{K}, \quad a = \frac{C}{K}, \quad g = \frac{G}{K}, \quad h = \frac{H}{K} \quad \text{and} \quad \phi = \Phi K,$$

the reduced system is given by

$$\frac{dn}{dt} = r(1 - e\sin 2\pi t)n - rn^2 - \frac{gn^2}{n^2 + h^2} - \frac{anp}{n + d} + n\frac{dw}{dt},$$

$$\frac{dp}{dt} = s(1 - e\sin 2\pi t)p - s\frac{p^2}{N},$$

$$Y_t \sim \text{Pois}(\phi n_t). \tag{3.6}$$

While Turchin and Ellner (2000) implicitly re-scaled the simulations from the model in order to match their means with that of the observed data, we formally estimate the scaling parameter ϕ.

Turchin and Ellner (2000) fitted the model by using a method which they call Non Linear Forecasting (NLF), which is an instance of Simulated Quasi-Maximum Likelihood (SQML) method (Smith, 1993). One of the drawbacks of their estimation procedure is that it does not take into account the fact that trapping data provides noisy estimates of voles density. Another issue is that their method could not be used to estimate parameters that affect the variance of conditional distributions $p(n_t|n_{t-1}, n_{t-2}, \dots)$, but not their mean (Turchin and Ellner, 2000).

3.4.1 Description of data and priors

While Turchin and Ellner (2000) consider several datasets, here we focus on the time series concerning voles abundance (mainly *Clethrionomys rufocanus*) in Kilpisjarvi, Finland. The data, shown in Figure 3-8, consists of 90 data points collected during the springs (mid-June) (triangles) and autumns (September) (stars) of each year, between 1952 and 1997. Each data point represents the number of voles trapped in a specific trapping season, divided by the number of hundred trap-nights used in that season. After 1980 the number of trap-nights was fixed to around 1000, but in earlier years this number is not available: it varied from a minimum of 500 to more than a thousand (Perry, 2000). This correction for the sampling effort implies that, if the number of the trapped voles in each season is approximately Poisson distributed, the trapping index is not.

We have dealt with this problem by multiplying the data in Figure 3-8 by 10 and by rounding each data-point to the nearest integer. This solution should give near-exact results for data collected after 1980, and a good approximation for all data-points representing a considerable population, thanks to the normal approximation to the Poisson distribution.

A useful source of prior information is represented by Turchin and Hanski (1997), where life history and data from short experiments were used to estimate the parameters

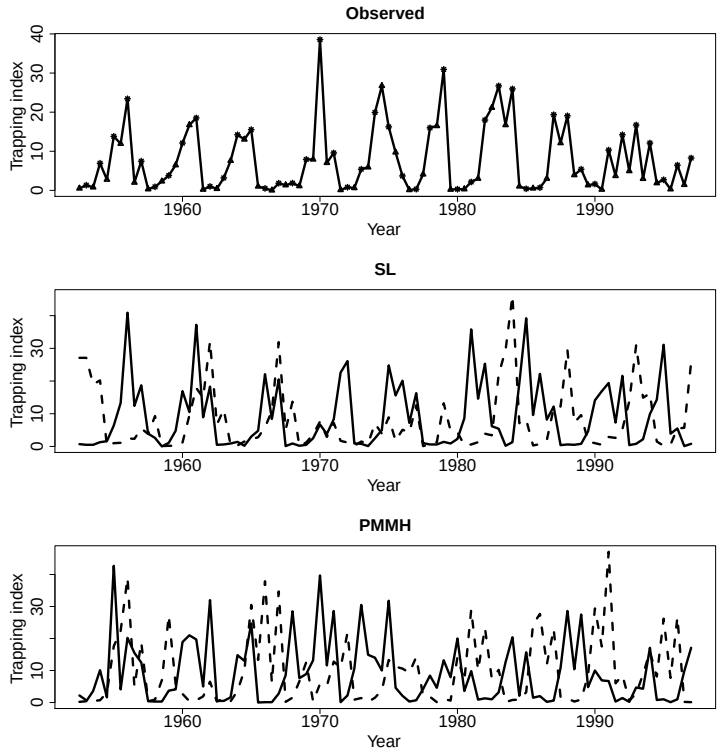

Figure 3-8: *Top: observed voles trapping index in Kilpisjarvi, between 1952 and 1997. Middle and bottom: two realization (solid and dashed) of model 3.5, using parameters equal to the posterior means given by SLMH and PMMH.*

of model (3.6). We report the prior distributions for each parameter in Table 3.4. The expected values of the prior distributions has been chosen on the basis of the remarks of Turchin and Hanski (1997), and we refer the reader to this reference for further details. The specific distributions and variabilities used for the priors have been chosen based on an attempt at quantifying the remarks of Turchin and Hanski (1997) regarding their confidence in their independently derived estimates. Admittedly, this process entails a certain degree of arbitrariness. No prior information was available for ϕ and σ, hence we have used improper uniform priors for both parameters. For SLMH we used the set of 17 summary statistics reported in Appendix B.3.

3.4.2 Comparison using simulated data

In order to verify the accuracy of SLMH and PMMH for this prey-predator model, we have simulated 24 datasets of length $T = 90$, using parameters values $r = 4.5$, $e = 0.8$, $g = 0.2$, $h = 0.15$, $a = 8$, $d = 0.06$, $s = 1$, $\sigma = 1.5$ and $\phi = 100$. We have then estimated the parameters with both methods, using 2.5×10^4 MCMC iteration, the first 5×10^3 of which were discarded as burn-in period, and 10^3 simulation from the model at each step. All the chains were initialized at the same parameter values. The resulting Root Mean Squared Errors (RMSEs) and variance-to-squared-bias ratios are reported in Table 3.5. While the RMSEs are quite similar for most parameters, the Table suggests that PMMH gives more accurate estimates for the scaling parameter ϕ and possibly for the generalist predation rate g. Indeed, SLMH estimates of ϕ are biased

Parameter	Prior distribution
r	$N(\mu = 5, \sigma = 1)$
e	$N(\mu = 1, \sigma = 1)$
g	$\text{Exp}(\lambda = 7)$
h	$\text{Gamma}(\kappa = 4, \theta = 40)$
a	$N(\mu = 15, \sigma = 15)$
d	$N(\mu = 0.04, \sigma = 0.04)$
s	$N(\mu = 1.25, \sigma = 0.5)$
σ	$\text{Unif}(0.5, \infty)$
ϕ	$\text{Unif}(0, \infty)$

Table 3.4: *Priors used for the voles-weasels model.*

downward and are around ten times more variable than the estimates obtained with PMMH. In the case of g the significance of the t-test should not be over-interpreted, given that it is attributable to PMMH achieving almost zero error on a single run.

3.4.3 Results from the Kilpisjarvi dataset

We fitted the Kilpisjarvi dataset using 1.5×10^5 MCMC iteration, of which the first 10^4 were discarded as burn-in period. At each step we used 10^3 simulations from the model (SLMH) or particles (PMMH). The resulting posterior means are reported in Table 3.6, while the marginal posterior densities of the parameters as shown in Figure 3-9.

SLMH and PMMH give similar estimates for most parameters, with substantial differences only for σ and ϕ. Indeed, PMMH's estimate of the former parameter is much higher than that obtained using SLMH. Interestingly, we encontered a similar pattern in Section 3.2, when fitting the blowfly model of Wood (2010) to Nicholson's experimental datasets (Nicholson, 1954, 1957). In that context, the process noise estimates were much higher under PMMH than under SLMH, on all datasets. This biased PMMH's estimates of the remaining parameters towards stability, particularly on two of the datasets. As we will show later in this section, this stabilizing effect of high process noise estimates on the dynamics is less noticeable here.

Figure 3-8 compares the observed data with trajectories simulated from model (3.5), using parameters equal to the posterior means given by SLMH and PMMH. Both methods seem to produce dynamics that are qualitatively similar to the observed ones, with the paths simulated using PMMH's estimates being slightly more irregular,

Parameter	RMSE SLMH	RMSE PMMH	P-value	Best
r	0.33(3.3)	0.25(9.9)	0.49	PMMH
e	0.19(0.1)	0.2(0.1)	0.78	SLMH
g	0.09(0.2)	0.08(0.5)	0.05	PMMH
h	0.04(0.2)	0.03(0.4)	0.15	PMMH
a	2.12(1.3)	1.97(1)	0.48	PMMH
d	0.02(0.5)	0.02(0.6)	0.57	SLMH
s	0.07(18.6)	0.08(10.9)	0.22	SLMH
σ	1.97(2.5)	0.71(2.1)	0.36	PMMH
ϕ	16.04(3.9)	4.85(7.4)	< 0.001	PMMH

Table 3.5: *RMSEs and variance-to-squared-bias ratios (in brackets) for SLMH and PMMH. P-values for differences in log-squared errors have been calculated using t-tests.*

which is attributable to the higher process noise estimate.

One of the main scientific questions model (3.5) was meant to address was whether the observed dynamic in voles densities can be classified as chaotic. To answer this question, we have randomly sampled 10^3 parameters sets from posteriors samples obtained by SLMH and PMMH. We have then used each parameters set to simulate a trajectory from the deterministic skeleton of model (3.5) for 10^5 months, which were discarded in order to let the system leave the transient, and used additional 10^4 months of simulation to estimate the maximal Lyapunov exponent as in Wolf et al. (1985). By doing this, we obtained the two approximate posterior densities of the Lyapunov exponent shown in Figure 3-10. Notice that the posterior produced by PMMH is slightly more skewed to the left relatively to that obtained with SLMH, which suggests that the system dynamics are estimated to be more stable under the former methods. Together with the high estimate of σ^2, this confirms the tendency of PMMH to inflate the noise and to bias the estimated dynamics toward stability. While this effect was very pronounced under the blowfly model studied described in Section 3.2, in this case it is very mild. Indeed, the median Lyapunov exponent is equal to -6×10^{-4} for SLMH and -0.015 for PMMH. These estimates are very close to each other and to the one (-0.02) reported by Turchin et al. (2003) for this dataset, and provide more model-based evidence supporting the hypothesis that this system lives on the edge of chaos.

	r	e	g	h	a
SLMH	4.85(0.63)	0.78(0.12)	0.11(0.11)	0.1(0.05)	8.0(3.3)
PMMH	5.11(0.7)	0.84(0.14)	0.14(0.11)	0.1(0.05)	6.3(2.1)
	d	s	σ	ϕ	
SLMH	0.07(0.03)	1.04(0.21)	8.4(2.3)	270.5(63.5)	
PMMH	0.08(0.03)	1.04(0.23)	14.8(1.7)	184.2(26.9)	

Table 3.6: *Estimated posterior means (standard deviations) for model 3.5.*

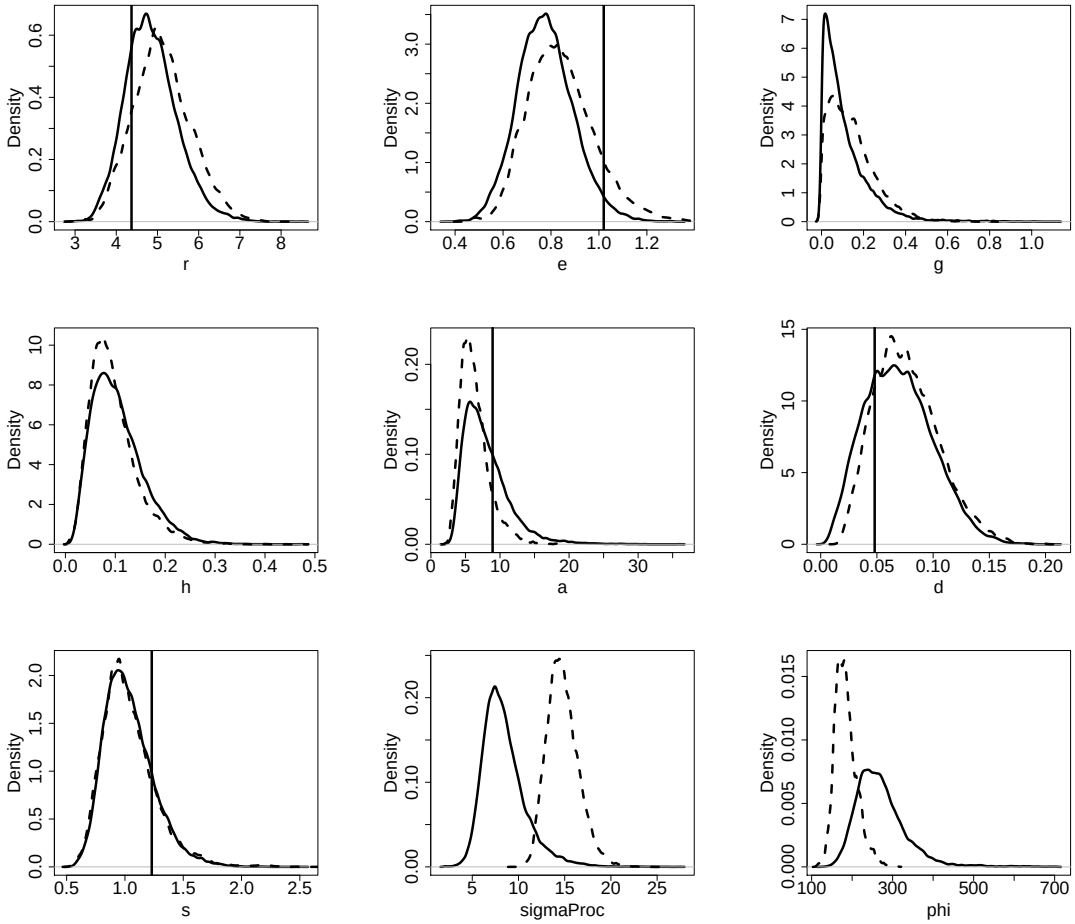

Figure 3-9: *Marginal posterior densities for voles model using SLMH (black) and PMMH (broke). The vertical lines correspond to estimates reported by Turchin et al. (2003), obtained using NLF (available only for 5 parameters).*

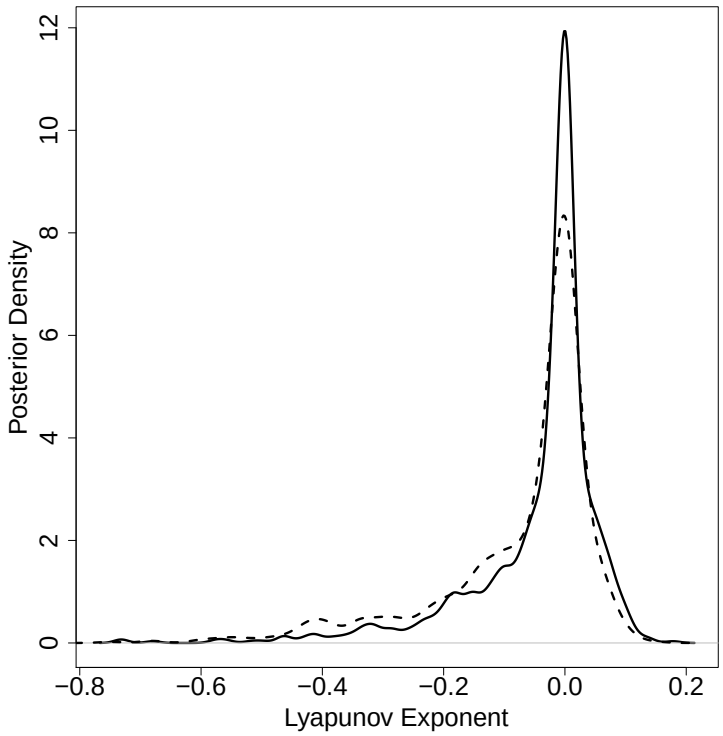

Figure 3-10: *Approximate posterior densities of Lyapunov exponents for SLMH (black) and PMMH (broken).*

3.5 Conclusions

Information reduction methods, such as ABC and SL, have become popular tools for dealing with complex phylogeographic (Hickerson et al., 2010), phylogenetic (Rabosky, 2009) and individual based (Hartig et al., 2014) models, but they do not seem to have been equally successful for dynamical SSMs of ecological interest. The main reasons for this might be that particle filters represent an obvious alternative, and that at the moment it is not clear whether information reduction methods can outperform them along any dimension of the inferential process. In fact, particle filters have the important advantage of using the full data, $\boldsymbol{y}^0$, thus avoiding both the information loss and the issue of choosing the summary statistics. On the other hand, this use of all the data makes filtering more susceptible to model mis-specification problems, in which failures to capture the data generating mechanism exactly can have a substantial negative impact on inference.

The robustness properties of methods based on summary or "intermediate" statistics, in particularly the protection they can offer against model mis-specification and outliers, has been widely recognized and exploited in econometrics, but it seems to have attracted less attention in the wider statistical community (Jiang et al., 2004). The blowflies example in Section 3.2, highlights the robustness of information reduction methods in the context of highly non-linear SSMs. Indeed, careless application of PMMH would have classified the dynamics of the system as nearly-underdamped under two of Nicholson's datasets, with the corresponding simulations from the model being

clearly inconsistent with the data (see Figure 3-1). On the contrary, SLMH reliably classifies the dynamics as cyclic. In this example using a fat-tailed observation density mitigated the problem, but we argue that these results have deeper practical implications. Model 3.2 has sufficient flexibility to reproduce the main features (quantified by the summary statistics) of Nicholson's datasets, as demonstrated by Figure 3-1. On the other hand, the model struggles to explain certain nuances of Nicholson's datasets, which leads to most of the particle filter's importance weights being close to zero at several time steps, as illustrated by Figure 3-3. This suggests that, in situations in which the model has a clear scientific interpretation, but lacks the ability to explain the observed dynamics in all their complexity, focusing on some salient features of the data might be a reasonable approach. Conversely, if the model is believed to be an accurate description of the system under study, or if it is meant to be used for the purpose of state estimation or forecasting, then it is compelling to fit it using the full data.

Another lesson learned from the blowflies example is that, for particle-filtering-based methods to work properly, a good initialization is often indispensable. This is because these methods are generally based on some form of importance sampling, hence when the initial estimates are far from the best fitting parameters most of the importance weights go to zero (particle depletion). In this context, methods based on information reduction can be useful, because they are robust to bad initializations. Methods that can provide reliable initial estimates, to be fed to more accurate but less robust methods, are of high practical value, but often under-represented in the literature. Exceptions are Lavine et al. (2013) who, in the context of pertussis epidemics, use SLMH to initialize a IF algorithm and Owen et al. (2014), who proposes to initialize PMMH using the output of preliminary ABC runs.

One recurrent theme in the examples presented in this chapter and in Chapter 2 is that using summary statistics might lead to a loss of accuracy in parameter estimation. Mild losses of accuracy are often acceptable when parameter estimation is not the main focus of analysis, but the aim is, for example, to determine whether the dynamics of the system are stable, oscillatory or chaotic, as in the blowflies and the voles examples. On the other hand, when dealing with models that are weakly identified even under the full data, as in Section 3.3, any further loss of information can lead to unreliable estimates. Hence, an important drawback of information reduction methods is that, in the absence of a benchmark, quantifying inferential inaccuracies require running simulation studies, which can be prohibitively expensive for complex models, such as those presented in Section 3.3. While in all the examples presented here a benchmark (PMMH) was available, this not always the case.

From a computational point of view, SLMH and PMMH performed similarly under the models considered in this chapter. For instance, a single point-wise estimate of $p(\boldsymbol{y}^0|\boldsymbol{\theta})$ or $p_{SL}(\boldsymbol{y}^0|\boldsymbol{\theta})$, under the voles model and on a single 2.50GHz Intel i7-4710MQ CPU, costs around 1.55 and 1.35 seconds, when 10^3 particles or simulated statistics vectors are used. This time difference is marginal, and probably highly dependent on implementation details. However, it is worth pointing out that is it much easier to parallelize the computation of $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ than that of $\hat{p}(\boldsymbol{y}^0|\boldsymbol{\theta})$. This is because of SIR's resampling step, which breaks the parallelisms at each time-step t (see Algorithm 1 in Chapter 2). For a review of parallelization strategies for the resampling step, see Li et al. (2015). A possibly simpler solution is to compute several estimates $\hat{p}_1(\boldsymbol{y}^0|\boldsymbol{\theta}), \ldots, \hat{p}_C(\boldsymbol{y}^0|\boldsymbol{\theta})$ in parallel, by running SIR with a fraction of the total number of particles M

on each of the C cores, and then average them at each PMMH step to obtain a single estimate of $p(\boldsymbol{y}^0|\boldsymbol{\theta})$.

Summary statistics selection is, in our opinion, an open problem, as many approaches proposed in the literature require the user to specify an initial set of summary statistics which can then be refined upon (see for example Blum et al. (2013), Fearnhead and Prangle (2012) or Nunes and Balding (2010)). While some fairly general approaches exist (Drovandi et al., 2014), finding a set of initial statistics under which the model is identifiable is, at the time of writing, a time consuming, problem dependent and largely non-automated process. In the context of models with several hidden states, devising summary statistics is particularly difficult, because these have to capture the relation between all the states, while being based only on (noisy proxies of) a subset of them. The two-path cholera model is a perfect example of this problem: out of seven state variables only one, the number of infected, is observed with noise.

Taken together our results lead us to some very practical conclusions. When faced with a real non-linear dynamic system for which good models are available, one should ideally use a state space method for final parameter estimation, combined with a minimum tuning information reduction approach for exploration of alternative model structures, initialization and checking of conclusions. Using state space methods alone may bias conclusions towards noise driven stable dynamics, while using information reduction alone may lead to inference that is less precise than it could be. If the model is only attempting to explain some features of the system, and not every detail of the data then information reduction is probably essential.

CHAPTER 4

FAST APPROXIMATE INFERENCE FOR INTRACTABLE MODELS

In chapters 2 and 3 we used the Metropolis-Hastings implementation of Synthetic Likelihood, which we refered to as Synthetic Likelihood Metropolis-Hastings (SLMH). An important issue with SLMH is its computational cost. Indeed, a new estimate, $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, has to generated by simulation at each MH step. This can be very expensive for complex models. In addition, the mixing of the sampler is typically quite poor, due to the noise in the synthetic likelihood estimates.

Here we propose an alternative algorithm, which provides parameter estimates by maximizing the synthetic likelihood, rather than by sampling the parameters posterior distribution.

4.1 Introduction

In this chapter we propose more computationally efficient approaches to statistical inference with Synthetic Likelihood. Gutmann and Corander (2015), Wilkinson (2014) and Meeds and Welling (2014), propose to reduce the computational costs of simulation-based approximate methods, which include Synthetic Likelihood, using Gaussian Processes (GPs). In particular, Gutmann and Corander (2015) and Wilkinson (2014) use GPs to smooth pointwise (synthetic) likelihood estimates, to obtain an approximation the likelihood function. The aim is limiting the number of likelihood estimates, which are generally expensive to compute. Meeds and Welling (2014) use GPs to approximate how the first two moments of the summary statistics vary with model parameters, thus obtaining a smooth proxy to the synthetic likelihood. In this chapter we do not use GPs, but we adopt a gradient-based approach. In particular, we aim at maximizing the synthetic likelihood function efficiently, using local estimates of its gradient and Hessian to set-up a stochastic Newton-Raphson optimization algorithm. The approach we use to estimate gradient and Hessian of the synthetic likelihood, and the way we use them to maximize this function are, to our best knowledge, novel. Our solution is specific to Synthetic Likelihood and it explictly exploits the parametric Gaussian assumption on the distribution of the statistics. Thus our proposal is different from the Bayesian Optimization in Likelihood-Free inference (BOLFI) of Gutmann and Corander (2015),

which could be used in conjunction with either parametric or non-parametric likelihood estimators. The approaches proposed by Wilkinson (2014) and Meeds and Welling (2014) aim at approximating the main mass of the posterior density using Gaussian Processes, hence they are different from the current proposal and from BOLFI, which aim at approximating the synthetic likelihood function in the vicinity of its mode.

By approximately maximizing the synthetic likelihood function, the optimization approach proposed here provides point estimates of the unknown parameters. As part of the optimization, the approach also produces estimates of the Hessian matrix in the vicinity of the (synthetic) MLE. These estimates could clearly be used to derive asymptotic information about parameter uncertainty. As pointed out in Section 1.3, the algorithm proposed in this chapter has reached its current form only recently, hence we leave the (computationally intensive) task of verifying the quality of its parameter uncertainty estimates to future work.

The rest of the chapter is structured as follows. In Section 4.2 we explain how the synthetic likelihood can maximized efficiently, by modelling the mean and covariance matrix of the summary statistics using local regressions. Section 4.3 describes a version of the Generalized Method of Moments that is very similar to SL, and explains how the optimizer described in Section 4.2 can be used in the context of this method. We illustrate the performance of proposed algorithms on the three simple examples presented in Section 4.4. Finally, Section 4.5 briefly outlines some possible extensions of the proposed approach, while Section 4.6 discusses the results obtained so far.

4.2 Maximizing the Synthetic likelihood

To simplify the notation, in this chapter we will often indicate $\log p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ with $l(\boldsymbol{\theta})$. Similarly, we will often use the notation $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in place of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.

As mentioned previously, the SLMH algorithm is quite simple to implement, but is quite computationally inefficient, because at each MH step the synthetic likelihood has to estimated by simulation. This can be expensive when the model is complex and/or when the number of statistics used is large. Indeed, the number of entries in $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is $O(d^2)$, so if d is increased the number of simulations N has to be increased too, in order to keep the variance of the estimated likelihood constant. Doucet et al. (2012) provides results regarding how the variance of the unbiased likelihood estimates affects the performance of the resulting MH sampler. Their results are directly relevant to SL, provided that the bias-corrected synthetic likelihood estimator, presented in Section 2.4.3, is being used.

To limit the number of simulations, we propose a stochastic optimization algorithm which aims at maximizing the synthetic likelihood function efficiently.

4.2.1 Approximating gradient and Hessian through local regressions

One possible way of maximizing $l(\boldsymbol{\theta})$ could be to set up a Gradient Descent (GD) or Newton-Raphson (NR) algorithm, using finite differences to estimate gradient $\nabla l(\boldsymbol{\theta})$ and Hessian $\nabla^2 l(\boldsymbol{\theta})$ of the synthetic likelihood. However, such algorithms would need substantial modifications and tuning, relative to the deterministic case, in order to guarantee convergence when only noisy estimates of $l(\boldsymbol{\theta})$ are available. An example of this approach is the Simultaneous Perturbation Stochastic Approximation, described

by Spall (2000) and employed by Ehrlich et al. (2013) in the context of ABC inference for Hidden Markov Models (HMMs) with intractable likelihoods.

In this work we propose a different approach, which exploits the Gaussian assumption made by Synthetic Likelihood. In particular, notice that the gradient and Hessian of $\log l(\boldsymbol{\theta})$ are

$$\begin{aligned}\nabla l(\boldsymbol{\theta})_k &= \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_k} \\ &= \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{s}^0 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_k}\right),\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}\nabla^2 l(\boldsymbol{\theta})_{kl} &= \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} \\ &= \frac{\partial^2 \boldsymbol{\mu}^T}{\partial \theta_k \partial \theta_l}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) - \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) - \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_l} \\ &- \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_l}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) - (\mathbf{s}^0 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) \\ &+ (\mathbf{s}^0 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_k \partial \theta_l}\boldsymbol{\Sigma}^{-1}(\mathbf{s}^0 - \boldsymbol{\mu}) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_k}\right) \\ &- \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_k \partial \theta_l}\right),\end{aligned} \tag{4.2}$$

where $k, l = 1, \ldots, p$ and p is the dimension of $\boldsymbol{\theta}$.

In general, both $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are unknown functions. Algorithm 2, below, details how local linear regressions can be used to estimate the first and second order derivatives of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ wrt $\boldsymbol{\theta}$, at a position $\boldsymbol{\theta}^0$ in the parameter space. These estimates can then be plugged into (4.1) and (4.2), to provide estimates of $\nabla l(\boldsymbol{\theta})|_{\boldsymbol{\theta}^0}$ and $\nabla^2 l(\boldsymbol{\theta})|_{\boldsymbol{\theta}^0}$, respectively.

Algorithm 2 deserves some comments. The accuracy of local models (4.3) and (4.4) depends on choice of bandwidth matrix $\boldsymbol{P}$. We briefly discuss this issue in Section 4.2.2. Even though $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is not diagonal in general, the linear models in step 3 can be fitted independently by Ordinary Least Squares without loss of efficiency, because all the regressions include the same explanatory variables (Davidson and MacKinnon, 1993).

In step 4 we model only the marginal variances of the summary statistics, while assuming that the correlation structure is constant. Originally, we intended to model each elements $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ independently, but this lead to estimates that often violated positive-definiteness constraints. An alternative, but much more elaborate, approach would be to model the whole covariance matrix using, for instance, the unconstrained parametrization of Pourahmadi (1999). While that approach might lead to more accurate estimates, we fear that it would slow down the optimization drastically.

According to Ruppert et al. (1997), assuming that the variances vary linearly at a local level is reasonable in most situation. Empirically, we have found that making this assumption improved the stability of the estimates, relatively to higher order polynomials.

Algorithm 2 Estimating $\nabla l(\boldsymbol{\theta})|_{\boldsymbol{\theta}^0}$ and $\nabla^2 l(\boldsymbol{\theta})|_{\boldsymbol{\theta}^0}$

1: Simulate N of parameter vectors from Gaussian density

$$\boldsymbol{\theta}_i' \sim N(\boldsymbol{\theta}^0, \boldsymbol{P}), \quad \text{for } \ i = 1, \ldots, N,$$

where $\boldsymbol{P}$ is an user-defined covariance matrix, determining the domain of the local model.

2: For each of the N parameter vectors, $\boldsymbol{\theta}_i'$, simulate a dataset, $\boldsymbol{y}_i$, from the model and transform it to vector of statistics, $\mathbf{s}_i = S(\boldsymbol{y}_i)$.

3: Define $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i' - \boldsymbol{\theta}^0$ and fit one quadratic regressions for each of the d statistics

$$s_{ij} = \mu_j + \sum_k \alpha_{jk}\theta_{ik} + \frac{1}{2}\sum_k \beta_{jk}\theta_{ik}^2 + \sum_{k<l} \gamma_{jkl}\theta_{ik}\theta_{il} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N\{0, \Sigma(\boldsymbol{\theta}_i)\}, \tag{4.3}$$

where s_{ij} is the j-th element of vector $\mathbf{s}_i$. These d regressions aim at approximating $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, locally. The estimated regression coefficients provide us with

$$E(s_j|\boldsymbol{\theta}^0) = \mu_j\big|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\mu}_j,$$

$$\left.\frac{\partial \mu_j}{\partial \theta_k}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\alpha}_{jk}, \quad \left.\frac{\partial^2 \mu_j}{\partial \theta_k^2}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\beta}_{jk} \ \text{ and } \ \left.\frac{\partial^2 \mu_j}{\partial \theta_k \partial \theta_l}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\gamma}_{jkl} \ (k < l).$$

4: The residuals of the previous regressions can be used to estimate the local behaviour of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Define $\mathbf{D}_i = \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T$, where $\boldsymbol{\epsilon}_i^T = (\epsilon_{i1}, \ldots, \epsilon_{id})$, and regress the diagonal elements of $\mathbf{D_i}$ non-linearly on the parameters

$$D_{i,jj} = \exp\left(\phi_j + \sum_k \nu_{jk}\theta_{ik}\right) z_{ij}, \quad z_{ij} \sim \chi^2(1), \quad \text{for } \ j = 1, \ldots, d. \tag{4.4}$$

This provides the estimates

$$E(D_{jj}|\boldsymbol{\theta}^0) = \Sigma_{jj}\Big|_{\boldsymbol{\theta}=\mathbf{0}} \approx \exp\left(\hat{\phi}_j\right) \ \text{ and } \ \left.\frac{\partial \Sigma_{jj}}{\partial \theta_k}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\nu}_{jk}\exp\left(\hat{\phi}_j\right).$$

Now, let $\boldsymbol{\Psi} = \Psi(\boldsymbol{\theta})$ be the correlation matrix of the statistics, and assume that

$$\frac{\partial \Psi_{jl}}{\partial \theta_k} \approx 0, \quad \text{for every} \quad j, l \in 1, \ldots, d, \ \text{ and } \ k \in 1, \ldots, p, \tag{4.5}$$

which, together with (4.4), implies that

$$\left.\frac{\partial \Sigma_{jl}}{\partial \theta_k}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \Psi_{jl}\left.\frac{\partial}{\partial \theta_k}\left(\Sigma_{jj}\Sigma_{ll}\right)^{\frac{1}{2}}\right|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\Psi}_{jl}\exp\left\{\frac{1}{2}(\hat{\phi}_j + \hat{\phi}_l)\right\}\left(\frac{\hat{\nu}_{jk} + \hat{\nu}_{lk}}{2}\right). \tag{4.6}$$

In addition

$$\boldsymbol{\Sigma}\Big|_{\boldsymbol{\theta}=\mathbf{0}} \approx \hat{\boldsymbol{\Phi}}\,\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Phi}},$$

where

$$\hat{\boldsymbol{\Psi}} = \operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\operatorname{diag}(\hat{\boldsymbol{\Sigma}})^{-\frac{1}{2}}, \quad \hat{\boldsymbol{\Sigma}} = \frac{\sum_i \mathbf{D}_i}{N-1},$$

and $\hat{\boldsymbol{\Phi}}$ is a $d \times d$ diagonal matrix such that $\hat{\Phi}_{jj} = \exp\left(\hat{\phi}_j/2\right)$, for $j = 1, \ldots, d$.

Model (4.4) can be fitted by maximizing, wrt $\{\phi_j, \boldsymbol{\nu}_j\}$, the log-likelihood

$$\sum_{i=1}^{N} \log p(D_{i,jj}|\phi_j, \boldsymbol{\nu}_j, \boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\log p(z_{ij}|\phi_j, \boldsymbol{\nu}_j, \boldsymbol{\theta}) - \left\{ \phi_j + \sum_{k=1}^{p} \nu_{jk}\theta_{ik} \right\} \right],$$

where $p(z_{ij}|\phi_j, \boldsymbol{\nu}_j, \boldsymbol{\theta})$ is the density of a $\chi^2(1)$ r.v., while the second term on the r.h.s. is the log-Jacobian of the transformation. Notice that $p(D_{i,jj}|\boldsymbol{\phi}_j, \boldsymbol{\nu}_j, \boldsymbol{\theta})$ can be re-expressed as the density of a Gamma-distributed r.v. with shape $1/2$ and scale $2\exp(\phi_j + \sum_k \nu_{jk}\theta_{ik})$. Hence, in practice, it is convenient to fit (4.4) as a Gamma-distributed Generalized Linear Model (GLM) with log-link function. Finally, if the estimates obtained in step 4 suggest diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ might be strongly dependent on the parameters, model (4.3) can be re-fitted using Weighted Least Squares.

4.2.2 A Stochastic Newton-Raphson algorithm

Assuming that the estimates provided by Algorithm 2 are reliable, they can be used to maximize the synthetic likelihood. In particular, we propose a stochastic Newton-Raphson scheme, whose main recursion is similar to that of the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm of Spall (2000). For $k = 1, 2, \ldots$ we use the following iteration

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - a_k \boldsymbol{A}_k^{-1} \hat{\boldsymbol{G}}_k,$$

$$\bar{\boldsymbol{H}}_k = \frac{b_k}{b_k + 1} \bar{\boldsymbol{H}}_{k-1} + \frac{1}{b_k + 1} \hat{\boldsymbol{H}}_k,$$

where $\hat{\boldsymbol{G}}_k$ and $\hat{\boldsymbol{H}}_k$ are estimates of $\nabla l(\boldsymbol{\theta})|_{\boldsymbol{\theta}_k}$ and $-\nabla^2 l(\boldsymbol{\theta})|_{\boldsymbol{\theta}_k}$, obtained by plugging the estimates of the derivatives of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, given by Algorithm 2 with $\boldsymbol{\theta}^0 = \boldsymbol{\theta}_k$, in (4.1) and (4.2). Also, a_k is a positive decreasing gain sequence

$$a_k = \frac{a_0}{k^\eta}, \quad \eta \in [0.602, 1],$$

where the lower and upper bounds for η are respectively its slowest possible (in order to guarantee convergence) and asymptotically optimal values (Spall, 2005), while b_k is a positive integer. In order to assure the positive definiteness, symmetry and stability of $\boldsymbol{A}_k$ we calculate it as follow

$$A_k = (\bar{\boldsymbol{H}}_k \bar{\boldsymbol{H}}_k)^{1/2} + \delta \boldsymbol{I},$$

where the square root indicates the positive semidefinite matrix square root, δ is a small constant and $\boldsymbol{I}$ is the identity matrix. Notice that $-\boldsymbol{U}\hat{\boldsymbol{G}}_k$ is assured to be a descent direction of $-l(\boldsymbol{\theta})$ for any positive definite matrix $\boldsymbol{U}$, hence, for optimization purposes, it is not necessary for $\boldsymbol{A}_k$ to be an extremely accurate estimate of $-\nabla^2 l(\boldsymbol{\theta})|_{\boldsymbol{\theta}_k}$.

At each iteration of this Newton-Raphson scheme, we have to simulate N parameters from $N(\boldsymbol{\theta}_k, \boldsymbol{P}_k)$, which is the first step of Algorithm 2. The issues involved in the choice of $\boldsymbol{P}_k$ are related to those described by Fan et al. (1998) in a smoothing context. The authors consider the use of local regression for scatter-plot smoothing, and what they name the "bandwidth" plays a role similar to that of $\boldsymbol{P}_k$ in our framework. In both contexts, as we increase the bandwidth or the diagonal entries of $\boldsymbol{P}_k$, the local regressions become more biased, because higher order terms become more

important. On the other hand, if we narrow too much the neighbourhood of $\boldsymbol{\theta}_k$ within which we simulate, the variance of the estimated gradient and Hessian will increase, and the noise will overwhelm the information available as such a local level. We have found that a good trade-off can been achieved by adopting the intuitively appealing choice $\boldsymbol{P}_k = \boldsymbol{A}_k^{-1}$. In particular, this leads to $\boldsymbol{P}_k \approx -\{\nabla^2 l(\boldsymbol{\theta})|_{\boldsymbol{\theta}_k}\}^{-1}$, which implies that $\boldsymbol{P}_k$ will be an approximation to the Fisher information matrix, if $\boldsymbol{\theta}_k$ is close to the maximizer of $l(\boldsymbol{\theta})$.

In the following we will refer to the Newton-Raphson procedure introduced in this section as the Maximum Synthetic Likelihood (MSL) algorithm.

4.3 Continuous Updating Generalized Method of Moments

We now discuss the relation between SL and a particular version of the Method of Moments. In particular, let us consider the objective function

$$f(\boldsymbol{\theta}) = (\mathbf{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}), \tag{4.7}$$

which, apart from the missing log-determinant of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, is proportional to the synthetic log-likelihood, $\log p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. The following estimator

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}), \tag{4.8}$$

is analogous to what Hansen et al. (1996) calls the Continuous Updating Generalized Method of Moments (CUGMM). In Section 4.3.1 we discuss the asymptotic properties of this estimator.

4.3.1 Asymptotic properties of CUGMM

The asymptotic properties of a more general version of estimator (4.8) have been studied by Pakes and Pollard (1989), who proved the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$. Under the assumption that $\boldsymbol{S}^0$ is asymptotically normal, their results guarantee that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}\big\{\mathbf{0}, (\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{\Gamma}_{\boldsymbol{\theta}_0})^{-1}\big\}, \tag{4.9}$$

where

$$(\Gamma_{\boldsymbol{\theta}_0})_{ij} = \left.\frac{\partial \mu_i}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},$$

and n is the number of underlying observations, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, used to calculate $\boldsymbol{S}_0$. We are particularly interested in this result, because asymptotic normality does not generally hold for the MSL estimator, as shown by Wood (2010). Hence, in the Appendix C, we assume the consistency of $\hat{\boldsymbol{\theta}}$ and we use it to provide an informal derivation of (4.9).

The striking aspect of (4.9) is that the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_n$ does not depend on $\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}} / \partial \theta_k$. However, this quantity appears in the gradient of (4.7)

$$\nabla f(\boldsymbol{\theta})_k = \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}^T}{\partial \theta_k} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \frac{1}{2} (\mathbf{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_k} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}), \tag{4.10}$$

hence it might still be needed in order to compute (4.8), using a gradient-based optimizer. As we will show in the following section, this might not be not necessary as long as the initialization of the optimizer is good enough, in a sense to be clarified shortly.

4.3.2 Practical optimization

Consider the approximate gradient

$$\tilde{\nabla} f(\boldsymbol{\theta})_k = \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}^T}{\partial \theta_k} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{s}^0 - \boldsymbol{\mu}_\theta),$$

which is obtained by imposing $\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}} / \partial \theta_k = \mathbf{0}$ in (4.10). In addition, make the following assumption

$$\exists\, \hat{\boldsymbol{\theta}} \;:\; \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{s}^0, \tag{4.11}$$

which implies that the model is able match the expected value of the statistics with the observed statistics vector, for some parameter value $\hat{\boldsymbol{\theta}}$. Notice that $\hat{\boldsymbol{\theta}}$ is a stationary point of (4.7), the CUGMM objective. Then

$$\exists\; r > 0 \;\;:\;\; \nabla f(\boldsymbol{\theta})^T \tilde{\nabla} f(\boldsymbol{\theta}) > 0 \;\; \forall\, \boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|| < r, \tag{4.12}$$

that is, $\tilde{\nabla} \hat{f}(\boldsymbol{\theta})$ will provide a descent direction on (4.7), provided that $\boldsymbol{\theta}$ is sufficiently close to the local minimum $\hat{\boldsymbol{\theta}}$. The proof of (4.12) is very simple, in fact

$$\nabla f(\boldsymbol{\theta})^T \tilde{\nabla} f(\boldsymbol{\theta}) = \left\{ \tilde{\nabla} f(\boldsymbol{\theta}) + O(||\mathbf{s}^0 - \boldsymbol{\mu}_\theta||^2) \right\}^T \tilde{\nabla} f(\boldsymbol{\theta}) = ||\tilde{\nabla} f(\boldsymbol{\theta})||^2 + O(||\mathbf{s}^0 - \boldsymbol{\mu}_\theta||^3),$$

which tends to zero from above as $\boldsymbol{\theta} \to \hat{\boldsymbol{\theta}}$, because of (4.11) and of the fact that $\nabla f(\boldsymbol{\theta})$ is $O(||\mathbf{s}^0 - \boldsymbol{\mu}_\theta||)$.

Assumption (4.11) is quite strong, and it will probably hold only approximately in an applied setting. However, given a parameter estimate $\hat{\boldsymbol{\theta}}$, is it straighforward to check, by simulation, how well this assumption holds.

To summarize, the derivatives $\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}} / \partial \theta_k$ do not appear in the distributional result (4.9) and are not essential for minimizing (4.7), at least under assumption (4.11) and in the vicinity of the minimizer $\hat{\boldsymbol{\theta}}$. These facts have important practical implications, if we are interested in using Algorithm 2 to estimate gradient and Hessian of the CUGMM objective (4.7). Indeed, in the following we impose $\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}} / \partial \theta_k = \mathbf{0}$, when using Algorithm 2 in conjunction with CUGMM, which allows us to skip step 4 of that algorithm.

4.4 Examples

In this section we test MSL, CUGMM and SLMH on three simple examples.

4.4.1 Exponential distribution

As a first example we consider a d-dimensional vector $\boldsymbol{X}$, such that

$$X_i \sim \text{Exp}(\lambda_i), \quad \text{for } \; i = 1, \ldots, d.$$

Hence, each entry of $\boldsymbol{X}$ is marginally distributed according to an exponential distribution with rate λ_i. In order to estimate $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_d\}$, we used the sample mean as a summary statistic

$$\boldsymbol{s} = \frac{1}{m} \sum_{k=1}^{d} \boldsymbol{X}_k,$$

where m is the number of observed or simulated random vectors.

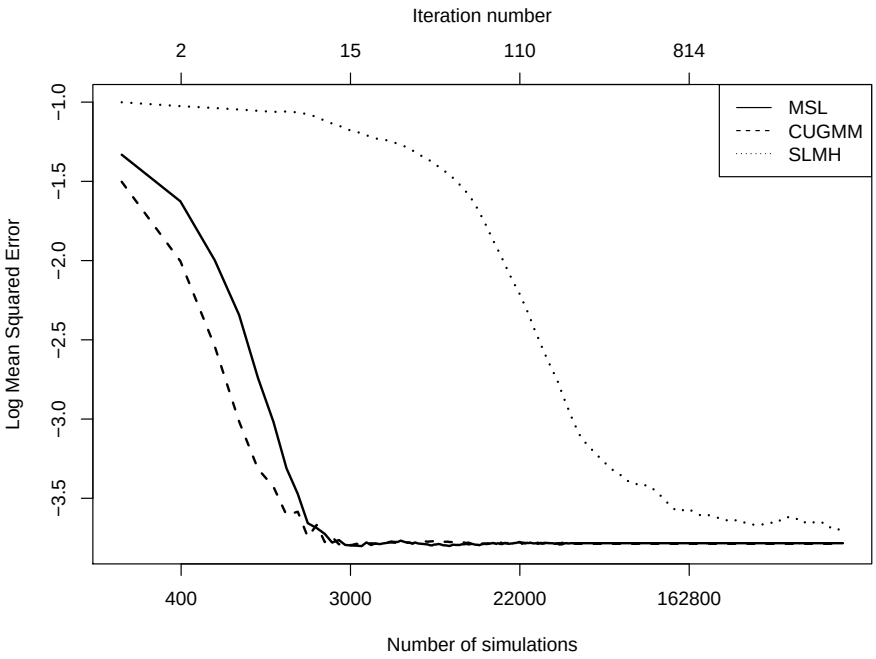

Figure 4-1: *Log-MSE for $\boldsymbol{\lambda}$ as a function of the number of iterations and of simulated statistics vector for MSL, CUGMM and SLMH. Notice that for MSL and CUGMM we used only 200 iterations, hence their log-MSE is depicted as constant after that.*

We estimated $\boldsymbol{\lambda}$ using MSL, CUGMM and SLMH. In particular, we chose $d = 10$, $\lambda_i = 2$, for $i = 1, \ldots, d$, and $m = 200$. Under MSL and CUGMM we used $N = 200$ simulated statistics to estimate gradient and Hessian at each step, and we used 200 optimization steps. Under SLMH we used $N = 200$ simulations to estimate the synthetic likelihood at each step, 5000 MCMC iterations and flat priors. We repeated the estimation 40 times, with the initial values of each run being randomly simulated from a uniform distribution on $[1, 3]$.

Figure 4-1, shows how the log Mean Squared Error (log-MSE), averaged over each of the 10 dimension and over the 40 runs, changes with the number of simulations, for each method. For MSL and CUGMM we used the latest position of the optimizer as point estimate of $\boldsymbol{\lambda}$, while for SLMH we used the sample mean of the most recent half of the chain. Notice that the log-MSE is reduced much faster using MSL and CUGMM, rather than SLMH. In particular, the first two methods have effectively converged after around 15 iterations, while the log-MSE of SLMH is still higher at 5000 iterations. Notice that CUGMM seems to converge slightly faster than MSL in early iterations, possibly because the latter is affected by higher variability brought about by the estimation of the derivatives of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ wrt $\boldsymbol{\theta}$.

4.4.2 Stable distribution

As a second example, we consider the 4-parameter, $\boldsymbol{\theta} = \{\alpha, \beta, \gamma, \delta\}$, α-stable family of distributions. Here, parameter α is the index of stability or characteristic exponent, β determines skewness, while γ and δ control respectively scale and location. As detailed by Nolan (2001), performing Maximum Likelihood inference on this model is difficult,

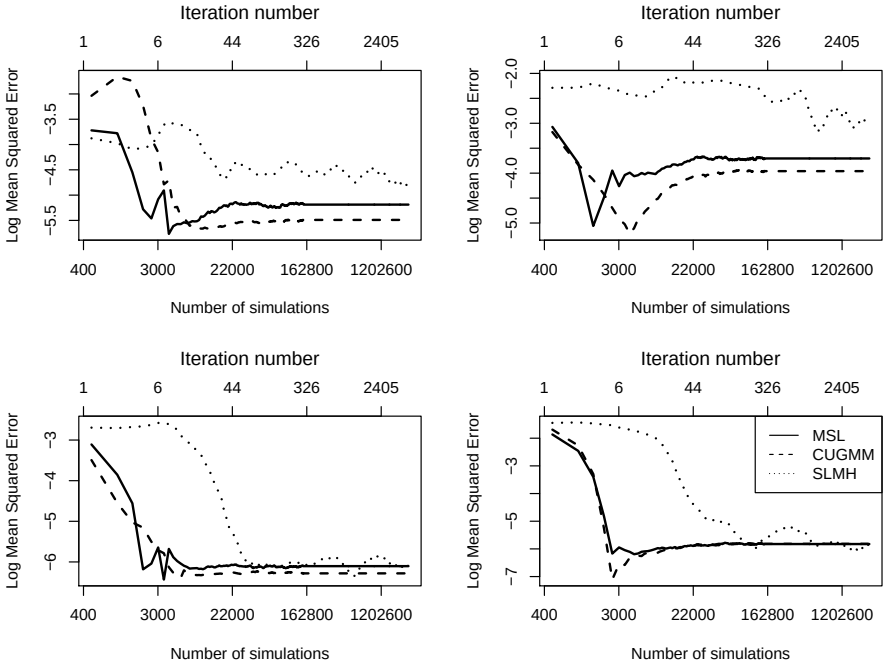

Figure 4-2: *Clock-wise from top-left: log-MSEs for α, β, γ and δ, as functions of the number of iterations and of simulated statistics vector for MSL, CUGMM and SLMH. Notice that for MSL and CUGMM we used only 300 iterations, hence their log-MSEs are depicted as constant after that.*

hence Rubio and Johansen (2013) considered the use of ABC methods to approximate the MLE.

The data consisted of 1000 random vectors, simulated from the model using $\alpha = 1.5$, $\beta = 0.1$, $\gamma = 1$ and $\delta = 2$. As summary statistic we used 15 empirical quantiles, corresponding to cumulative probabilities equally spaced between 0.1 and 0.9. All methods where initialized at $\alpha = 1.7$, $\beta = -0.2$, $\gamma = 1.3$ and $\delta = 1.5$. For SLMH we used flat priors for each of the parameters, $N = 500$ simulated summary statistics to estimate the synthetic likelihood at each step and 5000 MCMC iterations. For MSL and CUGMM we used $N = 500$ simulated statistics at each iteration and 300 optimization steps. We repeated the whole process 12 times.

The log-MSEs, for each method and parameter, are plotted in Figure 4-2 as functions of the number of iterations and of simulated statistics vectors. The point estimates for each method were obtained as in Section 4.4.1. MSL and CUGMM seem to have converge after around 50 iterations, and both algorithms decrease the log-MSE faster than SLMH, particularly for parameter α and γ. Notice that, in early iterations, the behaviour of MSL seems quite irregular and noisy. In addition, MSL does lightly worse than CUGMM, in terms of log-MSE, for parameters α and γ. In a run not shown here, we have verified that this problem can be solved by increasing the number of simulations used at each step of MSL to $N = 1000$.

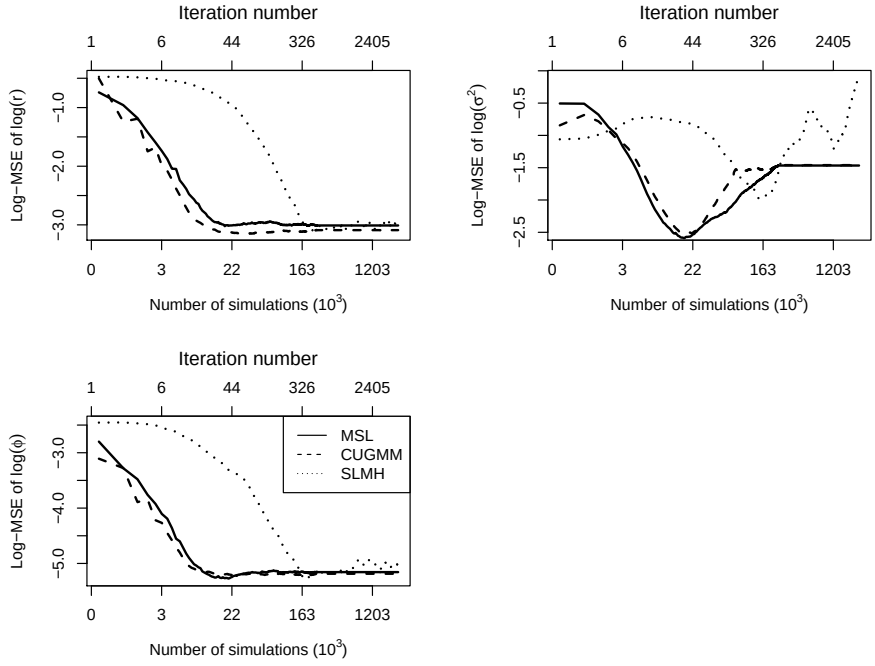

Figure 4-3: *Clock from top-left: log-MSEs for* $\log r$, $\log \sigma^2$, $\log \phi$ *and* δ *as functions of the number of iterations and of simulated statistics vector for CUGMM and SLMH. Notice that for CUGMM we used only 500 iterations, hence its log-MSE is depicted as constant after that.*

4.4.3 Ricker map

As a further example, we considered the Ricker map. In particular, we simulated 24 path from the model, using parameter values $\log r = 3.8$, $\log \sigma^2 = -1.2$ and $\log \phi = 2.3$. To fit the model with MSL, CUGMM and SLMH, we used the 13 summary statistics of Wood (2010) and the initial values $\log r = 3$, $\sigma^2 = -0.6$ and $\phi = 2.6$. For each method, we used 500 simulations from the model to estimate the likelihood (SLMH) or its derivatives (MSL and CUGMM). For MSE and CUGMM we used 500 iterations of the optimizer, while for SLMH we used 5000 iterations. We used flat priors for each of the parameters.

Figure (4-3) shows the log-MSE for each parameter and method. MSL and CUGMM reduce the log-MSE of $\log r$ and $\log \phi$ much faster than SLMH, but all algorithm struggle to estimate $\log \sigma^2$, which is barely identifiable using the statistics of Wood (2010). The performance of MSL and CUGMM is essentially identical in this setting.

4.5 Possible extensions

Here we discuss some preliminary work regarding possible extensions of MSL and CUGMM.

4.5.1 Additional regression step

A possible disadvantage of MSL and, to a lesser degree, CUGMM is that it they do not scale well with the number of summary statistics, d. In particular, estimating

the gradient and Hessian of CUGMM's objective function requires running d linear regressions. This is fairly cheap computationally but, in the case of MSL, it is also necessary to fit d Generalized Linear Models to estimate how $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ varies with $\boldsymbol{\theta}$.

In order to improve the scalability of the algorithm w.r.t. the number of summary statistics, it might be possible to introduce an additional regression step. In particular, similarly to what has been suggested by Fearnhead and Prangle (2012), we propose to regress the parameters on the summary statistics, which have been simulated respectively during step 2 and 1 of Algorithm 2. In practice, we introduce an additional step (say $2a$) which entails fitting the following linear models

$$\boldsymbol{\theta}_i = \boldsymbol{\beta}\boldsymbol{s}_i + \boldsymbol{\epsilon}_i \tag{4.13}$$

where $\boldsymbol{\beta}$ is an $p \times d$ matrix of regression coefficients, the errors $\boldsymbol{\epsilon}_i$ are independently distributed with mean zero and i is the index of the ith simulated pair of parameters and statistics. This new step requires the computation of a linear regression for each parameters, but this is the only additional computation that it entails.

After having computed the regressions, we propose to use the fitted values $\hat{\boldsymbol{\theta}}$ as our summary statistics, to be used in the subsequent steps of Algorithm 2. The rest of the algorithm is left unchanged. Assuming that the additional regressions are reasonably accurate and stable, this new step should allows us to reduce sharply the number of summary statistics used, thus reducing the computational cost of the subsequent steps.

Obviously, we have to verify that the accuracy of the additional regression step, but the way this is implemented in MSL or CUGMM should make it more accurate that in the case of Fearnhead and Prangle (2012). This is because we are going to re-compute $\hat{\boldsymbol{\beta}}$ at each step of the Newton-Raphson scheme, using the parameters and statistics generated during steps 1 and 2 of Algorithm 2. A linear relation between parameters and statistics should be more accurate on this local scale, than in the wider training set used by Fearnhead and Prangle (2012).

As an example we used MSL, with this additional regression step, to fit the α-stable distribution using 100 quantiles, corresponding to cumulative probabilities equally spaced in $[0.1, 0.9]$, as summary statistics. We used 500 simulated statistics to estimate gradient and Hessian at each step. Figure 4-4 shows the convergence plots of 8 separate runs. All the parameters seem to be well identified. We tried to use MSL, without the additional regression, in this setting, but the resulting estimates (not shown) were too noisy. The algorithm was also much more expensive to run: 100 GLMs had to be fitted at each step, in order to estimate the derivatives of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ wrt $\boldsymbol{\theta}$. When the additional regression was used, only four GLMs, one for each parameter, need to be fitted.

An additional benefit of regression (4.13) is that the resulting summary statistics, $\hat{\boldsymbol{\theta}}$, are generally close to normally distributed. To demonstrate this, we transformed the 100 summary statistics used to fit the α-stable model, to disrupt their normality. In particular, we used the cube of each quantile. Figure 4-5 shows the distribution of one of the transformed quantiles, which is very far from normal. Then, we simulated 10^4 parameters and summary statistics vectors, and we used them to fit regression (4.13). Figure 4-6 shows the distribution of the resulting predicted values $\hat{\boldsymbol{\theta}} = \{\hat{\log}\,\alpha, \hat{\beta}, \hat{\log}\,\gamma, \hat{\delta}\}$. Even though many of the original statistics were highly non-normal, the statistics provided by using regression (4.13) are very close to normally distributed. This result is probably a consequence of the fact the the new statistics $\hat{\boldsymbol{\theta}}$ are linear combinations of the original ones. Even though the latter can be highly correlated, some form of Central Limit Theorem might be in action here.

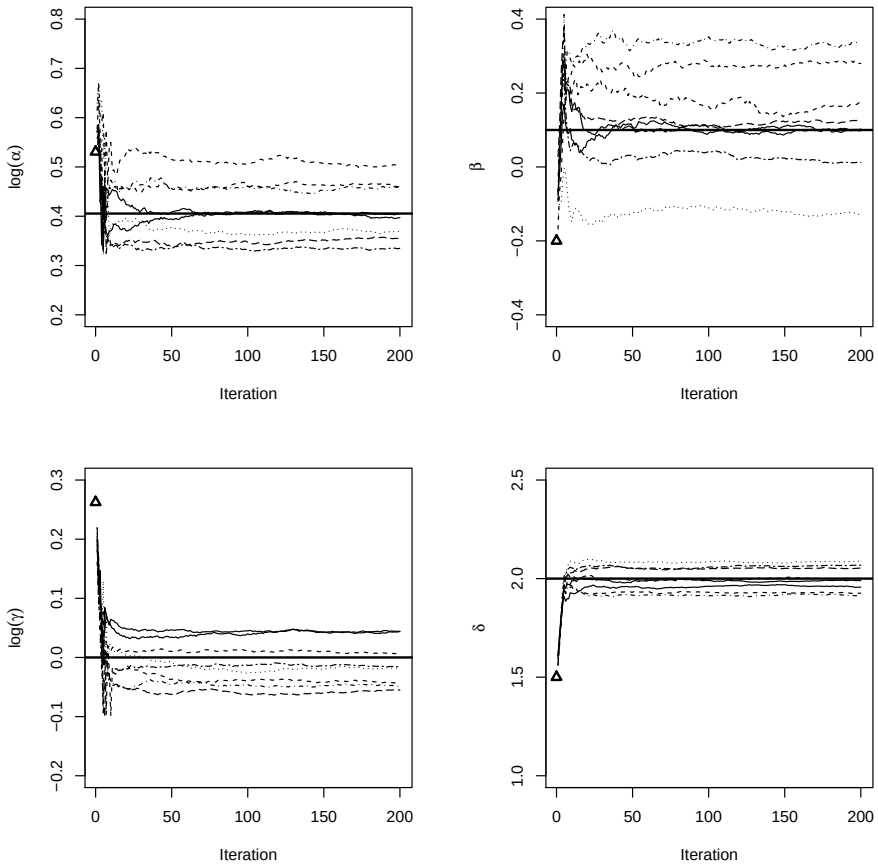

Figure 4-4: *Clock from top-left: convergence plots for* $\log \alpha$, β, $\log \gamma$ *and* α *using MSL with the four summary statistics obtained using regression (4.13). The triangles indicate the initialization used for each of the parameters.*

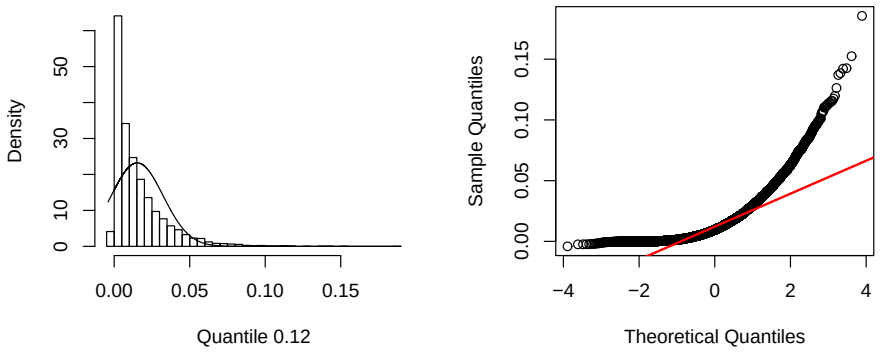

Figure 4-5: *Diagnostic plots to compare the normal approximation with the empirical distributions of the cube of the quantile corresponding to cumulative probability 0.12.*

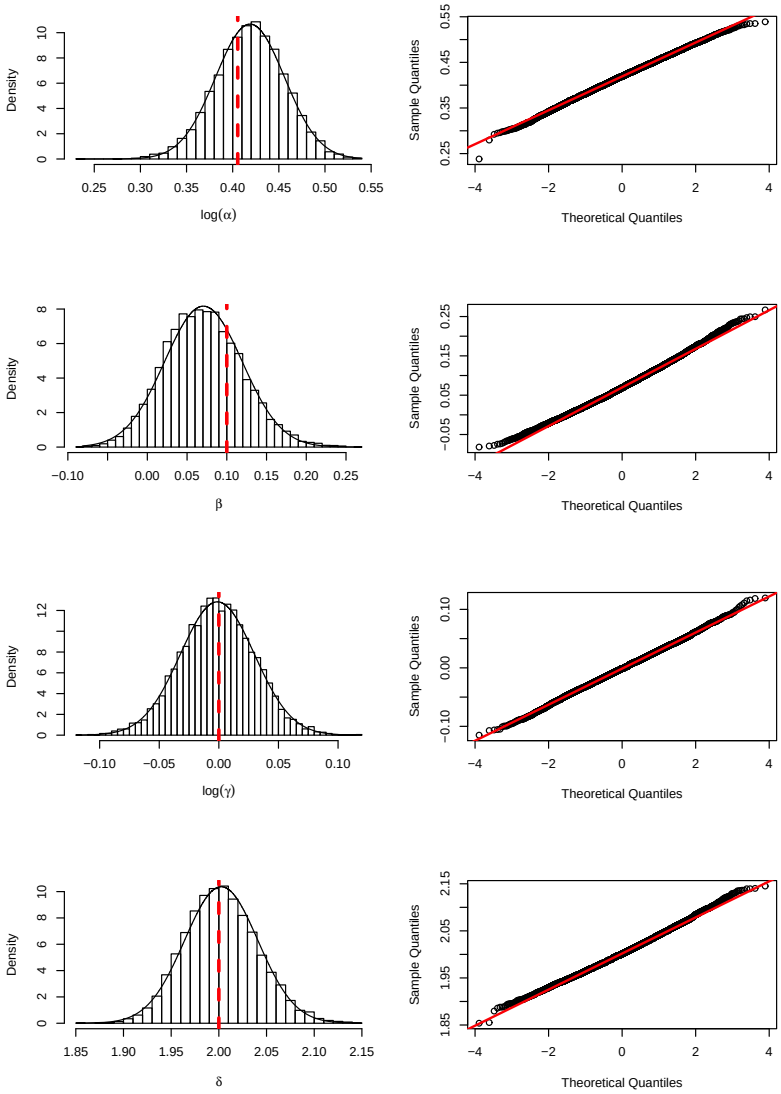

Figure 4-6: *Diagnostic plots to compare the normal approximation with the empirical distributions of* $\log \alpha$, $\hat{\beta}$, $\log \gamma$ *and* $\hat{\delta}$, *using the summary statistics proposed by Fearnhead and Prangle (2012). The vertical dashed lines indicate the current position in the parameter space, around which we are simulating parameters and statistics.*

4.5.2 A smoothing approach to Synthetic Likelihood

The MSL and CUGMM algorithms provide a sequence of estimates, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k$, where k is the number of iteration of the optimization routine. These are obtained by estimating the gradient and Hessian of the objective function using Algorithm 2, which is based on a local quadratic model for $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and a local linear model for the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. The accuracy of these regressions depends critically on $\boldsymbol{P}_k$, the covariance matrix determining the size of the local regression. Indeed, given that $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ can potentially be highly non-linear, $\boldsymbol{P}_k$ determines the bias-variance trade-off involved in fitting model (4.3) and (4.4).

For all the examples reported in Section 4.4, we chose $\boldsymbol{P}_k = \boldsymbol{A}_k^{-1}$, as described in Section 4.2.2. This choice is simple and intuitive, but it might be sub-optimal. Indeed, it would be preferable to select $\boldsymbol{P}_k$ by optimizing some appropriate criterion. However, it would be impossibly expensive to select $\boldsymbol{P}_k$ at each step by, for instance, cross-validation. For this reason, we propose a scheme that aims at modelling $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and, potentially, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, more accurately, by post-processing the output of MSL or CUGMM.

Assume that all the parameter vectors, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$, and summary statistics vectors, $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_M$, simulated while running MSL, have been stored. We are interested in approximating $\boldsymbol{\mu}_{\boldsymbol{\theta}} = E(\boldsymbol{S}|\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \text{cov}(\boldsymbol{S}|\boldsymbol{\theta})$, using all the M samples available so far. While one possibility is to use local linear regressions, as in Algorithm 2, here we consider another approach. The reasons for this choice will be detailed later.

Regardless of the method used to obtain smooth estimates of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, the result of the fitting will be the approximate functions $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Given that SL is based on the assumption that the statistics are approximately normally distributed, the value of the synthetic likelihood at any location of the parameter space is entirely determined by the relation between the first two moments of the summary statistics and the parameters. This entails that $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ are sufficient to define an approximation to the synthetic likelihood function, which we indicate with $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = p_{SL}(\boldsymbol{s}^0|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$. Notice that this is a global approximation, because it uses all the parameter values and summary statistics simulated during the MSL run. Assuming that MSL successfully converged to the vicinity of the mode of the synthetic likelihood, most of the parameter values simulated by MSL will be located around this mode. Hence, we can expect $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ to be a more accurate approximation to $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ close its mode, than in its tails.

Before explaining how $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ can be exploited, we detail how we approximate $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. For the mean vector, we use additive models of the form

$$E(S_j|\boldsymbol{\theta}) = \mu_j + \sum_k \alpha_{jk}\theta_k + \sum_k \beta_{jk}\theta_k^2 + \sum_k f_{jk}(\theta_k) + \sum_{k<l} \gamma_{jkl}\theta_k\theta_l, \tag{4.14}$$

where $j = 1, \ldots, d$ and $k, l = 1, \ldots, p$, while $f_{jk}(\cdot)$ are are unknown smooth function, with unknown degree of smoothness. In particular, they are represented by penalized regression splines, as described by Wood (2006). The advantage of this approach is that model (4.14) can be fitted to very large datasets, using the parallelization techniques described in Wood et al. (2015). This is very important in the current context, because M is generally fairly large and model (4.14) has to be fitted to each of the d summary statistics. In addition, the smoothness of $f_{jk}(\cdot)$ is estimated using the computationally efficient techniques described in Wood et al. (2015). Most importantly, all of the above is implemented by the *mgcv* R package (Wood, 2001). Having modelled $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, it is still

necessary to choose how to model $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. For the sake of simplicity, we propose to use model (4.4) for the marginal variances, while keeping the correlation structure constant, as done in Algorithm 2.

After $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ have been estimated, the approximate synthetic likelihood function, $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, is fully defined. $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ can be considered to be a likelihood to all effects, with the key advantage that evaluating it at any location, $\boldsymbol{\theta}$, does not require any additional simulation. Hence, $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ can be exploited in a number of ways. For instance, it can be maximized wrt $\boldsymbol{\theta}$, using a standard deterministic optimizer, to obtain Maximum Synthetic Likelihood estimates, which might be more accurate than those obtained by MSL. In the following we illustrate how $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ can be used to obtain approximate synthetic likelihood profiles.

To provide a simple example, we consider the α-stable distribution of Section 4.4.2. We fit a dataset, simulated using the same parameter values as in Section 4.4.2, using MSL. In particular, we use 10^3 simulated statistics vector per step, 100 optimization steps and the same 15 summary statistics as in Section 4.4.2. All the 10^5 statistics and parameter vectors simulated by MSL were stored. Of these, only those simulated after the 70th MSL iteration were used to fit models (4.14) and (4.4). Early iterations were discarded to reduce the computational cost of fitting model (4.14).

The black lines in Figure 4-7 represent likelihood profiles for each of the parameters, obtained using the $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. Hence, computing these curves did not require any additional simulation. In constrast, computing these profiles, using MSL alone, is extremely expensive. For example, the grey crosses in Figure 4-7 have been obtained using MSL, without using $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. Each single estimate requires fixing the value of the parameter being profiled, and running MSL to estimate the remaining parameters. Using $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ is much cheaper computationally and leads to likelihood profiles that are remarkably accurate in this example.

4.6 Conclusions

In this chapter we showed how the relation between first two moments of the summary statistics and model parameters can be approximated, using local regressions. We explained how these local regressions can be exploited by two algorithms, MSL and CUGMM, with the aim of producing computationally cheap parameter estimates. Further, we argued that, under CUGMM, the covariance of the summary statistics can be considered to be constant, without compromizing the optimization procedure.

In Section 4.4 we demonstrated that MSL and CUGMM outperform SLMH, in terms of MSE reduction as a function of the number of simulations, using three simple examples. Even though CUGMM performed slightly better than MSL in the α-stable example, a more thorough comparison between these two methods is needed. In addition, in Section 4.4 we have evaluated MSL, CUGMM and SLMH in terms of the point estimates they produce. However, all these methods provide also estimates of parameter uncertainty. Hence, it would be useful to verify the accuracy of these estimates, for example in terms of interval coverage.

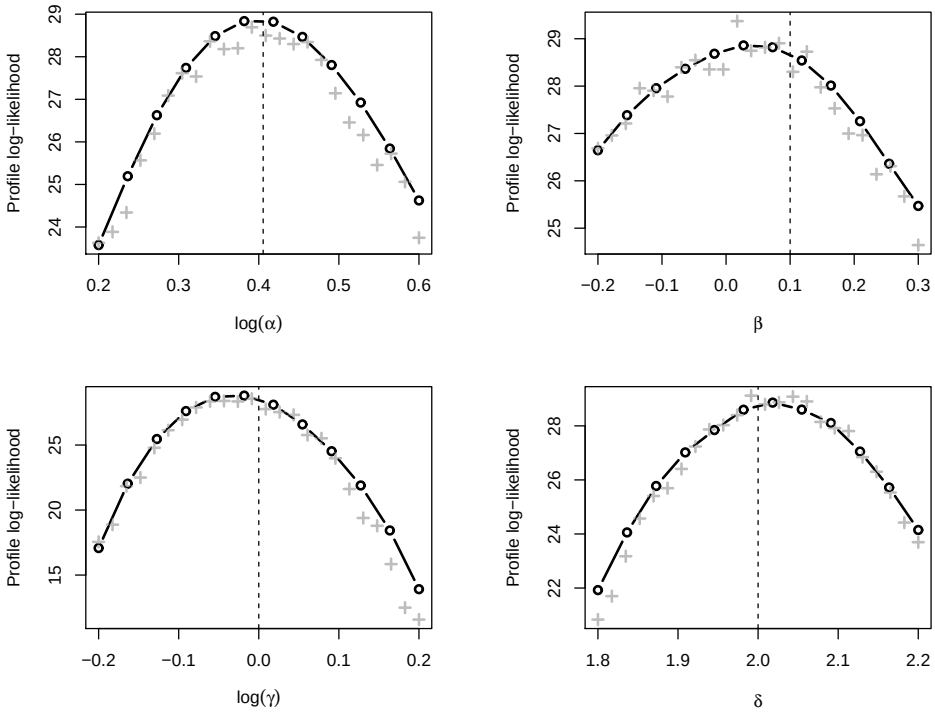

Figure 4-7: *Likelihood profiles for each parameter of the α-stable distribution, using $\tilde{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$. The grey crosses correspond to profiles obtained by re-running MSL for each value of the parameter being profiled. The vertical lines correspond to the true parameter values.*

CHAPTER 5

AN EMPIRICAL SADDLEPOINT APPROXIMATION FOR INTRACTABLE LIKELIHOODS

An important advantage of SL is that it requires less tuning than some alternative approaches, such as ABC methods. However, SL relies on the assumption that the summary statistics are approximately normally distributed. In this chapter we relax this assumption by proposing a novel flexible density estimator: the Extended Empirical Saddlepoint approximation. By illustrating its performance through two examples, we show that this estimator is able to capture large departure from normality, while being scalable to high dimensions.

5.1 Introduction

SL, as described by (Wood, 2010), uses a multivariate Gaussian density to approximate the distribution of the summary statistics. Under this distributional assumption, a pointwise estimate of the synthetic likelihood at $\boldsymbol{\theta}$ can be obtained using Algorithm 3. This procedure has already been described in previous chapters, but we report it also here for ease of reference. As discussed in Chapters 2 and 3, one advantage of

Algorithm 3 Estimating $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$

1: Simulate datasets $\boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_m$ from the model $p(\boldsymbol{y}|\boldsymbol{\theta})$.
2: Transform each dataset $\boldsymbol{Y}_i$ to a vector of summary statistics $\boldsymbol{S}_i = S(\boldsymbol{Y}_i)$.
3: Calculate sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the simulated statistics, possibly robustly.
4: Estimate the synthetic likelihood

$$\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}}|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) \right\},$$

where d is the number of summary statistics used.

SL, over most ABC methods, is that it does not require the user to choose a tolerance or an acceptance threshold and that the summary statistics are scaled automatically and dynamically by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. In addition, Blum (2010) showed that the convergence rate

of ABC methods degrades rapidly with d. This curse of dimensionality, brought about by the non-parametric nature of ABC, forces practitioners to use dimension reduction or statistics selection techniques, such as those described by Blum et al. (2013). SL is less sensitive to the number of statistics used, due to the parametric likelihood approximation.

We were led to propose saddlepoint approximations, among other multivariate density estimators, as a promising alternative to the Gaussian approximation, by the following considerations. In a multivariate setting the choice of density estimators is quite restricted. Indeed, non-parametric (kernel) estimators typically have convergence rates that are prohibitively low, even in moderate dimensions. For instance, multivariate kernel estimators typically converge at rate $O(n^{-\frac{2}{4+d}})$ (Scott, 2009). In contrast, Empirical Saddlepoint approximations converge at the parametric rate $O(n^{-\frac{1}{2}})$, regardless of d. In addition, while saddlepoint approximations are derived from asymptotic expansions, they are often very accurate even in small samples and, in contrast to Edgeworth approximations, they are strictly positive and do not show polynomial-like waves in the tails. Further, their empirical version provides a close approximation to the density of widely used statistics, such as M- (Ronchetti and Welsh, 1994) and L-estimators (Easton and Ronchetti, 1986).

The above properties of SL are not without cost. In fact, although the Central Limit Theorem assures asymptotic normality of many classes of statistics, improving the quality of the normal approximation is not easy in a multivariate setting. Finding a suitable normalizing transformation is particularly challenging in this context, because such transformation would need to ensure approximate normality across the parameter space. This motivates the main contribution of this work: we relax the multivariate normality assumption, while maintaining the ease-of-use and scalability of SL. We achieve this by proposing a flexible density estimator, namely an Extended Empirical Saddlepoint approximation.

5.2 Saddlepoint approximations

The following discussion is valid beyond the context of SL, hence we temporarily suppress the dependencies on $\boldsymbol{\theta}$. We restore them in Section 5.4, which describes how the proposed density estimator can be used within SL.

Saddlepoint expansions were introduced into the statistical literature by Daniels (1954) and can be used to approximate the density function of a random variable, starting from its moment or cumulant generating function. When $\boldsymbol{S}$ is a continuous random vector its probability density function, $p(\boldsymbol{s})$, is associated with the moment generating function

$$M(\boldsymbol{\lambda}) = E(e^{\boldsymbol{\lambda}^T \boldsymbol{S}}) = \int_{-\infty}^{+\infty} e^{\boldsymbol{\lambda}^T \boldsymbol{s}}\, p(\boldsymbol{s})\, d\boldsymbol{s},$$

while the cumulant generating function is defined as $K(\boldsymbol{\lambda}) = \log M(\boldsymbol{\lambda})$. In the following we assume that $M(\boldsymbol{\lambda})$ exists for $\boldsymbol{\lambda} \in I$, where I is a nonvanishing subset of R^d containing the origin. If $\boldsymbol{S}$ is a discrete random vector, the generating functions are obtained by substituting the integrals with summations over the support of $\boldsymbol{S}$.

Saddlepoint approximations rely on the one-to-one correspondence between the cumulant generating function and the probability density function of $\boldsymbol{S}$. For a continuous

$\boldsymbol{S}$, the saddlepoint density is

$$\hat{p}(\boldsymbol{s}) = \frac{1}{(2\pi)^{\frac{d}{2}} |K''(\hat{\boldsymbol{\lambda}})|^{\frac{1}{2}}} e^{K(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}^T \boldsymbol{s}},$$

where $\hat{\boldsymbol{\lambda}}$ is such that

$$K'(\hat{\boldsymbol{\lambda}}) = \boldsymbol{s}. \tag{5.1}$$

Condition (5.1) is often called the saddlepoint equation. The saddlepoint density is defined only on the interior $J_{V_{\boldsymbol{s}}}$ of the support $V_{\boldsymbol{s}}$ of the original density $p(\boldsymbol{s})$. Another important property of $\hat{p}(\boldsymbol{s})$ is that it is generally improper. A proper density can be obtained through normalization

$$\bar{p}(\boldsymbol{s}) = \frac{\hat{p}(\boldsymbol{s})}{\int_{J_{V_{\boldsymbol{s}}}} \hat{p}(\boldsymbol{s}) d\,\boldsymbol{s}}.$$

For a discrete $\boldsymbol{S}$ analogous results hold and $\bar{p}(\boldsymbol{s})$ should be interpreted as an approximation to $\text{pr}(\boldsymbol{S} = \boldsymbol{s})$. For a comprehensive introduction to saddlepoint approximations, see Butler (2007).

5.2.1 Empirical Saddlepoint approximation

Suppose that the analytic form of $K(\boldsymbol{\lambda})$ is unknown, as it generally is for simulation-based methods such as SL. If we can simulate from $p(\boldsymbol{s})$, then it is possible to estimate $K(\boldsymbol{\lambda})$ using the estimator proposed by Davison and Hinkley (1988)

$$\hat{K}_m(\boldsymbol{\lambda}) = \log \hat{M}_m(\boldsymbol{\lambda}) = \log \left(\frac{1}{m} \sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i} \right), \tag{5.2}$$

where m is the number of simulations used. Derivatives estimates of $\hat{K}(\boldsymbol{\lambda})$ are

$$\hat{K}'_m(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i} \boldsymbol{s}_i}{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}, \qquad \hat{K}''_m(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i} \boldsymbol{s}_i \boldsymbol{s}_i^T}{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}} - \hat{K}'_m(\boldsymbol{\lambda}) \hat{K}'_m(\boldsymbol{\lambda})^T.$$

These can be used to obtain an Empirical Saddlepoint approximation

$$\hat{p}_m(\boldsymbol{s}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\hat{K}''_m(\hat{\boldsymbol{\lambda}}_m)|^{\frac{1}{2}}} e^{\hat{K}_m(\hat{\boldsymbol{\lambda}}_m) - \hat{\boldsymbol{\lambda}}_m^T \boldsymbol{s}}, \tag{5.3}$$

where $\hat{\boldsymbol{\lambda}}_m$ is the solution of

$$\hat{K}'_m(\hat{\boldsymbol{\lambda}}_m) = \boldsymbol{s}. \tag{5.4}$$

Notice that $\hat{K}'(\hat{\boldsymbol{\lambda}})$ is a convex combination of the simulated vectors $\boldsymbol{s}_i$, hence (5.4) has no solution if $\boldsymbol{s}$ falls outside the convex hull of the $\boldsymbol{s}_i$s. This limitation is addressed in Section 5.3.

Feuerverger (1989) provides asymptotic results regarding how well $\hat{p}_m(\boldsymbol{s})$ approximates $\hat{p}(\boldsymbol{s})$ in a univariate setting. In Appendix D.1 we show how these carry over to the current multivariate setting. In particular, $\hat{p}_m(\boldsymbol{s})$ converges to $\hat{p}(\boldsymbol{s})$ at parametric rate $O(m^{-1/2})$ for $\boldsymbol{\lambda} \in I/2$, where $I/2$ is the subset of I such that $\boldsymbol{\lambda} \in I/2$ if $2\boldsymbol{\lambda} \in I$, while the convergence is slower outside this region. Regardless of the distribution of $\boldsymbol{S}$, $\boldsymbol{s} = \boldsymbol{\mu} = E(\boldsymbol{S})$ corresponds to $\boldsymbol{\lambda} = 0 \in I/2$, hence it might be advantageous to think of $K'(I/2)$ as a region approximately centred around $\boldsymbol{\mu}$. In Section 5.3 we build upon this interpretation.

5.3 Extended Empirical Saddlepoint approximation

The aim of this work is to use the flexibility of the Empirical Saddlepoint approximation to estimate densities for which the normal approximation is poor. The asymptotic results of Feuerverger (1989) suggest that the saddlepoint approximation should perform reasonably well in the central part of the distribution, while its accuracy decreases in the tails. More importantly, as stated in Section 5.2.1, the empirical saddlepoint equation (5.4) has a solution only if $\boldsymbol{s}$ lies inside the convex hull of the simulated data, so the resulting Empirical Saddlepoint density is not defined outside this subset of R^d. This is problematic in the context of SL because, whether we wish to estimate the unknown parameters by Maximum Likelihood or Markov Chain Monte Carlo, we cannot generally expect $\boldsymbol{s}^0$ to fall inside the convex hull of the simulated statistics in early iterations. In addition, if the model of interest is unable to generate summary statistics that are close to the observed ones, its inadequacy should ideally be quantified by a low, rather than an undefined, value of the synthetic likelihood. Hence, we need a remedy that allows us to solve (5.4) for any $\boldsymbol{s} = \boldsymbol{s}^0$.

To motivate our solution, notice that solving (5.4) is equivalent to minimizing

$$\{\hat{K}(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{s}\}^2,$$

which would be guaranteed to have a unique minimum, if strong convexity held. That is, if

$$\exists\, \epsilon \in R^+ \text{ such that } \boldsymbol{z}^T \hat{K}''(\boldsymbol{\lambda})\boldsymbol{z} > \epsilon ||\boldsymbol{z}||^2, \quad \forall\ \boldsymbol{\lambda}, \boldsymbol{z} \in R^d \text{ such that } ||\boldsymbol{z}|| > 0, \quad (5.5)$$

then (5.4) could be solved for any $\boldsymbol{s}$. Unfortunately, the following proposition states that this in not the case.

Proposition 5.1. *$\hat{K}(\boldsymbol{\lambda})$ is strictly, but not strongly, convex.*

Proof. See Appendix D.2. □

However, the fact that $\hat{K}(\boldsymbol{\lambda})$ is strictly convex assures that tilting this estimator with a strongly convex function will produce a modified estimator that is strongly convex itself, so that (5.4) could be solved for any $\boldsymbol{s}$. Therefore, we propose to use a modified estimator

$$\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s}) = g(\boldsymbol{s}, \gamma)\hat{K}_m(\boldsymbol{\lambda}) + \{1 - g(\boldsymbol{s}, \gamma)\}\hat{G}_m(\boldsymbol{\lambda}), \quad (5.6)$$

where $\hat{G}_m(\boldsymbol{\lambda})$ is a strongly convex function, while $g(\boldsymbol{s}, \gamma)$ is a function of $\boldsymbol{s}$, parametrized by γ, which determines the mix between the two functions. Furthermore, we require

$$g(\boldsymbol{s}, \gamma) : R^d \to [0, 1], \qquad \lim_{||\boldsymbol{s} - \hat{\boldsymbol{\mu}}|| \to \infty} g(\boldsymbol{s}, \gamma) = 0. \quad (5.7)$$

A natural choice for $\hat{G}_m(\boldsymbol{\lambda})$ is the parametric estimator of $K(\boldsymbol{\lambda})$

$$\hat{G}_m(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \hat{\boldsymbol{\mu}} + \frac{1}{2}\boldsymbol{\lambda}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\lambda}, \quad (5.8)$$

which is unbiased and consistent for multivariate normal random variables. This solution is related to that of Wang (1992), who modified the truncated estimator of Easton and Ronchetti (1986), and to the proposal of Bartolucci (2007), in the context of Empirical Likelihood (Owen, 2001). We refer to the density obtained by using estimator (5.6) within (5.3) as the Extended Empirical Saddlepoint approximation (ESA). In Section 5.3.1 we propose a particular form for $g(\boldsymbol{s}, \gamma)$.

5.3.1 Choice of mixture function $g(\boldsymbol{s}, \gamma)$

In the following we base our choice of $g(\boldsymbol{s}, \gamma)$ on the relative MSE performance of estimators (5.2) and (5.8), under normality of $\boldsymbol{S}$. Firstly notice that, irrespective of the distribution of $\boldsymbol{S}$, $\hat{M}(\boldsymbol{\lambda})$ is unbiased. If $\boldsymbol{S}$ is normally distributed, $e^{\boldsymbol{\lambda}^T \boldsymbol{S}}$ follows a log-normal distribution and

$$M(\boldsymbol{\lambda}) = e^{\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}}, \quad \text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\} = \frac{1}{m}\big(e^{\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}} - 1\big) e^{2\boldsymbol{\mu} + \boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}},$$

with the saddlepoint equation (5.1) being solved by

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{s} - \boldsymbol{\mu}). \tag{5.9}$$

In order to approximate the MSE of (5.2) as a function of $\boldsymbol{\lambda}$, we firstly approximate its expected value by Taylor expansion around $M(\boldsymbol{\lambda})$

$$\begin{aligned} E\big\{\hat{K}(\boldsymbol{\lambda})\big\} &= E\Big[\log M(\boldsymbol{\lambda}) + \tfrac{1}{M(\boldsymbol{\lambda})}\big\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\big\} - \tfrac{1}{2M(\boldsymbol{\lambda})^2}\big\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\big\}^2 + \cdots\Big] \\ &= \log M(\boldsymbol{\lambda}) - \tfrac{1}{2M(\boldsymbol{\lambda})^2}\text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\} + O(m^{-2}). \end{aligned}$$

Similarly we have that

$$\begin{aligned} E\big\{\hat{K}(\boldsymbol{\lambda})^2\big\} &= E\Big[\big\{\log M(\boldsymbol{\lambda})\big\}^2 + \tfrac{2\log\{M(\boldsymbol{\lambda})\}}{M(\boldsymbol{\lambda})}\big\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\big\} \\ &+ \Big\{\tfrac{1}{M(\boldsymbol{\lambda})^2} - \tfrac{\log M(\boldsymbol{\lambda})}{M(\boldsymbol{\lambda})^2}\Big\}\big\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\big\}^2 + \cdots\Big] \\ &= \big\{\log M(\boldsymbol{\lambda})\big\}^2 + \Big\{\tfrac{1}{M(\boldsymbol{\lambda})^2} - \tfrac{\log M(\boldsymbol{\lambda})}{M(\boldsymbol{\lambda})^2}\Big\}\text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\} + O(m^{-2}), \end{aligned}$$

hence

$$\begin{aligned} \text{var}\{\hat{K}(\boldsymbol{\lambda})\} &= E\big\{\hat{K}(\boldsymbol{\lambda})^2\big\} - E\big\{\hat{K}(\boldsymbol{\lambda})\big\}^2 \\ &= \tfrac{1}{M(\boldsymbol{\lambda})^2}\text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\} - \tfrac{1}{4M(\boldsymbol{\lambda})^4}\Big[\text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\}\Big]^2 + O(m^{-2}). \end{aligned}$$

Finally

$$\begin{aligned} \text{MSE}\{\hat{K}(\boldsymbol{\lambda})\} &= \text{Bias}\{\hat{K}(\boldsymbol{\lambda})\}^2 + \text{var}\{\hat{K}(\boldsymbol{\lambda})\} \\ &= \tfrac{1}{M(\boldsymbol{\lambda})^2}\text{var}\big\{\hat{M}(\boldsymbol{\lambda})\big\} + O(m^{-2}) \\ &= \tfrac{1}{m}\big(e^{\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}} - 1\big) + O(m^{-2}) \\ &= \tfrac{1}{m}\big\{e^{(\boldsymbol{s}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})} - 1\big\} + O(m^{-2}), \end{aligned} \tag{5.10}$$

where the last equality holds due to (5.9). The $O(m^{-2})$ term in (5.10) derives from

$$\begin{aligned} E\Big[\Big\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\Big\}^3\Big] &= E\Big[\Big\{\tfrac{1}{m}\textstyle\sum_{i=1}^m e^{\boldsymbol{\lambda}^T \boldsymbol{S}_i} - E(e^{\boldsymbol{\lambda}^T \boldsymbol{S}})\Big\}^3\Big] \\ &= \tfrac{1}{m^3}\textstyle\sum_{i=1}^m E\Big[\big\{e^{\boldsymbol{\lambda}^T \boldsymbol{S}_i} - E(e^{\boldsymbol{\lambda}^T \boldsymbol{S}})\big\}^3\Big] \\ &= \tfrac{1}{m^2}\mu_3\big(e^{\boldsymbol{\lambda}^T \boldsymbol{S}}\big), \end{aligned}$$

where $\mu_3(X)$ is the third central moment of a random variable X and the second equality is justified by independence.

Estimator (5.8) is unbiased and consistent, if $\boldsymbol{S}$ is normally distributed, hence

$$\text{MSE}\{\hat{G}_m(\boldsymbol{\lambda})\} = \text{var}\{\hat{G}_m(\boldsymbol{\lambda})\} = \boldsymbol{\lambda}^T \text{var}(\hat{\boldsymbol{\mu}})\boldsymbol{\lambda} + \frac{1}{4}\text{var}\Big(\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}\Big),$$

due to the independence between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ for normally distributed random variables (Basu's theorem). In addition, as m goes to infinity we have, from Rencher and Christensen (2012), that

$$(m-1)\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{m}(\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})^T \to \boldsymbol{W}, \quad \text{where} \quad \boldsymbol{W} \sim \text{Wishart}(\boldsymbol{\Sigma}, m-1),$$

and from Rao (2009)

$$\boldsymbol{\lambda}^T\boldsymbol{W}\boldsymbol{\lambda} \sim \tau^2 Q, \;\; \text{where} \;\; \tau^2 = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} \;\; \text{and} \;\; Q \sim \chi^2_{m-1},$$

hence, by using (5.9), we obtain

$$\begin{aligned} m\text{MSE}\{\hat{G}_m(\boldsymbol{\lambda})\} &\to \hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}} + \tfrac{m}{2(m-1)}(\hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}})^2 \\ &\to \hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}}\big(1 + \tfrac{1}{2}\hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}}\big) \\ &= (\boldsymbol{s}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\big\{1 + \tfrac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\big\}. \end{aligned}$$

Having derived approximations to the MSEs of (5.2) and (5.8), we propose to base the choice of mixture function on their relative sizes

$$g(\boldsymbol{s},\gamma) = \left[\frac{(\boldsymbol{s}-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{s}-\hat{\boldsymbol{\mu}})\big\{1 + \frac{1}{2}(\boldsymbol{s}-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{s}-\hat{\boldsymbol{\mu}})\big\} + 1}{\exp\big\{(\boldsymbol{s}-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{s}-\hat{\boldsymbol{\mu}})\big\}}\right]^{\gamma} \quad \text{with} \quad \gamma > 0. \tag{5.11}$$

Here γ is a tuning parameter, which determines the rate at which $g(\boldsymbol{s},\gamma)$ varies from 1 to 0 as the distance between $\boldsymbol{s}$ and $\hat{\boldsymbol{\mu}}$ increases. Function (5.11) fulfils requirement (5.7).

Our choice (5.11) has two main shortcomings: it is based on a normality assumption for $\boldsymbol{S}$ and, most importantly, it does not take the sample size m into account. In regard to the former issue: using higher moments of the simulated statistics to determine (5.11) might be attractive, but our experience suggests that this would result in very unstable estimates. In Section 5.3.2 we describe a selection procedure for γ which addresses the second problem.

5.3.2 Selecting γ by cross-validation

The choice of γ is critical for the performance of our method, and at first sight it not clear on what principle this choice should be based. We interpret γ as a complexity-controlling parameter, which determines the balance between two density estimators: the empirical saddlepoint, which is characterized by higher flexibility and variance, and

Algorithm 4 Cross-validation with nested normalization

1: Create a grid of possible values of γ.
2: Simulate m random vectors $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$ from the true density $p(\boldsymbol{s})$.
3: For each γ_i

- Estimate the normalizing constant of $\hat{p}_m(\boldsymbol{s}, \gamma_i)$ by importance sampling, that is

$$\bar{p}_m(\boldsymbol{s}, \gamma_i) = \frac{\hat{p}_m(\boldsymbol{s}, \gamma_i)}{\hat{z}_m(\gamma_i)}.$$

  where

$$\hat{z}_m(\gamma_i) = \frac{1}{l} \sum_{i=1}^{l} \frac{\hat{p}_m(\boldsymbol{S}_i)}{q(\boldsymbol{S}_i)}, \qquad \boldsymbol{S}_i \sim q(\boldsymbol{s}), \quad \text{for } i = 1, \ldots, l.$$

  A reasonably efficient importance density $q(\boldsymbol{s})$ can be obtained by fitting a multivariate normal density to $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$.

- For each of the k folds

  - Calculate the negative log-likelihood of the validation data using $\bar{p}_m(\boldsymbol{s}, \gamma_i)$ fitted to the training data.

4: Select the value γ_i that minimizes the negative validation log-likelihood, averaged across the K folds.

the normal distribution, which generally has higher bias, but lower variance. Hence, we propose to select γ by k-fold cross-validation, as detailed in the Algorithm 4.

In Appendix D.3 we show that, as m and $l \to \infty$, Algorithm 4 consistently selects the value of γ which minimizes the Kullback-Leibler divergence between $\bar{p}(\boldsymbol{s}, \gamma)$ and $p(\boldsymbol{s})$. Saddlepoint approximations are exact for Gaussian densities (Butler, 2007), hence the Gaussian case is recovered as $\gamma \to \infty$.

5.4 Use within Synthetic Likelihood

This section describes how the proposed density estimator can be used within the context of SL. To obtain an initial estimate, $\boldsymbol{\theta}_I$, of the unknown parameters it is reasonable to maximize the synthetic likelihood based on the Gaussian approximation, which is less computationally expensive. Then γ can be selected using Algorithm 4, with $p(\boldsymbol{s}) = p(\boldsymbol{s}|\boldsymbol{\theta}_I)$. Given γ, pointwise estimates of the synthetic likelihood can be based on the new density estimator by using a procedure analogous to Algorithm 3, which we describe in Appendix D.6.1.

Assuming that m, the number of summary statistics simulated from $p(\boldsymbol{s}|\boldsymbol{\theta})$, is much larger than d, the cost of evaluating the Gaussian synthetic likelihood is $O(md^2)$, which is the cost of obtaining $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Calculating $\hat{K}''(\boldsymbol{\lambda})$ has the same complexity, but solving the empirical saddlepoint equation (5.4) numerically implies that $\hat{K}''(\boldsymbol{\lambda})$ will be evaluated at several values of $\boldsymbol{\lambda}$. The proposal described in Section 5.3 assures that the underlying root finding problem is strongly convex, hence few iterations of a Newton-Raphson algorithm are generally sufficient to solve (5.4) with high accuracy. The computational

cost of a synthetic likelihood estimate is then $\mathrm{O}(lmd^2)$, if the normalizing constant is estimated using l importance samples.

Choosing m and l involves trading off the accuracy of the density estimator with the computational effort. The solution to this problem depends on the inferential approach followed. If the pointwise estimates of $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ are used within a Metropolis Hastings algorithm, than the theoretically motivated guidelines provided by Doucet et al. (2012) should be applicable. Under the Gaussian version of SL, Meeds and Welling (2014) choose m adaptively by modelling the distribution of the acceptance ratio. It might be possible to devise a similar approach to selecting m and l under the new density estimator. We are not aware of analogous results in the context of maximizing noisy likelihood functions, which is the approach we follow in Section 5.5 and 5.6. Our experience suggests that the choice of m and l is not critically important, provided that a robust stochastic optimizer is used and that convergence is assessed, possibly by running parallel optimizations.

Let $\boldsymbol{\theta}_0$ be the true parameter vector. In the following we prove that maximizing the synthetic likelihood leads to consistent parameter estimates, under either the Gaussian or the new density estimator. We start by making the following assumptions.

Assumption 5.2. *The summary statistics depend on a set of underlying observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, and have mean and covariance matrix*

$$\boldsymbol{\mu}^n_{\boldsymbol{\theta}} = E(\boldsymbol{S}_n|\,\boldsymbol{\theta}), \quad \boldsymbol{\Sigma}^n_{\boldsymbol{\theta}} = E\big\{(\boldsymbol{S}_n - \boldsymbol{\mu}^n_{\boldsymbol{\theta}})(\boldsymbol{S}_n - \boldsymbol{\mu}^n_{\boldsymbol{\theta}})^T|\,\boldsymbol{\theta}\big\}.$$

where $\boldsymbol{S}_n = S(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. In addition there exists $\delta > 0$ such that, for any $\boldsymbol{\theta}$

$$\lim_{n\to\infty} \boldsymbol{\mu}^n_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} \quad \textit{and} \quad \lim_{n\to\infty} n^{\delta}\boldsymbol{\Sigma}^n_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}.$$

Assumption 5.3. *$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mu(\boldsymbol{\theta})$ is one to one.*

Theorem 5.4. *If Assumption 5.2 and 5.3 hold, the synthetic likelihood, $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, based on the Gaussian density is asymptotically maximal at $\boldsymbol{\theta}_0$, as m, l, and $n \to \infty$.*

Proof 5.5. *See Appendix D.4.*

To prove consistency under the new density estimator, we make an additional assumption.

Assumption 5.6. *For every n, the moment generating function of $\boldsymbol{S}_n$ exists for $\boldsymbol{\lambda} \in I$, where I is a nonvanishing subset of R^d containing the origin.*

Assumption 5.7. *Let $\hat{\gamma}^n_{\boldsymbol{\theta}_I}$ be the output of Algorithm 4, corresponding to simulation effort m, l and sample size n. As m, l and $n \to \infty$, $\hat{\gamma}^n_{\boldsymbol{\theta}_I}$ converges to $\gamma_{\boldsymbol{\theta}_I} > 0$, for any initialization $\boldsymbol{\theta}_I$.*

Theorem 5.8. *Under Assumptions 5.2-5.7, the synthetic likelihood, $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, based on the ESA density is asymptotically maximal at $\boldsymbol{\theta}_0$, as m, l and $n \to \infty$.*

Proof 5.9. *See Appendix D.5.*

In Section 5.5 we illustrate the performance of ESA on a simple example, while in Section 5.6 we use it to fit a complex ecological model.

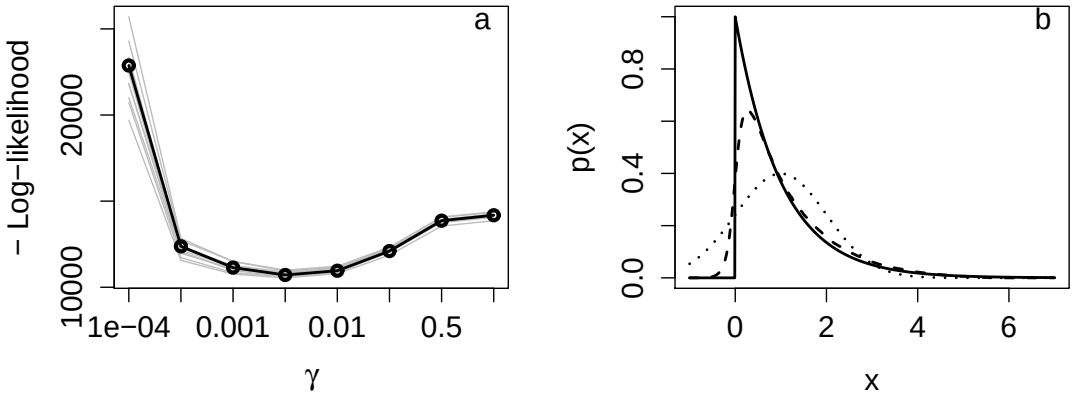

Figure 5-1: *a: Curves from 10-fold cross-validation, the black line is their average. b: True Exp*(1) *density (black), ESA (dashed) and normal (dotted) approximation.*

5.5 Multivariate shifted exponential distribution

Consider a d-dimensional random vector $\boldsymbol{X}$, where each marginal follows a shifted exponential distribution

$$X_k \sim \theta_k + \text{Exp}(\beta), \quad \text{for } k = 1, \dots, d. \tag{5.12}$$

The plot in Figure 5-1a contains the results of a 10-fold cross-validation run, obtained using $d = 10$, $l = 10^3$, $m = 10^4$, $\beta = 0.5$ and $\theta_1 = \cdots = \theta_d = 0$. The cross-validation curve is minimized by $\gamma = 5 \times 10^{-3}$, and the plot in Figure 5-1b shows the true and approximate marginal densities of one component X_k. The ESA approximation to the marginal, obtained by marginalizing the d-dimensional fit, is clearly more accurate than a normal density.

To demonstrate the usefulness of ESA in the context of SL, we use it to estimate the shifts $\theta_1, \dots, \theta_d$, all of which have been fixed to 1. In particular, we choose $\boldsymbol{s}^0 = \boldsymbol{x}$, where a single vector $\boldsymbol{x}$ has been simulated from (5.12), and we maximize the resulting synthetic likelihood, using either the Gaussian or the new density estimator. Notice that if we take $\boldsymbol{x}$, the true Maximum Likelihood Estimate, as the reference point estimate, the bias of the Gaussian estimates is $1/\beta$. By averaging the squared errors across the 10 dimensions, we obtain MSEs equal to 3.8 and 0.56, using the normal and the ESA approximation respectively. In an analogous 20-dimensional run, using $m = 5 \times 10^4$, the MSE was reduced from 4.1 to 1.26. P-values from t-test for differences in log-absolute errors were around 10^{-6} in both runs.

5.6 Formind forest model

5.6.1 The model

To test our proposal in a realistic setting, we consider Formind, an individual-based model describing the main natural processes driving forests dynamics. Here we describe its basic features, while we refer to Hartig et al. (2014) for further details.

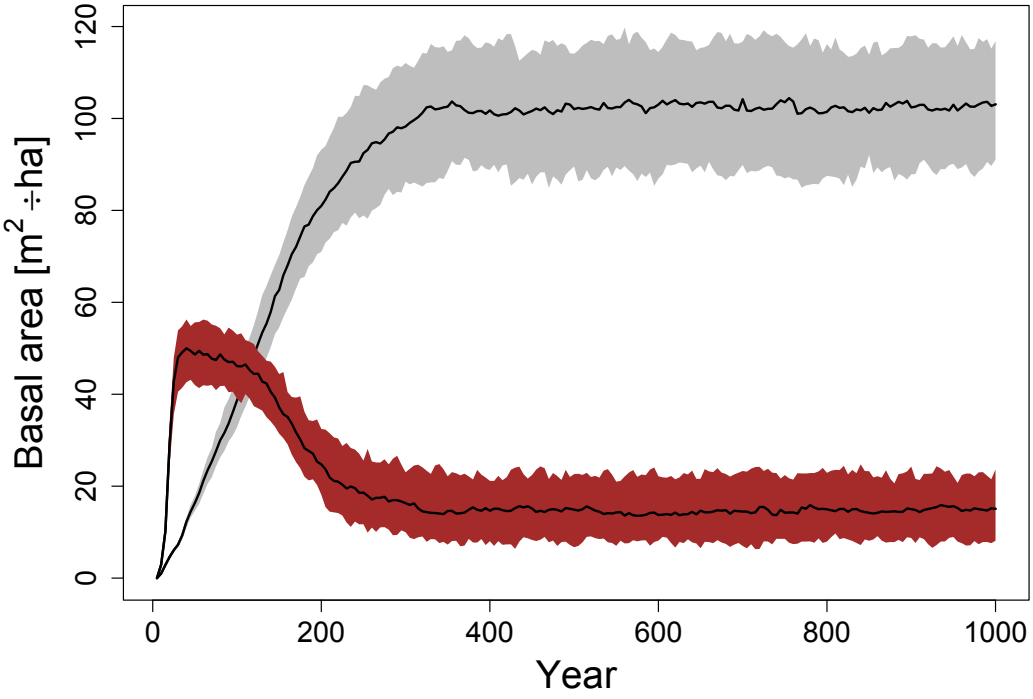

Figure 5-2: *Simulated total basal area of pioneers (brown) and late successionals (grey).*

The model describes the growth and population dynamics of tree individuals in a simulation area that is divided in 20×20m patches, with individual trees being assigned explicitly to one patch. Tree species with similar characteristics are grouped into Plant Functional Types (PFTs). A constant input of seeds deposits on average s_j seeds of the j-th PFT per hectare per year. The main factor determining both seed establishment and growth is the light climate in the patch. For example, pioneer types will establish only in patches relatively free of overshadowing trees, while late successional trees are able to grow below a dense canopy. Trees are subject to a baseline mortality rates m_j, which is specific to each PFT.

Figure 5-2 represents a typical model output. The two curves represent the dynamics of the median and 90% quantiles of the total basal area of stems on one hectare

$$\mathrm{b}_j = \sum_{k=1}^{N_j} 2\pi d_{jk}^2, \quad N_j = \text{number of trees in the } j\text{-th PFT},$$

of pioneers (brown) and late successionals (grey), obtained using 100 model runs. In the first years of simulation pioneer overgrow late successional trees, due to their higher growth rate and to the favourable light conditions. As time passes late successionals keep growing below the canopy, while pioneers are affected by higher baseline mortality rate and their seedlings struggle to establish in the shade. Finally, a dynamic equilibrium is reached, where both PFTs coexist.

In the context of Formind, the need for approximate simulation-based methods comes from the complexity of this model. Indeed, Formind was developed with a focus

Parameter	True Value	Normal	ESA	Scale	P-value
μ_{pio}	5	4.7 (1.4)	5.4 (0.7)	10^{-2}	0.002
μ_{suc}	5	9.3 (6.5)	6.1 (1.6)	10^{-3}	0.003
s_{pio}	80	108.4 (41.1)	91.6 (26.2)	1	0.07
s_{suc}	20	31.6 (15.7)	23.2 (4.7)	1	0.003

Table 5.1: *The first three columns contain true parameters, means and Root MSEs (in parentheses) of the estimates obtained using the normal and the ESA estimator. These values should be scaled using the factors contained in the fourth column. P-values for differences in log-absolute errors have been calculated using t-tests.*

on ecological plausibility, rather than statistical tractability, and most of its submodels describe highly non-linear biological processes, containing one or more sources of randomness. Most importantly, the raw output of Formind is the collection of all the characteristics of individual trees in the simulations area, which obviously do not correspond to individuals present in the actual survey data. Hence, it is necessary to work with summary statistics.

5.6.2 Simulation Results

We consider two PFTs, pioneer and late successional, and we reduce the model output to 6 summary statistics. In particular, to verify whether then new density estimator can deal with large departures from normality, we used the following transformed statistics

$$S_{jk} = \alpha_{jk}^{\frac{C_{jk}-\psi_{jk}}{\sigma_{jk}}}, \quad \text{for } j \in \{1,2\},\ k \in \{1,2,3\},$$

where C_{jk} is the number of trees of the j-th PFT falling in the k-th diameter class, while α_{jk}, ψ_{jk}, and σ_{jk} are constants, whose values are reported in Appendix D.6.3. The diameter categories used for each PFT correspond to trees with small, medium or large diameters.

We simulated 24 datasets from the model and estimated the baseline mortality rates and seed input intensities of the two PFTs by maximizing the synthetic likelihood, using both the normal and the ESA approximations. In both cases, we used $l = 10^3$ and $m = 10^4$ simulated summary statistics. When the ESA was used, γ was fixed to 5.5×10^{-3}, chosen using Algorithm 4. Table 5.1 reports the true parameters, together with the means and Root MSEs of the estimates, from the normal or the ESA approximations. See Appendix D.6.2 and D.6.3 for more details about the optimization setting.

Using the ESA, rather than the normal approximation, leads to lower MSEs for all model parameters. The plots in Figure 5-3 compare the marginal distributions of the summary statistics, simulated from the model using the true parameter values,

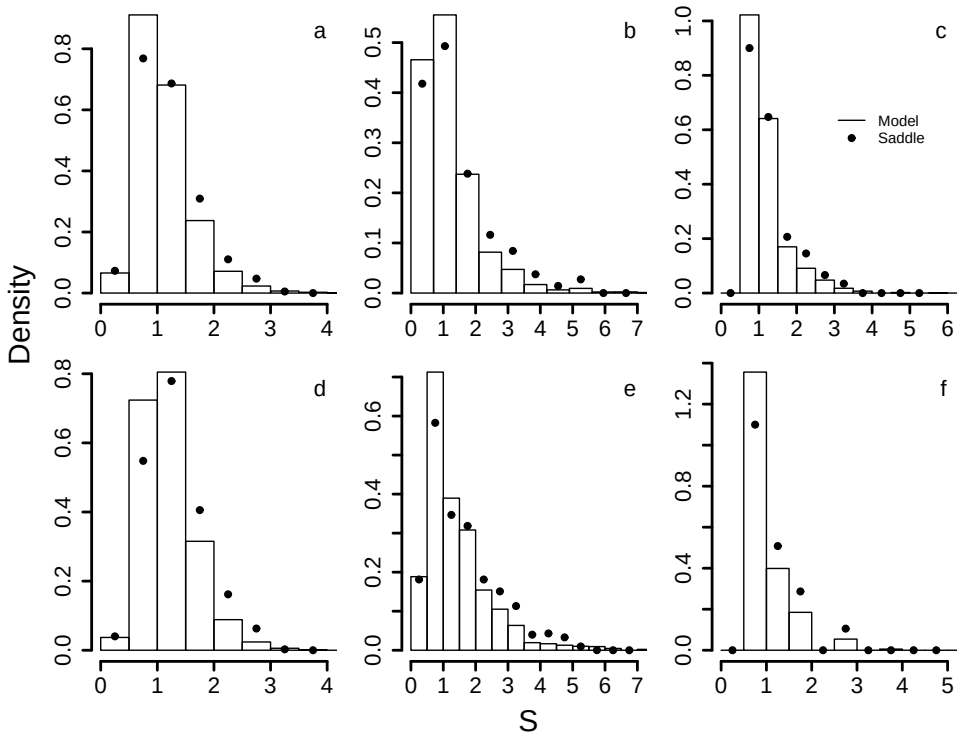

Figure 5-3: *Marginal distributions of summary statistics corresponding to small, medium and large pioneers (a, b, c) and successionals (d, e, f).*

with those obtained by simulating random vectors from ESA, fitted to the simulated statistics using the same values of γ and m used during the optimization. ESA gives a good fit to the marginal distributions of the summary statistics, all of which are far from normal.

5.7 Conclusions

We described a flexible density estimator and we illustrated its use in the context of models with intractable likelihoods. We have shown that ESA scales well with the number of dimensions and that it is able to model densities for which a normal approximation is clearly inadequate.

The proposed density estimator requires little tuning, because its only parameter, γ, can be selected using standard statistical tools, such as cross-validation. In the context of SL, and of approximate methods in general, this is an important feature, since it allows practitioners to focus on identifying informative summary statistics, rather than on other aspects of the inferential procedure. An alternative approach would be to select γ by optimizing the accuracy of the resulting parameter estimates, instead of the predictive performance of the density estimator. In this work we have not followed this approach, because a naive implementation of this idea would be computationally expensive for any fairly complex model.

From a practical point of view, the computational efficiency of SL, relatively to

ABC methods, is of critical importance. Indeed, producing pointwise estimates of the synthetic likelihood might seem unnecessarily expensive, if the aim is obtaining an approximate sample from $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$. On the other hand, while ABC methods target an approximation to $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$ directly, their accuracy is inversely proportional to the acceptance rate of the sampler. Gutmann and Corander (2015), Wilkinson (2014) and Meeds and Welling (2014) proposed using Gaussian Processes to increase the computational efficiency of SL and ABC methods. The first two proposals, being based on pointwise likelihood estimates, could be used in conjunction with ESA. The MSL algorithm, described in Chapter 4, and the proposal of Meeds and Welling (2014) model only the first two moments of the simulated statistics, hence it is not clear whether either approach could be modified to take higher moments into account, as the new density estimator does.

In this chapter we do not make any distributional assumption on the summary statistics, apart from (5.6), and we present ESA as a non-parametric density estimator, as suggested by Feuerverger (1989). However, previous contributions in the literature consider the use of Empirical Saddlepoint approximations for particular classes of statistics such as M-estimators (Monti and Ronchetti, 1993; Ronchetti and Welsh, 1994) and L-statistics (Easton and Ronchetti, 1986). It would be interesting to study the asymptotic performance of ESA for such statistics.

CHAPTER 6

CONCLUSIONS AND FURTHER WORK

In Chapter 2 we illustrated some of the difficulties that might be encountered when working with highly non-linear ecological and epidemiological State Space Models (SSMs). In particular, we showed that the likelihood function can be highly multimodal when the dynamics are far from stable and the process noise is low. Through simulated experiments, we demonstrated that approaches based on information reduction, such as SL and ABC, can deal with this issue better than methods based on particle filtering, such as PMMH and IF. In particular, we illustrated that, by using summary statistics that discard the phase of the systems, SL can provided stable synthetic likelihood estimates, even when the SIR's likelihood estimates are extremely noisy. However, this robustness does not come without cost. When the process noise is sufficient to smooth out the likelihood function, PMMH and IF consistently outperform ABC and SL in terms of accuracy in parameter estimation.

In Chapter 3 we continued the comparison work started in Chapter 2, by restricting our attention to SLMH and PMMH and by considering more realistic examples. When simulated data was used, the results of this chapter confirmed, though less dramatically than in Chapter 2, that full likelihood methods are generally more accurate than methods based on information reduction. However, new aspects of the robustness of information reduction methods were highlighted. In particular, the blowfly example (Section 3.2) clearly showed that SLMH is much more robust to bad initializations that PMMH, even when simulated data is used. When Nicholson's experimental datasets were used to fit the model, this example also showed that information reduction methods can be more robust than particle filters to model mis-specification and outliers. In particular, PMMH failed to classify the system's dynamics as cyclic on two of the datasets, which is attributable to the model's failure to explain few idiosyncracies of these datasets. PMMH's estimates seemed to be slightly biased toward stable dynamics also in the voles example (Section 3.4), when real data was used to fit the model.

Overall, the results obtained in Chapter 2 and 3 suggest that information reduction and state space methods should be used in conjunction, even when working with highly non-linear SSMs meant to reproduce the full data. In particular, information reduction methods can be used in support of state space methods, by providing initial parameter estimates and a robust benchmark, when model mis-specification and outliers might be present. Future work might be aimed at verifying, possibly theoretically, whether ABC

and SL share some of the robustness properties of other, more traditional, methods based on summary statistics, such as those described in Jiang et al. (2004).

In Chapter 4 we described how the synthetic likelihood can be maximized efficiently, by using local linear models and GLMs to describe how the first two moments of the summary statistics vary with the parameters. We also described how the resulting algorithm, MSL, is closely related to another approach, CUGMM, which can produce parameter estimates by employing the same local models. Both MSL and CUGMM outperformed SLMH in all the examples considered in Chapter 4, in terms of rapidity of MSE reduction as a function of the number of simulated statistics. Further, we showed that the CUGMM estimator has two important properties. First, the parameter estimates are asymptotically normal, which is not the case for MSL (Wood, 2010). Second, the derivatives of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ with respect to the $\boldsymbol{\theta}$ are not needed to compute the asymptotic covariance of $\hat{\boldsymbol{\theta}}$ and they are not essential to minimize CUGMM's objective function, provided that the optimizer's initialization is close enough to a local minimum. This property seems to be very important from a practical point of view, because modelling the covariance matrix of the statistics is challenging. The main issue is that any model for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ needs to preserve its positive definiteness. In MSL, we dealt with this requirement by modelling only the marginal variances of the statistics using GLMs, while considering the correlation structure constant. Given that MLS and CUGMM performed similarly in the examples described in Chapter 4, further comparison work is needed. In particular, it is possible that, from a practical point of view, CUGMM should generally be preferred, due to its lower computational costs and to the asymptotic normality of its parameter estimates. A promising direction for future research would be to use the derivatives estimates produced by the Algorithm 2 to create an adaptive transition kernel for SLMH. In particular, it should be possible to set up a version of the Riemannian Metropolis Adjusted Langevin Algorithm of Girolami and Calderhead (2011), which uses the estimated derivatives to adapt to the local geometrical structure of the synthetic likelihood.

In Chapter 5 we described a new density estimator, ESA, and we described how it can be used within the context of SL, in place of a Gaussian density. Through simulated experiments, we demonstrated that ESA scales well with the number of statistics, while being able to capture sizeable departures from normality. Probably the main drawback of ESA is that it is unnormalized. While we proposed to estimate its normalizing constant by Importance Sampling, more efficient approaches should be possible. For instance, it would be interesting to verify whether control variates could be used to estimate the normalizing constant more accurately. In particular, Mira et al. (2013) proposed a variance reduction scheme which uses control variates that are functions of the gradient of the target density. Given that the gradient of ESA can be calculated analytically, this approach might be directly applicable.

APPENDIX A

DETAILS OF THE COMPARISON ON SIMPLE CHAOTIC MAPS

A.1 Discretized SSM

The likelihood of a simple SSM can be written in the following form

$$p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}) = p(\boldsymbol{y}_1|\boldsymbol{\theta}) \prod_{t=2}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}, \boldsymbol{\theta}),$$

and, if m is the number of discrete levels of the hidden state, then

$$p(\boldsymbol{y}_1|\boldsymbol{\theta}) = \sum_{i=1}^{m} p(\boldsymbol{y}_1|\boldsymbol{n}_1^i, \boldsymbol{\theta}) p(\boldsymbol{n}_1^i|\boldsymbol{\theta}),$$

and

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}, \boldsymbol{\theta}) &= \sum_{i=1}^{m} p(\boldsymbol{y}_t|\boldsymbol{n}_t^i, \boldsymbol{\theta}) p(\boldsymbol{n}_t^i|\boldsymbol{y}_{1:t-1}, \boldsymbol{\theta}) \\ &= \sum_{i=1}^{m} p(\boldsymbol{y}_t|\boldsymbol{n}_t^i, \boldsymbol{\theta}) \sum_{j=1}^{m} p(\boldsymbol{n}_t^i|\boldsymbol{n}_{t-1}^j, \boldsymbol{\theta}) p(\boldsymbol{n}_{t-1}^j|\boldsymbol{y}_{1:t-1}, \boldsymbol{\theta}), \end{aligned}$$

where

$$\begin{aligned} p(\boldsymbol{n}_{t-1}^j|\boldsymbol{y}_{1:t-1}, \boldsymbol{\theta}) &= \frac{\sum_{k=1}^{m} p(\boldsymbol{y}_{1:t-1}, \boldsymbol{n}_{t-1}^j, \boldsymbol{n}_{t-2}^k|\boldsymbol{\theta})}{p(\boldsymbol{y}_{1:t-1}|\boldsymbol{\theta})} \\ &= p(\boldsymbol{y}_{t-1}|\boldsymbol{n}_{t-1}^j, \boldsymbol{\theta}) \sum_{k=1}^{m} p(\boldsymbol{n}_{t-1}^j|\boldsymbol{n}_{t-2}^k, \boldsymbol{\theta}) p(\boldsymbol{n}_{t-2}^k|\boldsymbol{y}_{1:t-2}, \boldsymbol{\theta}) \frac{p(\boldsymbol{y}_{1:t-2}|\boldsymbol{\theta})}{p(\boldsymbol{y}_{1:t-1}|\boldsymbol{\theta})} \\ &= p(\boldsymbol{y}_{t-1}|\boldsymbol{n}_{t-1}^j, \boldsymbol{\theta}) \sum_{k=1}^{m} p(\boldsymbol{n}_{t-1}^j|\boldsymbol{n}_{t-2}^k, \boldsymbol{\theta}) \frac{p(\boldsymbol{n}_{t-2}^k|\boldsymbol{y}_{1:t-2}, \boldsymbol{\theta})}{p(\boldsymbol{y}_{t-1}|\boldsymbol{y}_{1:t-2}, \boldsymbol{\theta})}. \end{aligned}$$

These formulas can be used to calculate the likelihood of a discrete SSM exactly.

A.2 Computational details

To fit the models described in this work we used the *synlik* (Fasiolo and Wood, 2014), *EasyABC* (Jabot et al., 2013) and *pomp* (King et al., 2014) R-packages. The first two provide implementations of SL and ABC respectively, while we used *pomp* to run the IF and PMMH algorithms.

The data was simulated using the following parameter values:

- Generalized Ricker: $r = 44.7$, $\theta = 1$, $\sigma = 0.3$, $\phi = 10$.
- Pennycuick: $r = 58$, $a = 0.1$, $\sigma = 0.3$, $\phi = 1$.
- Maynard-Smith: $r = 18$, $b = 6$, $\sigma = 0.4$, $\phi = 24$.
- Varley: $r = 15$, $b = 5.5$, $c = 1$, $\sigma = 0.45$, $\phi = 20$.

For SL and ABC-MCMC we used the set of 13 summary statistics proposed by Wood (2010):

- the autocovariances of the path $y_{1:T}$ up to lag 5;
- the mean population $\bar{y}$;
- the number of zeros observed;
- the coefficients of the regression

$$y_{t+1}^{0.3} = \beta_1 y_t^{0.3} + \beta_2 y_t^{0.6} + z_t;$$

- the coefficients of a cubic regression of the ordered differences $y_t - y_{t-1}$ on their observed values.

Tables A.1 to A.5 contain the limits of the uniform priors (or box constraints under IF) and initial values used for each model and parameter.

	Initial	Lower	Upper
r	2.80	2.00	5.00
σ	-2.30	-3.00	-0.22
ϕ	1.79	1.61	3.00

Table A.1: *Prior boundaries for Ricker*

Tables A.6 to A.11 contain the root median squared errors (MSE) and coverage frequencies for each parameter of the five models considered, using each method. The last row indicates which method achieved the lowest mean squared error, for each model parameter.

	Initial	Lower	Upper
r	2.80	2.00	5.00
θ	0.41	-0.69	0.41
σ	-2.30	-3.00	-0.22
ϕ	1.79	1.61	3.00

Table A.2: *Prior boundaries for Generalized Ricker*

	Initial	Lower	Upper
r	3.69	2.50	5.00
a	-1.20	-4.61	-0.69
σ	-0.69	-3.00	-0.22

Table A.3: *Prior boundaries for Pennycuick*

	Initial	Lower	Upper
r	2.30	1.50	4.00
b	2.20	0.69	2.30
σ	-0.69	-3.00	-0.22
ϕ	2.64	2.30	3.56

Table A.4: *Prior boundaries for Maynard-Smith*

	Initial	Lower	Upper
r	2.30	1.50	4.00
b	2.01	0.69	2.30
C	0.69	-2.30	0.69
σ	-1.61	-3.00	-0.22
ϕ	2.71	2.30	3.40

Table A.5: *Prior boundaries for Varley*

	r	σ	ϕ
SLMH	0.11(0.9)	0.34(0.92)	0.05(0.88)
SLMH_R	0.12(0.9)	0.34(0.92)	0.05(0.88)
ABC-MCMC	0.14(0.96)	0.2(1)	0.04(1)
IF	0.11(-)	0.28(-)	0.03(-)
PMMH	0.1(1)	0.21(1)	0.02(1)
Best	PMMH	ABC-MCMC	PMMH

Table A.6: *RMSEs(coverage) for Ricker*

	r	θ	σ	ϕ
SLMH	0.24(0.92)	0.06(0.98)	0.4(0.86)	0.17(0.96)
SLMH_R	0.23(0.96)	0.06(1)	0.41(0.92)	0.17(0.98)
ABC-MCMC	0.16(0.98)	0.04(1)	0.16(1)	0.13(1)
IF	0.13(-)	0.03(-)	0.3(-)	0.1(-)
PMMH	0.12(0.94)	0.03(1)	0.23(0.98)	0.11(0.98)
Best	PMMH	IF	ABC-MCMC	IF

Table A.7: *RMSEs(coverage) for Generalized Ricker*

	r	a	σ
SLMH	0.14(0.9)	0.05(0.94)	0.34(0.98)
SLMH_R	0.15(0.9)	0.04(0.94)	0.34(1)
ABC-MCMC	0.14(1)	0.07(1)	0.14(1)
IF	0.11(-)	0.03(-)	0.26(-)
PMMH	0.1(0.92)	0.02(0.98)	0.19(0.92)
Best	PMMH	PMMH	ABC-MCMC

Table A.8: *RMSEs(coverage) for Pennycuick*

	r	b	σ	ϕ
SLMH	0.13(0.92)	0.25(1)	0.43(0.88)	0.24(1)
SLMH_R	0.13(0.94)	0.2(1)	0.44(0.88)	0.22(1)
ABC-MCMC	0.11(1)	0.25(1)	0.17(1)	0.23(1)
IF	0.12(-)	0.45(-)	0.29(-)	0.48(-)
PMMH	0.09(0.98)	0.13(1)	0.23(0.96)	0.12(1)
Best	PMMH	PMMH	ABC-MCMC	PMMH

Table A.9: *RMSEs(coverage) for Hassell*

	r	b	σ	ϕ
SLMH	0.16(0.9)	0.07(0.88)	0.61(0.78)	0.12(0.94)
SLMH_R	0.15(0.96)	0.06(0.9)	0.67(0.92)	0.1(1)
ABC-MCMC	0.19(0.94)	0.06(1)	0.27(1)	0.09(1)
IF	0.11(-)	0.04(-)	0.26(-)	0.06(-)
PMMH	0.09(1)	0.04(1)	0.15(1)	0.05(1)
Best	PMMH	PMMH	PMMH	PMMH

Table A.10: *RMSEs(coverage) for Maynard-Smith*

	r	b	C	σ	ϕ
SLMH	0.16(0.96)	0.07(0.92)	0.16(1)	0.87(0.76)	0.1(0.92)
SLMH_R	0.16(0.98)	0.07(0.96)	0.17(1)	0.8(0.88)	0.11(0.94)
ABC-MCMC	0.17(0.98)	0.06(1)	0.17(1)	0.32(1)	0.07(1)
IF	0.1(-)	0.05(-)	0.12(-)	0.34(-)	0.07(-)
PMMH	0.1(0.96)	0.04(0.96)	0.08(1)	0.2(0.94)	0.06(0.96)
Best	IF	PMMH	PMMH	PMMH	PMMH

Table A.11: *RMSEs(coverage) for Varley*

APPENDIX B

DETAILS OF THE REAL DATA EXAMPLES

B.1 Blowfly Model

Assuming that the population abundance $\boldsymbol{n}_{1:T}$ is perfectly observed, the likelihood of model 3.2 can be estimated as follows. Firstly notice that

$$\begin{aligned} p\{\boldsymbol{n}_{(\tau+1):T}\} &= p\{n_T|\boldsymbol{n}_{1:(T-1)}\}p\{n_{T-1}|\boldsymbol{n}_{1:(T-2)}\}\cdots p\{n_{\tau+1}|\boldsymbol{n}_{1:\tau}\}, \\ &= p(n_T|n_{T-1}, n_{T-\tau})p(n_{T-1}|n_{T-2}, n_{T-1-\tau})\cdots p(n_{\tau+1}|n_\tau, n_1). \end{aligned}$$

where, to simplify the notation, we have dropped the dependence on model parameters. Then, consider a single likelihood component

$$p(n_t|n_{t-1}, n_{t-\tau}) = \int p_r(n_t - s_t|n_{t-1}, n_{t-\tau})p_s(s_t|n_{t-1})ds_t,$$

where $\tau + 1 \leq t \leq T$, while p_r and p_s indicate the conditional densities of r_t and s_t, respectively. This integral can be estimated using

$$\frac{1}{M_1}\sum_{i=1}^{M_1} p_r(n_t - s_t^i|n_{t-1}, n_{t-\tau}),$$

where $s_t^i \sim p_s(s_t|n_{t-1})$. Evaluating each term of the above sum requires solving an additional integral, in fact

$$p_r(n_t - s_t^i|n_{t-1}, n_{t-\tau}) = \int p_r(n_t - s_t^i|n_{t-1}, n_{t-\tau}, e_t)p(e_t)de_t,$$

which can be approximated by

$$\frac{1}{M_2}\sum_{j=1}^{M_2} p(n_t - s_t^i|n_{t-1}, n_{t-\tau}, e_t^j),$$

where e_t^j is a Gamma distributed random variable, with unit mean and variance equal to σ_p^2.

To fit this model with SL, we used the set of 16 summary statistics proposed by Wood (2010):

Parameter	Prior
δ	Unif$(0.09, 0.4)$
P	Unif$(3, 12)$
N_0	Unif$(150, 800)$
σ_p^2	Unif$(0.01, 1)$
τ	Unif$(5, 25)$
σ_d^2	Unif$(0.01, 1)$

Table B.1: *Priors used for the blowfly model in the simulated setting.*

- the autocovariances of the path $n_{1:T}$ up to lag 11;
- the mean population $\bar{n}$;
- the difference between mean and median population $\bar{n} - \tilde{m}$;
- the number of zeros observed;
- the coefficients of the regression

$$n_{t+1} = \beta_1 n_t + \beta_2 n_t^2 + \beta_3 n_t^3 + \beta_4 n_{t-6} + \beta_5 n_{t-6}^2 + z_t;$$

- the coefficients of a cubic regression of the ordered differences $n_t - n_{t-1}$ on their observed values.
- the number of turning points.

The priors used when fitting the simulated datasets are reported in Table B.1. Table B.2 reports the priors used when fitting Nicholson's datasets. Notice that for τ we have used a non-uniform prior, based on information reported by Gurney et al. (1980) concerning biologically plausible values of this delay parameter.

B.2 Cholera Model

One thing to notice about model (5.3) is that cholera-related deaths

$$D_t = \frac{I_t^o m}{\gamma + \delta + m},$$

are not offset by an equal number of births in the susceptible compartment S_{t+1}. Beside not making sense biologically, this would introduce a strong feedback mechanism during

Parameter	Prior
δ	Unif$(0.02, 1)$
P	Unif$(3, 30)$
N_0	Unif$(10, 1000)$
σ_p^2	Unif$(0.01, 5)$
τ	Norm$(14, 5)$
σ_d^2	Unif$(0.01, 5)$

Table B.2: *Priors used for the blowfly model when fitting Nicholson's datasets.*

epidemics. To offset this downward bias on total population, we tilt the number of births at each step as follows

$$B^*_{t+1} = B_{t+1} + \bar{D}\Delta t$$

where $\bar{D}$ is the monthly average of the observed number of deaths during the whole period and Δt is the time step used. B^*_t is then used in place of B_t in (5.3). With this choice the sum of the number individuals in each compartment does not match the official census, but we have verified that the mismatch is minimal.

Let d_t be the number of cholera-related deaths during the t-month, and define $r_t = d_t^{1/5}$. For SLMH we used the following set of 26 summary statistics:

- the coefficients (intercept excluded) of the regression

$$r_t = \alpha_1 + \alpha_2 t + \sum_{i=1}^{4} \alpha_{3i} sin(\psi_i 2\pi t) + \alpha_{4i} cos(\psi_i 2\pi t) + z_t;$$

  where $\psi_1 = 0.12$, $\psi_2 = 1$, $\psi_3 = 2$, and $\psi_4 = 3$; Let e_t be the t-residual of such regression;

- the autocovariances of $e_{1:T}$ at lag 2, 6, and 11;

- the mean $\bar{d}$ and variance $\text{Var}(d)$ of the number of deaths;

- the scaled difference between mean and median number of deaths $(\bar{d} - \tilde{d})/\text{Var}(d)$;

- the coefficients of the auto-regression

$$e_{t+1} = \beta_1 e_t + \beta_2 e_{t-2} + \beta_3 e_{t-3} + \beta_4 e_{t-4} + \beta_{10} e_{t-10} + z_t;$$

- the coefficients of a cubic regression of the ordered differences $e_t - e_{t-1}$ on their observed values;

Parameter	Prior
γ	Unif$(1, 365)$
ϵ	Unif$(0.1, 60)$
c	Unif$(0, 1)$
ρ	Unif$(1, 60)$
m	Unif$(0, 140)$
e^{β}	N$(0, 1000)$
$e^{\beta_1}, \ldots, e^{\beta_6}$	N$(0, 1000)$
$e^{\omega_1}, \ldots, e^{\omega_6}$	N$(0, 1000)$
σ	Unif$(0, 1)$
τ	Unif$(0, 1)$

Table B.3: *Priors used for the the Cholera model.*

- the number of turning points in $d_{1:T}$;
- the median and inter-quartile range of $e_{1:T}$.

Table B.3 reports the prior distributions used.

Calculating the AICs reported in the main text was not straightforward, because the joint posterior distributions of the parameters are far from normal for each model, hence the posterior mean is inadequate as a point estimate. In addition, for both SLMH and PMMH the (synthetic) likelihood is estimated with noise, which makes finding good point estimates more difficult. To work around this issue, for each model and method, we restricted our attention to parameters corresponding to likelihood estimates above the 99th quantile and we have re-estimated the likelihood at each of those parameter values, using a 2×10^4 particles or simulations from the model. Given that these estimates had very low noise, we have used the parameter vector corresponding to highest likelihood estimate as a proxy for the MLE. Finally, we re-estimated the likelihood at the MLE using 5×10^4 simulations, and we have used it to estimate the AIC.

B.3 Voles model

For SLMH we used the following set of 17 summary statistics:

- autocovariances of $n_1, \ldots, n_T$ up to lag 5;

- mean population $\bar{n}$;
- difference between mean and median population $\bar{n} - \tilde{n}$;
- coefficients $\beta_1, \ldots, \beta_5$ of the regression

$$n_{t+1} = \beta_1 n_t + \beta_2 n_t^2 + \beta_3 n_{t-6} + \beta_4 n_{t-6}^2 + \beta_5 n_{t-6}^3 + z_t;$$

- coefficients of a cubic regression of the ordered differences $n_t - n_{t-1}$ on their observed values.
- number of turning points, $\#n$.

APPENDIX C

PROOF OF ASYMPTOTIC NORMALITY OF THE CUGMM ESTIMATOR

Assume that the observed summary statistics depend on an underlying set of observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. In the following, we use the subscript n to denote quantities that depend on the sample size. In addition, assume that, as $n \to \infty$, the summary statistics are asymptotically normal distributed, that is

$$\sqrt{n}\{\boldsymbol{S}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}n}\} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{C.1}$$

for any $\boldsymbol{\theta}$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}n}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}n}$ are the expected value and variance of $\boldsymbol{S}_n$ at $\boldsymbol{\theta}$. Notice that $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ are the limiting values of $\boldsymbol{\mu}_{\boldsymbol{\theta}n}$ and $n\boldsymbol{\Sigma}_{\boldsymbol{\theta}n}$, respectively, and assume that

$$\begin{aligned} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}n}^T}{\partial \theta_k} &\to \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}}^T}{\partial \theta_k}, \qquad \frac{\partial^2 \boldsymbol{\mu}_{\boldsymbol{\theta}n}^T}{\partial \theta_k \partial \theta_l} \to \frac{\partial^2 \boldsymbol{\mu}_{\boldsymbol{\theta}}^T}{\partial \theta_k \partial \theta_l}, \\ \frac{\partial n\boldsymbol{\Sigma}_{\boldsymbol{\theta}n}}{\partial \theta_k} &\to \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_k}, \qquad \frac{\partial^2 n\boldsymbol{\Sigma}_{\boldsymbol{\theta}n}}{\partial \theta_k \partial \theta_l} \to \frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}}{\partial \theta_k \partial \theta_l}, \end{aligned} \tag{C.2}$$

for every $k, l = 1, \ldots, p$.

Let $\hat{\boldsymbol{\theta}}_n$ be the minimizer of $f_n(\boldsymbol{\theta})$, while indicate with $\boldsymbol{\theta}_0$ the true parameter vector. We consider the Taylor expansion

$$\nabla f_n(\boldsymbol{\theta}_0) = \left[\int_0^1 \nabla^2 f_n\{\hat{\boldsymbol{\theta}}_n + \alpha(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)\}\, d\alpha \right](\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n),$$

which implies that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left[\int_0^1 \frac{1}{n}\nabla^2 f_n\{\hat{\boldsymbol{\theta}}_n + \alpha(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)\}\, d\alpha \right]^{-1} \frac{1}{\sqrt{n}} \nabla f_n(\boldsymbol{\theta}_0).$$

We have

$$\begin{aligned} \frac{1}{2\sqrt{n}} \frac{\partial f_n(\boldsymbol{\theta}_0)}{\partial \theta_k} &= \frac{1}{\sqrt{n}} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_0 n}^T}{\partial \theta_k} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}^{-1} (\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_0 n}) + \frac{1}{2\sqrt{n}} (\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_0 n})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}}{\partial \theta_k} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}^{-1} (\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_0 n}) \\ &= \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_0 n}^T}{\partial \theta_k} \boldsymbol{z}_n + \frac{1}{2\sqrt{n}} \boldsymbol{z}_n^T \frac{\partial n\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}}{\partial \theta_k} \boldsymbol{z}_n, \end{aligned}$$

where $E(\boldsymbol{z}_n) = 0$ and $Var(\boldsymbol{z}_n) = n\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0 n}^{-1}$. Assumptions (C.1) and (C.2) imply that

$$\frac{1}{2\sqrt{n}}\frac{\partial f_n(\boldsymbol{\theta}_0)}{\partial \theta_k} \xrightarrow{p} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_0}^T}{\partial \theta_k}\boldsymbol{z}, \quad \text{with} \quad \boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}),$$

hence

$$\frac{1}{2\sqrt{n}}\nabla f_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}), \tag{C.3}$$

where

$$(\Gamma_{\boldsymbol{\theta}_0})_{ij} = \left.\frac{\partial \mu_i}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

To simplify the notation, let us define $\boldsymbol{\theta}_n^\alpha = \hat{\boldsymbol{\theta}}_n + \alpha(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)$. We have that

$$\begin{aligned}
\frac{1}{2n}\frac{\partial^2 f_n(\boldsymbol{\theta}_n^\alpha)}{\partial\theta_k\partial\theta_l} = \frac{1}{n}\Bigg\{ & \frac{\partial^2 \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}^T}{\partial\theta_k\partial\theta_l}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}) - \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}^T}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_l}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}) \\
& - \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}^T}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_l} - \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}^T}{\partial\theta_l}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}) \\
& - \frac{1}{2}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_l}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}) \\
& + \frac{1}{2}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_k\partial\theta_l}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}(\mathbf{S}_n^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n})\Bigg\} \\
= & \frac{1}{n}\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}^T}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n^\alpha n}^{-1}\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_n^\alpha n}}{\partial\theta_l} + O(n^{-\frac{1}{2}}) \xrightarrow{p} \frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_0}^T}{\partial\theta_k}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\theta}_0}}{\partial\theta_l},
\end{aligned}$$

due to (C.1), (C.2) and to the assumed consistency of $\hat{\boldsymbol{\theta}}_n$ and hence of $\boldsymbol{\theta}_n^\alpha$, for $\alpha \in [0,1]$. This implies that

$$\int_0^1 \frac{1}{2n}\nabla^2 f_n\{\hat{\boldsymbol{\theta}}_n + \alpha(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)\}\, d\alpha = \int_0^1 \frac{1}{2n}\nabla^2 f_n(\boldsymbol{\theta}_n^\alpha)\, d\alpha \xrightarrow{p} \int_0^1 \Gamma_{\boldsymbol{\theta}_0}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\, d\alpha = \Gamma_{\boldsymbol{\theta}_0}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0},$$

which, together with C.3, leads to

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}\big\{\mathbf{0}, (\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0})^{-1}\big\}.$$

APPENDIX D

EMPIRICAL SADDLEPOINT APPROXIMATIONS

D.1 Asymptotics of the multivariate empirical saddlepoint approximation

Here we follow Feuerverger (1989) but develop the results in a multivariate setting, and with some changes in notation. For $\boldsymbol{\lambda} \in I$, $\hat{M}_m(\boldsymbol{\lambda})$ converges to $M(\boldsymbol{\lambda})$ almost surely. This convergence is uniform and extends to $\hat{K}_m(\boldsymbol{\lambda})$:

$$\sup_{\boldsymbol{\lambda} \in I} |\hat{M}_m(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})| \to 0, \tag{D.1}$$

$$\sup_{\boldsymbol{\lambda} \in I} |\hat{K}_m(\boldsymbol{\lambda}) - K(\boldsymbol{\lambda})| \to 0. \tag{D.2}$$

Proof: Due to the Strong Law of Large Numbers $\hat{M}_m(\boldsymbol{\lambda})$ converges to $M(\boldsymbol{\lambda})$ almost surely, for all $\boldsymbol{\lambda}$ in any countable collection $\{\boldsymbol{\lambda}_i\}$. In addition $\hat{M}_m(\boldsymbol{\lambda})$ and $M(\boldsymbol{\lambda})$ are both convex functions and, for such functions, convergence on dense subsets implies uniform convergence on compact subsets (Roberts and Varberg, 1973). This proves (D.1), while (D.2) follows by continuity of the logarithm.

For $\boldsymbol{\lambda}$ in the interior of I, these results extend to derivatives of both $\hat{M}_m(\boldsymbol{\lambda})$ and $\hat{K}_m(\boldsymbol{\lambda})$:

$$\sup_{\boldsymbol{\lambda} \in \mathrm{int}(I)} |D^i \hat{M}_m(\boldsymbol{\lambda}) - D^i M(\boldsymbol{\lambda})| \to 0, \tag{D.3}$$

$$\sup_{\boldsymbol{\lambda} \in \mathrm{int}(I)} |D^i \hat{K}_m(\boldsymbol{\lambda}) - D^i K(\boldsymbol{\lambda})| \to 0, \tag{D.4}$$

where $i = \{i_1, \ldots, i_d\}$ and:

$$D^i M(\boldsymbol{\lambda}) = \frac{\partial^k M(\boldsymbol{\lambda})}{\partial \lambda_1^{i_1} \cdots \partial \lambda_d^{i_d}}, \quad \text{with} \quad \sum_{z=1}^{K} i_z = k \in N.$$

Proof: $D^i M(\boldsymbol{\lambda})$ is finite only for $\boldsymbol{\lambda} \in \mathrm{int}(I)$. If all the elements of i are even, then $D^i \hat{M}_m(\boldsymbol{\lambda})$ and $D^i M(\boldsymbol{\lambda})$ are convex and (D.3) follows as before. Otherwise, indicate with $\boldsymbol{\lambda}^o$ the elements of $\boldsymbol{\lambda}$ for which the corresponding element of i is odd. If there is an even number of components of $\boldsymbol{\lambda}^o$ which are negative, $D^i M(\boldsymbol{\lambda})$ is still convex, otherwise

$-D^i M(\boldsymbol{\lambda})$ is. Applying the uniform convergence argument for convex functions to the two sub-cases proves (D.3). In addition, $D^i K(\boldsymbol{\lambda})$ has the form $P(\boldsymbol{\lambda})/M(\boldsymbol{\lambda})^{2^k}$ with $P(\boldsymbol{\lambda})$ being a polynomial function of $D^l K(\boldsymbol{\lambda})$, where l belongs to the set of all d-dimensional vector such that:

$$l_j \in N, \quad \sum_{j=1}^{d} l_j \leq k \quad \text{for} \quad j = 1, \ldots, d.$$

Given that an analogous argument holds for $D^i \hat{K}_m(\boldsymbol{\lambda})$, (D.4) is proved by continuity.

After noticing that $\hat{M}_m(\boldsymbol{\lambda})$ and its derivatives are unbiased estimators of $M(\boldsymbol{\lambda})$ and its corresponding derivatives, it is straightforward to show that:

$$m \operatorname{Cov}\left\{D^i \hat{M}_m(\boldsymbol{\lambda}_1), D^j \hat{M}_m(\boldsymbol{\lambda}_2)\right\} = D^{i+j} M(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2) - D^i M(\boldsymbol{\lambda}_1) D^j M(\boldsymbol{\lambda}_2),$$

for $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ such that $\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 \in I$. This entails that, if we define $I/2$ to be the subset of I such that $\boldsymbol{\lambda} \in I/2$ if $2\boldsymbol{\lambda} \in I$, than $\hat{M}_m(\boldsymbol{\lambda})$ is a $\sqrt{m}$-consistent estimator of $M(\boldsymbol{\lambda})$, for $\boldsymbol{\lambda} \in I/2$. An analogous, but asymptotic, result for $\hat{K}_m(\boldsymbol{\lambda})$ is the following:

$$m \operatorname{Cov}\left\{D^i \hat{K}_m(\boldsymbol{\lambda}_1), D^j \hat{K}_m(\boldsymbol{\lambda}_2)\right\} \to D^{i+j}\left\{\frac{M(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)}{M(\boldsymbol{\lambda}_1) M(\boldsymbol{\lambda}_2)} - 1\right\},$$

where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are further restricted to the interior of $I/2$ if any of the elements of i or j is greater than zero. Finally, after noticing that on $I/2$:

$$\hat{\boldsymbol{\lambda}} = \hat{K'}^{-1}(\boldsymbol{x}) = \boldsymbol{\lambda} + O(m^{-\frac{1}{2}}),$$

we have that:

$$\begin{aligned}
\tfrac{\hat{p}_m(\boldsymbol{s})}{\hat{p}(\boldsymbol{s})} &= \tfrac{|K''(\boldsymbol{\lambda})|}{|\hat{K}''_m(\hat{\boldsymbol{\lambda}})|} \exp\left[\left\{\hat{K}_m(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}^T \hat{K}'_m(\hat{\boldsymbol{\lambda}})\right\} - \left\{K(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T K'(\boldsymbol{\lambda})\right\}\right] \\
&= \tfrac{|K''(\boldsymbol{\lambda})|}{|K''(\boldsymbol{\lambda})| + O(m^{-\frac{1}{2}})} \exp\left\{O(m^{-1/2})\right\} \\
&= 1 + O(m^{-\frac{1}{2}}),
\end{aligned}$$

by Taylor expansions, which are justified by the differentiability of all the functions involved. See Feuerverger (1989) for more details.

D.2 Proof of Proposition 1

Define

$$w_i = \frac{e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}{\sum_{i=1}^m e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}, \quad \bar{\boldsymbol{s}} = \hat{K}'(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^m w_i \boldsymbol{s}_i}{\sum_{i=1}^m w_i}, \quad \mathrm{i} = 1, \ldots, m, \tag{D.5}$$

and notice that $\hat{K}''(\boldsymbol{\lambda})$ is positive semi-definite

$$\boldsymbol{z}^T \hat{K}''(\boldsymbol{\lambda}) \boldsymbol{z} = \boldsymbol{z}^T \sum_{i=1}^m w_i (\boldsymbol{s}_i - \bar{\boldsymbol{s}})(\boldsymbol{s}_i - \bar{\boldsymbol{s}})^T \boldsymbol{z} = \sum_{i=1}^m w_i \boldsymbol{z}^T (\boldsymbol{s}_i - \bar{\boldsymbol{s}})(\boldsymbol{s}_i - \bar{\boldsymbol{s}})^T \boldsymbol{z} = \sum_{i=1}^m w_i \left\{\boldsymbol{z}^T (\boldsymbol{s}_i - \bar{\boldsymbol{s}})\right\}^2 \geq 0,$$

for all $\boldsymbol{z} \in R^d$ such that $||\boldsymbol{z}|| > 0$. In addition, define $\boldsymbol{q}_i = \boldsymbol{s}_i - \bar{\boldsymbol{s}}$ and assume that

$$r = \operatorname{rank}\left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m\right] = d. \tag{D.6}$$

Then $\hat{K}''(\boldsymbol{\lambda})$ is positive definite and $\hat{K}(\boldsymbol{\lambda})$ is strictly convex. In fact, suppose that there exists a non-zero vector $\boldsymbol{z}$ such that $\boldsymbol{z}^T\hat{K}''(\boldsymbol{\lambda})\boldsymbol{z} = 0$, which implies $\boldsymbol{z}^T\boldsymbol{q}_i = 0$ for $i = 1, \ldots, m$. Given that $\boldsymbol{z}$ can be expressed as a linear combination of $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$, this would imply that

$$\boldsymbol{z}^T\boldsymbol{z} = (b_1\boldsymbol{q}_1 + \cdots + b_m\boldsymbol{q}_m)^T\boldsymbol{z} = 0,$$

which contradicts the fact that $\boldsymbol{z}$ is a non-zero vector. Now, define

$$J \subset \{1, \ldots, m\} \text{ such that } \boldsymbol{\lambda}^T\boldsymbol{s}_i = \alpha > 0 \text{ for all } i \in J, \quad \boldsymbol{\lambda}^T\boldsymbol{s}_i < \alpha \text{ for all } i \notin J,$$

examination of (D.5) shows that

$$\lim_{c\to\infty} w_i = \frac{\lim_{c\to\infty} e^{c(\boldsymbol{\lambda}^T\boldsymbol{s}_i - \boldsymbol{\lambda}^T\boldsymbol{s}_j)}}{\lim_{c\to\infty}\sum_{k=1}^m e^{c(\boldsymbol{\lambda}^T\boldsymbol{s}_k - \boldsymbol{\lambda}^T\boldsymbol{s}_j)}} = \frac{0}{\text{Card}(J)} = 0, \quad \text{for all } i,\, j \text{ such that } j \in J,\, i \notin J,$$

$$\lim_{c\to\infty} w_i = \frac{\lim_{c\to\infty} e^{c(\boldsymbol{\lambda}^T\boldsymbol{s}_i - \boldsymbol{\lambda}^T\boldsymbol{s}_j)}}{\lim_{c\to\infty}\sum_{k=1}^m e^{c(\boldsymbol{\lambda}^T\boldsymbol{s}_k - \boldsymbol{\lambda}^T\boldsymbol{s}_j)}} = \frac{1}{\text{Card}(J)}, \quad \text{for all } i,\, j \text{ such that } i,\, j \in J.$$

Hence

$$\lim_{c\to\infty} \bar{\boldsymbol{s}} = \lim_{c\to\infty} \hat{K}'(c\boldsymbol{\lambda}) = \lim_{c\to\infty} \sum_{i=1}^m w_i\boldsymbol{s}_i = \frac{1}{\text{Card}(J)}\sum_{i\in J}^m \boldsymbol{s}_i,$$

and

$$\lim_{c\to\infty} \boldsymbol{\lambda}^T\boldsymbol{q}_i = \lim_{c\to\infty} \boldsymbol{\lambda}^T(\boldsymbol{s}_i - \bar{\boldsymbol{s}}) = \boldsymbol{\lambda}^T\left\{\boldsymbol{s}_i - \frac{1}{\text{Card}(J)}\sum_{i\in J}\boldsymbol{s}_i\right\} = \alpha - \alpha = 0, \quad \text{for all } i \in J.$$

Finally, we choose $\boldsymbol{z} = \boldsymbol{\lambda}$ and obtain

$$\lim_{c\to\infty} \boldsymbol{\lambda}^T\hat{K}''(c\boldsymbol{\lambda})\boldsymbol{\lambda} = \sum_{i=1}^m \lim_{c\to\infty} w_i \lim_{c\to\infty} \left(\boldsymbol{\lambda}^T\boldsymbol{q}_i\right)^2 = \frac{1}{\text{Card}(J)}\sum_{i\in J}\lim_{c\to\infty}\left(\boldsymbol{\lambda}^T\boldsymbol{q}_i\right)^2 = 0,$$

which implies that $\hat{K}(\boldsymbol{\lambda})$ is not strongly convex.

D.3 Optimality of the cross-validated Extended Empirical Saddlepoint

Let $p(\boldsymbol{s}|\boldsymbol{\theta})$ be the true density of the statistics and $\hat{p}_{SL}(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)$ be the ESA density. Assume that we have a training set of size m, a test set of size n_T and that we have used l simulations to normalize the density estimator. In this section we prove that, as m, n_T and $l \to \infty$, Algorithm 4 consistently selects the value of γ which minimizes the Kullback-Leibler divergence between $\hat{p}_{SL}(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)$ and $p(\boldsymbol{s}|\boldsymbol{\theta})$. When two folds are used, cross-validation (Algorithm 4) selects γ as follows

$$\hat{\gamma} = \underset{\gamma}{\text{argmin}}\left\{-\frac{1}{n_T}\sum_{i=1}^{n_T}\log\hat{p}_{SL}(\boldsymbol{s}_i|\boldsymbol{\theta}, \gamma)\right\} \quad \text{with} \quad \boldsymbol{s}_i \sim p(\boldsymbol{s}|\boldsymbol{\theta}),$$

but the Weak Law of Large Numbers implies that

$$\begin{aligned}\operatorname*{plim}_{m,l,n_T\to\infty} -\tfrac{1}{n_T}\textstyle\sum_{i=1}^{n_T}\log\hat{p}_{SL}(\boldsymbol{s}_i|\boldsymbol{\theta},\gamma) &= -\textstyle\int\log p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\gamma)p(\boldsymbol{s}|\boldsymbol{\theta})ds \\ &\propto \textstyle\int\big\{\log p(\boldsymbol{s}|\boldsymbol{\theta})p(\boldsymbol{s}|\boldsymbol{\theta}) - \log p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\gamma)p(\boldsymbol{s}|\boldsymbol{\theta})\big\}ds \\ &= \textstyle\int\log\frac{p(\boldsymbol{s}|\boldsymbol{\theta})}{p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\gamma)}p(\boldsymbol{s}|\boldsymbol{\theta})ds \\ &= \mathrm{KL}\Big\{p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\gamma), p(\boldsymbol{s}|\boldsymbol{\theta})\Big\}.\end{aligned}$$

Hence $p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\hat{\gamma})$ is the member of the $p_{SL}(\boldsymbol{s}|\boldsymbol{\theta},\gamma)$ family with minimal Kullback-Leibler distance from $p(\boldsymbol{s}|\boldsymbol{\theta})$. This result can easily be extended to k-fold cross-validation ($k > 2$).

D.4 Proof of Theorem 5.4

By the Weak Law of Large Numbers $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n$ converge to $\boldsymbol{\mu}_{\boldsymbol{\theta}}^n$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n$, as $m \to \infty$. Then

$$n^{-\delta}\log p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) \propto -(\boldsymbol{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}^n)^T\left(n^\delta\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n\right)^{-1}(\boldsymbol{s}^0 - \boldsymbol{\mu}_{\boldsymbol{\theta}}^n) - n^{-\delta}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n|.$$

Assumption 5.2 implies that

$$n^{-\delta}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n| = n^{-\delta}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n n^{-\delta}n^\delta| = n^{-\delta}\big(\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n n^\delta| - d\delta\log n\big) = O(n^{-\delta}\log n),$$

so

$$\operatorname*{plim}_{n\to\infty} n^{-\delta}\log p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) \propto -(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}}),$$

$\boldsymbol{\mu}_{\boldsymbol{\theta}_0}$ being the asymptotic mean vector at true parameters $\boldsymbol{\theta}_0$. If Assumption 5.3 holds

$$\operatorname*{argmax}_{\boldsymbol{\theta}}\left\{-(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})\right\} = \boldsymbol{\theta}_0,$$

which implies the consistency of SL under a Gaussian density estimator.

D.5 Proof of Theorem 5.8

Here we indicate with p_G and p_S the synthetic likelihoods based respectively on the Gaussian and on the ESA approximation. Taylor expansions lead to

$$\log\hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) = \log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}) + O\big\{e^{-\hat{\gamma}_{\boldsymbol{\theta}}^n(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)^T\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n\right)^{-1}(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)}\big\}.$$

By multiplying both sides by $n^{-\delta}$, we obtain

$$n^{-\delta}\log\hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) = n^{-\delta}\log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}) + O\big\{n^{-\delta}e^{-\hat{\gamma}_{\boldsymbol{\theta}}^n n^\delta(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)^T\left(n^\delta\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n\right)^{-1}(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)}\big\},$$

and Assumption 5.2 and 5.7, together with the Weak Law of Large Numbers, imply

$$\operatorname*{plim}_{m,l,n\to\infty} n^{-\delta}\log\hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) = \operatorname*{plim}_{n\to\infty}\big\{n^{-\delta}\log p_G(\boldsymbol{s}^0|\boldsymbol{\theta}) + O(n^{-\delta}e^{-\gamma_{\boldsymbol{\theta}}^n n^\delta})\big\} = \operatorname*{plim}_{n\to\infty} n^{-\delta}\log p_G(\boldsymbol{s}^0|\boldsymbol{\theta}).$$

Consistency follows from Theorem 5.4.

D.6 Practical implementation

D.6.1 Saddlepoint version of Algorithm 3

In this section we illustrate how a pointwise synthetic likelihood estimate can be obtained using the new density estimator, rather than a Gaussian density.

Algorithm 5 Estimating $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ using the Extended Empirical Saddlepoint approximation

1: Simulate datasets $\boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_m$ from the model $p(\boldsymbol{Y}|\boldsymbol{\theta})$.
2: Transform each dataset $\boldsymbol{Y}_i$ to a vector of summary statistics $\boldsymbol{S}_i = S(\boldsymbol{Y}_i)$.
3: Calculate sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the simulated statistics.
4: Estimate the synthetic likelihood

$$\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \hat{p}_m(\boldsymbol{s}^0, \gamma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\hat{K}''_m(\hat{\boldsymbol{\lambda}}_m, \gamma, \boldsymbol{s}^0)|^{\frac{1}{2}}} e^{\hat{K}_m(\hat{\boldsymbol{\lambda}}_m, \gamma, \boldsymbol{s}^0) - \hat{\boldsymbol{\lambda}}_m^T \boldsymbol{s}^0},$$

where $\hat{\boldsymbol{\lambda}}_m$ is the solution of the empirical saddlepoint equation

$$\hat{K}'_m(\hat{\boldsymbol{\lambda}}_m, \gamma, \boldsymbol{s}^0) = \boldsymbol{s}^0,$$

while $\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s})$ is given by equation (5.2) in Chapter 5.
5: Normalize $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ by importance sampling

$$\bar{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \frac{\hat{p}_m(\boldsymbol{s}^0, \gamma)}{\hat{z}_m(\gamma)},$$

where

$$\hat{z}_m(\gamma) = \frac{1}{l} \sum_{i=1}^{l} \frac{\hat{p}_m(\boldsymbol{S}_i, \gamma)}{q(\boldsymbol{S}_i)}, \qquad \boldsymbol{S}_i \sim q(\boldsymbol{s}), \quad \text{for } i = 1, \ldots, l.$$

A reasonably efficient importance density $q(\boldsymbol{s})$ is a Gaussian density with mean vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$.

D.6.2 Maximizing the synthetic likelihood

To maximize the synthetic likelihood we have used a special case of the Iterated Filtering procedure, firstly proposed by Ionides et al. (2006). Very briefly, suppose that $\hat{\boldsymbol{\theta}}_k$ is the estimate of the unknown parameters at the k-th step of the optimization routine. This estimate is updated as follows:

1. Simulate N parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from a user-defined density $p(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k)$ such that

$$E(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k) = \hat{\boldsymbol{\theta}}_k, \quad \text{var}(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k) = \sigma_k^2 \boldsymbol{\Sigma} \quad \text{and} \quad E(||\boldsymbol{\theta}_{k+1} - \hat{\boldsymbol{\theta}}_k||^{3/2}) = o(\sigma_k^2), \tag{D.7}$$

where σ_k^2 is a cooling schedule and $\boldsymbol{\Sigma}$ is a covariance matrix.

2. For each $\boldsymbol{\theta}_i$, obtain an estimate $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)$ of the synthetic likelihood, using either the multivariate normal density or the normalized ESA.

3. Update the estimate

$$\hat{\boldsymbol{\theta}}_{k+1} = \frac{\sum_{i=1}^{N} \boldsymbol{\theta}_i \hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)}{\sum_{i=1}^{N} \hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)}.$$

The convergence properties of this procedure have been studied, in the context of Hidden Markov Models, firstly by Ionides et al. (2006) and more in details by Ionides et al. (2011). Doucet et al. (2013) explicitly pointed out that it can be used as a general likelihood optimizer. While those papers considered situations where the likelihood ($p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ in our context) can be evaluated exactly, we have verified empirically that the algorithm works well also when the likelihood is estimated with Monte Carlo error. For both the shifted exponential and the Formind example we used the following cooling schedule

$$\sigma_k^2 = \sigma_0^{2k}, \quad \sigma_0^2 = 0.95.$$

In the shifted exponential example we performed 4 separate runs of the optimizer, using either the normal or the ESA approximation, in both the 10 and 20-dimensional setting.

D.6.3 Formind settings

The summary statistic were obtained using the following constants

$$\alpha_{1,1} = \alpha_{1,3} = \alpha_{2,1} = \alpha_{2,3} = 1.5,\ \alpha_{1,2} = 2, \alpha_{2,2} = 2$$

while ψ_{jk} and σ_{jk} were estimates of mean and standard deviations of C_{jk}, obtained by simulating tree counts at the true parameters. The 24 datasets were simulated from Formind using the same parameter values as in Table 1 in the supplementary material of Hartig et al. (2014). The chosen tree classes correspond to diameters at breast height $d < 0.2m$, $0.2m \leq d < 0.6m$, $d \geq 0.6m$ for pioneer and $d < 0.5m$, $0.5m \leq d < 1.4m$, $d \geq 1.4m$ for late successional trees. To generate the datasets the model was run for 10^5 years, and the final statistics vector was selected. The $m = 10^4$ summary statistics simulated to estimate $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ have been generated by simulating the model for 5.1×10^4 years, where the first 10^3 years of simulation were discarded to avoid the transient, and by storing a vector of statistics every 5 years.

Starting from initial values $\mu_{pio} = 0.03$, $\mu_{suc} = 0.003$, $s_{pio} = 120$ and $s_{suc} = 40$, we ran the optimizations using $N = 24$ and 100 iterations. The estimates reported in Table 1 in the main text were obtained by using the averages of the last 10 iterations of each optimization run as point estimates. The whole experiment took around 10 days on a quad-core Intel i7 3.6 GHz processor.

BIBLIOGRAPHY

Anderson, C. N. K., Hsieh, C., Sandin, S. A., Hewitt, R., Hollowed, A., Beddington, J., May, R. M., and Sugihara, G. (2008). Why fishing magnifies fluctuations in fish abundance. *Nature*, 452(7189):835–839.

Andrieu, C. and Doucet, A. (2003). Online expectation-maximization type algorithms for parameter estimation in general state space models. In *ICASSP (6)*, pages 69–72. Citeseer.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C., Doucet, A., and Tadic, V. B. (2005). On-line parameter estimation in general state-space models. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, pages 332–337. IEEE.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *Annals of Statistics*, 37(2):697–725.

Bartolucci, F. (2007). A penalized version of the empirical likelihood ratio for the population mean. *Statistics & probability letters*, 77(1):104–110.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics*, pages 377–380.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Berliner, L. M. (1992). Statistics, probability and chaos. *Statistical Science*, pages 69–90.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in northwest india: Data analysis via partially observed stochastic differential equation models driven by levy noise. *Journal of the American Statistical Association*, 106(494):440–451.

Blum, M., Nunes, M., Prangle, D., and Sisson, S. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, 28(2):189–208.

Blum, M. G. (2010). Approximate bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491).

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge University Press.

Cai, T. T., Liang, T., and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, 137:161–172.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500.

Cash, R. A., Music, S. I., Libonati, J. P., Craig, J. P., Pierce, N. F., and Hornick, R. B. (1974). Response of man to infection with vibrio cholerae. ii. protection from illness afforded by previous disease and vaccine. *Journal of Infectous Diseases*, 130(4):325–333.

Chan, K.-S. and Tong, H. (2001). *Chaos: a statistical perspective*. Springer Science & Business Media.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 25(4):631–650.

Davidson, R. and MacKinnon, J. G. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, 75(3):417–431.

Dean, T. A., Singh, S. S., Jasra, A., and Peters, G. W. (2011). Parameter estimation for hidden markov models with intractable likelihoods. *arXiv preprint arXiv:1103.5399*.

Desharnais, R. A., Costantino, R., Cushing, J., Henson, S. M., and Dennis, B. (2001). Chaos and population control of insect outbreaks. *Ecology Letters*, 4(3):229–235.

Diggle, P. J. and Gratton, R. J. (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227.

Dislich, C., Günter, S., Homeier, J., Schröder, B., and Huth, A. (2009). Simulating forest dynamics of a tropical montane forest in south ecuador. *Erdkunde*, 63(4):347–364.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208.

Doucet, A., Jacob, P. E., and Rubenthaler, S. (2013). Derivative-free estimation of the score vector and observed information matrix with application to state-space models. *arXiv preprint arXiv:1304.5768*.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704.

Doucet, A., Pitt, M., Deligiannidis, G., and Kohn, R. (2012). Efficient implementation of markov chain monte carlo when using an unbiased likelihood estimator. *arXiv preprint arXiv:1210.1871*.

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2014). Bayesian indirect inference using a parametric auxiliary model. *To appear in Statistical Science*.

Earn, D. J., Rohani, P., and Grenfell, B. T. (1998). Persistence, chaos and synchrony in ecology and epidemiology. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1390):7–10.

Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association*, 81(394):420–430.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ehrlich, E., Jasra, A., and Kantas, N. (2013). Gradient free parameter estimation for hidden markov models with intractable likelihoods. *Methodology and Computing in Applied Probability*, pages 1–35.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608.

Farnsworth, K. D., Thygesen, U. H., Ditlevsen, S., and King, N. J. (2007). How to estimate scavenger fish abundance using baited camera data. *Marine Ecology Progress Series*, 350:223.

Fasiolo, M. and Wood, S. (2014). *An introduction to synlik. R package version 0.1.1.*

Fearnhead, P. and Prangle, D. (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.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika*, 76(3):457–464.

Geweke, J. and Tanizaki, H. (2001). Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational Statistics & Data Analysis*, 37(2):151–170.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Glass, R. I., Becker, S., Huq, M. I., Stoll, B. J., Khan, M., Merson, M. H., Lee, J. V., and Black, R. E. (1982). Endemic cholera in rural bangladesh, 1966–1980. *American. J. Epidemiology.*, 116(6):959–970.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proc. F (Radar and Signal Process.)*, volume 140, pages 107–113. IET.

Grenfell, B. (1992). Chance and chaos in measles dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 383–398.

Grenfell, B., Kleczkowski, A., Gilligan, C., and Bolker, B. (1995). Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Statistical Methods in Medical Research*, 4(2):160–183.

Grenfell, B. T., Bjørnstad, O. N., and Finkenstädt, B. F. (2002). Dynamics of measles epidemics: scaling noise, determinism, and predictability with the tsir model. *Ecological Monograph*, 72(2):185–202.

Gurney, W., Blythe, S., and Nisbet, R. (1980). Nicholson's blowflies revisited. *Nature*, 287:17–21.

Gutmann, M. U. and Corander, J. (2015). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *arXiv preprint arXiv:1501.03291*.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. (2011). Statistical inference for stochastic simulation models–theory and application. *Ecology Letters*, 14(8):816–827.

Hartig, F., Dislich, C., Wiegand, T., and Huth, A. (2014). Technical note: Approximate bayesian parameterization of a process-based tropical forest model. *Biogeosciences*, 11:1261–1272.

Hartlap, J., Simon, P., and Schneider, P. (2007). Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix. *Astronomy & Astrophysics*, 464(1):399–404.

He, D., Ionides, E. L., and King, A. A. (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.

Hickerson, M. J., Carstens, B. C., Cavender-Bares, J., Crandall, K. A., Graham, C. H., Johnson, J. B., Rissler, L., Victoriano, P. F., and Yoder, A. D. (2010). Phylogeographys past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, 54(1):291–301.

Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. (2011). Iterated filtering. *The Annals of Statistics*, 39(3):1776–1802.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443.

Jabot, F., Faure, T., and Dumoulin, N. (2013). Easyabc: performing efficient approximate bayesian computation sampling schemes using r. *Methods in Ecology and Evolution*, 4(7):684–687.

Jiang, W., Turnbull, B., et al. (2004). The indirect method: inference based on intermediate statisticsa synthesis and examples. *Statistical Science*, 19(2):239–263.

Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology*, 86(11):2874–2880.

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

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J. M., and Chopin, N. (2014). On particle methods for parameter estimation in state-space models. *arXiv preprint arXiv:1412.8695*.

Kausrud, K. L., Mysterud, A., Steen, H., Vik, J. O., Østbye, E., Cazelles, B., Framstad, E., Eikeset, A. M., Mysterud, I., Solhøy, T., et al. (2008). Linking climate change to lemming cycles. *Nature*, 456(7218):93–97.

Kendall, B. E., Ellner, S. P., McCauley, E., Wood, S. N., Briggs, C. J., Murdoch, W. W., and Turchin, P. (2005). Population cycles in the pine looper moth: Dynamical tests of mechanistic hypotheses. *Ecological Monograph*, 75(2):259–276.

King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S. P., Ferrari, M. J., Kendall, B. E., Lavine, M., Nguyen, D., Reuman, D. C., Wearing, H., and Wood, S. N. (2014). *pomp: Statistical inference for partially observed Markov processes (R package)*.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454(7206):877–880.

King, R. (2014). Statistical ecology. *Annual Review of Statistics and Its Application*, 1(1):401–426.

Kitagawa, G. (1993). A self-organising state-space model. *Journal American Statistical Association*, 93:1203–1215.

Klaas, M., De Freitas, N., and Doucet, A. (2012). Toward practical n2 monte carlo: The marginal particle filter. *arXiv preprint arXiv:1207.1396*.

Koelle, K., Rodó, X., Pascual, M., Yunus, M., and Mostafa, G. (2005). Refractory periods and climate forcing in cholera dynamics. *Nature*, 436(7051):696–700.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden markov models and extensions. *Ecology*, 93(11):2336–2342.

Lavine, J. S., King, A. A., Andreasen, V., and Bjørnstad, O. N. (2013). Immune boosting explains regime-shifts in prevaccine-era pertussis dynamics. *PloS One*, 8(8):e72086.

Li, T., Bolic, M., and Djuric, P. M. (2015). Resampling methods for particle filtering: Classification, implementation, and strategies. *Signal Processing Magazine, IEEE*, 32(3):70–86.

Lin, L., Liu, K., and Sloan, J. (2000). A noisy monte carlo algorithm. *Physical Review D*, 61(7):074505.

Liu, J. and West, M. (2001). *Combined parameter and state estimation in simulation-based filtering*. Springer.

Malik, S. and Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2):190–209.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467.

Meeds, E. and Welling, M. (2014). Gps-abc: Gaussian process surrogate approximate bayesian computation. *arXiv preprint arXiv:1401.2838*.

Mira, A., Solgi, R., and Imparato, D. (2013). Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662.

Monti, A. C. and Ronchetti, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate m-estimators. *Biometrika*, 80(2):329–338.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9):2436–2445.

Nemeth, C., Fearnhead, P., and Mihaylova, L. (2013). Particle approximations of the score and observed information matrix for parameter estimation in state space models with linear computational cost. *arXiv preprint arXiv:1306.0735*.

Nicholson, A. J. (1954). An outline of the dynamics of animal populations. *Australian Journal of Zoology*, 2(1):9–65.

Nicholson, A. J. (1957). The self-adjustment of populations to change. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 22, pages 153–173. Cold Spring Harbor Laboratory Press.

Niemi, J. and West, M. (2010). Adaptive mixture modeling metropolis methods for bayesian analysis of nonlinear state-space models. *Journal of Computational and Graphical Statistics*, 19(2):260–280.

Nolan, J. P. (2001). Maximum likelihood estimation and diagnostics for stable distributions. In *Lévy processes*, pages 379–400. Springer.

Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Stat. Appl. Genet. Mol. Biol*, 9(1).

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Owen, J., Wilkinson, D. J., and Gillespie, C. S. (2014). Likelihood free inference for markov processes: a comparison. *arXiv preprint arXiv:1410.0524*.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057.

Pan, G. and Zhou, W. (2011). Central limit theorem for hotelling's t 2 statistic under large dimension. *The Annals of Applied Probability*, pages 1860–1910.

Perry, J. N. (2000). *Chaos in real data: the analysis of non-linear dynamics from short ecological time series*, volume 27. Springer.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599.

Polson, N. G., Stroud, J. R., and Müller, P. (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):413–428.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

Rabosky, D. L. (2009). Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Systematic biology*, 58(6):629–640.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796.

Rao, C. R. (2009). *Linear statistical inference and its applications*, volume 22. John Wiley & Sons.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of multivariate analysis*, volume 709. John Wiley & Sons.

Roberts, A. W. and Varberg, D. E. (1973). Convex functions.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the metropolis–hastings algorithm. *Biometrika*, 88(3):603–621.

Ronchetti, E. and Welsh, A. H. (1994). Empirical saddlepoint approximations for multivariate m-estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(2):313–326.

Rubio, F. J. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654.

Ruppert, D., Wand, M. P., Holst, U., and Hösjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273.

Scott, D. W. (2009). *Multivariate density estimation: theory, practice, and visualization*, volume 383. Wiley. com.

Sherlock, C., Thiery, A. H., Roberts, G. O., Rosenthal, J. S., et al. (2014). On the efficiency of pseudo-marginal random walk metropolis algorithms. *The Annals of Statistics*, 43(1):238–275.

Silk, D., Filippi, S., and Stumpf, M. P. (2013). Optimizing threshold-schedules for sequential approximate bayesian computation: applications to molecular systems. *Statistical applications in genetics and molecular biology*, 12(5):603–618.

Smith, A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1):S63–S84.

Smith, W. and Hocking, R. (1972). Algorithm as 53: Wishart variate generator. *Applied Statistics*, pages 341–345.

Spall, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *Automatic Control, IEEE Transactions on*, 45(10):1839–1853.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. Wiley. com.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Turchin, P. and Ellner, S. P. (2000). Living on the edge of chaos: population dynamics of fennoscandian voles. *Ecology*, 81(11):3099–3116.

Turchin, P. and Hanski, I. (1997). An empirically based model for latitudinal gradient in vole population dynamics. *American Naturalist*, 149(5):842–874.

Turchin, P., Wood, S. N., Ellner, S. P., Kendall, B. E., Murdoch, W. W., Fischlin, A., Casas, J., McCauley, E., and Briggs, C. J. (2003). Dynamical effects of plant quality and parasitism on population cycles of larch budmoth. *Ecology*, 84(5):1207–1214.

Wang, S. (1992). General saddlepoint approximations in the bootstrap. *Statistics & probability letters*, 13(1):61–66.

Wilkinson, R. D. (2014). Accelerating abc methods using gaussian processes. *arXiv preprint arXiv:1401.1436*.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. N. (2001). mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Yang, G.-J., Bradshaw, C. J., Whelan, P. I., and Brook, B. W. (2008). Importance of endogenous feedback controlling the long-term abundance of tropical mosquito species. *Population Ecology*, 50(3):293–305.