Statistical inference for epidemic processes in a homogeneous community (Part IV of the book Stochastic Epidemic Models and Inference)

Viet-Chi Tran, Catherine Laredo

Statistical inference for epidemic processes in a homogeneous community

Catherine Larédo (with Viet Chi Tran for Chapter 4)

April 16, 2020

This document is the Part IV of the book *Stochastic Epidemic Models with Inference* edited by Tom Britton and Etienne Pardoux [20]. It is written by Catherine Larédo, with the contribution of Viet Chi Tran for the Chapter 4.

Contents

Introduction

Mathematical modeling of epidemic spread and estimation of key parameters from data provided much insight in the understanding of public health problems related to infectious diseases. These models are naturally parametric models, where the present parameters rule the evolution of the epidemics under study.

Multidimensional continuous-time Markov jump processes $(\mathscr{Z}(t))$ on $\mathbb{Z}^p$ form a usual set-up for modeling epidemics on the basis of compartmental approaches as for instance the *SIR*-like (Susceptible-Infectious-Removed) epidemics (see Part I of these notes and also [2], [36], [84]). However, when facing incomplete epidemic data, inference based on $(\mathscr{Z}(t))$ is not easy to be achieved.

There are different situations where missing data are present. One situation concerns Hidden Markov Models, which are in most cases Markov processes observed with noise. It corresponds for epidemics to the fact that the exact status of all the individuals within a population are not observed, or that detecting the status has some noise (see [23] for instance). Another situation comes from the fact that observations are performed at discrete times. They can also be aggregated (e.g. number of infected per day). A third case, for multidimensional processes, is that some coordinates cannot be observed in practice. While the statistical inference has a longstanding theory for complete data, this is no longer true for many cases that occur in practice. Many methods have been proposed to fill this gap starting from the Expectation-Maximization algorithm ([35], [91]) up to various Bayesian methods ([26], [107]), Monte Carlo methods ([52], [105]), based on particle filtering ([42], [43]), Approximate Bayesian Computation methods ([9], [15], [115], [121]), maximum iterating filtering ([71]), Sequential Monte Carlo or Particle MCMC ([3], [38]), see also the R package POMP ([90]). Nevertheless, these methods do not completely circumvent the issues related to incomplete data. Indeed, as summarized in [19], there are some limitations in practice due to the size of missing data and to the various tuning parameters to be adjusted.

The aim of this part is to provide some tools to estimate the parameters ruling the epidemic dynamics on the basis of available data. We begin with a chapter about inferential methodology for stochastic processes which is not specific to applications to epidemics but is the backbone of the various inference methods detailed in the next chapters of this part.

The methods used to build estimators are linked with the precise nature of the observations, each kind of observations generating a different statistical problem. We detail these facts in the first chapter. We have intentionally omitted in this chapter the additional problem of noisy observations, which often occurs in practice. This is another layer which comes on top. It entails Hidden Markov Models and State space Models (see [23] or [125]) and also the R-package Pomp ([90]).

Chapter 2 is devoted to the statistical inference for Markov chains. Indeed, discrete time Markov chains models are interesting here because many questions that arise for more complex epidemic models can be illustrated in this set-up.

We had rather focus here on parametric inference since epidemic models always include in their dynamics parameters that need to be estimated in order to derive predictions. At the early stage of an outbreak, a good approximation for the epidemic dynamics is to consider that the population of Susceptible is infinite and that Infected individuals evolve according to a branching process (see Part I, Section **??** of these notes). We also present in this chapter some classical statistical results in this domain.

As detailed in Part I, Chapter **??**, epidemics in a close population of size N are naturally modeled by pure jump processes $(\mathscr{Z}^N(t))$. However, inference for such models requires that all the jumps (i.e. times of infection and recovery for the *SIR* model) are observed. Since these data are rarely available in practice, statistical methods

often rely on data augmentation, which allows us to complete the data and add in the analysis all the missing jumps. For moderate to large populations, the complexity increases rapidly, becoming the source of additional problems. Various approaches were developed during the last years to deal with partially observed epidemics. Data augmentation and likelihood-free methods such as the Approximate Bayesian Computation (ABC) opened some of the most promising pathways for improvement (see e.g. [18], [102]). Nevertheless, these methods do not completely circumvent the issues related to incomplete data. As stated also in [19], [28], there are some limitations in practice, due to the size of missing data and to the various tuning parameters to be adjusted (see also [2], [106]).

In this context, it appears that diffusion processes satisfactorily approximating epidemic dynamics can be profitably used for inference of model parameters for epidemic data, due to their analytical power (see e.g. [46], [110]). More precisely, when normalized by N, $(Z^N(t) = N^{-1}\mathscr{Z}^N(t))$ satisfies an ODE as the population size N goes to infinity and moreover, in the first part of these notes, it is proved that the Wasserstein L_1-distance between $(Z^N(t))$ and a multidimensional diffusion process with diffusion coefficient proportional to $1/\sqrt{N}$ is of order $o(N^{-1/2})$ on a finite interval $[0,T]$ (see Part I, Sections **??** and **??**). Hence, in the case of a major outbreak in a large community, epidemic dynamics can be described using multidimensional diffusion processes $(X^N(t))_{t\geq 0}$ with a small diffusion coefficient proportional to $1/\sqrt{N}$. We detail in Chapter 3 the parametric inference for epidemic dynamics described using multidimensional diffusion processes $(X^N(t))_{t\geq 0}$ with a small diffusion coefficient proportional to $1/\sqrt{N}$ based on discrete observations. Since epidemics are usually observed over limited time periods, we consider the parametric inference based on observations of the epidemic dynamics on a fixed interval $[0,T]$.

The last chapter is devoted to the inference for the continuous time *SIR* model. We present several algorithms which address the problem of incomplete data in this set-up: Expectation-Maximization algorithm, Monte Carlo methods and Approximate Bayesian Computation methods. Finally, all the classical statistical results required for this part are detailed in the Appendix.

Chapter 1

Observations and Asymptotic Frameworks

Multidimensional continuous-time Markov jump processes $(\mathscr{Z}(t))$ on $\mathbb{Z}^p$ form a usual set-up for modeling epidemics on the basis of compartmental approaches as for instance the *SIR*-like (Susceptible-Infectious-Removed) epidemics (see Part I of these notes and also [2], [36], [84]). However, when facing incomplete epidemic data, inference based on $(\mathscr{Z}(t))$ is not easy to be achieved.

Assume that a stochastic process $(\mathscr{Z}(t), t \in [0,T])$ models the epidemic dynamics with parameters associated with this process (transition kernels depending on a parameter θ for Markov chains, drift and diffusion coefficients for a diffusion process, infinitesimal generator for a Markov pure jump process). The observed process corresponds to the value θ_0 of this parameter. This value θ_0 is called the true (unknown) parameter value. Our concern here is the estimation of θ_0 from the observations that are available and the study of their properties. The methods used to build estimators are linked with the precise nature of the observations, each kind of observations generating a different statistical problem. We detail these facts in the next sections. We have intentionally omitted in this chapter the additional problem of noisy observations, which often occurs in practice. This is another layer which comes on top. It entails Hidden Markov Models and State space Models (see [23] or [125]) and also the R-package Pomp ([90]).

1.1 Various kinds of observations and asymptotic frameworks

As developed in Part I of these notes, the epidemic dynamics is modeled by a stochastic process $(\mathscr{Z}(t))$ defined on $[0,T]$ with values in $\mathbb{R}^p$, which describes at each time t the number of individuals in each of the p health states (e.g. $p = 3$ for the *SIR* model). Inference for epidemic models is complicated by the fact that collected observations usually do not contain all the information on the whole path of $(\mathscr{Z}(t), 0 \leq t \leq T)$. Moreover, the inference method relies on an asymptotic framework which allows us to control the properties of estimators. We detail here in a general set-up these facts, which are not specific to the inference for epidemic dynamics, but rely on general properties of inference for stochastic processes, this knowledge being useful for applications to epidemics.

1.1.1 Observations

Historically, continuous observation of $(\mathscr{Z}(t), 0 \leq t \leq T)$ was systematically assumed in the literature concerning the statistics of continuous time stochastic processes (see [69], [97], [98]). It is justified by the property that theoretical results can be obtained. However, many various cases can occur in practice. Among them, including the complete case, the more frequent are

Case (a). Continuous observation of $(\mathscr{Z}(t))$ on $[0,T]$.
Case (b). Discrete observations: $(\mathscr{Z}(t_1), \ldots, \mathscr{Z}(t_n))$ with $0 \leq t_1 < t_2 < \cdots < t_n \leq T$.
Case (c). Aggregated observations $(J_0, \ldots, J_{n-1})$ with $J_i = \int_{t_i}^{t_{i+1}} \mathscr{Z}(s)ds$.
Case (d). Model with latent variables: Some coordinates of $(\mathscr{Z}(t), t \in [0,T])$ are unobserved.

Case (a) corresponds to complete data. For the *SIR* epidemics, it means that the times of infection and recovery are observed for each individual in the population. Case (b) corresponds to the fact that observations are made at

successive known times (one observation per day or per week during the epidemic outburst (see [12], [27], [18], [28]). Case (c) occurs in epidemics when the available observations are the number of Infected individuals and Removed per week for instance. Case (d) deals with the fact that, in routinely collected observations of epidemic models, one or several model variables are unobserved (or latent) (see e.g. [23], [42] for general references and [18], [19], [71], [72], [107], [121] for applications to epidemics).

1.1.2 Various asymptotic frameworks

Taking into account an asymptotic framework is necessary to study and compare the properties of different estimators. It is also a preliminary step for the study of non-asymptotic properties. While for i.i.d. observations, the natural asymptotic framework is that the number n of observations goes to infinity, for stochastic processes various approaches are used according to the model properties or to the available observations. Two different situations need to be considered according to the time interval of observation $[0,T]$, where T either goes to infinity or is fixed.

1.1.2.1 Increasing time of observation $[0,T]$ with $T \to \infty$

If $(\mathscr{Z}(t))$ on $[0,T]$ is continuously observed, a general theory is available for ergodic processes and for stationary mixing processes. Inference can also be performed for some special models but does no longer rely on a general theory. This occurs for supercritical branching processes and for the explosive $AR(1)$ process.

Let us consider the case of discrete observations of a continuous time process with regular sampling Δ. The observations are: $(\mathscr{Z}(t_1), \mathscr{Z}(t_2), \ldots, \mathscr{Z}(t_n))$ with $t_i = i\Delta$ and $T = n\Delta$.
Two distinct cases arise from the study of parametric inference for diffusion processes
(1) The sampling interval Δ is fixed ($T = n\Delta$ and $n \to \infty$).
(2) The sampling interval $\Delta = \Delta_n \to 0$ with $T = n\Delta_n \to \infty$ as $n \to \infty$.
Since the likelihood is not explicit and difficult to compute, it raises many theoretical problems. References for the inference in these cases are Kessler [86], [87] followed by many others [88].

In practice, when a sampling interval Δ is present in the data collecting, it might be important to take it explicitly into account. Deciding whether Δ is small or not depends more on the time scale than on its precise value. However this parameter Δ explicitly enters in the estimators, and some estimators with apparently good properties for Δ fixed might explode for small Δ. It corresponds in theory to different rates of convergence for the various coordinates of the unknown parameter θ as $n \to \infty$. This typically occurs for discrete observations of a diffusion process (see Section 1.2).

1.1.2.2 Fixed observation time $[0,T]$

Several asymptotic frameworks are used.

(1) *Discrete observations on [0,T] with $T = n\Delta_n$ fixed*
The sampling interval $\Delta_n \to 0$ while the number of observations n tends to infinity.
For diffusion processes, only parameters in the diffusion coefficient can be estimated (see [49], [74]).

(2) *Observation of k i.i.d. sample paths of $(\mathscr{Z}^i(t), 0 \le t \le T)$, $i = 1, \ldots k$ with $k \to \infty$.*
Observations of $(\mathscr{Z}^i(t))$ can be continuous or discrete.This framework is relevant for panel data which describe for instance the dynamics of several epidemics in different locations. It allows us to include covariates or additional random effects in the model. The assumption is that the number of paths k goes to infinity (see e.g [60]).

(3) *Presence of a "Small parameter" $\varepsilon > 0 : (\mathscr{Z}^\varepsilon(t), 0 \le t \le T)$, and $\varepsilon \to 0$.*
Inference is studied in the set-up of a family of stochastic models $(\mathscr{Z}^\varepsilon(t), 0 \le t \le T)$ depending on a parameter $\varepsilon > 0$. Such a family of processes naturally appears in the theory of "Small perturbations of dynamical systems", where $(X^\varepsilon(t))$ denotes a diffusion process with small diffusion coefficient $\varepsilon\sigma(\cdot)$ (see e.g. [45]). The presence of a small parameter occurs in the study of epidemics in large closed populations of size N, when they are density dependent. The small parameter ε is associated to the population size N by the relation $\varepsilon = 1/\sqrt{N}$ leading to the

family of processes $\mathscr{Z}^\varepsilon(t) = \varepsilon^2 \mathscr{Z}(t)$ (normalization by the population size of the process). From a probability perspective, we refer to Part I, Sections **??** and **??** (see also [40, Chapter 8]). For statistical purposes, we investigate in Chapter 3 of this part the asymptotic framework "$\varepsilon \to 0$" and, for discrete observations, the cases where the sampling interval Δ can be fixed or $\Delta = \Delta_n \to 0$.

(4) *Asymptotics on the initial population number.*
It consists in assuming that one coordinate of $(\mathscr{Z}(t))$ at time 0 satisfies that $\mathscr{Z}^i(0) = M \to \infty$. The parametric inference for the continuous time *SIR* model is performed in this framework (see the results recalled in Section 4.2 or [2]). This is also used for subcritical branching processes where the initial number of ancestors goes to ∞ (see e.g.[60]).

1.1.3 Various estimation methods

As pointed out in the introduction of this part, we are mainly concerned by the problem of parametric inference. There exist several estimation methods.

Maximum Likelihood Estimation
This entails that one can compute the likelihood of the observation. For a continuously observed process, this is generally possible, but for a discrete time observation of a continuous-time process or for other kinds of incomplete observations, it is often intractable. This opens the whole domain of stochastic algorithms which aim at completing the data in order to estimate parameters with Maximum Likelihood methods. In particular, the well-known Expectation-Maximisation algorithm ([35]) and other related algorithms (see e.g. [3], [91], [107]) are based on the likelihood. For regular statistical models, Maximum Likelihood Estimators (MLE) are consistent and efficient (best theoretical variance).

Minimum Contrast Estimation or Estimating Functions
When it is difficult to use the accurate (exact) likelihood, pseudo-likelihoods (contrast functions; approximate likelihoods,..), or pseudo -score functions (approximations of the score function, estimating functions) are often used. When they are well designed, these methods lead to consistent estimators converging at the right rate. They might loose the efficiency property of MLE in regular statistical models (see e.g. [124] for the general theory and [32], [68] for stochastic processes).

Empirical and non-parametric Methods
This comprises all the methods that rely on limit theorems (such as the ergodic theorem) associated with various functionals of the observations. Among these methods, we can refer to Moments methods and Generalized Moment Methods (see e.g. [124] for the general theory and [65] for discrete observation of continuous-time Markov processes).

Algorithmic Methods
Many methods have been developed to perform estimation for incomplete data. It is difficult to be exhaustive. Let us quote [3], [38] for Particle Markov Monte Carlo methods; [10], [15], [17], [115], [121] for Approximate Bayesian Computation; [26], [102] for Bayesian MCMC; [71], [90] for iterated filtering and the R-package POMP. In the last chapter of this part, MCMC and ABC methods are detailed for the *SIR* model.

1.2 An example illustrating the inference in these various situations

Let us investigate here the consequences of these various situations for the statistical inference on a simple stochastic model for describing a population dynamics: the AR(1) model which is a simple model for describing dynamics in discrete time, its continuous time description corresponding to the Ornstein–Uhlenbeck diffusion process. Besides studying a simplified population model, the main interest of this example lies in the property that computations are explicit for the various inference approaches listed in the previous section.

1.2.1 A simple model for population dynamics: AR(1)

The AR(1) model is a classical model for describing population dynamics in discrete time. On $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space, let (ε_i) be a sequence of i.i.d. random variables on $\mathbb{R}$ with distribution $\mathscr{N}(0,1)$. Consider the autoregressive process on $\mathbb{R}$ defined, for $i \geq 0$,

$$X_{i+1} = aX_i + \gamma\varepsilon_{i+1}, \quad X_0 = x_0. \tag{1.2.1}$$

In order to compare this model with its continuous time version, the Ornstein–Uhlenbeck diffusion process, we assume that $a > 0$ and that x_0 is deterministic and known. The observations are $(X_i, i = 1, \ldots, n)$ and the unknown parameters $(a, \gamma) \in (0, +\infty)^2$. The distribution $\mathbb{P}^n_{a,\gamma}$ of the n-tuple $(X_1, \ldots, X_n)$ is easy to compute, since the random variables $(X_i - aX_{i-1}, i = 1, \ldots, n)$ are independent and identically distributed $\mathscr{N}(0, \gamma^2)$. If λ_n denotes the Lebesgue measure on $\mathbb{R}^n$, then

$$\frac{d\mathbb{P}^n_{a,\gamma}}{d\lambda_n}(x_i, i = 1, \ldots, n) = \frac{1}{(\gamma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\gamma^2}\sum_{i=1}^{n}(x_i - ax_{i-1})^2\right).$$

Hence, the loglikelihood function is

$$\log L_n(a, \gamma) = \ell_n(a, \gamma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\gamma^2 - \frac{1}{2\gamma^2}\sum_{i=1}^{n}(X_i - aX_{i-1})^2. \tag{1.2.2}$$

The maximum likelihood estimators are

$$\hat{a}_n = \frac{\sum_{i=1}^n X_{i-1}X_i}{\sum_{i=1}^n X_{i-1}^2}; \quad \hat{\gamma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{a}_n X_{i-1})^2. \tag{1.2.3}$$

The properties of $(\hat{a}_n, \hat{\gamma}_n^2)$ can be studied as $n \to \infty$: $(\hat{a}_n, \hat{\gamma}_n^2)$ is strongly consistent:

$$(\hat{a}_n, \hat{\gamma}_n^2) \to (a, \gamma^2) \text{ a. s. under } \mathbb{P}_{a,\gamma} \text{ as } n \to \infty.$$

The rates of convergence differ according to the probabilistic properties of (X_i).
(1) If $0 < a < 1$, (X_i) is a Harris recurrent Markov chain with stationary distribution
$\mu_{a,\gamma}(dx) = \mathscr{N}(0, \frac{\gamma^2}{1-a^2})$. The estimators $\hat{a}_n, \hat{\gamma}_n^2$ are asymptotically independent and satisfy

$$\begin{pmatrix} \sqrt{n}(\hat{a}_n - a) \\ \sqrt{n}(\hat{\gamma}_n^2 - \gamma^2) \end{pmatrix} \to \mathscr{N}_2\left(0, \begin{pmatrix} 1-a^2 & 0 \\ 0 & 2\gamma^4 \end{pmatrix}\right). \tag{1.2.4}$$

(2) If $a = 1$, (X_i) is a null recurrent random walk and $n(\hat{a}_n - 1)$ converges to a non-Gaussian distribution, while $\hat{\gamma}_n^2$ has the properties of Case (1).
(3) If $a > 1$ and $x_0 = 0$, (X_i) is explosive. One can prove that $a^n(\hat{a}_n - a)$ converges to a random variable $Y = \eta Z$, where η, Z are two independent random variables, $Z \sim \mathscr{N}(0,1)$ and η is an explicit positive random variable. The estimator $\hat{\gamma}_n^2$ keeps the properties of Case (1).

1.2.2 Ornstein–Uhlenbeck diffusion process with increasing observation time

This section is based on Chapter 1 of [48] where all the statistical inference is detailed. It is presented here as a starting point for problems that arise when dealing with epidemic data. In order to investigate the various situations detailed in Section 1.1, let us now consider the continuous time version of the $AR(1)$ population model, the Ornstein–Uhlenbeck diffusion process defined by the stochastic differential equation

$$d\xi_t = \theta\xi_t dt + \sigma dW_t; \; \xi_0 = x_0. \tag{1.2.5}$$

where $(W_t, t \geq 0)$ denotes a standard Brownian motion on $(\Omega, \mathscr{F}, P)$, and x_0 is either deterministic or is a random variable independent of (W_t). Then, $(\xi_t, t \geq 0)$ is a diffusion process on $\mathbb{R}$ with continuous sample paths. This equation can be solved, setting $Y_t = e^{-\theta t}\xi_t$, so that

$$\xi_t = x_0 e^{\theta t} + e^{\theta t}\int_0^t e^{-\theta s} dW_s. \tag{1.2.6}$$

Let us first consider the case where (ξ_t) is observed with regular sampling intervals Δ. The observations $(\xi_{t_i}; i = 1, \ldots, n)$ with $t_i = i\Delta$ satisfy

$$\xi_{t_{i+1}} = e^{\theta\Delta}\xi_{t_i} + \sigma e^{\theta(i+1)\Delta}\int_{i\Delta}^{(i+1)\Delta} e^{-\theta s} dW_s. \tag{1.2.7}$$

Hence, $(\xi_{t_{i+1}} - e^{\theta\Delta}\xi_{t_i})$ is independent of $\mathscr{F}_{t_i}$, where $\mathscr{F}_t = \sigma(\xi_0, W_s, s \leq t)$ and the sequence $(\xi_{t_i}, i \geq 0)$ is the autoregressive model $AR(1)$ defined in (1.2.1) setting

$$X_i = \xi_{t_i}, \quad a = e^{\theta\Delta}, \quad \gamma^2 = \frac{\sigma^2}{2\theta}(e^{2\theta\Delta} - 1), \tag{1.2.8}$$

since the random variables $((\sigma e^{\theta(i+1)\Delta}\int_{i\Delta}^{(i+1)\Delta} e^{-\theta s} dW_s), 1 \leq i \leq n)$ are independent Gaussian $\mathscr{N}(0, \gamma^2)$.

Cases (1), (2), (3) of the $AR(1)$ are respectively $\{\theta < 0\}$, $\{\theta = 0\}$ and $\{\theta > 0\}$.

Case (a) Continuous observation on $[0, T]$.

Let us first start with the parametric inference associated with the complete observation of (ξ_t) on $[0, T]$. The space of observations is $(C_T, \mathscr{C}_T)$, the space of continuous functions from $[0, T]$ into $\mathbb{R}$ and $\mathscr{C}_T$ is the Borel σ-algebra. associated with the topology of uniform convergence on $[0, T]$. Let $\mathbb{P}_{\theta,\sigma^2}$ denote the probability distribution on $(C_T, \mathscr{C}_T)$ of the observation $(\xi_t, 0 \leq t \leq T)$ satisfying (1.2.5) . It is well known that if $\sigma^2 \neq \tau^2$, the distributions $\mathbb{P}_{\theta,\sigma^2}$ and $\mathbb{P}_{\theta,\tau^2}$ are singular on $(C_T, \mathscr{C}_T)$ (see e.g. [97]). Indeed, the quadratic variations of (ξ_t) satisfy, as $\Delta_n = t_i - t_{i-1} \to 0$,

$$\sum_{i=1}^{n}(\xi_{t_i} - \xi_{t_{i-1}})^2 \to \sigma^2 T \text{ in } \mathbb{P}_{\theta,\sigma^2}\text{-probability.}$$

Therefore, the set $A = \{\omega, \sum_{i=1}^{n}(\xi_{t_i} - \xi_{t_{i-1}})^2 \to \sigma^2 T\}$ satisfies $\mathbb{P}_{\theta,\sigma^2}(A) = 1$ and $\mathbb{P}_{\theta,\tau^2}(A) = 0$ for $\tau^2 \neq .\sigma^2$.
A statistical consequence is that the diffusion coefficient is identified when (ξ_t) is continuously observed.

We assume that σ is fixed and known and omit it in this section. Let $\mathbb{P}_{0,\sigma^2} = \mathbb{P}_0$ the distribution associated with $\theta = 0$ (i.e $d\xi_t = \sigma dW_t$). The Girsanov formula gives an expression of the likelihood function on $[0, T]$,

$$L_T(\theta) = \frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}(\xi_t, 0 \leq t \leq T) = \exp\left(\frac{\theta}{\sigma^2}\int_0^T \xi_t \, d\xi_t - \frac{\theta^2}{2\sigma^2}\int_0^T \xi_t^2 dt\right). \tag{1.2.9}$$

Substituting (ξ_t) by its expression in (1.2.5), the MLE is

$$\hat{\theta}_T = \frac{\int_0^T \xi_t d\xi_t}{\int_0^T \xi_t^2 dt} = \theta + \sigma\frac{\int_0^T \xi_t dW_t}{\int_0^T \xi_t^2 dt}. \tag{1.2.10}$$

The estimator $\hat{\theta}_T$ defined in (1.2.10) reads as

$$\hat{\theta}_T = \theta + \frac{M_T}{\langle M\rangle_T} \text{ with } M_t = \frac{1}{\sigma}\int_0^t \xi_s dW_s. \tag{1.2.11}$$

where (M_t) is a $(\mathscr{F}_t)$-martingale in L^2 with angle bracket $\langle M\rangle_t$ (i.e. the process such that $(M_t^2 - \langle M\rangle_t)$ is a martingale). Noting that $\langle M\rangle_T \to \infty$ as $T \to \infty$, the law of large numbers yields that $\frac{M_T}{\langle M\rangle_T} \to 0$. Hence the MLE defined

by (1.2.10) is consistent. As for the $AR(1)$- model, the rate of convergence of $\hat{\theta}_T$ to θ depends on the properties of (M_t). Three different cases can be listed as $T \to \infty$:
(1) $\{\theta < 0\}$: (ξ_t) is a positive recurrent process with stationary distribution $\mathcal{N}(0, \frac{\sigma^2}{2|\theta|})$ and $\sqrt{T}(\hat{\theta}_T - \theta) \to_{\mathcal{L}} \mathcal{N}(0, 2|\theta|)$.
(2) $\{\theta = 0\}$: (ξ_t) is a null recurrent diffusion; $T\hat{\theta}_T$ converges to a fixed distribution.
(3) $\{\theta > 0\}$: (ξ_t) is a transient diffusion; $e^{\theta T}(\hat{\theta}_T - \theta)$ converges in distribution to $Y = \eta\, Z$, where η, Z are two independent random variables, $Z \sim \mathcal{N}(0,1)$ and η is an explicit a positive random variable.

Case (b)-1 Discrete observations with sampling interval Δ fixed.
Let $t_i = i\Delta, T = n\Delta$ and assume that the number of observations $n \to \infty$.
Using (1.2.8), $(X_i = \xi_{t_i})$ is an AR(1) with $a = e^{\theta\Delta}, \gamma^2 = \sigma^2 v(\theta)$ with $v(\theta) = \frac{1}{2\theta}(e^{2\theta\Delta} - 1)$.

Let $\phi_\Delta : (0,+\infty)^2 \to \mathbb{R} \times (0,+\infty)$

$$\phi_\Delta \,:\, m = \begin{pmatrix} a \\ \gamma^2 \end{pmatrix} \to \begin{pmatrix} \theta = \frac{\log a}{\Delta} \\ \sigma^2 = \frac{a^2-1}{2\log a}\Delta\gamma^2 \end{pmatrix}.$$

This is a C^1-diffeomorphism and the MLE for θ and σ^2 can be deduced from $(\hat{a}_n, \hat{\gamma}^2_n)$ obtained in Section 1.2.1. This yields

$$\hat{\theta}_n = \frac{1}{\Delta}\log\left(\frac{\sum_{i=1}^n X_{i-1}X_i}{\sum_{i=1}^n X_{i-1}^2}\right); \quad \hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n (X_i - \exp(\hat{\theta}_n\Delta)\, X_{i-1})^2.$$

These two estimators inherit the asymptotic properties of the maximum likelihood estimators $(\hat{a}_n, \hat{\gamma}^2_n)$ obtained in Subsection 1.2.1, their asymptotic variance is obtained using Theorem A.1.1 stated in the Appendix, Section A.1.2 (see also [124], Theorem 3.1). Therefore, $(\hat{\theta}_n, \hat{\sigma}^2_n)$ is consistent and, using that $a_n(\hat{m}_n - m)$ converges to a random variable Y yields

$$a_n \begin{pmatrix} \hat{\theta}_n - \theta \\ \hat{\sigma}^2_n - \sigma^2 \end{pmatrix} \to_{\mathcal{L}} \nabla_x \phi_\Delta(m) Y, \tag{1.2.12}$$

where a_n is respectively for Cases (1), (2), (3) the matrix

$$\begin{pmatrix} \sqrt{n} & 0 \\ 0 & \sqrt{n} \end{pmatrix}, \quad \begin{pmatrix} n & 0 \\ 0 & \sqrt{n} \end{pmatrix}, \quad \begin{pmatrix} e^{n\Delta\theta} & 0 \\ 0 & \sqrt{n} \end{pmatrix}.$$

In particular, for Case (1) where $Y \sim \mathcal{N}_2(0, \Sigma)$, the limit distribution $\mathcal{N}_2(0, \nabla_x\phi_\Delta(m)\Sigma(\nabla_x\phi(m))^*)$ where Σ is the matrix obtained in (1.2.4).

Looking precisely at the theoretical asymptotic variance of $\hat{\theta}_n$ obtained in (1.2.12), we can observe that, for small Δ, this variance is $\frac{2|\theta|}{\Delta}$ and therefore explodes. It corresponds to the property that $\sqrt{n}$ is not the right rate of convergence of θ for small Δ.

Case (b)-2 Discrete observations with sampling interval $\Delta = \Delta_n \to 0$
We just detail Case (1), which corresponds to the ergodic Ornstein–Uhlenbeck process, first studied in [86]. Under the condition $n\Delta_n^2 \to 0$, the estimators $\hat{\theta}_n, \hat{\sigma}^2_n$ are consistent and converge at different rates under $\mathbb{P}_\theta$,

$$\begin{pmatrix} \sqrt{n\Delta_n}(\hat{\theta}_n - \theta) \\ \sqrt{n}(\hat{\sigma}^2_n - \sigma^2) \end{pmatrix} \overset{\mathcal{L}}{\to} \mathcal{N}_2\left(0, \begin{pmatrix} 2|\theta| & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right). \tag{1.2.13}$$

Case (c)-1 Aggregated observations on intervals $[i\Delta, (i+1)\Delta]$ with Δ fixed.
Assume now that the available observations are aggregated data on successive intervals, (J_i) with

$$J_i = \int_{t_i}^{t_{i+1}} \xi_s\, ds. \tag{1.2.14}$$

The inference problem has first been studied by [56], [55] for an ergodic stationary diffusion process. It entails that $\theta < 0$ and that X_0 is random, independent of $(W_t, t \geq 0)$, distributed according to the stationary distribution of (ξ_t), $\mathcal{N}(0, \frac{\sigma^2}{2|\theta|})$.

The process $(J_i)_{i\geq 0}$ is a non-Markovian strictly stationary centered Gaussian process. Using (1.2.6) and (1.2.14), J_i and J_{i+1} are linked by the relation

$$J_{i+1} - e^{\theta\Delta} J_i = \frac{\sigma}{\theta} \int_{i\Delta}^{(i+1)\Delta} (e^{\theta\Delta} - e^{\theta((i+1)\Delta - s)}) dW_s + \frac{\sigma}{\theta} \int_{(i+1)\Delta}^{(i+2)\Delta} (e^{\theta((i+2)\Delta - s)} - 1) dW_s. \tag{1.2.15}$$

Hence, for all $i \geq 1$, $(J_{i+1} - e^{\theta\Delta} J_i)$ is independent of $(J_0, \ldots, J_{i-1})$ and (J_i) possesses the structure of an ARMA(1,1) process, for which the statistical inference is derived with other tools. Indeed,

$$\mathrm{Var}(J_i) = \sigma^2 r_0(\theta) \quad ; \quad \mathrm{Cov}(J_i, J_j) = \sigma^2 r_{i-j}(\theta) \quad \text{with}$$
$$r_0(\theta) = \frac{1}{\theta^2}\left(\Delta + \frac{1 - e^{\theta\Delta}}{\theta}\right) \quad ; \quad r_k(\theta) = -\frac{1}{2\theta^3} e^{-\theta\Delta} (e^{\theta\Delta} - 1)^2 \, e^{\theta\Delta|k|} \text{ if } k \neq 0.$$

Its spectral density has also an explicit expression, $f_{\theta,\sigma^2}(\lambda) = \sigma^2 f_\theta(\lambda)$.

The likelihood function is known theoretically but its exact expression is intractable. Instead of the exact likelihood, a well-known method to derive estimators is to use the Whittle contrast $U_n(\theta, \sigma^2)$ which provides efficient estimators. It is based on the periodogram: if j denotes now the complex number $j^2 = -1$,

$$U_n(\theta, \sigma^2) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log f_{\theta,\sigma^2}(\lambda) + \frac{I_n(\lambda)}{f_{\theta,\sigma^2}(\lambda)} \right) d\lambda, \text{ with } I_n(\lambda) = \frac{1}{n} |\sum_{k=0}^{n-1} J_k e^{-jk\lambda}|^2.$$

The estimators are then defined as any solution of $U_n(\tilde{\theta}_n, \tilde{\sigma}_n^2) = \inf_{\theta,\sigma^2} U_n(\theta, \sigma^2)$. This yields consistent and asymptotically Gaussian estimators at rate $\sqrt{n}$.

Case (c)-2 Aggregated observations on intervals $[i\Delta, (i+1)\Delta]$ with $\Delta = \Delta_n \to 0$.

Let us now consider the case of $\Delta = \Delta_n \to 0, T = n\Delta_n \to \infty$ as $n \to \infty$. Let $J_{i,n} = \int_{i\Delta_n}^{(i+1)\Delta_n} \xi_s ds$. Assume that $\theta < 0$. The diffusion is positive recurrent with stationary measure $\mu_{\theta,\sigma^2}(dx) \sim \mathcal{N}(0, \frac{\sigma^2}{2|\theta|})$. The following two convergences hold in probability (see [56]).

$$\frac{1}{n} \sum_{i=0}^{n-1} (\Delta_n^{-1} J_{i+1,n} - \Delta_n^{-1} J_{i,n})^2 \quad \to \quad \frac{2}{3}\sigma^2, \text{ while}$$
$$\frac{1}{n} \sum_{i=0}^{n-1} (\xi_{(i+1)\Delta_n} - \xi_{i\Delta_n})^2 \quad \to \quad \sigma^2.$$

Hence, for small Δ_n, the heuristics $\frac{1}{\Delta_n} J_{i,n} \sim \xi_{i\Delta_n}$ is too rough and does not yield good statistical results. The two processes corresponding to these two kinds of observations are structurally distinct: $(\xi_{i\Delta_n})$ is an AR(1) process while $(\frac{1}{\Delta_n} J_{i,n})$ is ARMA(1,1). Ignoring this can lead to biased estimators.

1.2.3 Ornstein–Uhlenbeck diffusion with fixed observation time

Case (a) Continuous observation on $[0, T]$

As in Section 1.2.2 Case **(a)**, the parameter σ^2 is identified from the continuous observation of (ξ_t). Therefore we assume that σ^2 is known. The expression for the likelihood (1.2.9) holds. We get that, without additional assumptions, as for instance the presence of a small parameter ε, the MLE given in (1.2.10) $\hat{\theta}_T$ has a fixed distribution. On a fixed time interval, parameters in the drift term of a diffusion cannot be consistently estimated.

Case (b)-1 Discrete observations with fixed sampling Δ

The number of observations n is fixed. Without additional assumptions, neither θ nor σ^2 can be consistently estimated.

Case (b)-2 Discrete observations with sampling $\Delta_n \to 0$

Let $\Delta = \Delta_n = T/n \to 0$ as $n \to \infty$. Equation (1.2.7) holds and (1.2.2) is the likelihood. The maximum likelihood estimator $\hat{\theta}_n$ satisfies

$$\hat{\theta}_n = \frac{1}{\Delta_n} \log\left(1 + \Delta_n \frac{\sum_{i=1}^n \xi_{t_{i-1}}(\xi_{t_i} - \xi_{t_{i-1}})}{\Delta_n \sum_{i=1}^n \xi_{t_{i-1}}^2}\right). \tag{1.2.16}$$

Since $t_i = i\frac{T}{n}$, using the property of stochastic integrals and the Lebesgue integral yields that, under $\mathbb{P}_\theta$,

$$\sum_{i=1}^n \xi_{t_{i-1}}(\xi_{t_i} - \xi_{t_{i-1}}) \to \int_0^T \xi_s d\xi_s \text{ in probability}; \quad \sum_{i=1}^n \Delta_n\, \xi_{t_{i-1}}^2 \to \int_0^T \xi_s^2 ds \text{ a.s.}$$

Therefore, as $n \to \infty$, $\hat{\theta}_n$ converges to the random variable $\theta_T = \frac{\int_0^T \xi_s d\xi_s}{\int_0^T \xi_s^2 ds}$. Hence $\hat{\theta}_n$ is not consistent. Note that θ_T is precisely the MLE for θ obtained for continuous observation, which possesses good properties only if $T \to \infty$. The story is different for the estimation of σ^2. The normalized quadratic variations of (ξ_t) is a consistent estimator of σ^2 and $\sum(\xi_{t_i} - \xi_{t_{i-1}})^2 \to \sigma^2 T$ in probability. Moreover,

$$\tilde{\sigma}^2 = \frac{1}{T}\sum_{i=1}^n (\xi_{t_i} - \xi_{t_{i-1}})^2 \text{ satisfies that } \sqrt{n}(\tilde{\sigma}^2 - \sigma^2) \to \mathcal{N}(0, 2\sigma^4). \tag{1.2.17}$$

Note that this result holds whatever the value of θ.

Case (c)-1 Aggregated observations on intervals $[i\Delta, (i+1)\Delta]$ with Δ fixed

As in Case (b)-1, θ and σ^2 cannot be consistently estimated.

Case (c)-2 Aggregated observations on intervals $[i\Delta, (i+1)\Delta]$ with $\Delta = \Delta_n \to 0$

This has been studied in [56]. Then, as $\Delta_n \to 0$, in probability,

$$\sum_{i=0}^{n-1} (\Delta_n^{-1} J_{i+1,n} - \Delta_n^{-1} J_{i,n})^2 \to \frac{2}{3}\sigma^2 T \text{ while } \sum_{i=0}^{n-1} (\xi_{(i+1)\Delta_n} - \xi_{i\Delta_n})^2 \to \sigma^2 T.$$

Here again, the heuristics $\frac{1}{\Delta_n} J_{i,n} \sim \xi_{i\Delta_n}$ is too rough and does not yield good statistical results.

1.2.4 Ornstein–Uhlenbeck diffusion with small diffusion coefficient

This asymptotic framework is "$\varepsilon \to 0$". It naturally occurs for diffusion approximations of epidemic processes. The equation under study is now

$$d\xi_t = \theta \xi_t dt + \varepsilon\sigma dW_t \;\; \xi_0 = x_0. \tag{1.2.18}$$

We detail the results for fixed observation time $[0, T]$.

Case (a) Continuous observation on $[0, T]$

As before, we assume that σ^2 is known and omit it. Let $\mathbb{P}_\theta^\varepsilon$ the distribution on $(C_T, \mathscr{C}_T)$ of (ξ_t) satisfying (1.2.18). The likelihood is now

$$L_{T,\varepsilon}(\theta) = \frac{d\mathbb{P}_\theta^\varepsilon}{dP_0^\varepsilon}(\xi_s, 0 \le s \le T) = \exp\left(\frac{\theta}{\varepsilon^2\sigma^2}\int_0^T \xi_s\, d\xi_s - \frac{\theta^2}{2\varepsilon^2\sigma^2}\int_0^T \xi_s^2 ds\right). \tag{1.2.19}$$

$$\hat{\theta}_{T,\varepsilon} = \theta + \varepsilon\sigma \frac{\int_0^T \xi_t dW_t}{\int_0^T \xi_t^2 dt}. \tag{1.2.20}$$

Therefore $\hat{\theta}_{T,\varepsilon} \to \theta$ in probability under P_θ^ε as $\varepsilon \to 0$. Moreover, using results of [92]),

$$\varepsilon^{-1}(\hat{\theta}_{T,\varepsilon} - \theta) \to_{\mathscr{L}} \mathscr{N}(0, \tau^2), \quad \text{with } \tau^2 = \frac{2\theta\sigma^2}{x_0^2(e^{2\theta T} - 1)}.$$

Case (b)-1 Discrete observations with fixed sampling interval Δ

If Δ is fixed, only θ can be consistently estimated (see [61]). This is detailed in Chapter 3, Section 3.5. Setting $a = e^{\theta\Delta}$, and $X_i = \xi_{i\Delta}$, then, using (1.2.1),

$$X_i = aX_{i-1} + \varepsilon\gamma\eta_i, \quad \text{where} \quad \gamma^2 = \frac{e^{2\theta\Delta} - 1}{2\theta}\sigma^2,$$

and (η_i) i.i.d. $\mathscr{N}(0,1)$ random variables. Using (1.2.2) and (1.2.3) yields

$$\hat{a}_{\varepsilon,\Delta} = a + \varepsilon\gamma\frac{\sum_{i=1}^n X_{i-1}\eta_i}{\sum_{i=1}^n X_{i-1}^2}.$$

Therefore, as $\varepsilon \to 0$, $\hat{a}_{\varepsilon,\Delta}$ is consistent and

$$\varepsilon^{-1}(\hat{a}_{\varepsilon,\Delta} - a) \to \mathscr{N}(0, V_\Delta), \quad \text{with } V_\Delta = \gamma^2\frac{e^{2\theta\Delta} - 1}{x_0^2(e^{2\theta T} - 1)} = \sigma^2\frac{(e^{2\theta\Delta} - 1)^2}{2x_0^2\theta(e^{2\theta T} - 1)}.$$

Note that for small Δ, $V_\Delta \sim \frac{\Delta}{x_0^2(e^{2\theta T} - 1)}\sigma^2$.

Case (b)-2 Discrete observations with sampling $\Delta = \Delta_n \to 0$

This was first studied in [57], [119] and is detailed in Chapter 3. Let $T = n\Delta_n$ (the number of observations $n \to \infty$ as $\Delta_n \to 0$). Both θ and σ can be estimated from discrete observations. One can prove that they converge at different rates: under $\mathbb{P}_\theta$ as $\varepsilon \to 0, n \to \infty$,

$$\begin{pmatrix} \varepsilon^{-1}(\hat{\theta}_{\varepsilon,n} - \theta) \\ \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \end{pmatrix} \to \mathscr{N}_2\left(0, \begin{pmatrix} \frac{2\theta\sigma^2}{x_0^2(e^{2\theta T} - 1)} & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right). \tag{1.2.21}$$

1.2.5 Conclusions

This detailed example based on the Ornstein–Uhlenbeck diffusion studied under various asymptotic frameworks and various kinds of observations shows that, before estimating parameters ruling the process under study, one has to carefully consider how the available observations are obtained from the process and to study their properties. Some approximations are relevant and keep good statistical properties, while other ones lead to estimators which are not even consistent.

Chapter 2

Inference for Markov Chain Epidemic Models

In order to present an overview of the statistical problems, we first detail the statistical inference for Markov chains. Indeed, discrete time Markov chains models are interesting here because many questions that can arise for more complex models can be illustrated in this set-up. Moreover, continuous-time stochastic models are often observed in practice at discrete times, which might sum up to a Markov chain model. Therefore, this point of view allows us to illustrate some classical statistical methods for stochastic models used in epidemics. We have rather focus here on parametric inference since epidemic models always include in their dynamics parameters that need to be estimated in order to derive predictions. A recap on parametric inference for Markov chains is given in the Appendix, Section A.2.1, together with some notations and basic definitions. We apply in this chapter these results on some classical stochastic models used in epidemics (see Part I, Chapter **??** and also [2], [36]).

2.1 Markov chains with countable state space

Markov chain models occur when assuming that a latent period of fixed length follows the receipt of infection by any susceptible. According to the epidemic model, the state space of the Markov chain can be finite if the epidemics takes place in a fixed finite population, countable (birth and death processes, branching processes, open Markov Models detailed in Part I, Chapter **??** of these notes), or continuous (see e.g. the simple AR(1) dynamic model).

Let us first consider a Markov chain (X_n) with finite state space $E = \{0, \dots, N\}$ and transition matrix $(Q(i,j), i,j \in E)$. Assume that $X_0 = x_0$ is deterministic and known. Our aim is to estimate the transition matrix Q, which corresponds to $q = N(N+1)$ parameters since, for all $i \in E$, $\sum_{j=0}^{N} Q(i,j) = 1$.
Following the definitions recalled in Section A.2 in the Appendix, denote by $\mathbb{P}_Q$ the distribution on $(E^{\mathbb{N}}, \mathscr{E}^{\mathbb{N}})$ of (X_n) and $\mathscr{F}_n = \sigma(X_0, \dots, X_n)$. Let $\mu_n = \otimes_{k=1}^{n} \nu_k$ with $\nu_k(\cdot)$ the measure on E such that $\nu_k(i) = 1$ for $i \in E$.

For A a subset of E, let $\delta_A(\cdot)$ denote the Dirac function: $\delta_A(x) = 1$ if $x \in A$, $\delta_A(x) = 0$ if $x \notin A$. Define

$$N_n^{ij} = \sum_{k=1}^{n} \delta_{\{i,j\}}(X_{k-1}, X_k); \quad N_n^{i\cdot} = \sum_{k=1}^{n} \delta_{\{i\}}(X_{k-1}). \tag{2.1.1}$$

Using (2.1.1), the likelihood and the loglikelihood read as

$$L_n(Q) = \frac{d\mathbb{P}_Q}{d\mu_n}(X_k, k = 1, \dots, n) = \prod_{k=1}^{n} Q(X_{k-1}, X_k) = \prod_{i,j \in E} Q(i,j)^{N_{ij}^n}, \tag{2.1.2}$$

$$\ell_n(Q) = \sum_{i,j \in E} N_n^{ij} \log Q(i,j). \tag{2.1.3}$$

The computation of the Maximum Likelihood Estimator, $(\hat{Q}_n(i,j,),i,j\in E)$, corresponds to the maximization of $\ell_n(Q)$ under the $(N+1)$ constraints $\{\sum_{j=0}^N Q(i,j) \ -1=0\}$. This yields that

$$\hat{Q}_n(i,j)=\frac{N_n^{ij}}{N_n^{i.}}. \tag{2.1.4}$$

Since the random variables $(N_n^{ij}, i\neq j)$ are equal to the number of transitions from i to j up to time n and $N_n^{i.}$ is the time spent in state i up to time n, the estimators $\hat{Q}_n(i,j)$ are equal to the empirical estimates of the transitions.

To study the properties of the MLE, we assume

(H1) The Markov chain (X_n) with transition matrix Q is positive recurrent aperiodic on E.

Denote by $\lambda_Q(\cdot)$ the stationary distribution of (X_n). Then, the following holds.

Proposition 2.1.1. *Under (H1), the MLE $(\hat{Q}_n(i,j),i,j\in E)$ is strongly consistent and, under $\mathbb{P}_Q$,*

$$\sqrt{n}\left(\hat{Q}_n(i,j)-Q(i,j)\right)_{0\leq i\leq N,0\leq j\leq N-1}\to_{\mathscr{L}}\mathscr{N}_q(0,\Sigma) \text{ with } q=N(N+1),$$

$$\Sigma_{ij,ij}=\frac{Q(i,j)(1-Q(i,j))}{\lambda_Q(i)};\quad \Sigma_{ij,ij'}=-\frac{Q(i,j)Q(i,j')}{\lambda_Q(i)};\quad \Sigma_{ij,i'j'}=0 \text{ if } i'\neq i.$$

Proof. Under (H1), successive applications of the ergodic theorem yield that, almost surely under $\mathbb{P}_Q$, $\frac{1}{n}N_n^{ij}\to\lambda_Q(i)Q(i,j)$, $\frac{1}{n}N_n^{i.}\to\lambda_Q(i)$ so that $\hat{Q}_n(i,j)\to Q(i,j)$.
Let us study $(\hat{Q}_n(i,j)-Q(i,j))$. For $0\leq i\leq N, 0\leq j\leq N-1$, define

$$Y_k^{ij}=\left(\delta_{\{j\}}(X_k)-Q(i,j)\right)\delta_{\{i\}}(X_{k-1}),\quad M_n^{ij}=\sum_{k=1}^n Y_k^{ij}. \tag{2.1.5}$$

Then

$$\hat{Q}_n(i,j)-Q(i,j)=\frac{N_n^{ij}-Q(i,j)N_n^{i.}}{N_n^{i.}}=\frac{M_n^{ij}}{N_n^{i.}}=\frac{\sum_{k=1}^n Y_k^{ij}}{N_n^{i.}}. \tag{2.1.6}$$

Clearly, $E_Q(Y_k^{ij}|\mathscr{F}_{k-1})=0$ and (M_n^{ij}) is a centered $\mathscr{F}_n$-martingale with values in $\mathbb{R}^q$. Its angle bracket is the random matrix $\langle M\rangle_n$ with indices $(ij),(i'j')$

$$\langle M\rangle_n^{ij,i'j'}=\sum_{k=1}^n E_Q(Y_k^{ij}Y_i^{i'j'}|\mathscr{F}_{k-1}).$$

Straightforward computations yield that

$$\begin{aligned}
&E_Q(Y_k^{ij}Y_k^{ij}|\mathscr{F}_{k-1}))=Q(i,j)(1-Q(i,j))\delta_{\{i\}}(X_{k-1}),\\
&E_Q(Y_k^{ij}Y_k^{ij'}|\mathscr{F}_{k-1}))=-Q(i,j)Q(i,j')\delta_{\{i\}}(X_{k-1}) \text{ if } j'\neq j \text{ and}\\
&E_Q(Y_k^{ij}Y_i^{i'j'}|\mathscr{F}_{k-1})=0 \text{ if } i'\neq i.
\end{aligned}$$

Define the q-dimensional matrix J_Q by

$$\begin{aligned}
&J_Q(ij,ij)=Q(i,j)(1-Q(i,j))\lambda_Q(i),\\
&J_Q(ij,ij')=-Q(i,j)Q(i,j')\lambda_Q(i) \text{ for } j'\neq j \text{ and}\\
&J_Q(ij,i'j')=0 \text{ if } i'\neq i.
\end{aligned}$$

Then, the ergodic theorem yields that $\frac{1}{n}\langle M\rangle_n^{ij,i'j'}\to J_Q(ij,i'j')$ a.s. under $\mathbb{P}_Q$.
Applying the Central Limit Theorem for multidimensional martingales (see Appendix, Section A.4.2) yields that, under $\mathbb{P}_Q$, $\frac{1}{\sqrt{n}}M_n\to\mathscr{N}(0,J_Q)$ in distribution. Finally, using that $\frac{1}{n}N_n^{i.}\to\lambda_Q(i)$ a.s., an application of Slutsky's lemma to (2.1.6) achieves the proof of Proposition 2.1.1. □

2.1.1 Greenwood model

This is a basic model which was introduced by Greenwood [59] to study measles epidemics in United Kingdom. It is an *SIR* epidemic in a finite population of size N. The latent period is fixed and equal 1 with infectiousness confined to a single time point. At the moment of infectiousness of any given infective, the chance of contact with any specified susceptible, sufficient or adequate to transmit the infection is $p = 1-q$. Infected individuals are removed from the infection chain. At time 0, assume that the number of Susceptible S_0 and Infected I_0 verify $S_0 + I_0 = N$.
Denote by S_n, I_n the number of Susceptible and Infected at time n. Then, for all $n \geq 0$,

$$S_n = I_{n+1} + S_{n+1}, \tag{2.1.7}$$

and, at each generation the actual number of new cases has a Binomial distribution depending on the parameter p. In the Greenwood model, the chance of a susceptible of being infected depends only on the presence of some infectives and not on their actual number. Hence, if $I_n = 0$, the epidemic terminates immediately since there is no further infectives. If $I_n \geq 1$, the conditional distribution of I_{n+1} given the past $\mathscr{F}_n = \sigma((S_i, I_i), i = 0, \dots n)$ is

$$\mathscr{L}(I_{n+1}|\mathscr{F}_n) = \text{Bin}(S_n, p) \quad \text{and } S_{n+1} = S_n - I_{n+1}.$$

The process keeps going on up to the time where there is no longer Infected in the population. Noting that $\mathscr{F}_n = \sigma(S_i, i = 0, \dots n)$, (S_n) is a Markov chain on $\{0, \dots, S_0\}$ with transition matrix

$$Q_p(i,j) = \binom{i}{i-j} p^{i-j}(1-p)^j \text{ if } 0 \leq j \leq i \leq S_0; \quad Q_p(i,j) = 0 \text{ otherwise.} \tag{2.1.8}$$

Parametric inference
Assume that the successive numbers of Susceptible $(s_0, s_1, \dots, s_n)$ have been observed up to time n. In this model, (S_n) decreases with n, and extinction occurs after a geometric number of generations. Therefore, the inference framework is to assume that S_0 (hence N) $\to \infty$.
Let $\mathbb{P}_p$ the probability associated to the Markov chain with transition Q_p and initial condition s_0. The likelihood associated with parameter p and observations $(s_1, \dots, s_n)$ is, if $s_0 \geq s_1 \cdots \geq s_n$,

$$L_n(p; s_1, \dots, s_n) = \prod_{k=1}^{n} \mathbb{P}_p(S_k = s_k | S_{k-1} = s_{k-1}) = C(s_0, \dots, s_n) p^{s_0 - s_n} (1-p)^{\sum_{k=1}^n s_n}. \tag{2.1.9}$$

All the quantities independent of p have been gathered in the term $C(s_0, , \dots, s_n)$. They depend on the model and the observations, and therefore have no influence on the estimation of p. Elementary computations yield that the value of p that maximizes the likelihood is

$$\hat{p}_n = \frac{s_0 - s_n}{\sum_{k=0}^{n-1} s_k} = \frac{1}{\text{“mean time to infection”}}.$$

Another approach for estimating parameters of a stochastic process is the Conditional Least Squares (CLS) method. This is the analog of the traditional Least Squares method for i.i.d. observations. It is especially relevant when computing the likelihood is intractable. Noting that $\mathbb{E}_p(S_k|\mathscr{F}_{k-1}) = (1-p)S_{k-1}$, it reads as

$$U_n(p, S_1, \dots, S_n) = \sum_{k=1}^{n} \left(S_k - \mathbb{E}_p(S_k|\mathscr{F}_{k-1})\right)^2 = \sum_{k=1}^{n} \left(S_k - (1-p)S_{k-1}\right)^2. \tag{2.1.10}$$

The associated Conditional Least Squares estimator is

$$\tilde{p}_n = 1 - \frac{\sum_{k=1}^n s_{k-1} s_k}{\sum_{k=1}^n s_{k-1}^2}. \tag{2.1.11}$$

A concern in statistics is to answer the question: how does such an estimator (or other ones) behave according to the asymptotic framework (here $S_0 \to \infty$). Is one of these two estimators better?

2.1.2 Reed–Frost model

It is also a chain Binomial *SIR* model relevant to model the evolution of an ordinary influenza in a small group of individuals. The latent period is long with respect to a short infectious period and new infections occur at successive generations separated by latent periods. It is assumed that latent periods are equal to 1, contacts between Susceptibles and Infected are independent, and that the probability of contact between a Susceptible and an Infected is $p = 1-q$. Therefore the probability of a Susceptible escaping infection given I Infected is q^I, and if $\mathscr{F}_n = \sigma((S_0,I_0),\ldots,(S_n,I_n))$,

$$\mathscr{L}(I_{n+1}|\mathscr{F}_n) = \text{Bin}(S_n,p_n) \text{ with } p_n = 1-q^{I_n} \quad \text{and } S_{n+1} = S_n - I_{n+1}.$$

Then (S_n,I_n) is a Markov chain on $\mathbb{N}^2$ with probability transitions,

$$\begin{aligned} Q_q((s_n,i_n),(s_{n+1},i_{n+1})) &= \binom{s_n}{s_{n+1}} (q^{i_n})^{s_{n+1}} (1-q^{i_n})^{i_{n+1}} \text{ if } s_{n+1}+i_{n+1} = s_n, \\ &= 0 \text{ otherwise.} \end{aligned}$$

Parametric inference

Assume that the successive numbers of Susceptible and Infected have been observed up to time n and consider the estimation of $q = 1-p$. Denote $\mathbb{P}_q$ the probability associated with the Markov chain with transition Q_q and initial condition (s_0,i_0). Then, if $s_{k+1}+i_{k+1} = s_k$ for $k = 0,\ldots,n-1$,

$$L_n(q;(s_1,i_1\ldots,(s_n,i_n)) = \prod_{k=0}^{n-1} \binom{s_k}{s_{k+1}} (q^{i_k})^{s_{k+1}} (1-q^{i_k})^{i_{k+1}}. \tag{2.1.12}$$

Therefore, $\log L_n(q) = C((s_k,i_k)) + \sum_{k=0}^{n-1}(s_{k+1}i_k \log q + i_{k+1}\log(1-q^{i_k}))$.
Differentiating with respect to q yields

$$\frac{d\log L_n}{dq} = \frac{1}{q}\sum_{k=0}^{n-1} \frac{i_k}{1-q^{i_k}}(s_{k+1} - s_k q^{i_k}).$$

The maximum likelihood estimator $\hat{q}_n$ of q is a solution of the equation

$$\sum_{k=0}^{n-1} \frac{i_k}{1-q^{i_k}}(s_{k+1} - s_k q^{i_k}) = 0.$$

Its properties can be studied as the number of observations increases (implying that the initial population grows to infinity).

Here a problem which occurs in practice already appears in this simple model, the case of "Partially Observed Markov Processes": it corresponds to the fact that both coordinates (S_n,I_n) are not observed, but only the successive numbers of Infected individuals $(I_k, k = 0,\ldots,n-1)$ are available. In the special case of Hidden Markov Models (see the Appendix, Section A.1.2 for the definition of H.M.M.), the theory for inference is now well known ([23], [125]), while there is no general theory for partially observed Markov processes. Many methods and algorithms have been proposed to deal with it in practice (see e.g. [38], [43], [71]). For applications specific to epidemics, many authors have addressed this problem (see e.g. [26], [30], [67], [71], together with the development of packages (see R package POMP [89])

2.1.3 Birth and death chain with re-emerging

We consider now the example of an epidemic model with re-emerging in a large infinite population. It can be described by a birth and death chain on $\mathbb{N}$ with reflection at 0. This models for instance farm animals epidemics when infection can also be produced by the environment. Let p,q denote the birth rate and death rates with $\{0 < p,q < 1, p+q < 1\}$. We assume that $I_0 = i_0 \geq 1$ and that (I_n), the number of infected at time n, evolves as follows:

- if $k \geq 1$, then $\mathbb{P}(I_{n+1} = k+1 | I_n = k) = p$, $\mathbb{P}(I_{n+1} = k-1 | I_n = k) = q$, $\mathbb{P}(I_{n+1} = k | I_n = k) = 1-p-q$;

- if $k = 0$, then $\mathbb{P}(I_{n+1} = 1 | I_n = 0) = p$, $\mathbb{P}(I_{n+1} = 0 | I_n = 0) = 1-p$ (re-emerging probability).

The Markov chain (I_n) is irreducible aperiodic on $\mathbb{N}$ and, if $p < q$, (I_n) is positive recurrent with stationary distribution

$$\lambda_{(p,q)}(i) = \left(1 - \frac{p}{q}\right)\left(\frac{p}{q}\right)^i.$$

Parametric inference

Let $\Theta = \{(p,q), 0 < p < q < 1 \text{ with } p+q < 1\}$ and let θ_0 be the true parameter value. Assume that $I_0 = i_0 > 0$ is non-random and fixed and consider the estimation of $\theta = (p,q) \in \Theta$ based on the observation of the successive numbers of Infected up to time n.

Let $(Q_\theta(i,j), i,j \in \mathbb{N})$ denote the transition kernel (I_n):

- if $i \neq 0$, then $Q_\theta(i,j) = p\delta_{\{i+1\}}(j) + q\delta_{\{i-1\}}(j) + (1-p-q)\delta_{\{i\}}(j)$,

- if $i = 0$, then $Q_\theta(0,j) = p\delta_1(j) + (1-p)\delta_0(j)$.

Noting that for $j \neq \{i-1, i, i+1\}$, $N_n^{ij} = 0$, the loglikelihood $\ell_n(\theta)$ satisfies

$$\begin{aligned}\ell_n(\theta) &= \sum_{i,j \in N} N_n^{ij} \log Q_\theta(i,j) \\ &= B_n \log p + D_n \log q + R_n \log(1-p-q) + N_n^{0,0} \log(1-p), \text{ with} \\ &\qquad B_n = \sum_{i \geq 0} N_n^{i,i+1},\ D_n = \sum_{i \geq 1} N_n^{i,i-1},\ R_n = \sum_{i \geq 1} N_n^{i,i}. \end{aligned} \tag{2.1.13}$$

Since the Markov chain (I_n, I_{n+1}) is positive recurrent on $\mathbb{N}^2$ with stationary measure $\lambda_\theta(i)Q_\theta(i,j)$, we can study directly the limit behaviour of $\ell_n(\theta)$. Applying the ergodic theorem to (I_n, I_{n+1}) yields that, almost surely under $\mathbb{P}_{\theta_0}$,

$$\begin{aligned}\frac{1}{n}N_n^{i,i+1} &\to p_0 \lambda_{\theta_0}(i) \text{ for } i \geq 1, \\ \frac{1}{n}N_n^{i,i-1} &\to q_0 \lambda_{\theta_0}(i), \\ \frac{1}{n}N_n^{i,i} &\to r_0 \lambda_{\theta_0}(i), \\ \frac{1}{n}N_n^{0,0} &\to (1 - \frac{p_0}{q_0})(1-p_0).\end{aligned}$$

Therefore, using (2.1.13),

$$\begin{aligned}\frac{1}{n}B_n &\to p_0, \\ \frac{1}{n}D_n &\to q_0 \times \frac{p_0}{q_0} = p_0, \\ \frac{1}{n}R_n &\to \frac{r_0 p_0}{q_0}, \\ \frac{1}{n}N_n^{0,0} &\to (1 - \frac{p_0}{q_0})(1-p_0).\end{aligned}$$

Joining these results, under P_{θ_0}, as $n \to \infty$,

$$\frac{1}{n}\ell_n(\theta) \to p_0 \log p + p_0 \log q + \frac{r_0 p_0}{q_0} \log r + (1 - \frac{p_0}{q_0})(1-p_0)\log(1-p) := J(\theta_0, \theta).$$

We can check directly that $\theta \to J(\theta_0, \theta)$ possesses a unique global maximum at θ_0. The associated maximum likelihood estimator $\hat{\theta}_n$ is

$$\hat{p}_n = \frac{1}{n} B_n; \quad \hat{q}_n = \frac{B_n}{D_n + R_n}(1 - \frac{1}{n} B_n). \tag{2.1.14}$$

Successive applications of the ergodic theorem yield that $(\hat{p}_n, \hat{q}_n)$ converges $\mathbb{P}_{\theta_0}$ a.s. to (p_0, q_0).

To study the limit distribution of $(\hat{p}_n, \hat{q}_n)$, we use the results of Section A.2.1.1 in the Appendix. Let $Q = (Q(i,j))$ denote the (unnormalized) transition kernel on $\mathbb{N} \times \mathbb{N}$:

$$\begin{aligned} &Q(i, i+1) = 1 = Q(i,i) \text{ for } i \in \mathbb{N} \text{ and} \\ &Q(i, i-1) = 1 \text{ for } i \geq 1. \end{aligned}$$

According to (A.2.1), the family $(Q_\theta, \theta \in \Theta)$ is dominated by Q with associated function $f_\theta(i,j)$:

$$\begin{aligned} f_\theta(i, i+1) &= p \text{ for } i \geq 0, \\ f_\theta(i, i-1) &= q \text{ for } i \geq 1, \\ f_\theta(i,i) &= 1 - p - q, \\ f_\theta(0,0) &= 1 - p. \end{aligned}$$

Except the compactness assumption of Θ (only required for the consistency of the MLE), the Markov chain satisfies Assumptions (H1)–(H8) of Section A.2.1.1, Therefore, under $\mathbb{P}_{\theta_0}$,

$$\sqrt{n} \begin{pmatrix} \hat{p}_n - p_0 \\ \hat{q}_n - q_0 \end{pmatrix} \to_{\mathscr{L}} \mathscr{N}(0, I^{-1}(\theta_0)),$$

with, using Definition (A.2.5),

$$I(\theta_0) = \sum_{i \geq 0} \lambda_{\theta_0}(i) \sum_{j \geq 0} \frac{\nabla_\theta f_{\theta_0}(i,j) \nabla_\theta^* f_{\theta_0}(i,j)}{f_{\theta_0}(i,j)^2} Q_{\theta_0}(i,j).$$

Hence $I(\theta_0)$ can be explicitly computed: for $\theta = (p,q)$, we get

$$I(p,q) = \begin{pmatrix} \frac{r+p^2}{p(1-p)r} & \frac{p}{qr} \\ \frac{p}{qr} & \frac{p(1-p)}{rq^2} \end{pmatrix} \Rightarrow I^{-1}(p,q) = \begin{pmatrix} p(1-p) & -pq \\ -pq & \frac{q^2(p^2+r)}{p(1-p)} \end{pmatrix}. \tag{2.1.15}$$

2.1.4 Modeling an infection chain in an Intensive Care Unit

This example is taken from Chapter 4 of [36]. It aims at describing nosocomial infections (i.e. infections acquired in a hospital). The incidence of these infections is highest in an Intensive Care Unit, which is characterized by a small number of beds (about 10 beds at most) and rapid turnover of patients by way of admission and discharge. There are two routes for infection (colonization) for a patient.

- The endogenous route (α mechanism): bacteria are already present in a newly admitted patient but at low undetectable levels and resistant bacteria develop because of antibiotic treatments during the stay. Let $e^{-\alpha} = (1-a)$ the probability per individual per time unit of getting infected by this route.

- The exogenous route (β transmission): it models the probability of infection of a Susceptible by an Infected in the ICU per time unit, $e^{-\beta} = (1-b)$.

To describe the composition of the ICU in terms of Infected and Susceptible individuals on long time intervals, a Markov chain model can be used as follows. Each patient has probability d of being discharged by unit of time. Discharge and admission take place every day at noon; new admitted individuals are susceptible. Observations are obtained from a bookkeeping scheme that concerns the state of the ICU immediately after discharge (12h05).

Consider the simplest example, an ICU with two beds. It corresponds to three possible states: State 0 (both patients are Susceptible), State 1 (one Susceptible, one Infected) and 2 (both are Infected). Denote by X_n the composition of the ICU at time n. Let us compute according to $\theta = (a,b,d)$ the transition matrix Q_θ of (X_n). Introduce $\bar{X}_{n+1}$ the state of the ICU just before discharge (at 11h55) on the next day. If $X_n = 0$, $\mathbb{P}(\bar{X}_{n+1} = 0) = a^2$, $\mathbb{P}(\bar{X}_{n+1} = 1) = 2a(1-a)$ and $\mathbb{P}(\bar{X}_{n+1} = 2) = (1-a)^2$. If $X_n = 1$, $\mathbb{P}(\bar{X}_{n+1} = 1) = ab$, and $\mathbb{P}(\bar{X}_{n+1} = 2) = (1-ab)$. Finally, if $X_n = 2$, $\mathbb{P}(\bar{X}_{n+1} = 2) = 1$. This yields that, after discharge (12h05), $X_{n+1} = 0,1,2$ with respective probabilities,

$$Q_\theta = \begin{pmatrix} (a+(1-a)d)^2 & 2(1-a)(1-d)(a+(1-a)d) & (1-a)^2(1-d)^2 \\ abd+(1-ab)d^2 & 2(1-ab)d(1-d)+ab(1-d) & (1-ab)(1-d)^2 \\ d^2 & 2d(1-d) & (1-d)^2 \end{pmatrix}.$$

Let $\Theta = (0,1)^3$. Assume that the states (X_i) of the ICU after discharge have been observed up to time n. The maximum likelihood estimator of θ reads, using (2.1.1),

$$\begin{aligned} \ell_n(\theta) &= \sum_{k=1}^{n} \log Q_\theta(X_{k-1},X_k) = \sum_{i,j\in\{0,1,2\}} N_n^{ij} \log Q_\theta(i,j), \\ \hat{\theta}_n &= \operatorname{argsup}_{\theta\in\Theta} \ell_n(\theta). \end{aligned}$$

Since (X_n) is a positive recurrent Markov chain on $\{0,1,2\}$, we can apply the results stated in the Appendix, Section A.2. The MLE $\hat{\theta}_n$ is consistent and converges at rate $\sqrt{n}$ to a Gaussian law $\mathcal{N}_3(0, I^{-1}(\theta))$, where $I(\theta)$ is the Fisher information matrix defined in (A.2.5).

Assume now that there is no systematic control of the exact status of the patients after discharge, but that each patient is tested with probability p. Then, the observations are no longer (X_n), but (Y_n), with conditional transition matrix $(T_p(i,j) = P(Y_n = j | X_n = i), 0 \le i,j \le 2)$,

$$T_p = \begin{pmatrix} (1-p)^2 & 2p(1-p) & p^2 \\ p(1-p) & p^2+(1-p)^2 & p(1-p) \\ p^2 & 2p(1-p) & (1-p)^2 \end{pmatrix}.$$

If only (Y_n) is observed, we have to deal with a Hidden Markov Model (X_n, Y_n) as defined in the Appendix, Section A.1.2. The estimation of θ or (θ, p) has to take into account this additional noise to be efficient (see e.g [23]).

2.2 Two extensions to continuous state and continuous time Markov chain models

2.2.1 A simple model for population dynamics.

The $AR(1)$ model is a classical model of population dynamics with continuous state space and allows us to illustrate explicitly various inference questions. Consider the autoregressive process on $\mathbb{R}$s introduced in Section 1.2.1 defined by

$$X_0 = x_0; \quad \text{and for } i \ge 1, X_i = aX_{i-1} + \gamma\varepsilon_i,$$

where (ε_i) is a sequence of i.i.d. random variables on $\mathbb{R}$ with distribution $f_\theta(x)dx$, independent of X_0.
This is a Markov chain on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with transition kernel $Q_{\theta,a}(x,dy) = f_\theta(y-ax)dy$. If X_0 is known, choosing as dominating kernel the Lebesgue measure on $\mathbb{R}$, the likelihood reads as

$$L_n(a,\theta) = \prod_{i=1}^{n} f_\theta(X_i - aX_{i-1}).$$

The Gaussian $AR(1)$ corresponds to $\varepsilon_i \sim \mathcal{N}(0,\gamma^2)$:

$$Q_{a,\gamma^2}(x,dy) = \frac{1}{\gamma\sqrt{2\pi}} \exp -(\frac{1}{2\gamma^2}(y-ax)^2)dy, \text{ and}$$

$$\begin{aligned} L_n(a,\gamma^2) &= \prod_{i=1}^{n} \frac{1}{\gamma\sqrt{2\pi}} \exp(-\frac{1}{2\gamma^2}(X_i - aX_{i-1})^2), \\ \ell_n(a,\gamma^2) &= -(n/2)\log\gamma^2 - 1/(2\gamma^2)\sum_{i=1}^{n}(X_i - aX_{i-1})^2. \end{aligned}$$

The properties of the MLE have been presented in Chapter 1.

2.2.2 Continuous time Markov epidemic model

We just recall here results for the *SIR* Markov jump process (see Section 4.2). Assume that the jump process $(\mathscr{Z}^N(t))$ is continuously observed on $[0,T]$. Its dynamics is described by the two parameters (λ,γ). The Maximum Likelihood Estimator $(\hat{\lambda},\hat{\gamma})$ is explicit (see [2] or Section 4.2). Indeed, let (T_i) denote the successive jump times and set $J_i = 0$ if we have an infection and $J_i = 1$ if we have a recovery. Let $K_N(T) = \sum_{i\geq 0} 1_{T_i \leq T}$. Then

$$\begin{aligned} \hat{\lambda}_N &= \frac{1}{N}\frac{\sum_{i=1}^{K_N(T)}(1-J_i)}{\int_0^T S^N(t)I^N(t)dt} = \frac{1}{N}\frac{\#\text{ Infections}}{\int_0^T S^N(t)I^N(t)dt}, \\ \hat{\gamma}_N &= \frac{1}{N}\frac{\sum_{i=1}^{K_N(T)} J_i}{\int_0^T I^N(t)dt} = \frac{\#\text{ Recoveries}}{\text{“Mean infectious period”}}. \end{aligned}$$

As the population size N goes to infinity, $(\hat{\lambda}_N,\hat{\gamma}_N)$ is consistent and

$$\sqrt{N}\begin{pmatrix}\hat{\lambda}_N - \lambda \\ \hat{\gamma}_N - \gamma\end{pmatrix} \to \mathscr{N}_2\left(0, I^{-1}(\lambda,\gamma)\right), \text{ where } I(\lambda,\gamma) = \begin{pmatrix} \frac{\int_0^T s(t)i(t)dt}{\lambda} & 0 \\ 0 & \frac{\int_0^T i(t)dt}{\gamma}\end{pmatrix},$$

and $(s(t),i(t))$ is the solution of the ODE associated with the limit behaviour of the normalized process $(\mathscr{Z}^N(t)/N)$: $\frac{ds}{dt} = -\lambda s(t)i(t)$; $\frac{di}{dt} = \lambda s(t)i(t) - \gamma i(t)$.

The matrix $I(\lambda,\gamma)$ is the Fisher information matrix of this statistical model.

2.3 Inference for Branching processes

At the early stage of an outbreak, a good approximation for the epidemic dynamics is to consider that the population of Susceptible is infinite and that Infected individuals evolve according to a branching process (see Section **??** of Part I). We present here some classical statistical results in this domain. This Markov chain model is an example of non-ergodic processes which leads to different statistical results.

2.3.1 Notations and preliminary results

Some basic facts on discrete time branching processes (or Bienaymé–Galton–Watson processes) are given in Part I, Section **??** of these notes (see also the classical monographs on branching processes [6] or [79]). We complete these facts with some properties useful for the inference.

Consider an ancestor $Z_0 = 1$ has ξ_0 children according to an offspring law G defined by

$$\mathbb{P}(\xi_0 = k) = p_k, \;\; k \geq 0 \;\text{ and }\; \sum_{k\geq 0} p_k = 1.$$

Let $m = \mathbb{E}(\xi_0)$ and $g(s) = \mathbb{E}(s^{\xi_0})$. The i-th of those children has $\xi_{1,i}$ children, where the random variables $\{\xi_{k,i}, k \geq 0, i \geq 1\}$ are i.i.d. with distribution G. Let Z_n denote the number of individuals in generation n. Then,

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{n,i}. \tag{2.3.1}$$

Denote by $E = \{\exists\, n,\, Z_n = 0\}$ the set of extinction.
If $m \leq 1$ and if $p_1 \neq 1$, the process Z_n has a probability $q = 1$ of extinction.
If $m > 1$, the process is supercritical and has a probability of extinction $q < 1$, which is the smallest solution of the equation $g(s) = s$ on $[0,1]$. The set E^c is equal to $\{\omega, Z_n(\omega) \to \infty\}$.

This extinction probability is an important parameter for the early stages of an epidemic. It corresponds to the probability of a minor outbreak.
We complete the results given in Part I, Section **??**. Let $\mathscr{F}_n = \sigma(Z_0, \ldots, Z_n)$ and define $W_n = m^{-n} Z_n$. Then (W_n) is a $\mathscr{F}_n$-martingale.

Theorem 2.3.1. *Assume that $m > 1$ and that the offspring law G has finite variance σ^2. Then, there is a non-negative random variable W such that*

(i) $W_n \to W$ *as* $n \to \infty$ *a.s. and in* L^2.

(ii) $\{W > 0\} = \{Z_n \to \infty\} = E^c$ *and* $\{W = 0\} = \{\lim_n Z_n = 0\} = E$.

(iii) Moreover, $\mathbb{E}W = 1$, $\mathrm{var}(W) = \frac{\sigma^2}{m(m-1)}$.

Corollary 2.3.2. *If $m > 1$, then, almost surely*

$$\frac{1}{m^n} \sum_{i=1}^{n} Z_i \to \frac{m}{m-1} W; \quad \frac{1}{m^n} \sum_{i=1}^{n} Z_{i-1} \to \frac{1}{m-1} W. \tag{2.3.2}$$

Proof. We write $\sum_{i=1}^n Z_i = \sum_{i=1}^n m^i \frac{Z_i}{m^i}$. Using Theorem 2.3.1, $\frac{Z_n}{m^n} \to W$ a.s. An application of the Toeplitz lemma stated below and some algebra yield the two results.

Lemma 2.3.3. *(Toeplitz Lemma) Let (a_n) a sequence of non-negative real numbers and (x_n) a sequence on $\mathbb{R}$. If $\sum_{i=1}^n a_i \to \infty$ and if $(x_n) \to x \in \mathbb{R}$ as $n \to \infty$, then*

$$\frac{\sum_{i=1}^n a_i x_i}{\sum_i^n a_i} \to x \text{ as } n \to \infty.$$

□

Assume that the offspring distribution $G_\theta(\cdot)$ depends on a parameter θ with finite mean $m(\theta) > 1$ and finite variance $\sigma^2(\theta)$. Denote by $\mathbb{P}_\theta$ the law on $(\mathbb{N}^{\mathbb{N}}, \mathscr{B}(N^{\mathbb{N}}))$ of the branching process (Z_n) with offspring law $G_\theta(\cdot)$. Then $(Z_n, n \geq 0)$ is a Markov chain with state space $\mathbb{N}$, initial condition $Z_0 = 1$ and transition matrix,

$$Q_\theta(i,j) = G_\theta^{\star i}(j), \tag{2.3.3}$$

where $\star$ denotes the convolution product of two functions and $f^{\star i}$ is the i-fold convolution product of $f(\cdot)$. Let μ_i denote the measure $\mu_i(k) = 1$ for all $k \in \mathbb{N}$, $\lambda_n = \otimes_{i=1}^n \mu_i$. Then, the likelihood reads as

$$\frac{d\mathbb{P}_\theta}{d\lambda_n}(Z_0, \ldots, Z_n) = L_n(\theta) = \prod_{i=1}^{n} G_\theta^{\star Z_{i-1}}(Z_i); \quad \ell_n(\theta) = \sum_{i=1}^{n} \log(G_\theta^{\star Z_{i-1}}(Z_i)). \tag{2.3.4}$$

Under this expression, studying the likelihood for general offspring laws is intractable. We detail in the next section a framework where it is possible to study this likelihood, and in the next section another method based on Weighted Conditional Least Squares.

2.3.2 Inference when the offspring law belongs to an exponential family

Among parametric families of distributions, exponential families of distributions, widely used in statistics, provide here a nice framework to study this likelihood. A short recap is given in the Appendix Section A.1.2 (see e.g. the classical monograph [11] for the complete exposition).

Assume that the offspring law is a power series distribution:

$$p(k) = A(\zeta)^{-1}\, a_k\, \zeta^k, \quad \text{with } A(\zeta) = \sum_{k\geq 0} a_k \zeta^k. \tag{2.3.5}$$

Setting $\theta = \log\zeta$, $\Theta = \{\theta \in \mathbb{R},\, A(e^\theta) < \infty\}$, $h : k \to h(k) = a_k$ and $\phi : \theta \to \phi(\theta) = \log A(\log(e^\theta))$, we get that it is a special case of an exponential family of distributions on $\mathbb{N}$ with $T(X) = X$ and

$$p(\theta,k) = h(k)\exp(k\theta - \phi(\theta)). \tag{2.3.6}$$

The random variable X satisfies that

$$m(\theta) := \mathbb{E}_\theta(X) = \nabla_\theta \phi(\theta); \quad \sigma^2(\theta) := Var_\theta(X) = \nabla^2_\theta \phi(\theta). \tag{2.3.7}$$

Moreover, if $X_1, \ldots, X_n$ are i.i.d. with distribution (2.3.5), then

$$\mathbb{P}(X_1 + \cdots + X_n = k) = H(n,k)\exp(k\theta - n\phi(\theta)) \text{ where } H(n,k) = h^{*n}(k). \tag{2.3.8}$$

Therefore for offspring distributions satisfying (2.3.5) or (2.3.6), the transition kernel is

$$Q_\theta(i,k) = H(i,k)\exp(k\theta - i\phi(\theta)).$$

Let us note that several families of classical distributions on $\mathbb{N}$ are included in this set-up:

- Geometric distributions on $\mathbb{N}^*$ with parameter p (i.e. $P(X=k) = p(1-p)^{k-1}$):
  $\theta = \log(1-p)$; $h(k) = 1$ and $\phi(\theta) = \log\frac{e^\theta}{1-e^\theta}$.

- Binomial distributions $(\mathscr{B}(N,p), p \in (0,1))$ with N fixed:
  $\theta = \log\frac{p}{1-p}$, $h(k) = \frac{N!}{k!(N-k)!}$ and $\phi(\theta) = N\log(1+e^\theta)$.

- Poisson distributions $\mathscr{P}(\lambda)$: $\theta = \log\lambda$, $h(k) = \frac{1}{k!}$ and $\phi(\theta) = e^\theta$.

- Negative Binomial distributions $(\mathscr{NB}(r,p), p \in (0,1))$ with r fixed (i.e. $P(X=k) = \frac{\Gamma(k+r)}{\Gamma(r)k!}p^r(1-p)^k$:
  $\theta = \log(1-p)$, $h(k) = \frac{(k+r)\ldots(r+1)}{k!}$ and $\phi(\theta) = r\log(1-e^\theta)$).

Let us come back to the likelihood (2.3.4). Let $\theta_0 \in \Theta$ be the true value of the parameter. We assume

(A1) The offspring distribution G_θ belongs to an exponential power series family: For all $k \in \mathbb{N}$, $G_\theta(k) = h(k)\exp(k\theta - \phi(\theta))$.

(A2) Θ is a compact subset of $\{\theta, \sum_{k\geq 0} h(k)e^{\theta k} < \infty\}$, $\theta_0 \in \mathrm{Int}(\Theta)$.

(A3) For all $\theta \in \Theta$, $m(\theta) > 1$ and $\sigma^2(\theta)$ finite.

(A4) There exists a $\delta > 0$ such that $E(Y^{2+\delta}) = \mu_{2+\delta} < \infty$ where $Y \sim G_\theta$.

Consider the estimation of θ when the successive generation sizes $(Z_1, \dots Z_n)$ are observed. Under (A1)–(A3), the loglikelihood is, using (2.3.8),

$$\ell_n(\theta) = C(Z_0, \dots, Z_n) + \sum_{i=1}^{n} (\theta Z_i - \phi(\theta) Z_{i-1}), \tag{2.3.9}$$

with $C(Z_0, \dots, Z_n) = \sum_{i=1}^{n} \log H(Z_{i-1}, Z_i)$. The constant $C(Z_0, \dots, Z_n)$ depends only on the observations and brings no information on θ.

The M.L.E $\hat{\theta}_n$, defined as any solution of $\nabla_\theta \ell_n(\theta) = 0$, satisfies

$$\sum_{i=1}^{n} Z_i - \nabla_\theta \phi(\hat{\theta}_n) \sum_{i=1}^{n} Z_{i-1} = 0.$$

Using that $\nabla_\theta \phi(\theta) = m(\theta)$ (see (2.3.7)), $\hat{\theta}_n$ satisfies

$$m(\hat{\theta}_n) = \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} Z_{i-1}}. \tag{2.3.10}$$

By Theorem 2.3.1, $m(\theta_0)^{-n} Z_n$ converges a.s. and in L^2 under $\mathbb{P}_{\theta_0}$ to a random variable W such that $W > 0$ on E^c, the non-extinction set, which satisfies $\mathbb{P}_{\theta_0}(E^c) = 1 - q > 0$ under (A3).

Theorem 2.3.4. *Assume (A1)–(A4).Then, on E^c, $m(\hat{\theta}_n)$ satisfies*

- ***(i)*** *$m(\hat{\theta}_n) \to m(\theta_0)$ a.s. under $\mathbb{P}_{\theta_0}$.*
- ***(ii)*** *$m(\theta_0)^{n/2}(m(\hat{\theta}_n) - m(\theta_0)) \to_{\mathcal{L}} \sqrt{(m(\theta_0)-1)\sigma^2(\theta_0)}\, \eta^{-1} N$, where η, N are independent r.v.s, $N \sim \mathcal{N}(0,1)$, and η is the positive variable defined by $\eta^2 = W$ on E^c.*

Clearly, $m(\theta)$ is the parameter that is naturally estimated here.

Proof. Let us write

$$m(\hat{\theta}_n) = \frac{\frac{\sum_{i=1}^{n} Z_i}{m^n}}{\frac{\sum_{i=1}^{n} Z_{i-1}}{m^n}}.$$

Using Corollary 2.3.2, both terms of the above fraction converge a.s. so that $m(\hat{\theta}_n) \to m(\theta_0)$ a.s.
Let us prove (ii). The score function reads as

$$\nabla_\theta \ell_n(\theta) = \sum_{i=1}^{n} Z_i - m(\theta) \sum_{i=1}^{n} Z_{i-1} = \sum_{i=1}^{n} (Z_i - m(\theta) Z_{i-1}).$$

Under $\mathbb{P}_{\theta_0}$, $\nabla_\theta \ell_n(\theta_0)$ is a centered $\mathscr{F}_n$-martingale (M_n) with increments $X_i = Z_i - m(\theta_0) Z_{i-1}$. Conditionally on $\mathscr{F}_{i-1}$, X_i is the sum of Z_{i-1} independent centered random variables so that

$$\mathbb{E}_{\theta_0}(X_i^2 | \mathscr{F}_{i-1}) = \sigma^2(\theta_0) Z_{i-1}; \quad \langle M \rangle_n = \sigma^2(\theta_0) \sum_{i=1}^{n} Z_{i-1}.$$

Hence

$$s_n^2(\theta_0) = \mathbb{E}_{\theta_0}(\langle M \rangle_n) = \sigma^2(\theta_0) \sum_{i=1}^{n} m(\theta_0)^{i-1} = \sigma^2(\theta_0) \frac{m(\theta_0)^n - 1}{m(\theta_0) - 1}.$$

Therefore $s_n^2(\theta_0) \to \infty$ as $n \to \infty$ and

$$\frac{s_n^2(\theta_0)}{m(\theta_0)^n} \to \frac{\sigma^2(\theta_0)}{m(\theta_0) - 1}. \tag{2.3.11}$$

Let us check the conditions of the Central limit theorem for martingales (see A.4.1) recalled in the Appendix Under (A3), (M_n) is a square integrable centered $\mathscr{F}_n$-martingale such that $\mathbb{E}_{\theta_0}(\langle M_n\rangle) = s_n(\theta_0)^2 \to \infty$. Let us check (H2). We have

$$\frac{1}{s_n(\theta_0)^2}\langle M_n\rangle = \frac{m^n(\theta_0)}{s_n^2(\theta_0)}\frac{\sigma^2(\theta_0)}{m^n(\theta_0)}\sum_{i=1}^{n} Z_{i-1}.$$

Hence according to Corollary 2.3.2 and Theorem 2.3.1, $\frac{1}{s_n(\theta_0)^2}\langle M_n\rangle \to W$ in probability under $\mathbb{P}_{\theta_0}$ with $W > 0$ on E^c and $E_{\theta_0}(W) = 1$. Therefore we can set $W = \eta^2$ and obtain (H2).
It remains to check the asymptotic negligibility Assumption (H1'). We have, for $X_i = Z_i - m(\theta_0)Z_{i-1}$,

$$\mathbb{E}_{\theta_0}(|X_i|^{2+\delta}|\mathscr{F}_{i-1}) = Z_{i-1}\mathbb{E}_{\theta_0}(|Y - m(\theta_0)|^{2+\delta}).$$

Under (A4), using that $\mathbb{E}_{\theta_0}(|Y - m(\theta_0)|^{2+\delta}) \le C(\mu_{2+\delta} + m(\theta_0)^{2+\delta}) < \infty$ yields

$$\frac{1}{s_n^{2+\delta}}\sum_{i}^{n}\mathbb{E}_{\theta_0}(|X_i|^{2+\delta}|\mathscr{F}_{i-1}) = (\frac{1}{s_n^{\delta}})\mathbb{E}_{\theta_0}(|Y - m(\theta_0)|^{2+\delta})(\frac{1}{s_n^2}\sum_{i=1}^{n} Z_{i-1}). \tag{2.3.12}$$

Using Corollary 2.3.2 and (2.3.11) yields that the last term of (2.3.12) is bounded in probability under $\mathbb{P}_{\theta_0}$. Since $\delta > 0$, the first term of (2.3.12) tends to 0, which achieves the proof of (H1').
Therefore, we get that on the non-extinction set , under $\mathbb{P}_{\theta_0}$,

$$(\frac{M_n}{s_n}, \frac{\langle M\rangle_n}{s_n^2}) \to_{\mathscr{L}} (\eta N, \eta^2), \tag{2.3.13}$$

with η, N independent, $\eta = W^{1/2}$ and $N \sim \mathscr{N}(0,1)$.
To study the limit distribution of $\hat{m}_n$, we write

$$\hat{m}_n - m(\theta_0) = \frac{\sum_{i=1}^{n}(Z_i - m(\theta_0)Z_{i-1})}{\sum_{i=1}^{n} Z_{i-1}} = \sigma^2(\theta_0)\frac{M_n}{\langle M_n\rangle}.$$

This yields that

$$m(\theta_0)^{n/2}(\hat{m}_n - m(\theta_0)) = \sigma^2(\theta_0)\frac{m(\theta_0)^{n/2}}{s_n}\frac{\frac{M_n}{s_n}}{\frac{\langle M_n\rangle}{s_n^2}}.$$

Using (2.3.11) and (2.3.13) achieves the proof of (ii). □

Let us stress that here, contrary to the previous models, the Fisher information, $E_\theta\langle M\rangle_n = \sigma^2(\theta)\frac{m(\theta)^n - 1}{m(\theta) - 1}$ converges to infinity at a much faster rate than "usually" for $m(\theta) > 1$. Indeed, the information contained in $(Z_1, \ldots, Z_n)$ is of the same order as the information in the last observation Z_n. In that respect, the model is explosive in terms of growth of information.

Note that this result could be obtained using the MLE Heuristics presented in the Appendix, substituting $\sqrt{n}$ by s_n and using that
$\nabla_\theta^2 \ell_n(\theta) = -\nabla_\theta^2\phi(\theta)\sum_{i=1}^{n} Z_{i-1} = -\sigma^2(\theta)\sum_{i=1}^{n} Z_{i-1} = -\langle M(\theta)\rangle_n$.

Now we have estimated $m(\theta)$ instead of θ. To estimate θ, we just have to consider the application $\theta \to m(\theta)$. Assuming that there exists ϕ differentiable such that $\phi(y) = \theta = m^{-1}(y)$, an application of Theorem A.1.1 yields the result for θ.

2.3.3 Parametric inference for general Galton–Watson processes

We assume now that the offspring distribution $G(\cdot)$ has mean m and finite variance σ^2 and consider the Galton–Watson process with initial condition $Z_0 = 1$ and offspring distribution G. We assume

(B1) The offspring law G satisfies $m > 1$ and $\sigma^2 < \infty$.

(B2) The offspring law G has a finite moment of order 4: $E(Y^4) = \mu_4 < \infty$ where $Y \sim G$.

On the basis on the successive population sizes $(Z_1, \ldots, Z_n)$, we are concerned with the estimation of $\theta = (m, \sigma^2)$. Denote by $\mathbb{P}_\theta$ the distribution on $(\mathbb{N}^{\mathbb{N}}, \mathscr{B}(N^{\mathbb{N}}))$ of (Z_n). Under (B1), the branching process is supercritical ($m > 1$) and the non-extinction set E^c has a positive probability. Clearly, studying estimators based on (2.3.4) is intractable. Therefore, we had rather study estimators based on Conditional Least Square methods. The conditional mean and variance of Z_n with respect to $\mathscr{F}_{n-1}$ write

$$\mathbb{E}_\theta(Z_n|\mathscr{F}_{n-1}) = mZ_{n-1}; \quad Var_\theta(Z_n|\mathscr{F}_{n-1}) = \sigma^2 Z_{n-1}. \tag{2.3.14}$$

On the non-extinction set E^c, let us consider the contrast function (which is a weighted Conditional Least Square method):

$$U_n(\theta) = \sum_{i=1}^{n} \frac{1}{Z_{i-1}}(Z_i - mZ_{i-1})^2. \tag{2.3.15}$$

Note that $U_n(\theta)$ only depends on m and therefore σ^2 cannot be estimated using U_n.
Define $\tilde{m}_n$ as a solution of

$$U_n(\tilde{m}_n) = min_{\theta\in\Theta} U_n(\theta).$$

Hence it satisfies $\nabla_\theta U_n(\tilde{m}_n) = 0$, which yields

$$\tilde{m}_n = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_{i-1}}. \tag{2.3.16}$$

The simplest approach for estimating σ^2 is to use the residual variance:

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Z_{i-1}}(Z_i - \tilde{m}_n Z_{i-1})^2. \tag{2.3.17}$$

Then the following holds.

Theorem 2.3.5. *Assume (B1)–(B2). Then, on the non-extinction set E^c, the estimators $(\tilde{m}_n, \tilde{\sigma}_n^2)$ defined in* (2.3.16)–(2.3.17) *satisfy, as $n \to \infty$, under $\mathbb{P}_\theta$,*

- ***(i)*** *$\tilde{m}_n \to m$ almost surely.*
- ***(ii)*** *$m^{n/2}(\tilde{m}_n - m) \to_{\mathscr{L}} \sqrt{(m-1)\sigma^2}\ \eta^{-1}\ N$, where η, N are independent r.v.s, $N \sim \mathscr{N}(0,1)$, η is the positive variable defined by $\eta^2 = W$.*
- ***(iii)*** *$\tilde{\sigma}_n^2 \to \sigma^2$ in probability under $\mathbb{P}_\theta$.*
- ***(iv)*** *$\sqrt{n}(\tilde{\sigma}_n^2 - \sigma^2) \to_{\mathscr{L}} \mathscr{N}(0, 2\sigma^4)$.*

Proof. The study of the asymptotic properties of $\tilde{m}_n$ is similar to the previous section, since $\tilde{m}_n$ has the same expression with respect to the observations that $\hat{m}_n(\theta)$. The proofs of (iii) and (iv) are derived from [60], Chapter 3. Let us prove (iii). We have $\tilde{\sigma}_n^2 - \sigma^2 = \frac{1}{n}(A_n^1 + A_n^2 + A_n^3)$ with

$$A_n^2 = (m - \tilde{m}_n)^2 (\sum_{i=1}^{n} Z_{i-1}),$$

$$A_n^3 = 2(m - \tilde{m}_n) \sum_{i=1}^{n} (Z_i - mZ_{i-1}) \text{ and}$$

$$A_n^1 = \sum_{i=1}^{n} X_i \quad \text{with } X_i = \frac{1}{Z_{i-1}}(Z_i - mZ_{i-1})^2 - \sigma^2. \tag{2.3.18}$$

Let us study the first term A_n^1. It is a centered $\mathscr{F}_n$-martingale under $\mathbb{P}_\theta$. The computation of $\mathbb{E}_\theta(X_i^2|\mathscr{F}_{i-1})$ relies on the property that, for i.i.d. random variables Y_i with $\mathbb{E}(Y_i)=m, \mathrm{Var}Y_i=\sigma^2$ and finite fourth moment $\mathbb{E}(Y^4)=\mu_4$, $\bar{Y}=\frac{1}{n}\sum_{i=1}^n Y_i$ satisfies

$$\mathbb{E}(\bar{Y}-m)^4=\frac{3\sigma^4}{n^2}+\frac{1}{n^3}(\mu_4-3\sigma^4).$$

Hence on the non-extinction set E^c,

$$\langle A^1\rangle_n=\sum_{i=1}^n(2\sigma^4+\frac{1}{Z_{i-1}}(\mu_4-3\sigma^4)). \tag{2.3.19}$$

Hence $\mathrm{Var}(A_1^n)\leq 2n(\sigma^4+\mu^4)$. Therefore, applying a strong law of large numbers for martingales ([64], Theorem 2.18) yields $\frac{1}{n}A_n^1\to 0\,\mathbb{P}_\theta$-a.s.
The second term is $A_n^2=\left(m^n(\tilde{m}_n-m)^2\right)\left(\frac{1}{m^n}\sum_{i=1}^n Z_{i-1}\right)$. By (ii) and Corollary 2.3.2, we get that these two terms converge in distribution so that A_n^2 is bounded in probability.
Noting that $M_n=\sum_{i=1}^n(Z_{i-1}-mZ_{i-1})$ is the martingale studied in the previous section yields that $A_n^3=[m^{n/2}(m-\tilde{m}_n)][\frac{1}{m^{n/2}}M_n]$. By (ii) $[m^{n/2}(m-\tilde{m}_n)]$ converges in distribution. The CLT for (M_n) stated in (2.3.13) yields that $m^{n/2}M_n$ converges in distribution. Hence, A_n^3 is also bounded in probability. Joining these results yields that $\frac{1}{n}(A_n^1+A_n^2+A_n^3)\to 0$, which achieves the proof of (iii).

Let us prove (iv). The previous computations yield that $n^{-1/2}A_n^2$ and $n^{-1/2}A_n^3$ both converge to 0. The martingale (A_n^1) is centered square integrable and $s_n^2=E_\theta\langle M\rangle_n$ satisfies $\frac{1}{n}s_n^2\to 2\sigma^4$. Condition (H1') is satisfied assuming the existence of a moment of order $4+\delta$ with $\delta>0$ for the offspring law G. Therefore, the CLT for martingales (see Theorem A.4.1) yields that $\frac{1}{n}\langle M\rangle_n\to 2\sigma^4$ a.s. Joining these results achieves the proof of (iv). □

With similar arguments, one can prove the asymptotic independence of $(\tilde{m}_n,\tilde{\sigma}_n^2)$.

The extinction probability is an important parameter in many applications. In the early stages of an epidemic, the extinction probability corresponds to the probability of a minor outbreak. However, unless the extinction probability q is a function of m and σ^2 only, it cannot be consistently estimated observing the generation sizes. A parametric setting $(G_\theta,\theta\in\Theta)$ is required for the offspring law. Let $g(\theta,s)$ denote the generating function of G_θ and define

$$\tilde{q}_n=\inf\{s,g(s,\tilde{\theta}_n)=s\}.$$

Then, according to [60], under additional regularity assumptions, $\tilde{q}_n$ is consistent if $\tilde{\theta}_n$ is consistent, and converges at the same rate $m(\theta_0)^{n/2}$ as $\tilde{\theta}_n$.

2.3.4 Examples

Example 1. Let us consider the supercritical branching process with offspring law $\mathscr{P}oi(\lambda)$ with $\lambda>1$ and initial condition $Z_0=1$. Theorem 2.3.4 applies here and yields that, under $\mathbb{P}_{\lambda_0}$, on the non-extinction set E^c,

$$\begin{aligned}\hat{\lambda_n} &= \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_{i-1}}\to\lambda_0 \quad \text{a.s.},\\ \lambda_0^{n/2}(\hat{\lambda_n}-\lambda_0) &\to \sqrt{(\lambda_0-1)\lambda_0}\,\eta^{-1}N, \text{ with } \eta,N \text{ independent } N\sim\mathscr{N}(0,1),\eta^2=W,\end{aligned}$$

where $W>0$ on E^c, $EW=1, \mathrm{Var}W=\frac{1}{\lambda_0-1}$.

Example 2. Consider now the supercritical branching process with offspring law the Geometric distribution G on N^* with parameter p ($G(k)=p(1-p)^{k-1}, k\geq 1$). First, note that $\mathbb{P}_p(E)=0$ and if $Y\sim G$, $\mathbb{E}(Y)=1/p$ and $\mathrm{Var}Y=\frac{1-p}{p^2}$. Assume that $0<p<1$ and that $Z_0=1$. Theorem 2.3.4 yields that, under $\mathbb{P}_{p_0}$,

$$\frac{1}{\hat{p}_n} = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_{i-1}}\to\frac{1}{p_0} \quad \text{a. s.},$$

$$p_0{}^{-n/2}(\frac{1}{\hat{p}_n}-\frac{1}{p_0}) \quad \to \quad \sqrt{\frac{(1-p_0)^2}{p_0^3}}\,\eta^{-1}N \text{ with } \eta^2 = W, \mathbb{E}(W)=1, \text{Var}W=1.$$

To estimate p, an application of Theorem A.1.1 with $\phi(y)=1/y$ yields that, under $\mathbb{P}_{\theta_0}$,

$$\hat{p}_n = \frac{\sum_{i=1}^n Z_{i-1}}{\sum_{i=1}^n Z_i} \to p_0, \quad p_0^{-n/2}(\hat{p}_n - p_0) \to \sqrt{p_0}(1-p_0)\,\eta^{-1}W.$$

Example 3. Consider the general fractional linear branching process with offspring law G: $G(0)=a, G(k)=(1-a)p(1-p)^{k-1}, k\geq 1$. Then the mean offspring is $m=\frac{1-a}{p}$ and $\sigma^2=\frac{(1-a)}{p^2}(1-p+a)$. Assume that $m>1$, the extinction set has probability $q=\frac{a}{1-p}$. On E^c, m is estimated at rate $m^{n/2}$ while σ^2 is estimated at rate $\sqrt{n}$. Therefore, $\hat{q}_n$, which depends on $\hat{m}_n$ and $\hat{\sigma}_n^2$, is estimated at rate $\sqrt{n}$.

2.3.5 Variants of Branching processes

A large class of branching processes are used for modeling Epidemic dynamics. It encompasses subcritical or critical branching processes (a), branching processes with immigration (b), multitype branching processes with immigration, Crump-Mode-Jagers branching process, which are continuous time branching processes which are no longer Markov if the time between successive generations is not exponential (c).

Case (a) can be studied either assuming that the initial population size $\{Z_0 \to \infty\}$ or conditionally on late extinction (leading to quasi-stationary distributions). Cases (b) and (c) can be studied along similar lines than the ones in the previous section. Stating all these results is beyond the scope of these notes. We had rather choose to present accurately the simplest case, which already contains many problems arising in these other models.

Chapter 3

Inference Based on the Diffusion Approximation of Epidemic Models

3.1 Introduction

The contents of this chapter is mainly based on the three papers [61], [62] and [63].

In the first part of these notes, several mathematical models have been proposed to describe Epidemic dynamics in a closed homogeneous community. The properties of the stochastic *SEIR* model have been studied in the first part of these notes. Several mathematical formalisms were proposed to describe transitions of individuals between states: ODE/PDE ([36]), difference equations and continuous or discrete-time stochastic processes (see Part I, Sections of these notes and also [33], [36]), such as point processes, Pure jump processes, renewal processes, branching processes, diffusion processes. When data are available, key parameters can be estimated using these models through likelihood-based or M-estimation methods sometimes coupled to Bayesian methods (see e.g. [36]). However, these data are most often partially observed (e.g. infection and recovery dates are not observed for all individuals during the outbreak, not all the infectious individuals are reported) and also temporally and/or spatially aggregated. In this case, estimation via likelihood-based approaches is rarely straightforward, regardless to the mathematical formalism.

For instance, the natural modeling of epidemics by pure jump processes presents systematically the drawback that inference for such models requires that all the jumps are observed. Since these data are rarely available in practice, statistical methods rely on data augmentation in order to complete the data and add in the analysis all the missing jumps. For moderate to large populations, the complexity increases rapidly, becoming the source of additional problems. Various approaches were developed during the last years to deal with partially observed epidemics. Data augmentation and likelihood-free methods such as the Approximate Bayesian Computation (ABC) opened some of the most promising pathways for improvement (see e.g.[18], [102]). Nevertheless, these methods do not completely circumvent the issues related to incomplete data. As stated also in [28], [19], there are some limitations in practice, due to the size of missing data and to the various tuning parameters to be adjusted (see also [2], [106]). Moreover, identifiability issues are rarely addressed.

In this context, it appears that diffusion processes, satisfactorily approximating epidemic dynamics (see e.g. [46], [110]), can be profitably used for inference of model parameters from epidemiological data. In Part I, Sections **??** and **??**, the Markov jump process $(\mathscr{Z}^N(t))$ in a closed population of size N, when normalized by N, $(Z^N(t) = N^{-1}\mathscr{Z}^N(t))$ satisfies an ODE as the population size N goes to infinity. In Section **??**, it is proved the Wasserstein L_1-distance between $(Z^N(t))$ and a multidimensional diffusion process with diffusion coefficient proportional to $1/\sqrt{N}$ is of order $o(N^{-1/2})$ on a finite interval $[0,T]$. Hence, epidemic dynamics can be described using multidimensional diffusion processes $(X^N(t))_{t\geq 0}$ with a small diffusion coefficient proportional to $1/\sqrt{N}$. Since epidemics are usually observed over limited time periods, we consider in what follows the parametric inference based on observations of the epidemic dynamics on a fixed interval $[0,T]$. Let us stress that this approach

assumes a major outbreak in a large community.

Historically, statistics for diffusions were developed for continuously observed processes which renders possible getting an explicit formulation of the likelihood ([92], [97]). In this context, two asymptotics exist for estimating parameters in the drift coefficient of a diffusion continuously observed on a time interval $[0,T]$: $T \to \infty$ for recurrent diffusions and $\{T$ fixed and the diffusion coefficient tends to $0\}$. As mentioned above, in practice, epidemic data are not continuous, but partial, with various mechanisms underlying the missingness and leading to intractable likelihoods: trajectories can be discretely observed with a sampling interval (low frequency or high frequency observations, i.e. $n \to \infty$); discrete observations can correspond to integrated processes; some coordinates can be unobserved. Since the 1990s, statistical methods associated to the first two types of data have been developed (e.g. [49], [50], [55]), [87]). Recently proposed approaches for multidimensional diffusions are based on the filtering theory ([42], [51]). Concerning diffusions with small diffusion coefficient from discrete observations, it was first studied in [47], [57], [119], and more devoted to epidemic dynamics in [61], [62]. Statistical inference for diffusion processes entails some special features, that we recall for sake of clarity in A.3. It reveals that, in the context of discrete observations, it is important to distinguish parameters in the drift and parameters in the diffusion coefficients because they are not estimated at the same rate. We detail and extend here some recent work ([61], [62], [63]) where we focus on the parametric inference in the drift coefficient $b(\alpha, X^{\varepsilon}(t))$ and in the diffusion coefficient $\varepsilon\sigma(\beta, X^{\varepsilon}(t))$ of a multidimensional diffusion model $(X^{\varepsilon}(t))_{t\geq 0}$ with small diffusion coefficient, when it is observed at discrete times on a fixed time interval in the asymptotics $\varepsilon \to 0$.

Section 3.2 presents the diffusion approximation of the Markov jump process describing the epidemic dynamics starting from its Q- matrix and detail these approximations for several epidemic models studied in Part I of these notes, where another method is used to get these approximations (see Part I, Sections **??** and **??**). We then consider the parametric inference when the epidemic dynamics is observed at discrete times on a finite interval, which corresponds to one outbreak of the epidemics. The inference is studied for small sampling intervals (Section 3.4) and fixed sampling intervals (Section 3.5). On simulated data sets of two epidemic models, the *SIR* and the *SIRS* with seasonal forcing (see [84, Chapter 5]), we study the properties of our estimators based on discrete observations of these two jump Markov processes, and compare our results to the optimal inference for these jump processes, which is obtained when all the jumps (i.e. observations of all the times of infection and recovery within the population) are observed (Section 3.6).

It often occurs that in practice some components of the epidemics are not observed. In the *SIR* epidemics, the successive numbers of Susceptible for instance might be unobserved and the data consist of the successive increments of the number of Infected on each time interval. We study in Section 3.7 the inference when one coordinate of the process is observed at discrete times. We detail the results on two examples, the 2-dimensional Ornstein–Uhlenbeck diffusion process and the diffusion approximation of the *SIR*-model when only the successive numbers of Infected are available (Section 3.7.2.1). Finally, Section 3.7.2.2 is devoted to the estimation based on the real data set on Influenza epidemics, which is described by an *SIRS* epidemic model.

3.2 Diffusion approximation of jump processes modeling epidemics

This section starts from the definition of the stochastic epidemic model by a Pure jump Markov process $(\mathscr{Z}^N(t))$ on $\mathbb{Z}^d$ specified by its Q - matrix. We detail how to get the diffusion approximation of $(\mathscr{Z}^N(t))$ from this description, which is another way for getting the diffusion process obtained in Part I, Section **??** of these notes. Using limit theorems for stochastic processes, we characterize the limiting Gaussian process. Then, based on the theory of small perturbations of dynamical systems ([45]), we link the normalized process to a diffusion process with small diffusion coefficient. These approximations are then applied to *SIR*, *SEIR*, and *SIRS* models for epidemic dynamics.

3.2.1 Approximation scheme starting from the jump process Q-matrix

Let $(\mathscr{Z}^N(t))$ a multidimensional Markov jump process with state space $E \subset \mathbb{Z}^p$ which describes the epidemic dynamics in a closed population of size N, the integer "p" corresponding to the number of health states in the infection dynamics model.

This process is described by an initial distribution on E and a collection of non-negative functions $(\beta_j(t,\cdot) : E \to \mathbb{R}^+)$ indexed by $j \in \mathbb{Z}^p$, $j \neq (0,\dots,0)$, that satisfy,

$$\forall i \in E, 0 < \sum_{j\in\mathbb{Z}^p} \beta_j(t,i) = \beta(t,i) < \infty. \tag{3.2.1}$$

These functions are the transition rates of the process $(\mathscr{Z}^N(t))$ with $Q(t)$-matrix having as elements

$$q_{i,i+j}(t) = \beta_j(t,i) \text{ if } j \neq 0, \text{ and } q_{i,i}(t) = -\beta(t,i) \text{ for } i, i+j \in E. \tag{3.2.2}$$

Another useful description of $(\mathscr{Z}^N(t))$ is based on the joint distribution of its jump chain and holding times. The process stays in each state $i \in E$ during an exponential time $\mathscr{E}(\beta(t,i))$, and then jumps to the state $i+j$ according to a Markov chain (X_n) with transition probabilities $\mathbb{P}(X_{n+1} = i+j \mid X_n = i) = \beta_j(t,i)/\beta(t,i)$.

We consider the class of density dependent Markov jump processes $(\mathscr{Z}^N(t))$ which possess a limit behaviour when normalized by the population size N. Let us define the two sets

$$E = \{0,\dots,N\}^p \quad E^- = \{-N,\dots,N\}^p. \tag{3.2.3}$$

The state space of $(\mathscr{Z}^N(t))$ is E and its jumps belong to E^-.
From the original jump process $(\mathscr{Z}^N(t))$ on $E = \{0,\dots,N\}^p$, let

$$Z^N(t) = \frac{\mathscr{Z}^N(t)}{N} \text{ with state space } E_N = \{N^{-1}i, i \in E\}. \tag{3.2.4}$$

Its jumps are now $y = j/N$ and transition rates from $z \in E_N$ to $z + j/N$ at time t defined using (3.2.2),

$$q^N_{z,z+y}(t) = \beta_{Ny}(t, Nz). \tag{3.2.5}$$

Denote for $x = (x_1,\dots,x_p) \in \mathbb{R}^d$, $[x] = ([x_1],\dots,[x_p]) \in \mathbb{Z}^p$, where $[x_i]$ is the integer part of x_i.

We assume in the sequel that $(\mathscr{Z}^N(t))$ is density dependent, i.e. there exist a collection of functions $\beta_j : \mathbb{R}^+ \times [0,1]^p \to \mathbb{R}^+$ such that,

(H1) $\forall j$, $\forall z \in [0,1]^p$ $\frac{1}{N}\beta_j(t,[Nz]) \to \beta_j(t,z)$ as $N \to \infty$ locally uniformly in t.

(H2) $\forall j \in E^-$, $\beta_j(t,z) \in C^2(\mathbb{R}^+, [0,1]^p)$.

Then, define the two functions $b^N(t,z)$ and $b(t,z) : \mathbb{R}^+ \times [0,1]^p \to \mathbb{R}^p$ and the two $p \times p$ positive symmetric matrices Σ^N and Σ (with the notation $M^\star$ for the transposition of a matrix or of a column vector j in E),

$$b^N(t,z) = \frac{1}{N}\sum_{j\in E^-} \beta_j(t,[Nz])j; \quad b(t,z) = \sum_{j\in E^-} \beta_j(t,z)j; \tag{3.2.6}$$

$$\Sigma^N(t,z) = \frac{1}{N}\sum_{j\in E^-} \beta_j(t,[Nz])jj^\star; \quad \Sigma(t,z) = \sum_{j\in E^-} \beta_j(t,z)jj^\star. \tag{3.2.7}$$

Under (H1) the functions $b(t,z)$ and $\Sigma(t,z)$ are well defined and $b(t,z)$ is Lipschitz under (H2). Therefore, there exists a unique smooth solution $z(t)$ to the ODE

$$\frac{dz}{dt} = b(t,z(t))dt\,; \quad z(0) = x. \tag{3.2.8}$$

Let $\nabla_z b(t,z)$ denote the gradient of $b(t,z)$

$$\nabla_z b(t,z) = \Big(\frac{\partial b_i}{\partial z_j}(t,z)\Big)_{1\le i,j\le p}. \tag{3.2.9}$$

The resolvent matrix $\Phi(t,u)$ associated with (3.2.8) is defined as the solution

$$\frac{d\Phi}{dt}(t,s) = \nabla_z b(t,z(t))\Phi(t,s); \quad \Phi(s,s) = I_p. \tag{3.2.10}$$

Under (H1), (H2) the following holds: if $Z^N(0) \to x$ as $N \to \infty$, then, locally uniformly in t,

$$\forall t \geq 0, \lim_{N\to\infty} \| Z^N(t) - z(t) \| = 0 \text{ a.s.} \tag{3.2.11}$$

where $z(t)$ is solution of (3.2.8).
Let $(D, \mathscr{D})$ denote the space of "cadlag" functions $\{f : \mathbb{R}^+ \to \mathbb{R}^p\}$ endowed with the Skorokhod topology. Then,

$$\sqrt{N}(Z^N(t) - z(t))_{t\geq 0} \to (G(t))_{t\geq 0} \text{ in distribution in } (D, \mathscr{D}), \tag{3.2.12}$$

where $(G(t))$ is a centered p-dimensional Gaussian process with covariance matrix

$$\mathrm{Cov}(G(t), G(r)) = \int_0^{t\wedge r} \Phi(t,u)\Sigma(u,z(u))\ \Phi^\star(r,u)du. \tag{3.2.13}$$

The proofs of these results are given under a general form in Part I, Sections **??** and **??** of these notes, and based on this presentation in [61], [62].

Heuristically, there is an approach which yields the diffusion approximation of $(Z^N(t))$; it rests on an expansion of the generator $\mathscr{A}_N$ of $(Z^N(t))$ (3.2.4). For $f \in C^2(\mathbb{R}^+ \times \mathbb{R}^p, \mathbb{R})$, it reads as

$$\mathscr{A}_N f(t,z) = \sum_{j\in E^-} \beta_j(t,Nz)(f(t,z+\frac{j}{N}) - f(t,z)).$$

A Taylor expansion of $\mathscr{A}_N f(t,z)$ yields, using (H1), (H2) and (3.2.6), for $j = (j_1,\ldots,j_p)^* \in E^-$,

$$\begin{aligned}
\mathscr{A}_N f(t,z) &= \sum_{j\in E^-} N\beta_j(t,z)(f(t,z+\frac{j}{N}) - f(t,z)) + o(1/N) \\
&= (\nabla_z f(t,z))^\star\, b(t,z) + \frac{1}{2N}\left(\sum_{j\in E^-} \beta_j(t,z) \sum_{k,l=1}^d j_k\, j_l\, \nabla^2_{z_k z_l} f(t,z)\right) \\
&\qquad + o(1/N) \\
&= (\nabla_z f(t,z))^\star\, b(t,z) + \frac{1}{2N}\sum_{k,l=1}^d \Sigma_{kl}(t,z)\, \nabla^2_{z_k z_l} f(t,z) + o(1/N),
\end{aligned}$$

where the last equality is obtained using (3.2.7). The first two terms of the last expression correspond to the generator of a diffusion process on $\mathbb{R}^p$ with drift coefficient $b(t,\cdot)$ and diffusion matrix $\frac{1}{N}\Sigma(t,\cdot)$,

$$dX^N(t) = b(t,X^N(t))dt + \frac{1}{\sqrt{N}}\sigma(t,X^N(t))dB(t)\ ;\ X^N(0) = x, \tag{3.2.14}$$

where $(B(t)_{t\geq 0})$ is a Brownian motion on $\mathbb{R}^p$ defined on a probability space $(\Omega, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ independent of $X^N(0)$, and $\sigma(t,\cdot)$ is a square root of $\Sigma(t,\cdot)$: $\sigma(t,z)\ \sigma(t,z)^\star = \Sigma(t,z)$.

These approaches can be connected together a posteriori using the theory of random perturbations of dynamical systems ([7], [45]) and the following theorem.

Theorem 3.2.1. *Setting $\varepsilon = 1/\sqrt{N}$, the paths of $X^N(\cdot)$ satisfy, as $\varepsilon \to 0$,*

$$X^N(t) = X^\varepsilon(t) = z(t) + \varepsilon g(t) + \varepsilon^2 R^\varepsilon(t), \ \textit{with} \sup_{t\leq T} \| \varepsilon R^\varepsilon(t) \| \to 0 \textit{ in probability,} \tag{3.2.15}$$

where $z(t)$ is the solution of (3.2.8)*, $B(t)$ is a p-dimensional Brownian motion and $(g(t))$ is the process satisfying the SDE*

$$dg(t) = \nabla_z b(t,z(t))g(t)dt + \sigma(t,z(t))dB(t), \ g(0) = 0.$$

This stochastic differential equation can be solved explicitly and we get, using (3.2.10), that

$$g(t) = \int_0^t \Phi(t,s)\sigma(s,z(s))dB(s). \tag{3.2.16}$$

Hence, $(g(t))$ is a centered Gaussian process having the same covariance matrix (3.2.13) as the process $(G(t))$ defined in (3.2.12). Therefore, for $\varepsilon = 1/\sqrt{N}$, $\sqrt{N}(Z_t^N - z(t))_{t\geq 0}$ and $\varepsilon^{-1}\left(Z^\varepsilon(t) - z(t)\right)_{t\geq 0}$ converge to a Gaussian process having the same distribution.

It is moreover proved in Part I, Section **??** of these notes, that the Wasserstein L^1 distance between $(Z^N(t))$ and $(X^N(t))$ converges to 0.

3.2.2 Diffusion approximation of some epidemic models

3.2.2.1 The diffusion approximation applied to the SIR epidemic model

We apply first the generic method leading successively to $b(\cdot)$, $\Sigma(\cdot)$ and (X^N) described in 3.2.1 to the *SIR* model introduced iin Part I, Chapter **??** of these notes through the 2-dimensional continuous-time Markov jump process $\mathscr{Z}^N(t) = (S(t), I(t))$ to build the associated *SIR* diffusion process. Along to its initial state $\mathscr{Z}^N(0) = (S(0), I(0))$, the Markov jump process is characterized by two transitions, $(S,I) \xrightarrow{\frac{\lambda}{N} SI} (S-1, I+1)$ and $(S,I) \xrightarrow{\gamma I} (S, I-1)$. Parameters λ and $\gamma = 1/d$ represent the transmission rate and the recovery rate (or the inverse of the mean infection duration d), respectively. The rate $\lambda SI/N$ translates two main assumptions: the population mix homogeneously (same λ for each pair between one S and one I) and the transmission is proportional to the fraction of infectious individuals in the population, I/N (frequency-dependent formulation of the transmission term).

The diffusion approximation of the process$(\mathscr{Z}^N(t))$ describing the epidemic dynamics can be summarized by three steps. The original *SIR* jump process in a closed population has state space $\{0,\ldots,N\}^2$, the jumps j are $(-1,1)$ and $(0,-1)$ with transition rates,

$$q_{(S,I),(S-1,I+1)} = \lambda S\,\frac{I}{N} = \beta_{(-1,1)}(S,I); \quad q_{(S,I),(S,I-1)} = \gamma I = \beta_{(0,-1)}(S,I).$$

Normalizing $(\mathscr{Z}^N(t))$ by the population size N, we obtain, setting $z = (s,i) \in [0,1]^2$, as $N \to \infty$,

$$\frac{1}{N}\beta_{(-1,1)}([Nz]) \to \beta_{(-1,1)}(s,i) = \lambda si; \quad \frac{1}{N}\beta_{(0,-1)}([Nz]) \to \beta_{(0,-1)}(s,i) = \gamma i.$$

These two limiting functions clearly satisfy (H1)–(H2). Finally, the two functions given in (3.2.6), (3.2.7) are well defined and now depend on (λ,γ).
Set $\theta = (\lambda,\gamma)$ and denote by $b(\theta,z)$ and $\Sigma(\theta,z)$ the associated functions. We get

$$b(\theta,(s,i)) = \begin{pmatrix} -\lambda si \\ \lambda si - \gamma i \end{pmatrix}; \quad \Sigma(\theta,(s,i)) = \begin{pmatrix} \lambda si & -\lambda si \\ -\lambda si & \lambda si + \gamma i \end{pmatrix}. \tag{3.2.17}$$

Assume that $\mathscr{Z}^N(0)$ satisfies $(N^{-1}S(0), N^{-1}I(0)) \to x = (s_0, i_0)$ with $s_0 > 0$, $i_0 > 0$, $s_0 + i_0 \leq 1$ as $N \to \infty$. Then the associated ODE is, using (3.2.8),

$$\frac{ds}{dt} = -\lambda si\,; \quad \frac{di}{dt} = \lambda si - \gamma i\,; \quad (s(0), i(0)) = (s_0, i_0). \tag{3.2.18}$$

The diffusion approximation of the *SIR* epidemics obtained in (3.2.14) is the solution of the SDE

$$\begin{aligned} dS^N(t) &= -\lambda S^N(t)I^N(t)dt + \frac{1}{\sqrt{N}}\sqrt{\lambda S^N(t)I^N(t)}\,dB_1(t), \quad S^N(0) = s_0, \\ dI^N(t) &= (\lambda S^N(t)I^N(t) - \gamma I^N(t))dt - \frac{1}{\sqrt{N}}\left(\sqrt{\lambda S^N(t)I^N(t)}\,dB_1(t) - \sqrt{\gamma I^N(t)}\,dB_2(t)\right), \\ I^N(0) &= i_0, \end{aligned}$$

where $(B(t))$ is standard two-dimensional Brownian motion and $\sigma(\theta,z)$ corresponds to the Choleski decomposition of $\Sigma(\theta,z) = \sigma(\theta,z)\sigma^\star(\theta,z)$,

$$\sigma(\theta,(s,i)) = \begin{pmatrix} \sqrt{\lambda si} & 0 \\ -\sqrt{\lambda si} & \sqrt{\gamma i} \end{pmatrix}.$$

In order to visualize the influence of the population size N on the sample paths of the normalized jump process $Z_N(t) = \mathscr{Z}^N(t)/N$, several trajectories have been simulated using an *SIR* model with parameters $(\lambda,\gamma) = (0.5, 1/3)$, so that $R_0 = \lambda/\gamma = 1.5$. Results are displayed in Figure 3.2.1. We observe that, as the population size increases, the stochasticity of sample paths decreases. However, it still keeps a non-negligible stochasticity for a large population size ($N = 10000$). Since the peak of $I^N(t)$ is quite small (about 0.08 here), this can be explained by a moderate size of the ratio "signal over noise" even for large N (here of order $0.08/0.01$).

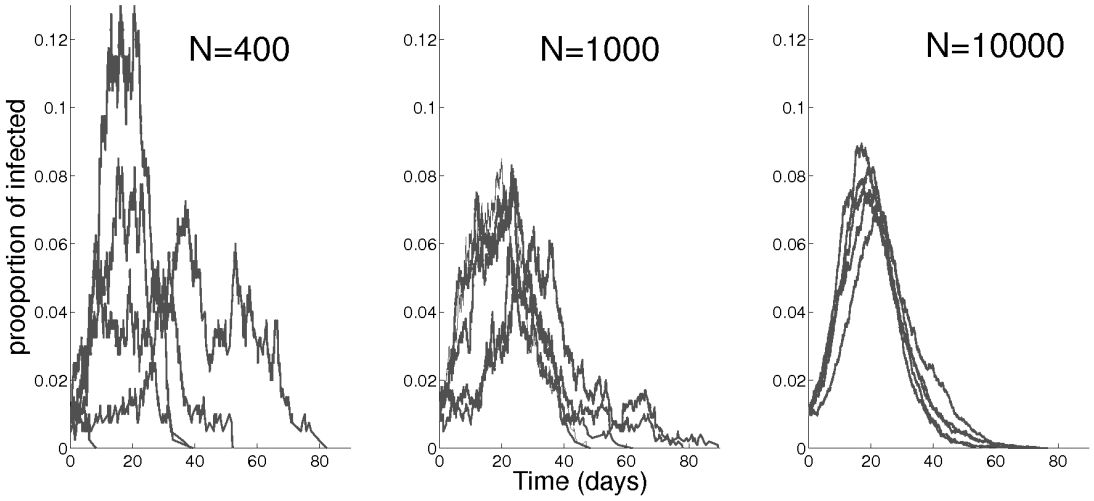


Figure 3.2.1: Five simulated trajectories of the proportion of infectious individuals over time using the *SIR* Markov jump process for $(s_0,i_0) = (0.99, 0.01)$ $(\lambda,\gamma) = (0.5, 1/3)$ and for each $N = \{400, 1000, 10000\}$ (from left to right).

3.2.2.2 The diffusion approximation applied to the *SIRS* epidemic model with seasonal forcing

Another important class of epidemics models is the *SIRS* model, which allows possible reinsertion of removed individuals into S class. The additional transition reads as $(S,I) \overset{\delta(N-S-I)}{\longrightarrow} (S+1,I)$, where δ is the average rate of immunity waning. To mimic recurrent epidemics, additional mechanisms need to be considered. Indeed, to avoid that successive epidemics cycles die out, one way is to introduce an external immigration flow. Hence, one possible model to describe recurrent epidemics is the *SIRS* model with seasonal transmission (at rate $\lambda(t)$), external immigration flow in the I class (at rate η) and, when the time-scale of study is large, demography (with birth and death rates equal to μ for a stable population of size N). Seasonality in transmission is captured using a time non-homogeneous transmission rate, expressed under a periodic form

$$\lambda(t) := \lambda_0(1 + \lambda_1 \sin(2\pi t/T_{per})) \tag{3.2.19}$$

where λ_0 is the baseline transition rate, λ_1 the intensity of the seasonal effect on transmission and T_{per} is introduced to model an annual or t seasonal trend (see [84], Chapter 5). Typically for modeling Influenza epidemics, we fixed it at $T = 365$.

Assuming again a constant population size, we obtain a new two-dimensional system with four transitions for the corresponding Markov jump process:

$$(S,I) \overset{\frac{\lambda(t)}{N}S(I+N\eta)}{\longrightarrow} (S-1,I+1) \quad ; \qquad (S,I) \overset{\mu S}{\longrightarrow} (S-1,I);$$
$$(S,I) \overset{(\gamma+\mu)I}{\longrightarrow} (S,I-1) \quad ; \qquad (S,I) \overset{\mu N+\delta(N-S-I)}{\longrightarrow} (S+1,I).$$

Figure 3.2.2 illustrates the dynamics of the *SIRS* model (in ODE formalism) which is forced using sinusoidal terms. In particular, given the parameter values we have chosen, we can notice two distinct regimes: one with annual cycles (top graph) and the other with biennial dynamics (middle graph). The qualitative changes in model dynamics are explored by modifying a control parameter or *bifurcation parameter* (here λ_1) and constructing a *bifurcation diagram*.

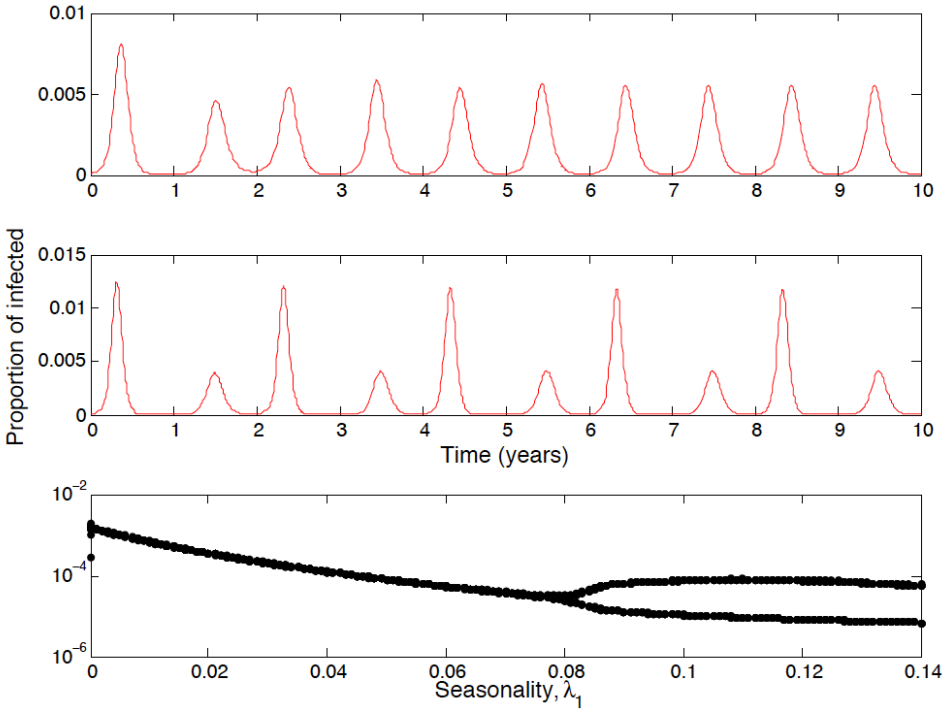


Figure 3.2.2: Proportion of infected individuals, $I(t)$, over time (top and middle panels) simulated using the ODE variant of the *SIRS* model with $N = 10^7$, $T_{per} = 365$, $\mu = 1/(50 \times T_{per})$, $\eta = 10^{-6}$, $(s_0, i_0) = (0.7, 10^{-4})$ and $(\lambda_0, \gamma, \delta) = (0.5, 1/3, 1/(2 \times 365))$. The top panel corresponds to $\lambda_1 = 0.05$, the middle panel to $\lambda_1 = 0.1$. The bottom panel represents the bifurcation diagram with respect to λ_1.

The diffusion approximation is built following the same generic scheme of Section 3.2.1 as for the *SIR* model in Section 3.2.2.1. The four jumps j corresponding to functions β_j are $j^* = (-1,1);(-1,0);(0,-1);(1,0)$ leading to

$$\beta_{(-1,1)}(t,S,I) = \frac{\lambda(t)}{N}S(I+N\eta), \quad \beta_{(0,-1)}(t,S,I) = (\gamma+\mu)S,$$
$$\beta(0,-1)(t,S,I) = (\gamma+\mu)S, \quad \beta_{(1,0)}(t,S,I) = \mu N + \delta(N-S-I)S.$$

The jump process is time-dependent and so we have to check (H1b)–(H2). Straightforward computations yield that they are satisfied since, for $(s,i) \in [0,1]^2$,

$$\beta_{(-1,1)}(t,(s,i)) = \lambda(t)s(i+\eta); \quad \beta_{(-1,0)}(t,(s,i)) = \mu s;$$

$$\beta_{(0,-1)}(t,(s,i)) = (\gamma+\mu)i; \qquad \beta_{(1,0)}(t,(s,i)) = \mu+\delta(1-s-i).$$

Finally, setting $\theta = (\lambda_0,\lambda_1,\gamma,\delta,\eta,\mu)$, the associated drift function $b(\theta,t,(s,i))$ and diffusion matrix $\Sigma(\theta,t,(s,i))$ are

$$b(\theta,t,(s,i)) = \begin{pmatrix} -\lambda(t)s(i+\eta)+\delta(1-s-i)+\mu(1-s) \\ \lambda(t)s(i+\eta)-(\gamma+\mu)i \end{pmatrix}, \tag{3.2.20}$$

$$\Sigma(\theta,t,(s,i)) = \begin{pmatrix} \lambda(t)s(i+\eta)+\delta(1-s-i)+\mu(1+s) & -\lambda(t)s(i+\eta) \\ -\lambda(t)s(i+\eta) & \lambda(t)s(i+\eta)+(\gamma+\mu)i \end{pmatrix}. \tag{3.2.21}$$

Therefore, the associated ODE is, using (3.2.20),

$$\begin{aligned} \frac{ds}{dt} &= -\lambda(t)s(i+\eta)+\delta(1-s-i)+\mu(1-s), \quad s(0)=s_0; \\ \frac{di}{dt} &= \lambda(t)s(i+\eta)-(\gamma+\mu)i, \quad i(0)=i_0. \end{aligned}$$

Choosing $\sigma(\theta,t,(s,i))$ such that $\sigma(\theta,t,(s,i))\sigma(\theta,t,(s,i))^\star = \Sigma(\theta,t,(s,i))$, we obtain that the approximating diffusion $X_N(t)$ satisfies

$$dX^N(t) = b(\theta,t,(S_N,I_N))dt + \frac{1}{\sqrt{N}}\sigma(\theta,t(S^N,I^N));\ X^N(0)=x. \tag{3.2.22}$$

3.2.2.3 A Minimal model for Ebola Transmission with temporal transition rate

According to [21], a basic model for Ebola dynamics consists in a *SEIR* model with temporal transmission rate. In a rough approximation, assuming homogeneous mixing in a size N community yields, setting $\mathscr{Z}^N(t) = (S,E,I)$,

$$\begin{aligned} &(S,E,I) \xrightarrow{\lambda(t)\frac{SI}{N}} (S-1,E+1,I); \\ &(S,E,I) \xrightarrow{\nu E} (S,E-1,I+1); \\ &(S,E,I) \xrightarrow{\gamma I} (S,E,I-1). \end{aligned}$$

The diffusion approximation has drift and diffusion matrix given by, for $z=(s,e,i)$,

$$b(\theta,t,z) = \begin{pmatrix} -\lambda(t)si \\ \lambda(t)si-\nu e \\ \nu e-\gamma i \end{pmatrix}; \quad \Sigma(\theta,t,z) = \begin{pmatrix} \lambda(t)si & -\lambda(t)si & 0 \\ -\lambda(t)si & \lambda(t)si+\nu e & -\nu e \\ 0 & -\nu e & \nu e+\gamma i \end{pmatrix}.$$

Two questions concerning the inference arise in this model: the non-parametric estimation of $\lambda(\cdot)$ and the presence of random effects since the dynamics are observed in different locations.

3.2.2.4 Two variants of the *SEIRS* model with demography

In Part I, Chapter **??** of these notes, an example of *SEIRS* model with demography is proposed (see Example **??**). Removed individuals loose their immunity at rate δ; there is an influx of susceptible at rate μN and individuals, whichever type, die at rate μ. Hence, 9 jumps are present in this model, for (s,e,i,r), which yields for $Z=(S,E,I)$,

$$(S,E,I) \xrightarrow{\lambda\frac{SI}{N}} (S-1,E+1,I), \quad (S,E,I) \xrightarrow{\nu E} (S,E-1,I+1),$$

$$(S,E,I) \xrightarrow{\mu N+\delta(N-S-E-I)} (S+1,E,I), \quad (S,E,I) \xrightarrow{\mu I+\gamma I} (S,E,I-1).$$

$$(S,E,I) \xrightarrow{\mu S} (S-1,E,I), \quad (S,E,I) \xrightarrow{\mu E} (S,E-1,I), \quad (S,E,I) \xrightarrow{\mu(N-S-E-I)} (S,E,I).$$

This yields, setting $z=(s,e,i)$ and $\theta=(\lambda,\nu,\gamma,\delta,\mu)$

$$b(\theta,z) = \begin{pmatrix} -\lambda si+\mu(1-s)+\delta(1-s-i-e) \\ \lambda si-(\mu+\nu)e \\ \nu e-(\gamma+\nu)i \end{pmatrix};$$

$$\Sigma(\theta,z) = \begin{pmatrix} \lambda si+\mu(1+s)+\delta(1-s-i-e) & -\lambda si & 0 \\ -\lambda si & \lambda si+(\mu+\nu)e & -\nu e \\ 0 & -\nu e & \nu e+(\gamma+\nu)i \end{pmatrix}.$$

3.3 Inference for discrete observations of diffusions on [0,T]

Our concern here is parametric inference for these models. Statistical inference for discretely observed diffusion processes present some specific properties (see Section A.3 in the Appendix) that lead us to consider distinct parameters in the drift coefficient (here α) and in the diffusion coefficient (β). The state space of the diffusion is $\mathbb{R}^p$, and the parameter set Θ is a subset of $\mathbb{R}^a\times\mathbb{R}^b$, with $\alpha\in\mathbb{R}^a,\beta\in\mathbb{R}^b$. For instance, the *SIR* diffusion approximation corresponds to $p=2$ and $\alpha=\beta=(\lambda,\gamma)$.

In order to deal with general epidemics, we consider time-dependent diffusion processes on $\mathbb{R}^p$ with small diffusion coefficient $\varepsilon=1/\sqrt{N}$ satisfying the stochastic differential equation (SDE):

$$dX(t)=b(\alpha,t,X(t))dt+\varepsilon\sigma(\beta,t,X(t))\,dB(t)\,;\quad X(0)=x, \tag{3.3.1}$$

where $(B(t)_{t\geq 0})$ is a p-dimensional Brownian motion defined on a probability space $(\Omega,(\mathscr{F}_t)_{t\geq 0},\mathbb{P})$, $b(\alpha,t,\cdot):\mathbb{R}^p\to\mathbb{R}^p$ and $\sigma(\beta,t,\cdot):\mathbb{R}^p\to\mathbb{R}^p\times\mathbb{R}^p$ and x is non-random fixed.

Since epidemic dynamics are usually observed at discrete times, our aim is to study the estimation of $\theta=(\alpha,\beta)$ based on the observations

$$(X(t_k),k=1\ldots n) \text{ with } t_k=k\Delta;\ T=n\Delta \quad (\text{sampling interval } \Delta). \tag{3.3.2}$$

For observations on a fixed time interval, $[0,T]$, there are distinct asymptotic results according to Δ.

(1) High frequency sampling $\Delta=\Delta_n\to 0$: The number of observations $n=T/\Delta_n$ goes to ∞ while $T=n\Delta_n$ is fixed. There is a double asymptotic framework: $\varepsilon\to 0$ and $\Delta\to 0$ (or $n=T/\Delta\to\infty$) simultaneously. Let us stress that we shall use both notations for this second asymptotics $n\to\infty$ or $\Delta\to 0$. Although it might be confusing, it is sometimes better to state results according to the number of observations and sometimes according to the sampling interval Δ.

(2) Low frequency sampling Δ is fixed: It leads to a finite number of observations $n=T/\Delta$. Results are obtained in the asymptotic framework $\varepsilon\to 0$.

At first glance, the low frequency sampling seems a priori a suitable framework for epidemic data. However, both high and low frequency observations could be appropriate in practice because the choice of the statistical framework depends more on the relative magnitudes between T, Δ and the population size N ($=\varepsilon^{-2}$) than on their accurate values.

From a statistical point of view, the sequence $(X(t_k))$ is a time-dependent Markov chain and therefore the likelihood depends on its transition probabilities. However, the link between the parameters present in the SDE and the transition probabilities of $(X(t_k))$ is generally not explicit, which leads to intractable likelihoods. This is a well known problem for discrete observations of diffusion processes. Alternative approaches based on M-estimators or contrast functions (see e.g. [124] for independent random variables, [88] for stochastic processes) have to be investigated (see also the recap presented in Section A.3 in the Appendix of this part).

After the statement of some preliminary results, we present successively the statistical inference for high frequency sampling, where the asymptotics is $\varepsilon=1/\sqrt{N}\to 0$, $\Delta_n=T/n\to 0$ (Section 3.4), and for the low frequency sampling, $\varepsilon=1/\sqrt{N}\to 0$, Δ fixed (Section 3.5).

3.3.1 Assumptions, notations and first results

Let θ_0 be the true value of the parameter and Θ the parameter set. Denote by $\mathscr{M}_p(\mathbb{R})$ the set of $p\times p$ matrices. We first assume that $b(\alpha,t,z)$ and $\sigma(\beta,t,z)$ are measurable in (t,z), Lipschitz continuous with respect to the second variable and satisfy a linear growth condition: for all $t\geq 0, z, z'\in\mathbb{R}^p$, there exists a global constant K such that

(S1): $\forall\theta\in\Theta, \parallel b(\alpha,t,z)-b(\alpha,t,z')\parallel+\parallel\sigma(\beta,t,z)-\sigma(\beta,t,z')\parallel\leq K\parallel z-z'\parallel$.

(S2): $\forall(\alpha,\beta)\in\Theta, \parallel b(\alpha;t,z)\parallel^2+\parallel\sigma(\beta;t,z)\parallel^2\leq K(1+\parallel z\parallel^2)$.

(S3): $\forall(\beta,t,z)$, $\Sigma(\beta;t,z)=\sigma(\beta;t,z)\sigma^\star(\beta;t,z)$ is non-singular.

Assumptions (S1)–(S3) are classical assumptions that ensure that, for all θ, (3.3.1) admits a unique strong solution (see e.g. [83, Chapter 5.2.B]).

Another set of assumptions is required for the inference:

(S4): $\Theta=K_a\times K_b$ is a compact set of $\mathbb{R}^{a+b}$, $\theta_0\in\mathrm{Int}(\Theta)$.

(S5): For all $t\geq 0$, $b(\alpha;t,z)\in C^3(K_a\times\mathbb{R}^+\times\mathbb{R}^p,\mathbb{R}^p)$ and $\sigma(\beta;t,z)\in C^2(K_b\times\mathbb{R}^+\times\mathscr{M}_p(\mathbb{R}))$.

(S6): $\alpha\neq\alpha'\Rightarrow b(t;\alpha,z(\alpha,t))\not\equiv b(t;\alpha',z(\alpha',t))$.

(S7): $\beta\neq\beta'\Rightarrow\Sigma(t;\beta,z(\alpha_0,t))\not\equiv\Sigma(t;\beta',z(\alpha_0,t))$.

Assumptions **(S4)–(S5)** are classical for the inference for diffusion processes. Usually, it is sufficient in **(S5)** to deal with C^2 functions. The additional differentiability condition comes from regularity conditions required on $\alpha\to\Phi(\alpha,t,s)$. Indeed, **(S5)** on $b(\alpha,t,z)$ ensures that the function $\Phi(\alpha,t,t_0)$ belongs to $C^2(K_a\times[0,T]^2,\mathscr{M}_p)$. Assumption **(S6)** is the usual identifiability assumption for a diffusion continuously observed on $[0,T]$ and **(S7)** is an identifiability assumption for parameters in the diffusion coefficient.

Since $(X(t))$ is a diffusion process on $(\Omega,(\mathscr{F}_t)_{t\geq0},\mathbb{P})$, the space of observations is $(C_T=C([0,T],\mathbb{R}^p),\mathscr{C}_T)$ where $\mathscr{C}_T$ is the Borel σ-algebra on $C([0,T],\mathbb{R}^p)$. Let $\mathbb{P}_\theta=\mathbb{P}_{\alpha,\beta}$ the probability distribution on $(C_T,\mathscr{C}_T)$ of $(X(t)),0\leq t\leq T)$ satisfying (3.3.1). Let $\mathscr{G}_k^n$ denote the σ-algebra $\sigma(X(s),s\leq\frac{kT}{n})$.

For $g(\theta,t,z):\Theta\times[0,T]\times\mathbb{R}^p\to\mathbb{R}^p$, $\nabla_z g(\cdot)$ is the $\mathscr{M}_p$ matrix

$$\nabla_z g(\cdot)=(\frac{\partial g_i}{\partial z_j}(\theta,t,z))_{1\leq i,j\leq p}\quad\text{and}\quad\nabla_\theta g(\cdot)=(\frac{\partial g_i}{\partial\theta_j}(\theta,t,z)) \tag{3.3.3}$$

If $z=z(\theta,t)$, then

$$\nabla_\theta\left(g(\theta,t,z(\theta,t))\right)=\nabla_\theta g(\cdot)+\nabla_z g(\cdot)\nabla_\theta z(\cdot). \tag{3.3.4}$$

Quantities are indexed by θ (resp. α or β) when they depend on both α,β (resp.α or β). Introducing the dependence with respect to t,θ in (3.2.8), (3.2.10), (3.2.16) yields,

$$\begin{array}{rl}\frac{\partial z}{\partial t}(\alpha,t)= & b(\alpha,t,z(\alpha,t));\quad z(\alpha,0)=x_0,\\ g(\alpha,\beta,t)= & \int_0^t\Phi(\alpha,t,u)\sigma(\beta,u,z(\alpha,u))dB(u),\quad\text{with }\Phi(\alpha,\cdot)\text{ such that}\\ \frac{\partial\Phi}{\partial t}(\alpha,t,u)= & \nabla_z b(\alpha,t,z(\alpha,t))\Phi(\alpha,t,u)\ ,\ \Phi(\alpha,u,u)=I_p.\end{array} \tag{3.3.5}$$

The expansion (3.2.15) holds for time-dependent diffusion processes.

Proposition 3.3.1. *Assume (S1)–(S5). Then, under $\mathbb{P}_\theta$, $(X(t),0\leq t\leq T)$ satisfies that, uniformly with respect to θ,*

$$\begin{array}{rl}X(t)= & z(\alpha,t)+\varepsilon g(\theta,t)+\varepsilon^2R^\varepsilon(\theta,t),\quad\textit{with}\\ lim_{\varepsilon\to0,r\to\infty} & \mathbb{P}_\theta(\sup_{t\leq T}\parallel R^\varepsilon(\theta,t)\parallel>r)=0\\ \sup_{t\leq T}\parallel R^\varepsilon(\theta,t)\parallel & \textit{has uniformly bounded moments.}\end{array} \tag{3.3.6}$$

Proposition 3.3.2. *Under (S1)–(S5), the process $(g(\theta,t))$ satisfies using (3.3.5)*

$$\forall s<t, \quad g(\theta,t)=\Phi(\alpha,t,s)g(\theta,s)+\int_s^t \Phi(\alpha,t,u)\sigma(\beta,u,z(\alpha,u))dB(u),$$

where the two terms of the r.h.s. above are independent random variables.

Proposition 3.3.3. *Assume (S1)–(S2). If moreover $b(\alpha,\cdot)$ and $\sigma(\beta,\cdot)$ have uniformly bounded derivatives, there exist constants only depending on T and θ such that*

(i) $\forall t\in[0,T]$, $\mathbb{E}_\theta(\|R^\varepsilon(\theta,t)\|^2<C_1$,

(ii) $\forall t\in[0,T]$, *as* $h\to 0$, $\mathbb{E}_\theta(\|R^\varepsilon(\theta,t+h)-R^\varepsilon(\theta,t)\|^2)<C_2h$.

We refer to [7], [45], and [57] for the proofs of these propositions for θ fixed. Assumption (S4) allows us to extend these results to $\theta\in\Theta$.

3.3.2 Preliminary results

Let us define using (3.3.5) the random variables,

$$B_k(\alpha,X)=X(t_k)-z(\alpha,t_k)-\Phi(\alpha,t_k,t_{k-1})\left[X(t_{k-1})-z(\alpha,t_{k-1})\right]. \tag{3.3.7}$$

Then the following holds.

Lemma 3.3.4. *Assume (S1)–(S4). Then, under $\mathbb{P}_\theta$, as $\varepsilon\to 0$,*

$$\begin{aligned} B_k(\alpha,X) &= \varepsilon\sqrt{\Delta}\,T_k(\theta)+\varepsilon^2 D_k^\varepsilon(\theta,\Delta) \quad \textit{with } \sup_k \mathbb{E}_\theta||D_k^\varepsilon(\theta,\Delta)||^2\le C\Delta, \textit{ and} \\ T_k(\theta) &= \frac{1}{\sqrt{\Delta}}\int_{t_{k-1}}^{t_k}\Phi(\alpha,t_k,u)\sigma(\beta,u,z(\alpha,u))dB(u), \end{aligned}$$

where C a constant independent of $\theta,\varepsilon,\Delta$.

Therefore, the random variables $T_k(\theta)$ are p-dimensional $\mathscr{G}_k^n$-measurable independent Gaussian random variables with covariance matrix

$$S_k(\alpha,\beta)=S_k(\theta)=\frac{1}{\Delta}\int_{t_{k-1}}^{t_k}\Phi(\alpha,t_k,s)\Sigma(\beta,s,z(\alpha,s))\Phi^\star(\alpha,t_k,s)ds. \tag{3.3.8}$$

Proof. Using Propositions 3.3.1 and 3.3.2 yields that

$$\begin{aligned} D_k^\varepsilon(\theta,\Delta) &= R^\varepsilon(\theta,t_k)-\Phi(\alpha,t_k,t_{k-1})R^\varepsilon(\theta,t_{k-1}) \\ &= R^\varepsilon(\theta,t_k)-R^\varepsilon(\theta,t_{k-1})-(\Phi(\alpha,t_k,t_{k-1})-I_p)R^\varepsilon(\theta,t_{k-1}) \\ &= R^\varepsilon(\theta,t_k)-R^\varepsilon(\theta,t_{k-1})-\Delta\nabla_z b(\alpha,t_{k-1},z(\alpha,t_{k-1}))R^\varepsilon(\theta,t_{k-1}) \\ &\quad +\Delta^2 O_P(1). \end{aligned}$$

An application of Proposition 3.3.3 together with (S4) yields the result. □

Define the two random matrices

$$\Sigma_k(\beta)=\Sigma(\beta,t_k,X(t_k)), \quad \sigma_k(\beta)=\sigma(\beta,t_k,X(t_k)). \tag{3.3.9}$$

Then, for small Δ, we have using (3.3.9)

Lemma 3.3.5. *Assume (S1)–(S5). Then, under $\mathbb{P}_\theta$, as $\varepsilon,\Delta\to 0$,*

$$||S_k(\theta)-\Sigma_{k-1}(\beta)||\le K\varepsilon\sup_{\theta,t\le T}||g(\theta,t)||+\Delta\sup_{\theta,z}||\nabla_z\Sigma(\beta,s,z)||\le\varepsilon C_1O_P(1)+C_2\Delta.$$

The proof is straightforward using (S1), (S5) and Proposition (3.3.6).

Let us now state some preliminary results on the random variables $B_k(\alpha, X)$ defined in (3.3.7) useful for the inference.

Under $\mathbb{P}_{\theta_0}$, Proposition 3.3.1 yields that $B_k(\alpha, X)$ converges to $B_k(\alpha, z(\alpha_0, \cdot))$ and that $B_k(\alpha_0, X)$ converges to $B_k(\alpha_0, z(\alpha_0, \cdot)) = 0$ a.s. Let us define on $[0, T]$

$$\Gamma(\alpha_0, \alpha, t) = b(\alpha_0, t, z(\alpha_0, t)) - b(\alpha, t, z(\alpha, t)) - \nabla_z b(\alpha, t, z(\alpha, t))(z(\alpha_0, t) - z(\alpha, t)). \tag{3.3.10}$$

The sequence $B_k(\alpha, z(\alpha_0, \cdot))$ satisfies:

Lemma 3.3.6. *Assume (S1), (S2), (S4). Then, as $\Delta \to 0$, there exists a constant C such that*

$$\frac{1}{\Delta} B_k(\alpha, z(\alpha_0, \cdot)) = \Gamma(\alpha_0, \alpha, t_{k-1}) + \Delta \|\alpha - \alpha_0\| \, r_k(\alpha_0, \alpha)$$

with $\sup_{k, \alpha \in K_a} \|r_k(\alpha_0, \alpha)\| \leq C$.

Proof. Using (3.3.7), (3.3.10) and that $\Phi(\alpha, t_k, t_{k-1}) = I_p + \Delta \nabla_z b(\alpha, t_{k-1}, z(\alpha, t_{k-1})) + \Delta^2 O(1)$, yields

$$\begin{aligned} B_k(\alpha, z(\alpha_0, \cdot)) &= \int_{t_{k-1}}^{t_k} (b(\alpha_0, s, z(\alpha_0, s)) - b(\alpha, s, z(\alpha, s)))ds \\ &\qquad + (I_p - \Phi(\alpha, t_k, t_{k-1}))\,(z(\alpha_0, t_{k-1}) - z(\alpha, t_{k-1})) \\ &= \Delta \Gamma(\alpha_0, \alpha, t_{k-1}) + \Delta^2 \|\alpha - \alpha_0\| \, r_k(\alpha_0, \alpha). \end{aligned}$$

Assumptions (S1), (S2) and (S4) ensure that the remainder term has order Δ^2 uniformly in k, α. □

Consider now the random variables $B_k(\alpha, X)$

Lemma 3.3.7. *Assume (S1)–(S5). Then, under $\mathbb{P}_{\theta_0}$, as $\varepsilon, \Delta \to 0$, the following holds for all $k \leq n$,*

$$\begin{aligned} \frac{1}{\Delta}(B_k(\alpha, X) - B_k(\alpha_0, X)) &= \frac{1}{\Delta} B_k(\alpha, z(\alpha_0, \cdot)) + \varepsilon ||\alpha - \alpha_0|| \eta_k(\alpha_0, \alpha, \varepsilon, \Delta) \\ &= \Gamma(\alpha_0, \alpha, t_{k-1}) + ||\alpha - \alpha_0|| (\Delta O(1) + \varepsilon O_P(1)), \end{aligned}$$

where $\eta_k = \eta_k(\alpha_0, \alpha, \varepsilon, \Delta)$ is $\mathscr{G}_{k-1}^n$-measurable and uniformly bounded in probability .

Proof. Using (3.3.6) and (3.3.7), we get that

$$\begin{aligned} &B_k(\alpha, X) - B_k(\alpha_0, X) = \\ &B_k(\alpha, z(\alpha_0, \cdot)) + \varepsilon(\Phi(\alpha_0, t_k, t_{k_1}) - \Phi(\alpha, t_k, t_{k_1}))(g(\theta_0, t_{k-1}) + \varepsilon R^\varepsilon(\theta_0, t_{k-1})). \end{aligned}$$

By (S1)–(S5),

$$\begin{aligned} &\left\|\Phi(\alpha_0, t_k, t_{k_1}) - \Phi(\alpha, t_k, t_{k_1})\right\| \\ &\quad \leq 2\Delta \left\|\nabla_z b(\alpha_0, , t_{k-1}, z(\alpha_0, t_{k-1})) - \nabla_z b(\alpha, , t_{k-1}, z(\alpha, t_{k-1}))\right\|. \end{aligned}$$

Now, this term is bounded by $K\Delta||\alpha - \alpha_0||$ since $(t, \alpha) \to \nabla_z b(\alpha, z(\alpha, t))$ is uniformly continuous on $[0, T] \times K_a$. Using now Proposition 3.3.1 yields that $(g(\theta_0, t_{k-1}) + \varepsilon R^\varepsilon(\theta_0, t_{k-1}))$ is bounded in $\mathbb{P}_{\theta_0}$-probability and $\mathscr{G}_{k-1}^n$-measurable. □

The next lemma concerns the properties of $B_k(\alpha_0, X)$.

Lemma 3.3.8. *Assume (S1)–(S5). Then, using* (3.3.9)*, as $\varepsilon, \Delta \to 0$, under $\mathbb{P}_{\theta_0}$,*

(i) *$B_k(\alpha_0,X) = \varepsilon\sigma_{k-1}(\beta_0)(B(t_k)-B(t_{k-1})) + E_k(\theta_0,\varepsilon,\Delta)$, where $E_k = E_k(\theta_0,\varepsilon,\Delta)$ satisfies that, for $m \geq 2$, $\mathbb{E}_{\theta_0}(||E_k||^m|\mathscr{G}^n_{k-1}) \leq C\varepsilon^m\Delta^m$.*

(ii) *If (V_k) is a sequence of $\mathscr{G}^n_{k-1}$-measurable random variables in $\mathbb{R}^p$ uniformly bounded in probability, then $\sum_{k=1}^n V_k^* B_k(\alpha_0,X) \to 0$ in probability.*

Proof. Let us first prove (i) and study the term E_k. We have $E_k = E_k^1 + E_k^2$ with

$$\begin{aligned}E_k^1 =& \int_{t_{k-1}}^{t_k} (b(\alpha_0,t,X(t)) - b(\alpha_0,t,z(\alpha_0,t)))\,dt \\ &+ (I_p - \Phi(\alpha_0,t_k,t_{k-1}))\,(X(t_{k-1}) - z(\alpha_0,t_{k-1})) \text{ and} \\ E_k^2 =& \varepsilon\int_{t_{k-1}}^{t_k} (\sigma(\beta_0,s,X(s)) - \sigma(\beta_0,s,X(t_{k-1})))\,dB(s).\end{aligned}$$

Set in (3.3.6), $R_1^\varepsilon(\theta,t) = g(\theta,t) + \varepsilon R^\varepsilon(\theta,t)$. Using that $(t,x) \to b(\alpha,t,x)$ is uniformly Lipschitz, we obtain,

$$\begin{aligned}\left\|E_k^1\right\| \leq& \Delta C \sup_{t\in[t_{k-1};t_k]} \|X(t) - z(\alpha_0,t)\| \\ &+ \Delta\varepsilon \left\| \left(\int_0^1 \nabla_z b(\alpha_0,z(\alpha_0,t)))\Phi(\alpha_0,t,t_{k-1})dt\right) R_1^\varepsilon(\theta_0,t_{k-1})\right\| \\ \leq& C'\varepsilon\Delta \sup_{t\in[t_{k-1};t_k]} \|R_1^\varepsilon(\theta_0,t)\|\end{aligned}$$

The proof for E_k^2 follows the sketch given in [57, Lemma 1]. We first prove this result based on the stronger condition Σ and b bounded. Then, similarly to [57, Proposition 1], this assumption can be relaxed. We use sequentially the Burkhölder–Davis–Gundy (see e.g. [83]) and Jensen inequalities to obtain

$$\begin{aligned}&\mathbb{E}_{\theta_0}(\left\|E_k^2\right\|^m|\mathscr{G}^n_{k-1}) \\ &\leq C\varepsilon^m\mathbb{E}_{\theta_0}\left(\left(\int_{t_{k-1}}^{t_k} \|\sigma(\beta_0,s,X(s)) - \sigma(\beta_0,t_{k-1},X(t_{k-1})\|^2\,ds\right)^{m/2}|\mathscr{G}^n_{k-1}\right) \qquad (3.3.11)\\ &\leq C\varepsilon^m\Delta^{m/2-1}\int_{t_{k-1}}^{t_k} \mathbb{E}_{\theta_0}\left(\|\sigma(\beta_0,s,X(s)) - \sigma(\beta_0,t_{k-1},X(t_{k-1}))\|^m|\mathscr{G}^n_{k-1}\right)|ds. \qquad (3.3.12)\end{aligned}$$

Then, using that $(t,x) \to \sigma(\beta,t,x)$ is Lipschitz, we obtain:

$$\begin{aligned}&\mathbb{E}^\varepsilon_{\theta_0}(\left\|E_k^2\right\|^m|\mathscr{G}^n_{k-1}) \leq C'\varepsilon^m\Delta^{m/2-1}\int_{t_{k-1}}^{t_k} \mathbb{E}^\varepsilon_{\theta_0}(\|X(s) - X(t_{k-1})\|^m)ds \\ &\leq C'\varepsilon^m\Delta^{m/2-1}\int_{t_{k-1}}^{t_k} \mathbb{E}^\varepsilon_{\theta_0}\left[\left\|\int_{t_{k-1}}^s (b(\alpha_0,u,X(u))du + \varepsilon\sigma(\beta_0,u,X(u))dB(u))\right\|\right]^m ds.\end{aligned}$$

Since b is bounded on $\mathscr{U}$, $\left\|\int_{t_{k-1}}^s b(\alpha_0,u,X(u))du\right\| \leq K|s-t_{k-1}|$ and Itô's isometry yields

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(\left\|\int_{t_{k-1}}^s \sigma(\beta_0,u,X(u))dB(u)\right\|^m\right) &\leq \mathbb{E}_{\theta_0}\left(\left\|\int_{t_{k-1}}^s \Sigma(\beta_0,u,X(u))du\right\|\right)^{m/2} \\ &\leq K|s-t_{k-1}|^{m/2}.\end{aligned}$$

Thus, $\mathbb{E}_{\theta_0}(\left\|E_k^2(\theta_0)\right\|^m|\mathscr{G}^n_{k-1}) \leq C''\varepsilon^m\Delta^{m/2-1}\int_{t_{k-1}}^{t_k}|s-t_{k-1}|^{m/2}ds \leq C'''\varepsilon^m\Delta^m$.

The proof of (ii) relies on the Lemma A.4.3 for triangular arrays stated in Section A.4. Set $\zeta_k^n = V_k^* B_k(\alpha_0,X)$. Using (i), we have

$$\mathbb{E}_{\theta_0}(\zeta_k^n|\mathscr{G}^n_{k-1}) = V_k^*\mathbb{E}^\varepsilon_{\theta_0}(E_k(\theta_0,\varepsilon,\Delta)|\mathscr{G}^n_{k-1})$$

with $\mathbb{E}^{\varepsilon}_{\theta_0}(\|E_k(\theta_0,\varepsilon,\Delta)\|\,|\mathscr{G}^n_{k-1}) \le C\varepsilon\Delta$.
Since $\sup_{k\le n}\|V_k\|$ is bounded in probability, $\sum_{k=1}^n \mathbb{E}_{\theta_0}(\zeta^n_k|\mathscr{G}^n_{k-1}) \le C\varepsilon T \to 0$.
Therefore condition (i) of Lemma A.4.3 is satisfied with $U=0$.
Now, $\mathbb{E}_{\theta_0}[(\zeta^n_k)^2|\mathscr{G}^n_{k-1})] = V_k^* \mathbb{E}_{\theta_0}(B_k(\alpha_0,X)B_k^*(\alpha_0,X)|\mathscr{G}^n_{k-1})V_k$.
Using (i) of Lemma 3.3.8 yields that

$$\begin{aligned}\mathbb{E}_{\theta_0}(B_k(\alpha_0,X)B_k^*(\alpha_0,X)|\mathscr{G}^n_{k-1}) &= \varepsilon^2\Delta\Sigma_{k-1}(\beta_0) + \mathbb{E}_{\theta_0}(E_kE_k^*|\mathscr{G}^n_{k-1}) \\ &\le K_1\varepsilon^2\Delta + C_2\varepsilon^2\Delta^2.\end{aligned}$$

Hence, $\sum_{k=1}^n \mathbb{E}_{\theta_0}((\zeta^n_k)^2|\mathscr{G}^n_{k-1}) \to 0$. Therefore, applying Lemma A.4.3 achieves the proof. □

A last lemma concerns the terms $(\nabla_{\alpha_i}B_k)$ for $i=1,\dots,a$.

Lemma 3.3.9. *Assume (S1)–(S6). Then, under $\mathbb{P}_{\theta_0}$, for all $i,j\le a$, for all $\alpha\in K_a$, as $\varepsilon,\Delta\to 0$,*

(i) $\frac{1}{\Delta}\nabla_{\alpha_i}B_k(\alpha,X) = -\nabla_{\alpha_i}b(\alpha,t_{k-1},z(\alpha,t_{k-1})) + M^i_k(\alpha)[(z(\alpha_0,t_{k-1})-z(\alpha,t_{k-1})) + \varepsilon Z^{\varepsilon}_{k-1}(\theta_0)] + \Delta O_P(1)$, *where $M^i_k(\alpha)$ are uniformly bounded matrices and $Z^{\varepsilon}_{k-1}(\theta_0)$ are $\mathscr{G}^n_{k-1}$-measurable r.v.s uniformly bounded in probability.*

(ii) For all $k\le n$, $\frac{1}{\Delta}\left\|\nabla^2_{\alpha_i,\alpha_j}B_k(\alpha,X)\right\|$ is bounded uniformly in probability.

Proof. We have, using (3.3.6) and (3.3.7),

$$\begin{aligned}B_k(\alpha,X) =& (X(t_k)-X(t_{k-1})) - (z(\alpha,t_k)-z(\alpha,t_{k-1})) \\ &+ (I_p - \Phi(\alpha,t_k,t_{k-1}))(X(t_{k-1})-z(\alpha,t_{k-1})).\end{aligned}$$

Therefore,

$$\nabla_{\alpha_i}B_k(\alpha,X) = E_{k,i} + \varepsilon\Delta M^i_k(\alpha)Z_{k-1}(\theta_0)$$

with $M^i_k(\alpha) = -\frac{1}{\Delta}\nabla_{\alpha_i}\Phi(\alpha,t_k,t_{k-1})$, $Z_{k-1}(\theta_0) = g(\theta_0,t_{k-1}) + \varepsilon R^{\varepsilon}(\theta_0,t_{k-1})$ and

$$\begin{aligned}E_{k,i} =& -\nabla_{\alpha_i}(z(\alpha,t_k)-z(\alpha,t_{k-1})) + (\Phi(\alpha,t_k,t_{k-1})-I_p)\nabla_{\alpha_i}z(\alpha,t_{k-1}) \\ &-\nabla_{\alpha_i}\Phi(\alpha,t_k,t_{k-1})(z(\alpha_0,t_{k-1})-z(\alpha,t_{k-1})).\end{aligned}$$

Proposition 3.3.1 yields the result for $Z_k(\theta_0)$.

Now, $\Phi(\alpha,t_k,t_{k-1}) = \exp\{\int_{t_{k-1}}^{t_k}\nabla_z b(\alpha,s,z(\alpha,s))ds\}$ so that

$$M^i_k(\alpha) = -\nabla_{\alpha_i}\nabla_z b(\alpha,t_{k-1},z(\alpha,t_{k-1})) + \Delta \mathrm{O}(1).$$

To study $E_{k,i}$, we use that $\Phi(\alpha,t_k,t_{k-1}) - I_p = \Delta\nabla_z b(\alpha,t_{k-1},z(\alpha,t_{k-1})) + \Delta^2 O(1)$ and $\nabla_{\alpha_i}\left(b(\alpha,t,z(\alpha,t))\right) = \nabla_{\alpha_i}b(\alpha,t,z(\alpha,t)) + \nabla_z b(\alpha$
Therefore $E_{k,i} = -\nabla_{\alpha_i}b(\alpha,t_{k-1},z(\alpha,t_{k-1})) + M^i_k(\alpha)(z(\alpha_0,t_{k-1})-z(\alpha,t_{k-1})) + \Delta\mathrm{O}(1)$.
This achieves the proof of (i).

Let us prove (ii). We have $\nabla^2_{\alpha_i\alpha_j}B_k(\alpha,X) = f^{ij}_k(\alpha_0,\alpha,\Delta) + \xi^{ij}_k(\theta_0,\alpha,\varepsilon,\Delta)$ with
$\xi^{ij}_k = (\nabla^2_{\alpha_i\alpha_j}\Phi(\alpha,t_k,t_{k-1}))[X(t_{k-1})-z(\alpha_0,t_{k-1})]$ and

$$\begin{aligned}f^{ij}_k(\alpha_0,\alpha,\Delta) =& -\left(\nabla^2_{\alpha_i\alpha_j}z(\alpha,t_k) - \Phi(\alpha,t_k,t_{k-1})\nabla^2_{\alpha_i\alpha_j}z(\alpha,t_{k-1})\right) \\ &+\nabla_{\alpha_i}\Phi(\alpha,t_k,t_{k-1})\nabla_{\alpha_j}z(\alpha,t_{k-1}) \\ &+\nabla_{\alpha_j}\Phi(\alpha,t_k,t_{k-1})\nabla_{\alpha_i}z(\alpha,t_{k-1}) \\ &-\nabla^2_{\alpha_i\alpha_j}\Phi(\alpha,t_k,t_{k-1})(z(\alpha,t_{k-1})-z(\alpha_0,t_{k-1})).\end{aligned}$$

The result is obtained using Proposition 3.3.1 and the property that $\frac{1}{\Delta}\|\nabla_{\alpha_i}\Phi(\cdot)\|$ and $\frac{1}{\Delta}\left\|\nabla^2_{\alpha_i,\alpha_j}\Phi(\cdot)\right\|$ are uniformly bounded. □

3.4 Inference based on high frequency observations on $[0,T]$

We assume that both ε and $\Delta = \Delta_n$ go to 0. The number of observations n goes to infinity. We study the estimation of $\theta = (\alpha, \beta)$ based on $(X(t_k), k = 1, \ldots, n)$.

Lemmas 3.3.4 and 3.3.5 yield that the random variables $\frac{1}{\varepsilon\sqrt{\Delta}} B_k(\alpha_0, X)$ are approximately conditionally independent centered Gaussian random variables in $\mathbb{R}^p$ with covariance approximated by $\Sigma_{k-1}(\beta_0)$. Analogously to [86] or [57], we introduce a contrast function, using definitions (3.3.7), (3.3.9),

$$\check{U}_{\varepsilon,\Delta}(\alpha,\beta) = \sum_{k=1}^{n} \log\det\Sigma_{k-1}(\beta) + \frac{1}{\varepsilon^2\Delta}\sum_{k=1}^{n} B_k(\alpha,X)^*\, \Sigma_{k-1}^{-1}(\beta)\, B_k(\alpha,X). \tag{3.4.1}$$

The minimum contrast estimators are defined as any solution of

$$(\check{\alpha}_{\varepsilon,\Delta}, \check{\beta}_{\varepsilon,\Delta}) = \underset{(\alpha,\beta)\in\Theta}{\operatorname{argmin}}\, \check{U}_{\varepsilon,\Delta}(\alpha,\beta). \tag{3.4.2}$$

3.4.1 Properties of the estimators

In what follows, we use to describe the asymptotics with respect to $\Delta = \Delta_n$ either $\Delta \to 0$ or $n \to \infty$. Indeed, it is more explicit to state results according to the number of observations n rather than in terms of the size of the sampling interval $\Delta = \Delta_n$. Results are obtained when $\varepsilon \to 0$ and $\Delta \to 0$ (or $n \to \infty$) simultaneously.

Define, for $\theta = (\alpha,\beta) \in \Theta$ with Θ a compact subset of $\mathbb{R}^a \times \mathbb{R}^b$, the matrices $I_b(\theta) = (I_b(\theta)_{ij}, 1 \le i,j \le a)$, $I_\sigma(\theta) = (I_\sigma(\theta)_{i,j}, 1 \le i,j \le b)$ and $I(\theta)$ by

$$(I_b(\theta))_{ij} = \int_0^T (\nabla_{\alpha_i} b(\alpha,t,z(\alpha,t)))^* \Sigma^{-1}(\beta,t,z(\alpha,t)) \nabla_{\alpha_j} b(\alpha,t,z(\alpha,t)) dt, \tag{3.4.3}$$

$$\begin{aligned} &(I_\sigma(\theta))_{i,j} = \\ &\frac{1}{2T}\int_0^T \operatorname{Tr}\Big(\nabla_{\beta_i}\Sigma(\beta,t,z(\alpha,t))\Sigma^{-1}(\beta,t,z(\alpha,t))\nabla_{\beta_j}\Sigma(\beta,t,z(\alpha,t))\Sigma^{-1}(\beta,t,z(\alpha,t))\Big)\, dt. \end{aligned} \tag{3.4.4}$$

$$I(\theta) = \begin{pmatrix} I_b(\theta)) & 0 \\ 0 & I_\sigma(\theta) \end{pmatrix}. \tag{3.4.5}$$

Recall that A^* denotes the transpose of a matrix A and $\operatorname{Tr}(A)$ its trace.

Theorem 3.4.1. *Assume that $(X(t))$ satisfying (3.3.1) is observed at times $t_k = k\Delta_n$ with $T = n\Delta_n$ fixed. Assume **(S1)–(S7)** and that $I_b(\theta_0)$ is non-singular. Then, as $\varepsilon \to 0, \Delta = \Delta_n \to 0$,*

(i) $(\check{\alpha}_{\varepsilon,\Delta}, \check{\beta}_{\varepsilon,\Delta}) \to (\alpha_0,\beta_0)$ in $\mathbb{P}_{\theta_0}$-probability.

(ii) If moreover $I_\sigma(\theta_0)$ is non-singular,

$$\begin{pmatrix} \varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0) \\ \sqrt{n}\left(\check{\beta}_{\varepsilon,\Delta} - \beta_0\right) \end{pmatrix} \to \mathcal{N}_{a+b}\left(0, \begin{pmatrix} I_b^{-1}(\alpha_0,\beta_0) & 0 \\ 0 & I_\sigma^{-1}(\alpha_0,\beta_0) \end{pmatrix}\right)$$

in distribution under $\mathbb{P}_{\theta_0}$.

Note that results are obtained without any condition linking ε and Δ (or n). Indeed, the previous results obtained in [57] require conditions linking ε and Δ that do not fit epidemic data, where generally the orders of magnitude for N and n satisfy $N >> n$ so that Δ is comparatively large with respect to $\varepsilon = 1/\sqrt{N}$. We proposed in [61] another method based on Theorem 3.2.1 which extends results obtained in [47], where the inference in the case $\sigma(\beta,x) \equiv 1$

was investigated for one-dimensional diffusions using expansion (3.2.15).

Since estimators of parameters in the drift (here α) and in the diffusion coefficient (here β) converge at distinct rates, ε^{-1} and $\sqrt{n} = \Delta^{-1/2}$ respectively, the study of asymptotic properties has to be performed according to the successive steps:

Step (1): Consistency of $\check{\alpha}_{\varepsilon,\Delta}$ (Proposition 3.4.2).

Step (2): Tightness of the sequence $\varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0)$ with respect to β (Proposition 3.4.3).

Step (3): Consistency of $\check{\beta}_{\varepsilon,\Delta}$ (Proposition 3.4.5).

Step (4): Asymptotic normality for both estimators (Theorem 3.4.1,(ii)).

The proof is technical and detailed in a separate section. Before this proof, let us state some comments.

3.4.1.1 Comments

(1) *Efficiency of estimators*: Note that the matrix $I_b(\theta)$ is equal to the Fisher information matrix associated to the estimation of α when $(X(t))$ is continuously observed on $[0,T]$ (see e.g. [92] and Section A.3 in the Appendix). Therefore $\check{\alpha}_{\varepsilon,\Delta}$ is efficient for this statistical model.

(2) *Comparison with estimation based on complete observation of the jump process* $(\mathscr{Z}^N(t))$: Coming back to epidemics, we can compare the estimation of the parameters of the pure jump process $(\mathscr{Z}^N(t))$ and $(Z^N(t) = \frac{1}{N}\mathscr{Z}^N(t))$ describing the epidemic dynamics in a finite population of size N and the estimators built from its diffusion approximation. Let us stress that there is a main difference between these two approaches. Statistical inference for $(\mathscr{Z}^N(t))$ is based on the observations of all the jumps, which implies here the observation of all the times of infection and recovery for each individual in the population, while for the diffusion $(X(t))$, we consider discrete observations $(X(t_k), k = 1,\dots,n)$.

(3) *Comparison of estimators for the SIR epidemic dynamics*: Assume that the jump process $(\mathscr{Z}^N(t))$ is continuously observed on $[0,T]$. Its dynamics is described by the two parameters (λ,γ). Set $Z^N(t) = (S^N(()), I^N(t))^*$, and assume that $Z^N(0) \to x_0 = (s_0, i_0)^*$, with $s_0 > 0, i_0 > 0$. Let $s(t) = s(\lambda_0,\gamma_0,t); i(t) = i(\lambda_0,\gamma_0,t)$ the solution of the corresponding ODE.

The Maximum Likelihood Estimator $(\hat{\lambda},\hat{\gamma})$ is explicit (see [2] or Chapter 4 of this part). Indeed, let (T_i) denote the successive jump times and set $J_i = 0$ if we have an infection and $J_i = 1$ if we have a recovery. Let $K_N(T) = \sum_{i\geq 0} 1_{T_i \leq T}$. Then

$$\hat{\lambda}_N = \frac{1}{N}\frac{\sum_{i=1}^{K_N(T)}(1-J_i)}{\int_0^T S^N(t)I^N(t)dt} = \frac{1}{N}\frac{\#\text{ Infections}}{\int_0^T S^N(t)I^N(t)dt},$$
$$\hat{\gamma}_N = \frac{1}{N}\frac{\sum_{i=1}^{K_N(T)} J_i}{\int_0^T I^N(t)dt} = \frac{\#\text{ Recoveries}}{\text{“Mean infectious period”}}.$$

As the population size N goes to infinity, $(\hat{\lambda}_N,\hat{\gamma}_N)$ is consistent and

$$\sqrt{N}\begin{pmatrix}\hat{\lambda}_N - \lambda \\ \hat{\gamma}_N - \gamma\end{pmatrix} \to \mathscr{N}_2\left(0, I^{-1}(\lambda,\gamma)\right), \text{ where } I(\lambda,\gamma) = \begin{pmatrix}\frac{\int_0^T s(t)i(t)dt}{\lambda} & 0 \\ 0 & \frac{\int_0^T i(t)dt}{\gamma}\end{pmatrix}.$$

The matrix $I(\lambda,\gamma)$ is the Fisher information matrix of this statistical model.

Consider now the *SIR* diffusion approximation $X(t)$ described in Section 3.2.2.1. We have

$$b(\theta,(s,i)) = \begin{pmatrix}-\lambda si \\ \lambda si - \gamma i\end{pmatrix}; \quad \Sigma(\theta,(s,i)) = \begin{pmatrix}\lambda si & -\lambda si \\ -\lambda si & \lambda si + \gamma i\end{pmatrix}. \tag{3.4.6}$$

Therefore,

$$\nabla_\theta b(\theta,(s,i)) = \begin{pmatrix} -si & 0 \\ si & -i \end{pmatrix}; \Sigma^{-1}(\theta,(s,i)) = \frac{1}{\lambda\gamma si}\begin{pmatrix} \lambda s+\gamma & \lambda s \\ \lambda s & \lambda s \end{pmatrix}.$$

The matrix $I_b(\theta)$ defined in (3.4.3) is

$$I_b(\lambda,\gamma) = \begin{pmatrix} \frac{1}{\lambda}\int_0^T s(t)i(t)dt & 0 \\ 0 & \frac{1}{\gamma}\int_0^T i(t)dt \end{pmatrix}.$$

Therefore, we obtain the same information matrix in both cases.

Consider the *SIRS* model with immunity waning δ. We have $\theta = (\lambda,\gamma,\delta)$ The diffusion approximation satisfies

$$b(\theta,(s,i)) = \begin{pmatrix} -\lambda si+\delta(1-s-i) \\ \lambda si-\gamma i \end{pmatrix}; \quad \Sigma(\theta,(s,i)) = \begin{pmatrix} \lambda si+\delta(1-s-i) & -\lambda si \\ -\lambda si & \lambda si+\gamma i \end{pmatrix}.$$

Hence,

$$\nabla_\theta b(\theta,s,i) = \begin{pmatrix} -si & 0 & (1-s-i) \\ si & -i & 0 \end{pmatrix},$$

$$I_b(\theta) = \int_0^T \nabla_\theta^* b(\theta,s(t),i(t))\Sigma^{-1}(\theta,s(t),i(t))\nabla_\theta b(\theta,s(t),i(t))dt.$$

Then $I_b(\theta)$ can be computed and compare to the Fisher information matrix derived from the statistical model corresponding to complete observation of the *SIRS* jump process.

3.4.2 Proof of Theorem 3.4.1

Recall the notations: for a matrix A, A^* the transposition of A, $\det(A)$ the determinant of A and $\mathrm{Tr}(A)$ the trace of A.

3.4.2.1 Step (1): Consistency of $\check{\alpha}_{\varepsilon,\Delta}$

Let us define, using (3.3.10),

$$K_1(\alpha_0,\alpha,\beta) = \int_0^T \Gamma(\alpha_0,\alpha,t)^*\Sigma^{-1}(\beta,t,z(\alpha_0,t))\Gamma(\alpha_0,\alpha,t)dt. \tag{3.4.7}$$

By Assumption (S4), if $\alpha \neq \alpha_0$, $b(\alpha,t,z(\alpha,t)) \not\equiv b(\alpha_0,t,z(\alpha_0))$, therefore the function $\Gamma(\alpha_0,\alpha,\cdot) \not\equiv 0$, which implies that $K_1(\alpha_0,\alpha,\beta)$ is non-negative and equal to 0 if and only if $\alpha = \alpha_0$.

The contrast function $\check{U}_{\varepsilon,\Delta}(\alpha,\beta)$ defined in (3.4.1) satisfies

Proposition 3.4.2. *Assume (S1)–(S6). Then, as $\varepsilon,\Delta \to 0$, the following convergences hold.*

(i) $\sup_{\theta\in\Theta}|\varepsilon^2\left(\check{U}_{\varepsilon,\Delta}(\alpha,\beta) - \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)\right) - K_1(\alpha_0,\alpha,\beta)| \to 0$ *in $\mathbb{P}_{\theta_0}$-probability.*

(ii) $\check{\alpha}_{\varepsilon,\Delta} \to \alpha_0$ *in probability under $\mathbb{P}_{\theta_0}$.*

Proof. Let us prove (i). We have, by (3.4.1) and (3.3.9),

$$\varepsilon^2(\check{U}_{\varepsilon,\Delta}(\alpha,\beta) - \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)) = T_1 + T_2$$

with

$$T_1 \quad = \quad 2\sum_{k=1}^n \frac{(B_k(\alpha,X) - B_k(\alpha_0,X))^*}{\Delta}\,\Sigma_{k-1}^{-1}(\beta)\,B_k(\alpha_0,X),$$

$$T_2 = \Delta \sum_{k=1}^{n} \frac{(B_k(\alpha,X)-B_k(\alpha_0,X))^*}{\Delta} \Sigma_{k-1}^{-1}(\beta) \frac{(B_k(\alpha,X)-B_k(\alpha_0,X))}{\Delta}.$$

By Lemma 3.3.7, $\frac{(B_k(\alpha,X)-B_k(\alpha_0,X))}{\Delta}$ is bounded, and (ii) of Lemma 3.3.8 yields that T_1 goes to 0 in $\mathbb{P}_{\theta_0}$-probability. Using now (3.3.10), we have by Lemma 3.3.7, setting $\zeta_k = \Delta r_k + \varepsilon \|\alpha-\alpha_0\| \eta_k$,

$$\begin{aligned} T_2 &= \Delta \sum_{k=1}^{n} (\Gamma(\alpha_0,\alpha,t_{k-1}) + \zeta_{k-1})^* \Sigma_{k-1}^{-1}(\beta) (\Gamma(\alpha_0,\alpha,t_{k-1}) + \zeta_{k-1}) \\ &= \Delta \sum_{k=1}^{n} \left(\Gamma(\alpha_0,\alpha,t_{k-1})^* \Sigma_{k-1}^{-1}(\beta) \Gamma(\alpha_0,\alpha,t_{k-1}) + R_k(\theta_0,\theta,\varepsilon,\Delta) \right). \end{aligned}$$

The first term of the above formula as a Riemann sum converges by Lemma 3.3.6 to the function $K_1(\alpha_0,\alpha,\beta)$ defined in (3.4.7) as $\Delta \to 0$. This convergence is uniform with respect to the parameters. The remainder term is

$$\begin{aligned} R_k(\theta_0,\theta,\varepsilon,\Delta) =& \Gamma(\alpha_0,\alpha,t_{k-1})^* (\Sigma_{k-1}^{-1}(\beta) - \Sigma^{-1}(\beta,t_{k-1},z(\alpha_0,t_{k-1}))) \Gamma(\alpha_0,\alpha,t_{k-1}) \\ &+ \Delta R_k^1(\theta_0,\theta,\varepsilon,\Delta) + \varepsilon R_k^2(\theta_0,\theta,\varepsilon,\Delta). \end{aligned}$$

Using Proposition 3.3.1 and Lemma 3.3.7, it is straightforward to get that $\sup_k \|R_k(\theta_0,\theta,\varepsilon,\Delta)\| \to 0$ in $\mathbb{P}_{\theta_0}$-probability uniformly with respect to θ. Hence, T_2 converges to $K_1(\alpha_0,\alpha,\beta)$ in $\mathbb{P}_{\theta_0}$-probability uniformly with respect to θ.
Let us prove (ii). Noting that, for all β, $K_1(\alpha_0,\alpha_0,\beta) = 0$, we have

$$\begin{aligned} 0 \le& K_1(\alpha_0,\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - K_1(\alpha_0,\alpha_0,\check{\beta}_{\varepsilon,\Delta}) \\ \le& [\varepsilon^2(\check{U}_{\varepsilon,\Delta}(\alpha,\check{\beta}_{\varepsilon,\Delta}) - \check{U}_{\varepsilon,\Delta}(\alpha_0,\check{\beta}_{\varepsilon,\Delta})) - K_1(\alpha_0,\alpha,\check{\beta}_{\varepsilon,\Delta})] \\ &+ [K_1(\alpha_0,\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - \varepsilon^2(\check{U}_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - \check{U}_{\varepsilon,\Delta}(\alpha_0,\check{\beta}_{\varepsilon,\Delta}))] \\ &+ \varepsilon^2[\check{U}_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - \check{U}_{\varepsilon,\Delta}(\alpha,\check{\beta}_{\varepsilon,\Delta})] \\ \le& 2 \sup_{\beta \in K_b} |\varepsilon^2[\check{U}_{\varepsilon,\Delta}(\alpha,\beta) - \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)] - K_1(\alpha_0,\alpha,\beta)|, \end{aligned}$$

where the last inequality is obtained using that the minimum of $\check{U}_{\varepsilon,\Delta}(\alpha,\beta)$ is attained at $(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta})$. By Proposition 3.4.2 (i), we finally get that

$$|K_1(\alpha_0,\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - K_1(\alpha_0,\alpha_0,\check{\beta}_{\varepsilon,\Delta})| \to 0,$$

which yields by Assumption **(S6)** that $\check{\alpha}_{\varepsilon,\Delta} \to \alpha_0$ in $\mathbb{P}_{\theta_0}$-probability as $\varepsilon,\Delta \to 0$. □

3.4.2.2 Step (2): Tightness of the sequence $\varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0)$

This step is crucial in the presence of different rates of convergence for α and β and concerns results that hold for all $\beta \in K_b$.

Proposition 3.4.3. *Assume **(S1)–(S4)** and that $I_b(\alpha_0,\beta_0)$ defined in (3.4.3) is non-singular. Then, as $\varepsilon,\Delta \to 0$, $\sup_{\beta \in K_b} \left\| \varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0) \right\|$ is bounded in $\mathbb{P}_{\theta_0}$-probability.*

Proof. Recall the notation: for f a twice differentiable real function, $\nabla_\alpha^2 f = (\frac{\partial^2 f}{\partial\alpha_i \partial\alpha_j})_{1\le i,j\le a}$.
Under (S5), $\check{U}_{\varepsilon,\Delta}(\alpha,\beta)$ is C^2 and a Taylor expansion of $\nabla_\alpha \check{U}_{\varepsilon,\Delta}$ at point $(\alpha_0,\check{\beta}_{\varepsilon,\Delta})$ w.r.t. α yields,

$$0 = \varepsilon \nabla_\alpha \check{U}_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) = \varepsilon \nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\check{\beta}_{\varepsilon,\Delta}) + \varepsilon^2 N_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) \frac{(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0)}{\varepsilon}, \tag{3.4.8}$$

$$\text{with} \quad N_{\varepsilon,\Delta}(\alpha,\beta) = \int_0^1 \nabla_\alpha^2 \check{U}_{\varepsilon,\Delta}(\alpha_0 + t(\alpha-\alpha_0),\beta) dt. \tag{3.4.9}$$

The proof relies on two properties: under $\mathbb{P}_{\theta_0}$, as $\varepsilon,\Delta\to 0$, for all $\beta\in K_b$, $(\varepsilon\nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta))$ converges in distribution to a Gaussian law and the sequence $\varepsilon^2\nabla^2_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)$ converges almost surely.

Let us study $-\varepsilon\nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)$. Define the $a\times a$ matrix

$$J(\theta_0,\beta)=\int_0^T (\nabla_\alpha b(\alpha_0,t,z(\alpha_0,t)))^*\Xi(\theta_0,\beta,t)\nabla_\alpha b(\alpha_0,t,z(\alpha_0,t))dt, \text{ with} \tag{3.4.10}$$

$$\Xi(\theta_0,\beta,t)=\Sigma^{-1}(\beta,t,z(\alpha_0,t))\Sigma(\beta_0,t,z(\alpha_0,t))\Sigma^{-1}(\beta,t,z(\alpha_0,t)). \tag{3.4.11}$$

The following holds.

Lemma 3.4.4. *Assume (S1)–(S5). Then, as $\varepsilon,\Delta\to 0$,*

$$-\varepsilon\nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)\to\mathcal{N}(0,4J(\theta_0,\beta)) \text{ in distribution under } \mathbb{P}_{\theta_0}.$$

Proof. We have, using the notations of Lemma 3.3.9 and setting

$$H^i_k(\alpha_0,\beta)=\Sigma^{-1}(\beta,t_{k-1},z(\alpha_0,t_{k-1}))\nabla_{\alpha_i} b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1})) \tag{3.4.12}$$

$$-\varepsilon\nabla_{\alpha_i}\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)=-\frac{2}{\varepsilon\Delta}\sum_{k=1}^n(\nabla_{\alpha_i}B_k(\alpha_0,X))^*\Sigma^{-1}_{k-1}(\beta)B_k(\alpha_0,X)=A^i_n+A^{\prime,i}_n+A^{\prime\prime,i}_n,$$

with

$$\begin{aligned}
A^i_n &= \frac{2}{\varepsilon}\sum_{k=1}^n H^i_k(\alpha_0,\beta)^*B_k(\alpha_0,X),\\
A^{\prime,i}_n &= -2\sum_{k=1}^n (M^i_k(\alpha_0)Z_{k-1}(\theta_0))^*\Sigma^{-1}_{k-1}(\beta)B_k(\alpha_0,X),\\
A^{\prime\prime,i}_n &= 2\sum_{k=1}^n \frac{\nabla_{\alpha_i}B_k(\alpha_0,X)}{\Delta}^*\frac{\Sigma^{-1}_{k-1}(\beta)-\Sigma^{-1}(\beta,t_{k-1},z(\alpha_0,t_{k-1}))}{\varepsilon}B_k(\alpha_0,X).
\end{aligned}$$

By Lemma 3.3.8 (ii), Lemma 3.3.9 and Theorem 3.3.1, $A^{\prime,i}_n$ and $A^{\prime\prime,i}_n$ tend to 0 in $\mathbb{P}_{\theta_0}$-probability as $\varepsilon,\Delta\to 0$. To study A^i_n, we write, using the notations of Lemma 3.3.4,

$$\begin{aligned}
B_k(\alpha_0,X)=\varepsilon\sqrt{\Delta}T_k(\theta_0)&+\varepsilon^2(R(\theta_0,t_k)-R(\theta_0,t_{k-1}))\\
&+\varepsilon^2(I_p-\Phi(\alpha_0,t_k,t_{k-1}))R(\theta_0,t_{k-1}).
\end{aligned} \tag{3.4.13}$$

Hence, $A^i_n=D^i_n+C^i_n+C^{\prime,i}_n$ where, using (3.4.12),

$$D^i_n=2\sqrt{\Delta}\sum_{k=1}^n (H^i_k(\alpha_0,\beta))^*T_k(\theta_0), \tag{3.4.14}$$

$$C^i_n=2\varepsilon\sum_{k=1}^n (H^i_k(\alpha_0,\beta))^*(R(\theta_0,t_k)-R(\theta_0,t_{k-1})) \text{ and}$$

$$C^{\prime,i}_n=2\varepsilon\Delta\sum_{k=1}^n (H^i_k(\alpha_0,\beta))^*\frac{1}{\Delta}(I_p-\Phi(\alpha_0,t_k,t_{k-1}))R(\theta_0,t_{k-1}).$$

Let us first study $C^{\prime,i}_n$. Noting that

$$\frac{1}{\Delta}(I_p-\Phi(\alpha_0,t_k,t_{k-1}))=\nabla_z b(\alpha_0,z(\alpha_0,t_{k-1}))+\Delta O(1),$$

we have $|C^{\prime,i}_n|\le\varepsilon nC(\theta_0)$, with $C(\theta_0)$ bounded in probability.

To study C_n^i, we first apply an Abel transform to the sequence and get

$$C_n^i = 2\varepsilon \sum_{k=1}^{n} (H_{k-1}^i(\alpha_0,\beta) - H_k^i(\alpha_0,\beta))^* R(\theta_0,t_{k-1}) + \varepsilon H_k^n(\alpha_0,\beta)^* R(\theta_0,t_n).$$

The continuity assumptions ensure that $\sup_{k\le n} \frac{1}{\Delta} \left\| H_{k-1}^i(\alpha_0,\beta) - H_k^i(\alpha_0,\beta)) \right\|$ is bounded. Hence $C_n^i \to 0$ since $\|R(\theta_0,t_k)\|$ is uniformly bounded.

It remains to study the main term $D_n = (D_n^i)_{1\le i\le a}$ defined in (3.4.14). Let

$$H_k(\alpha_0,\beta) = \Sigma_{k-1}^{-1}(\beta)\nabla_\alpha b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1})).$$

Then (D_n) is a multidimensional triangular array which reads as $D_n = \sum_{k=1}^n \zeta_k^n$ with $\zeta_k^n = \sqrt{\Delta} H_k(\alpha_0,\beta)^* T_k(\theta_0) \in \mathbb{R}^a$.

Note that D_n does not depend on ε and convergence results are obtained for $\Delta_n \to 0$. To apply to (D_n) a theorem of convergence in law for triangular arrays (Theorem A.4.2 in the Appendix or [74] Theorem 2.2.14), we have to prove that,

(i) $\sum_{k=1}^n \mathbb{E}_{\theta_0}(\zeta_k^n | \mathscr{G}_{k-1}^n) = 0$,

(ii) $\sum_{k=1}^n \mathbb{E}_{\theta_0}(\zeta_k^{n,i}\zeta_k^{n,j} | \mathscr{G}_{k-1}^n) \to J_{ij}(\theta_0,\beta)$ (see Definition 3.4.10 below),

(iii) $\sum_{k=1}^n \mathbb{E}_{\theta_0}((\zeta_k^{n,i})^4 | \mathscr{G}_{k-1}^n) \to 0$.

Since $T_k(\alpha_0)$ is centered, (i) is clearly satisfied. For (ii), consider for $1 \le i,j \le a$,

$$\begin{aligned}\mathbb{E}_{\theta_0}(\zeta_k^{n,i}\zeta_k^{n,j} | \mathscr{G}_{k-1}^n) &= \Delta H_k^i(\alpha_0,\beta)^* \mathbb{E}_{\theta_0}(T_k(\theta_0)T_k^*(\theta_0)) H_k^j(\alpha_0,\beta) \\ &= \Delta H_k^i(\alpha_0,\beta)^* S_k(\alpha_0,\beta_0) H_k^j(\alpha_0,\beta).\end{aligned}$$

Noting that $\|S_k(\theta_0) - \Sigma(\beta_0,t_{k-1},z(\alpha_0,t_{k-1}))\| \le C\Delta$ yields, using Definition 3.4.11,

$$\begin{aligned}&\mathbb{E}_{\theta_0}(\zeta_k^{n,i}\zeta_k^{n,j} | \mathscr{G}_{k-1}^n) \\ &= \Delta(\nabla_{\alpha_i} b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1})))^* \Xi(\theta_0,\beta,t_{k-1}) \nabla_{\alpha_j} b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1})) \\ &\quad + \Delta^2 O(1).\end{aligned}$$

Therefore, as a Riemann sum,

$$\begin{aligned}&\sum_{k=1}^n \mathbb{E}_{\theta_0}(\zeta_k^{n,i}\zeta_k^{n,j} | \mathscr{G}_{k-1}^n) \\ &\quad \to \int_0^T (\nabla_{\alpha_i} b(\alpha_0,t,z(\alpha_0,t)))^* \Xi(\theta_0,\beta,t) \nabla_{\alpha_j} b(\alpha_0,t,z(\alpha_0,t)) dt.\end{aligned}$$

Checking (iii) is easily obtained since $\mathbb{E}_{\theta_0}((\zeta_k^{n,i})^4 | \mathscr{G}_{k-1}^n) \le \Delta^2 \sup_{k,\beta} \|H_k(\alpha_0,\beta)\|$.
Joining these results achieves the proof of Lemma 3.4.4. □

Using (3.4.8) and 3.4.9, it remains to study the term

$$\varepsilon^2 \nabla_\alpha^2 \check{U}_{\varepsilon,\Delta}(\alpha_0 + t(\check{\alpha}_{\varepsilon,\Delta} - \alpha_0), \check{\beta}_{\varepsilon,\Delta})$$

We have $\varepsilon^2 \nabla_{\alpha_i\alpha_j}^2 \check{U}_{\varepsilon,\Delta}(\alpha,\beta) = \sum_{l=1}^4 A_l^{ij}$ with

$$A_1^{ij} = \frac{2}{\Delta} \sum_{k=1}^n (\nabla_{\alpha_i} B_k(\alpha_0))^* \Sigma_{k-1}^{-1}(\beta) \nabla_{\alpha_j} B_k(\alpha_0),$$

$$\begin{aligned}
A_2^{ij} &= \frac{2}{\Delta}\sum_{k=1}^{n}(\nabla_{\alpha_i}B_k(\alpha)-\nabla_{\alpha_i}B_k(\alpha_0))^*\Sigma_{k-1}^{-1}(\beta)(\nabla_{\alpha_j}B_k(\alpha)+\nabla_{\alpha_j}B_k(\alpha_0)),\\
A_3^{ij} &= 2\sum_{k=1}^{n}\frac{1}{\Delta}(\nabla^2_{\alpha_i\alpha_j}B_k(\alpha))^*\Sigma_{k-1}^{-1}(\beta)B_k(\alpha_0),\\
A_4^{ij} &= 2\Delta\sum_{k=1}^{n}\frac{1}{\Delta}(\nabla^2_{\alpha_i\alpha_j}B_k(\alpha,X))^*\Sigma_{k-1}^{-1}(\beta)\frac{1}{\Delta}(B_k(\alpha,X)-B_k(\alpha_0,X)).
\end{aligned}$$

By Lemmas 3.3.6, 3.3.9 and 3.3.7, A_2^{ij} and A_4^{ij} satisfy $\left\|A_l^{ij}\right\| \le CT\|\alpha-\alpha_0\|$. Lemma 3.3.9 (ii) and Lemma 3.3.8 (ii) yield that $A_3^{ij}\to 0$.

The main term A_1^{ij} satisfies, by Lemma 3.3.9 (i),

$$\begin{aligned}
A_1^{ij} =& 2\Delta\sum_{k=1}^{n}(\nabla_{\alpha_i}b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1})))^*\Sigma_{k-1}^{-1}(\beta)\nabla_{\alpha_j}b(\alpha_0,t_{k-1},z(\alpha_0,t_{k-1}))\\
&+\varepsilon O_P(1).
\end{aligned}$$

Theorem 3.3.1 yields that, under $\mathbb{P}_{\theta_0}$, $\Sigma_{k-1}^{-1}(\beta)=\Sigma^{-1}(\beta,t,z(\alpha_0,t))+\varepsilon O_P(1)$. Therefore, as a Riemann sum, we get, using (3.4.3), that $A_1^{ij}\to(I_b(\alpha_0,\beta))_{ij}$ in $\mathbb{P}_{\theta_0}$-probability as $\varepsilon,\Delta\to 0$. Joining these results, we get that, under $\mathbb{P}_{\theta_0}$, as $\varepsilon,\Delta\to 0$, for all β, $\varepsilon^2\nabla^2_\alpha\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)\to 2I_b(\alpha_0,\beta)$. Using now the consistency of $\check{\alpha}_{\varepsilon,\Delta}$ yields that

$$\sup_{\beta\in K_b}\left\|\varepsilon^2\nabla^2_\alpha\check{U}_{\varepsilon,\Delta}(\alpha_0+t(\check{\alpha}_{\varepsilon,\Delta}-\alpha_0),\beta)-\varepsilon^2\nabla^2_\alpha\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta)\right\|\le K\left\|\check{\alpha}_{\varepsilon,\Delta}-\alpha_0)\right\|. \tag{3.4.15}$$

Coming back to (3.4.8), it remains to prove that $N_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\beta)$ is non-singular. Under (S3), $\Sigma(\beta,t,z)$ is non-singular. Hence,

$$\begin{aligned}
\inf_{\beta\in K_b}\det & \left(\left[\int_0^T\nabla_{\alpha_i}b(\alpha_0,t,z(\alpha_0,t))^*\Sigma^{-1}(\beta,t,z(\alpha_0,t))\nabla_{\alpha_j}b(\alpha_0,t,z(\alpha_0,t)dt\right]_{1\le i,j\le a}\right)\\
\ge & \quad c\det\left(\left[\int_0^T\nabla_{\alpha_i}b(\alpha_0,t,z(\alpha_0,t))^*\nabla_{\alpha_j}b(\alpha_0,t,z(\alpha_0,t)dt\right]_{1\le i,j\le a}\right)>0.
\end{aligned}$$

Now, the consistency of $\check{\alpha}_{\varepsilon,\Delta}$ implies that, using (3.4.9), $\mathbb{P}^{\varepsilon}_{\theta_0}(\det(\varepsilon^2N_{\varepsilon,\Delta}(\check{\alpha},\beta))>0)$ tends to 1. Therefore (3.4.8) yields

$$\varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta}-\alpha_0)=-(\varepsilon^2N^{-1}_{\varepsilon,\Delta}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta})(\varepsilon\nabla_\alpha\check{U}_{\varepsilon,\Delta}(\alpha_0,\check{\beta}_{\varepsilon,\Delta}))$$

is tight. □

3.4.2.3 Step (3): consistency of $\check{\beta}_{\varepsilon,\Delta}$

Let us now study the estimation for the diffusion parameter. Set

$$\begin{aligned}
K_2(\alpha_0,\beta_0,\beta) &= \tfrac{1}{T}\int_0^T\mathrm{Tr}\left(\Sigma^{-1}(\beta,t,z(\alpha_0,t))\Sigma(\beta_0,t,z(\alpha_0,t))\right)dt\\
&- \tfrac{1}{T}\int_0^T\log\det\left(\Sigma^{-1}(\beta,t,z(\alpha_0,t))\Sigma(\beta_0,t,z(\alpha_0,t))\right)\,dt\,-p
\end{aligned} \tag{3.4.16}$$

Using the following inequality for invertible symmetric $p\times p$ matrices A, $\log\det A+p\le\mathrm{Tr}(A)$, $K_2(\alpha_0,\beta_0,\beta)\ge 0$ and $K_2(\alpha_0,\beta_0,\beta)=0$ if and only if

$$\{\forall t\in[0,T],\Sigma(\beta_0,t,z(\alpha_0,t))=\Sigma(\beta,t,z(\alpha_0,t)),$$

which implies $\beta=\beta_0$ by (S7).

Proposition 3.4.5. *Assume (S1)–(S7). Then, if $I_b(\alpha_0,\beta_0)$ defined in* (3.4.3) *is non-singular, the following holds in $\mathbb{P}_{\theta_0}$-probability, using* (3.4.1), (3.4.2) *and* (3.4.16)

(i) $\sup_{\beta\in K_b}|\frac{1}{n}\left(\check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\beta)-\check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\beta_0)\right)-K_2(\alpha_0,\beta_0,\beta)|\to 0$ *as* $\varepsilon,\Delta\to 0$.

(ii) $\check{\beta}_{\varepsilon,\Delta}\to\beta_0$ *as* $\varepsilon,\Delta\to 0$.

Proof. Let us first prove (i). Using (3.4.1) and (3.3.9),we get
$\frac{1}{n}\left(\check{U}_{\Delta,\varepsilon}(\alpha,\beta)-\check{U}_{\Delta,\varepsilon}(\alpha,\beta_0)\right)=A_1(\beta_0,\beta)+A_2(\alpha,\beta_0,\beta)$ with

$$A_1(\beta_0,\beta)=\frac{1}{n}\sum_{k=1}^{n} log\det(\Sigma_{k-1}(\beta)\Sigma_{k-1}^{-1}(\beta_0)), \tag{3.4.17}$$

$$A_2(\alpha,\beta_0,\beta)=\frac{1}{n\Delta\varepsilon^2}\sum_{k=1}^{n} B_k(\alpha,X)^*(\Sigma_{k-1}^{-1}(\beta)-\Sigma_{k-1}^{-1}(\beta_0))B_k(\alpha,X). \tag{3.4.18}$$

Using that, under **(S5)**, $z\to\log\left(\det\left[\Sigma(\beta,t,z)\Sigma^{-1}(\beta_0,t,z)\right]\right)$ is differentiable, an application of Proposition 3.3.1 yields that, under $\mathbb{P}_{\theta_0}$,

$$\begin{aligned}&A_1(\beta_0,\beta)\\&=\frac{\Delta}{T}\left(\sum_{k=1}^{n}\log\left(\det\left[\Sigma(\beta,t_{k-1},z(\alpha_0,t_{k-1}))\Sigma^{-1}(\beta_0,t_{k-1},z(\alpha_0,t_{k-1}))\right]\right)+\varepsilon R^{1,\varepsilon}_{\theta_0,\beta}(t_{k-1})\right),\end{aligned}$$

with $\left\|R^{1,\varepsilon}_{\alpha_0,\beta,\beta_0}\right\|$ uniformly bounded in probability. Hence, $A_1(\beta_0,\beta)$, as a Riemann sum, converges to $\frac{1}{T}\int_0^T\log\left(\det\left[\Sigma(\beta,t,z(\alpha_0,t))\Sigma^{-1}(\beta\right.\right.$
as $\varepsilon,\Delta\to 0$.
Applying Lemma 3.3.8 to $B_k(\alpha_0,X)$ and the notations therein yields

$$A_2(\theta_0,\theta)=\frac{\Delta}{T}\sum_{k=1}^{n} Z_k^* M_k Z_k+\sum_{i=1}^{4}\Lambda^i(\theta_0,\theta), \tag{3.4.19}$$

with

$$Z_k=\frac{1}{\sqrt{\Delta}}\left(B(t_k)-B(t_{k-1})\right),\ T_k=\Sigma_{k-1}^{-1}(\beta)-\Sigma_{k-1}^{-1}(\beta_0),\ M_k=\sigma_{k-1}^*(\beta_0)T_k\ \sigma_{k-1}(\beta_0),$$

and

$$\begin{aligned}\Lambda_1(\alpha,\theta_0)&=\frac{2\sqrt{\Delta}}{\varepsilon}\sum_{k=1}^{n}E_k^*T_k\,Z_k,\\\Lambda_2(\alpha,\theta_0)&=\frac{1}{T\varepsilon^2}\sum_{k=1}^{n}E_k^*E_k,\\\Lambda_3(\alpha,\theta_0)&=\frac{2}{T\varepsilon^2}\sum_{k=1}^{n}(B_k^*(\alpha,X)-B_k^*(\alpha_0,X))T_k\,B_k(\alpha_0,X),\ \text{and}\\\Lambda_4(\alpha,\theta_0)&=\frac{1}{T\varepsilon^2}\sum_{k=1}^{n}(B_k^*(\alpha,X)-B_k^*(\alpha_0,X))T_k\,(B_k(\alpha,X)-B_k(\alpha_0,X)).\end{aligned}$$

The random vectors Z_k are $\mathcal{N}(0,I_p)$ independent of $\mathcal{G}_{k-1}^n$ and M_k is $\mathcal{G}_{k-1}^n$-measurable. Using that for $Z\sim\mathcal{N}(0,I_p)$, $E(Z^*MZ)=\mathrm{Tr}(M)$, we get

$$\mathbb{E}^{\varepsilon}_{\theta_0}(Z_k^*\,M_k\,Z_k|\mathcal{G}_{k-1}^n)=\mathrm{Tr}(M_k)=\mathrm{Tr}\left(\Sigma_{k-1}^{-1}(\beta)\Sigma_{k-1}(\beta_0)-I_p\right).$$

Hence, the first term of $A_2(\alpha_0,\beta_0,\beta)$ converges to

$$\frac{1}{T}\int_0^T\mathrm{Tr}\left(\Sigma^{-1}(\beta,t,z(\alpha_0,t))\Sigma(\beta_0,t,z(\alpha_0,t))\right)dt-p.$$

It remains to study the other terms of $A_2(\alpha_0,\beta_0,\beta)$. To study Λ_1, let $\zeta_k^n = \frac{\sqrt{\Delta}}{\varepsilon} E_k^* T_k\, Z_k$.
We have, by Lemma 3.3.8 that, in $\mathbb{P}_{\theta_0}$-probability,

$$\mathbb{E}(\zeta_k^n|\mathscr{G}_{k-1}^n) \leq \frac{\sqrt{\Delta}}{\varepsilon} \sup \|T_k\| \left(\mathbb{E}(\|E_k\|^2 |\mathscr{G}_{k-1}^n)\right)^{1/2} \leq C\Delta^{3/2}, \text{ and}$$
$$\mathbb{E}((\zeta_k^n)^2|\mathscr{G}_{k-1}^n) \leq \frac{\Delta}{\varepsilon^2} \sup \|T_k\|^2 \Delta^2\varepsilon^2 \leq C\Delta^3.$$

Therefore, by Lemma A.4.3, $\Lambda_1(\alpha,\theta_0) \to 0$ in $\mathbb{P}_{\theta_0}$-probability as $\varepsilon,\Delta \to 0$. Similar arguments yield that $\Lambda_2(\alpha,\theta_0) \to 0$ in $\mathbb{P}_{\theta_0}$-probability.
For $\Lambda_3(\alpha,\theta_0)$, set $\zeta_k^n = \frac{1}{\varepsilon^2}(B_k^*(\alpha,X) - B_k^*(\alpha_0,X))T_k\, B_k(\alpha_0,X)$ Using Lemma 3.3.7 yields that

$$E(\zeta_k^n|\mathscr{G}_{k-1}^n) \leq \frac{\|\alpha-\alpha_0)\|}{\varepsilon}\Delta^2 O_P(1), \text{ and}$$
$$\mathbb{E}((\zeta_k^n)^2|\mathscr{G}_{k-1}^n) \leq \frac{\|\alpha-\alpha_0\|^2}{\varepsilon^2}\Delta^3 O_P(1),$$

so that $\sum \mathbb{E}(\zeta_k^n|\mathscr{G}_{k-1}^n) \leq \Delta \frac{\left\|\check{\alpha}_{\varepsilon,\Delta}-\alpha_0)\right\|}{\varepsilon}$. By Proposition 3.4.3, the sequence $(\varepsilon^{-1}\left\|\check{\alpha}_{\varepsilon,\Delta}-\alpha_0\right\|)$ is uniformly bounded in probability, so that $\sum \mathbb{E}(\zeta_k^n|\mathscr{G}_{k-1}^n) \to 0$ and $\sum \mathbb{E}((\zeta_k^n)^2|\mathscr{G}_{k-1}^n) \to 0$.

Another application of Lemma A.4.3 yields that $\Lambda_3(\check{\alpha}_{\varepsilon,\Delta},\theta_0) \to 0$. For Λ_4, the result is straightforward since $|\Lambda_4| \leq n\Delta^2(\frac{\left\|\check{\alpha}_{\varepsilon,\Delta}-\alpha_0\right\|}{\varepsilon})^2$. This achieves the proof of (i).
Let us study (ii). We have, using (3.4.16),

$$\begin{aligned} 0 &\leq K_2(\alpha_0,\beta_0,\check{\beta}_{\varepsilon,\Delta}) \leq [K_2(\alpha_0,\beta_0,\check{\beta}_{\varepsilon,\Delta}) - \frac{1}{n}(\check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - \check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\beta_0))] \\ &+ \frac{1}{n}(\check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) - \check{U}_{\Delta,\varepsilon}(\check{\alpha}_{\varepsilon,\Delta},\beta_0)). \end{aligned}$$

Noting that the last term of the above inequality is non-negative, (i) yields that $K_2(\alpha_0,\beta_0,\check{\beta}_{\varepsilon,\Delta}) \to 0$, which ensures, by Assumption (S5), that $\check{\beta}_{\varepsilon,\Delta} \to \beta_0$ in $\mathbb{P}_{\theta_0}$-probability. □

3.4.2.4 Step (4): Asymptotic normality

Let us now study the asymptotic properties of these estimators and achieve the proof of Theorem 3.4.1. Let us define for $\theta = (\alpha,\beta)$,

$$\Lambda_{\varepsilon,n}(\theta) = \begin{pmatrix} \varepsilon\nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha,\beta) \\ \frac{1}{\sqrt{n}}\nabla_\beta \check{U}_{\varepsilon,\Delta}(\alpha,\beta) \end{pmatrix} \quad \text{and} \tag{3.4.20}$$

$$D_{\varepsilon,n}(\theta) = \begin{pmatrix} \varepsilon^2\left(\nabla^2_{\alpha_i,\alpha_j}\check{U}_{\varepsilon,\Delta}(\theta)\right)_{1\leq i,j\leq a} & \frac{\varepsilon}{\sqrt{n}}\left(\nabla^2_{\alpha_i\beta_j}\check{U}_{\varepsilon,\Delta}(\theta)\right)_{1\leq i\leq a,1\leq j\leq b} \\ \frac{\varepsilon}{\sqrt{n}}\left(\nabla^2_{\alpha_i\beta_j}\check{U}_{\varepsilon,\Delta}(\theta)\right)_{1\leq i\leq a,1\leq j\leq b} & \frac{1}{n}\left(\nabla^2_{\beta_i\beta_j}\check{U}_{\varepsilon,\Delta}(\theta)\right)_{1\leq i,j\leq b} \end{pmatrix}. \tag{3.4.21}$$

A Taylor expansion at point θ_0 yields,

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \end{pmatrix} &= \Lambda_{\varepsilon,n}(\check{\alpha}_{\varepsilon,\Delta},\check{\beta}_{\varepsilon,\Delta}) \\ &= \Lambda_{\varepsilon,n}(\theta_0) + \int_0^1 D_{\varepsilon,n}(\theta_0 + t(\check{\theta}_{\varepsilon,\Delta}-\theta_0))dt \begin{pmatrix} \varepsilon^{-1}(\check{\alpha}_{\varepsilon,\Delta}-\alpha_0) \\ \sqrt{n}(\check{\beta}_{\varepsilon,\Delta}-\beta_0) \end{pmatrix}. \end{aligned} \tag{3.4.22}$$

Therefore, we have to prove that, under $\mathbb{P}_{\theta_0}$, as $\varepsilon,\Delta \to 0$ (or $n = \Delta^{-1/2} \to \infty$),

(i) $-\Lambda_{\varepsilon,n}(\theta_0) \to \mathcal{N}(0, 4I(\theta_0))$ in distribution,

(ii) $\sup_{t\in[0,1]} \left\| D_{\varepsilon,n}(\theta_0 + t(\check{\theta}_{\varepsilon,\Delta} - \theta_0)) - 2I(\theta_0) \right\| \to 0$ in probability.

Proof. Let us prove (i). We have that, for $1 \le i \le a$,

$$-\varepsilon \nabla_{\alpha_i} \check{U}_{\varepsilon,\Delta}(\alpha_0, \beta_0) = \sum_{k=1}^{n} \xi_k^i(\theta_0) \text{ with } \xi_k^i(\theta_0) = -\frac{2}{\varepsilon\Delta} B_k^*(\alpha)\Sigma_{k-1}^{-1}(\beta_0)\nabla_{\alpha_i} B_k(\alpha_0). \tag{3.4.23}$$

Using that, for a positive symmetric matrix $M(x)$,

$$\frac{d}{dx}(\log\det M(x)) = \mathrm{Tr}\left(M^{-1}(x)\frac{d}{dx}M(x)\right)$$

and (3.3.9), set

$$M_k^j(\beta) = \Sigma_k^{-1}(\beta)\nabla_{\beta_j}\Sigma_k(\beta). \tag{3.4.24}$$

Then $\frac{1}{\sqrt{n}}\nabla_{\beta_j}\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0) = \sum_{k=1}^n \eta_k^i(\theta_0)$ with

$$\eta_k^j(\theta_0) = \frac{1}{\sqrt{n}}[\mathrm{Tr}(M_{k-1}^j(\beta_0)) - \frac{1}{\varepsilon^2\Delta} B_k^*(\alpha_0) M_{k-1}^j(\beta_0)\Sigma_{k-1}^{-1}(\beta_0)B_k(\alpha_0, X)]. \tag{3.4.25}$$

The proof that $-\varepsilon\nabla_\alpha \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0)$ converges to the Gaussian distribution $\mathcal{N}(0, I_b(\theta_0))$ is obtained by substituting β with β_0 in the proof of Proposition 3.4.3.

Let us study $-\frac{1}{\sqrt{n}}\nabla_\beta \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0)$. Let us first prove

Lemma 3.4.6. *If M is a $\mathcal{G}_{k-1}^n$-measurable random matrix, then*

$$\frac{1}{\varepsilon^2\Delta}\mathbb{E}\left(B_k^*(\alpha_0)M\Sigma_{k-1}^{-1}(\beta_0)B_k(\alpha_0,X)|\mathcal{G}_{k-1}^n\right) = \mathrm{Tr}(M) + \Delta R_k(\varepsilon,\Delta) \tag{3.4.26}$$

with $\sup_k |R_k(\varepsilon,\Delta)|$ uniformly bounded in $\mathbb{P}_{\theta_0}$-probability.

Proof. Using Lemma 3.3.8,

$$\begin{aligned}
&\mathbb{E}(B_k^*(\alpha_0)M\Sigma_{k-1}^{-1}(\beta_0)B_k(\alpha_0)|\mathcal{G}_{k-1}^n) = \sum_{l,l'=1}^p \left(M\Sigma_{k-1}^{-1}(\beta_0)\right)_{ll'}\mathbb{E}(B_k^l(\alpha_0)B_k^{l'}(\alpha_0)|\mathcal{G}_{k-1}^n)\\
&= \varepsilon^2\Delta\sum_{l,l'=1}^p \left(M\Sigma_{k-1}^{-1}(\beta_0)\right)_{ll'}(\Sigma_{k-1}(\beta_0))_{l'l} + \sum_{l,l'=1}^p \left(M\Sigma_{k-1}^{-1}(\beta_0)\right)_{ll'}\mathbb{E}(E_k^l E_k^{l'}|\mathcal{G}_{k-1}^n)\\
&= \varepsilon^2\Delta\mathrm{Tr}(M) + R_k(\varepsilon,\Delta)
\end{aligned}$$

with $|R_k(\varepsilon,\Delta)| \le C\varepsilon^2\Delta^2$ in probability. □

Let us study the convergence of the triangular array $\sum_{k=1}^n \mathbb{E}(\xi_k^i(\theta_0))$. By Lemma 3.4.6, we have for $j \le b$,

$$\sum_{k=1}^n \mathbb{E}(\eta_k^j(\theta_0)|\mathcal{G}_{k-1}^n) = \frac{1}{\varepsilon^2\Delta\sqrt{n}}\sum_{k=1}^n R_k(\varepsilon,\Delta) \le \frac{CT}{\sqrt{n}} \to 0.$$

Consider now, for $j_1, j_2 \le b$, $\sum_{k=1}^n \mathbb{E}(\eta_k^{j_1}(\theta_0)\eta_k^{j_2}(\theta_0)|\mathcal{G}_{k-1}^n)$.

We have

$$\begin{aligned}
&\mathbb{E}(\eta_k^{j_1}(\theta_0)\eta_k^{j_2}(\theta_0)|\mathcal{G}_{k-1}^n)\\
&= \frac{1}{n}[\mathrm{Tr}(M_{k-1}^{j_1}(\beta_0)M_{k-1}^{j_2}(\beta_0)) - 2\mathrm{Tr}M_{k-1}^{j_1}(\beta_0))\mathrm{Tr}M_{k-1}^{j_2}(\beta_0) + C_k^{j_1,j_2}(\varepsilon,\Delta) + \Delta O_P(1)],
\end{aligned}$$

with

$$C_k^{j_1 j_2}(\varepsilon,\Delta) = \frac{1}{\varepsilon^4\Delta^2}\mathbb{E}\left(B_k^*(\alpha_0)M_{k-1}^{j_1}(\beta_0)\Sigma_{k-1}^{-1}(\beta_0)B_k(\alpha_0)B_k^*(\alpha_0)M_{k-1}^{j_2}(\beta_0)\Sigma_{k-1}^{-1}(\beta_0)B_k(\alpha_0)|\mathscr{G}_{k-1}^n\right).$$

Therefore, omitting the parameters when there is no ambiguity,

$$C_k^{j_1 j_2}(\varepsilon,\Delta) = \sum_{l_1,l_2,l_3,l_4}(M_{k-1}^{j_1}\Sigma_{k-1}^{-1})_{l_1 l_2}(M_{k-1}^{j_2}\Sigma_{k-1}^{-1})_{l_3 l_4}\mathbb{E}\left(B_k^{l_1}(\alpha_0)B_k^{l_2}(\alpha_0)B_k^{l_3}(\alpha_0)B_k^{l_4}(\alpha_0)|\mathscr{G}_{k-1}^n\right).$$

Based on the property that, if Z is a p-dimensional Gaussian random variable $\mathscr{N}(0,\Sigma)$, $E(Z_{l_1}Z_{l_2}Z_{l_3}Z_{l_4}) = \Sigma_{l_1 l_2}\Sigma_{l_3 l_4} + \Sigma_{l_1 l_3}\Sigma_{l_2 l_4} + \Sigma_{l_1 l_4}\Sigma$ we get that

$$C_k^{j_1 j_2}(\varepsilon,\Delta) = \left(\mathrm{Tr}(M_{k-1}^{j_1}M_{k-1}^{j_2}) + 2\mathrm{Tr}M_{k-1}^{j_1}\mathrm{Tr}M_{k-1}^{j_2} + \Delta O_P(1)\right).$$

Therefore $\sum_{k=1}^n \mathbb{E}(\eta_k^{j_1}(\theta_0)\eta_k^{j_2}(\theta_0)|\mathscr{G}_{k-1}^n) = \frac{2}{n}\sum_{k=1}^n \mathrm{Tr}(M_{k-1}^{j_1}M_{k-1}^{j_2}) + \Delta O_P(1)$.
Now, under $\mathbb{P}_{\theta_0}$, $M_k^j(\beta_0) = \Sigma^{-1}(\beta_0,t_k,z(\alpha_0,t_k))\nabla_{\beta_j}\Sigma(\beta_0,t_k,z(\alpha_0,t_k)) + \varepsilon O_P(1)$ so that, using (3.4.4), as $\varepsilon,\Delta\to 0$,

$$\sum_{k=1}^n \mathbb{E}(\eta_k^{j_1}(\theta_0)\eta_k^{j_2}(\theta_0)|\mathscr{G}_{k-1}^n) \to 4(I_\sigma(\theta_0))_{j_1 j_2}.$$

The proofs that $\sum_{k=1}^n \mathbb{E}(\left\|\eta_k^i(\theta_0)\right\|^4|\mathscr{G}_{k-1}^n)\to 0$, $\sum_{k=1}^n \mathbb{E}(\xi_k^i(\theta_0)\eta_k^j(\theta_0)|\mathscr{G}_{k-1}^n)\to 0$ are similar and omitted. Finally, applying the theorem of convergence in law for triangular arrays recalled in Section A.4 yields that $\sum_{k=1}^n \eta_k^i(\theta_0)\to\mathscr{N}(0,4I_\sigma(\theta_0))$. Joining these results achieves the proof of (i). □

It remains to study $D_{\varepsilon,n}(\theta)$ defined in (3.4.21).

Proof. We have already proved that

$$\sup_{t\in[0,1]}\left\|\varepsilon^2(\nabla^2_{\alpha_i,\alpha_j}\check{U}_{\varepsilon,\Delta}(\theta_0+t(\check{\theta}_{\varepsilon,\Delta}-\theta_0)) - 2(I_b(\theta_0))_{ij}\right\|\to 0$$

in probability. Consider now the term $\frac{1}{n}\nabla^2_{\beta_i,\beta_j}\check{U}_{\varepsilon,\Delta}(\alpha,\beta)$. It reads as

$$\begin{aligned}\nabla^2_{\beta_i\beta_j}\check{U}_{\varepsilon,\Delta}(\alpha,\beta) &= \sum_{k=1}^n\left(\mathrm{Tr}\left(\nabla_{\beta_i}M_{k-1}^j(\beta)\right) - \frac{1}{\varepsilon^2\Delta}B_k(\alpha)^*(\nabla_{\beta_i}M_{k-1}^j(\beta))\Sigma_{k-1}^{-1}(\beta)B_k(\alpha)\right)\\ &\quad + \frac{1}{\varepsilon^2\Delta}\sum_{k=1}^n B_k(\alpha)^*M_{k-1}^j(\beta)M_{k-1}^i(\beta)\Sigma_{k-1}^{-1}(\beta)B_k(\alpha).\end{aligned}$$

Let us define the matrices, for $1\le i,j\le b$,

$$M^i(\alpha,\beta,t) = \Sigma^{-1}(\beta,t,z(\alpha,t))\nabla_{\beta_i}\Sigma(\beta,t,z(\alpha,t)), \quad \text{and} \tag{3.4.27}$$

$$T_k^{ij}(\beta) = [M_k^j(\beta)M_k^i(\beta) - \nabla_{\beta_i}M_k^j(\beta)]\Sigma_{k-1}^{-1}(\beta). \tag{3.4.28}$$

Using (3.4.26) yields that the first term of $\nabla^2_{\beta_i\beta_j}\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0)$ is uniformly bounded in probability and that the second term satisfies $\sum_{k=1}^n(\mathrm{Tr}\left(M_{k-1}^j(\beta_0)M_{k-1}^i(\beta_0)\right) + \Delta O_P(1))$. Hence,

$$\frac{1}{n}\nabla^2_{\beta_i\beta_j}\check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0)\to -\frac{1}{T}\int_0^T \mathrm{Tr}(M^j(\alpha_0,\beta_0,t)M^i(\alpha_0,\beta_0,t))dt.$$

It remains to prove that, under $\mathbb{P}_{\theta_0}$,

$$\sup_{t\in[0,1]} \frac{1}{n} \left\| \nabla^2_{\beta_i\beta_j} \check{U}_{\varepsilon,\Delta}(\theta_t) - \nabla^2_{\beta_i\beta_j} \check{U}_{\varepsilon,\Delta}(\theta_0) \right\| \to 0$$

with $\theta_t = \theta_0 + t(\check{\theta}_{\varepsilon,\Delta} - \theta_0)$ and that the terms

$$\frac{\varepsilon}{\sqrt{n}} (\nabla^2_{\alpha_i\beta_j} \check{U}_{\varepsilon,\Delta}(\alpha,\beta) - \nabla^2_{\alpha_i\beta_j} \check{U}_{\varepsilon,\Delta}(\alpha_0,\beta_0)) \to 0.$$

These two proofs rely on similar tools and are omitted. □

3.5 Inference based on low frequency observations

Consider now the case where the sampling interval Δ is fixed and the time interval for observations is fixed. It follows that the number of observation points $n = T/\Delta$ is finite. We prove that only parameters in the drift function can be consistently estimated. This agrees with the previous results where the rate of estimation of parameter β in the diffusion coefficient is $\sqrt{n}$ in the high frequency set-up. Sometimes, when modeling epidemic dynamics, a parameter is added in the *SIR* model to take account of larger fluctuations, substituting the term $\sqrt{SI}$ by $(S(t)I(t))^a$ in the diffusion term. While in the "High frequency" set-up, this parameter a can be consistently estimated, this is no longer true for a fixed sampling interval.

In order to illustrate that β cannot be consistently estimated in this set-up, we study the inference on a simple example, the one-dimensional Brownian motion with drift on $[0,T]$.

3.5.1 Preliminary result on a simple example

Let us consider the estimation of (α,β) as $\varepsilon \to 0$ and $n = T/\Delta$ finite, for the process

$$dX(t) = \alpha dt + \varepsilon\beta dB(t); \quad X(0) = 0. \tag{3.5.1}$$

The observations are $(X(t_k), k = 1,\dots,n)$. The n random variables $(X(t_k) - X(t_{k-1}))$ are independent Gaussian with distribution $\mathcal{N}(\alpha\Delta, \varepsilon^2\beta^2\Delta)$. The likelihood is explicit and the maximum likelihood estimators are

$$\hat{\alpha}_\varepsilon = \frac{X(T)}{T}; \quad \hat{\beta}^2_\varepsilon = \frac{1}{n\Delta\varepsilon^2} \sum_{k=1}^{n} (X(t_k) - X(t_{k-1}) - \Delta\hat{\alpha}_{\varepsilon,\Delta})^2. \tag{3.5.2}$$

Under $\mathbb{P}_{\theta_0}$, $\hat{\alpha}_\varepsilon = \alpha_0 + \varepsilon\beta_0 \frac{B(T)}{T}$. Therefore, as $\varepsilon \to 0$, $\hat{\alpha}_\varepsilon \to \alpha_0$ and $\varepsilon^{-1}(\hat{\alpha}_\varepsilon - \alpha_0) = \beta_0 \frac{B(T)}{T}$ is a Gaussian random variable $\mathcal{N}(0, \frac{\beta_0^2}{T})$.

The MLE of β_0^2 is $\hat{\beta}^2_\varepsilon = \beta_0^2 (\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - \frac{1}{n}\frac{B(T)^2}{T})$, where $(Z_k, k = 1,\dots,n)$ are i.i.d. $\mathcal{N}(0,1)$.

Hence, since n is fixed, $\hat{\beta}^2_\varepsilon$ is a fixed random variable which does not depend on ε with expectation $\beta_0^2(1 - \frac{1}{n}) \neq \beta_0^2$, implying that it is a biased estimator of β_0^2.

This simple example shows that parameters in the diffusion coefficient cannot be estimated as $\varepsilon \to 0$.

3.5.2 Inference for diffusion approximations of epidemics

Considering equation (3.3.1), three cases might occur: β unknown; β known or $\Sigma(\beta,x) = \phi(\beta)\Sigma(x)$ (with $\phi(\beta)$ a known real function on $\mathbb{R}^+$); β present in the drift coefficient (e.g. $\beta = \varphi(\alpha)$ with φ a known function). This last case systematically occurs for the diffusion approximation of epidemic dynamics: the parameters ruling the jump process modeling the epidemic dynamics are both present in the drift and in the diffusion coefficients, i.e. $\beta \equiv \alpha$.

For example, the diffusion approximation of the *SIR*, we have, setting $\alpha = (\lambda, \gamma)$, that the drift term is $b(\alpha, z)$ and the diffusion term is $\Sigma(\alpha, z)$

Having in mind epidemics, we study here this case and assume that, under $\mathbb{P}_\alpha$,

$$dX(t) = b(\alpha, t, X(t))dt + \varepsilon\sigma(\alpha, t, X(t))dB(t), \quad X(0) = x. \tag{3.5.3}$$

The time interval is $[0, T]$, the sampling interval is Δ with $T = n\Delta$, and both T, Δ, n are fixed.

The observations consist of the n random variables $(X(t_k), k = 1, \ldots, n)$ with $t_k = k\Delta$. As in the previous section, the inference is based on the random variables $B_k(\alpha, X)$ defined in (3.3.7), which satisfy using Lemma 3.3.4

$$B_k(\alpha, X) = \varepsilon\sqrt{\Delta}T_k(\alpha) + \varepsilon^2 D_k^\varepsilon(\alpha), \text{ with } D_k^\varepsilon = R^\varepsilon(\alpha, t_k) - \Phi(\alpha, t_k, t_{k-1})R^\varepsilon(\alpha, t_{k-1}). \tag{3.5.4}$$

$$T_k(\alpha) = \frac{1}{\sqrt{\Delta}} \int_{t_{k-1}}^{t_k} \Phi(\alpha, t_k, u)\sigma(\alpha, u, z(\alpha, u))dB(u), \tag{3.5.5}$$

$$S_k(\alpha) = \frac{1}{\Delta} \int_{t_{k-1}}^{t_k} \Phi(\alpha, t_k, u)\Sigma(\alpha, u, z(\alpha, u))\Phi^*(\alpha, t_k, u)du. \tag{3.5.6}$$

This leads to define the contrast function depending now on $(X(t_1), \ldots, X(t_n))$,

$$\bar{U}_\varepsilon\left(\alpha, (X_{t_k})\right) = \bar{U}_\varepsilon(\alpha) = \sum_{k=1}^n \log\det S_k(\alpha) + \frac{1}{\varepsilon^2\Delta} \sum_{k=1}^n B_k^*(\alpha, X)S_k^{-1}(\alpha)B_k(\alpha, X). \tag{3.5.7}$$

Denote by α_0 the true value of the parameter and Θ the parameter set. We assume

(S4b): Θ a compact set of $\mathbb{R}^a$; $\alpha \in \text{Int}(\Theta)$.

(S5b): Assumption (S5) on $b(\alpha, t, z)$ and $\sigma(\alpha, t, z)$.

(S6b): $\alpha \neq \alpha_0 \Rightarrow \{\exists k,\ 1 \leq k \leq n,\ z(\alpha, t_k) \neq z(\alpha_0, t_k)\}$.

The estimator is defined as any solution of

$$\bar{\alpha}_\varepsilon = \underset{\alpha \in K_a}{\text{argmin}}\ \bar{U}_\varepsilon\left(\alpha, (X_{t_k})\right). \tag{3.5.8}$$

Let us study the properties of $\bar{\alpha}_\varepsilon$. For this, define, using (3.5.6), the $p \times a$ matrix $G_k(\alpha) = (G_k^1, \ldots, G_k^a)$ and the $a \times a$ matrix $M(\alpha)$,

$$M(\alpha) = \Delta \sum_{k=1}^n G_k(\alpha)^* S_k(\alpha)^{-1} G_k(\alpha), \text{ with} \tag{3.5.9}$$

$$G_k^i(\alpha) = \frac{1}{\Delta}(-\nabla_{\alpha_i} z(\alpha, t_k) + \Phi(\alpha, t_k, t_{k-1})\nabla_{\alpha_i} z(\alpha, t_{k-1})). \tag{3.5.10}$$

Then, the following holds

Theorem 3.5.1. *Assume (S1)–(S3), (S4b)–(S6b). Then, as $\varepsilon \to 0$, under $\mathbb{P}_{\alpha_0}$,*

(i) $\bar{\alpha}_\varepsilon \to \alpha_0$ in probability.

(ii) If moreover $M(\alpha_0)$ defined in (3.5.9) *is non-singular, then*

$$\varepsilon^{-1}(\bar{\alpha}_\varepsilon - \alpha_0) \to \mathcal{N}_a(0, M^{-1}(\alpha_0))$$

in distribution.

Proof. Let us first prove (i). Define, using (3.5.4), (3.5.6),

$$\bar{K}_\Delta(\alpha_0,\alpha) = \frac{1}{\Delta}\sum_{k=1}^{n} B_k^*(\alpha, z(\alpha_0,\cdot))S_k^{-1}(\alpha)B_k(\alpha, z(\alpha_0,\cdot)). \tag{3.5.11}$$

Since $B_k(\alpha_0, z(\alpha_0,\cdot)) = 0$, $\bar{K}_\Delta(\alpha_0,\alpha) \geq 0$ and $\bar{K}_\Delta(\alpha_0,\alpha_0) = 0$. Assume now that $\bar{K}_\Delta(\alpha_0,\alpha) = 0$. Then, for all $k \in \{1,..n\}$,

$$z(\alpha,t_k) - z(\alpha_0,t_k) = \Phi(\alpha,t_k,t_{k-1})(z(\alpha,t_{k-1}) - z(\alpha_0,t_{k-1})).$$

The matrix $\Phi(\alpha,t_k,t_{k-1})$ being non-singular, the identifiability Assumption **(S6b)** implies that $\alpha = \alpha_0$.
Since the sum in (3.5.7) is finite, we get, using (3.3.7) and Proposition 3.3.1, that $\sup_{\alpha\in K_a} |\varepsilon^2 \bar{U}_\varepsilon(\alpha) - \bar{K}_\Delta(\alpha_0,\alpha)| \to 0$ in $\mathbb{P}_{\theta_0}$-probability as $\varepsilon \to 0$. Therefore, we have

$$\begin{aligned} 0 &\leq \bar{K}_\Delta(\alpha_0,\bar{\alpha}_\varepsilon) - \bar{K}_\Delta(\alpha_0,\alpha_0) \\ &\leq 2\sup_{\alpha\in K_a} |\varepsilon^2 U_\varepsilon(\alpha) - \bar{K}_\Delta(\alpha_0,\alpha)| + \varepsilon^2 |U_\varepsilon(\bar{\alpha}) - U_\varepsilon(\alpha_0)| \\ &\leq 2\sup_{\alpha\in K_a} |\varepsilon^2 U_\varepsilon(\alpha) - \bar{K}_\Delta(\alpha_0,\alpha)|. \end{aligned}$$

Then the proof of (i) is achieved by means of the identifiability Assumption **(S6b)**.

Let us now prove (ii). To study the asymptotic properties of $\bar{\alpha}_\varepsilon$ as $\varepsilon \to 0$, we write, for $i,j \leq a$,

$$\begin{aligned} 0 &= \varepsilon \nabla_{\alpha_i} \bar{U}_\varepsilon(\bar{\alpha}_\varepsilon) \\ &= \varepsilon \nabla_{\alpha_i} \bar{U}_\varepsilon(\alpha_0) + \varepsilon^2 \sum_{j=1}^{a} \Big(\int_0^1 (\nabla^2_{\alpha_j\alpha_i} \bar{U}_\varepsilon(\alpha_0 + t(\bar{\alpha}_\varepsilon - \alpha_0))dt \Big) \Big(\frac{\bar{\alpha}_\varepsilon^j - \alpha_0^j}{\varepsilon} \Big). \end{aligned}$$

Consider first $\varepsilon\nabla_\alpha \bar{U}_\varepsilon(\alpha_0)$. Using (3.3.7) and (3.5.6), for $i = 1,\ldots,a$, it reads as

$$\begin{aligned} \varepsilon \nabla_{\alpha_i} \bar{U}_\varepsilon(\alpha_0) &= \varepsilon \sum_{k=1}^{n} \nabla_{\alpha_i} \log\det S_k(\alpha_0) + \frac{1}{\varepsilon\Delta}\sum_{k=1}^{n} B_k^*(\alpha_0)\, \nabla_{\alpha_i} S_k^{-1}(\alpha_0) B_k(\alpha_0) \\ &+ \frac{2}{\varepsilon\Delta}\sum_{k=1}^{n} (\nabla_{\alpha_i} B_k^*(\alpha_0))\, S_k^{-1}(\alpha_0) B_k(\alpha_0) = A_1^i(\alpha_0) + A_2^i(\alpha_0) + A_3^i(\alpha_0). \end{aligned}$$

Since $\nabla_{\alpha_i} \log(\det S_k(\alpha_0)) = \mathrm{Tr}(S_k^{-1}(\alpha_0)\nabla_{\alpha_i} S_k(\alpha_0))$, $A_1^i(\alpha_0)$ is well defined and, under the regularity assumptions, $A_1^i(\alpha_0) = n\varepsilon O(1)$, which goes to 0 as $\varepsilon \to 0$, n being fixed.

Applying Lemma 3.3.4 for the variables $T_k(\alpha_0)$, $D_k^\varepsilon(\alpha_0)$ yields that

$$\begin{aligned} A_2^i(\alpha_0) = \varepsilon &\sum_{k=1}^{n} T_k^*(\alpha_0)\nabla_{\alpha_i} S_k^{-1}(\alpha_0) T_k(\alpha_0) \\ &+ 2\frac{\varepsilon}{\sqrt{\Delta}}\sum_{k=1}^{n} T_k^*(\alpha_0)\nabla_{\alpha_i} S_k^{-1}(\alpha_0)(\varepsilon D_k^\varepsilon(\alpha_0)) \\ &+ \frac{\varepsilon}{\Delta}\sum_{k=1}^{n} (\varepsilon D_k^\varepsilon(\alpha_0))^* \nabla_{\alpha_i} S_k^{-1}(\alpha_0)(\varepsilon D_k^\varepsilon(\alpha_0)). \end{aligned}$$

It follows from Lemma 3.3.4, that $\sup_k \|\varepsilon D_k^\varepsilon(\alpha_0)\|$ is bounded. Therefore, $A_2^i(\alpha_0) \to 0$ in $\mathbb{P}_{\alpha_0}$-probability.

Let us study the main term $(A_3^i(\alpha)$ of $\varepsilon\nabla_{\alpha_i}\bar{U}_\varepsilon(\alpha_0)$.
Using Proposition 3.3.1 and (3.3.7), (3.5.10) yields that, under $\mathbb{P}_{\alpha_0}$,

$$\nabla_{\alpha_i} B_k(\alpha_0) = \Delta G_k^i(\alpha_0) - \varepsilon(\nabla_{\alpha_i}\Phi(\alpha_0,t_k,t_{k-1})(g(\alpha_0,t_{k-1}) + \varepsilon R^\varepsilon(\alpha_0,t_{k-1}))), \tag{3.5.12}$$

where $\sup_k \|\varepsilon R(\alpha,t_k)\|$ is uniformly bounded in probability. Therefore,

$$A_3^i(\alpha_0) = 2\sqrt{\Delta}\sum_{k=1}^{n} ((G_k^i(\alpha_0))^* S_k^{-1}(\alpha_0) T_k(\alpha_0) + \varepsilon R_k'(\alpha_0)),$$

with $R_k'(\alpha_0)$ uniformly bounded in probability. By Lemma 3.3.4, $(T_k(\alpha_0)), k = 1,\dots,n)$ are independent centered Gaussian random variables with covariance matrix $S_k(\alpha_0)$. We that $A_3(\alpha_0) = (A_3^1(\alpha_0),\dots,A_3(\alpha_0))^*$ converges to the Gaussian random variable $\mathcal{N}_a(0,4M(\alpha_0))$. Joining all these results yields that

$$-\varepsilon\nabla_\alpha \bar{U}_\varepsilon(\alpha_0) \to \mathcal{N}_a(0,4M(\alpha_0)) \text{ with}$$

$$M(\alpha_0) = (M(\alpha_0))_{ij} = \Delta\sum_{k=1}^n (G_k^i(\alpha_0))^* S_k^{-1}(\alpha_0) G_k^j(\alpha_0).$$

Consider $\varepsilon^2\nabla^2_{\alpha_j\alpha_i}\bar{U}_\varepsilon(\alpha)$. Similar computations yield that

$$\varepsilon^2\nabla^2_{\alpha_j\alpha_i}\bar{U}_\varepsilon(\alpha_0) = 2\Delta\sum_{k=1}^n (G_k^i(\alpha_0))^* S_k^{-1}(\alpha_0) G_k^j(\alpha_0) + n\varepsilon O_P(1).$$

Therefore, for all $1 \le i,j \le a$,

$$\varepsilon^2\nabla^2_{\alpha_i\alpha_j}\bar{U}_\varepsilon(\alpha_0) \to 2M_{ij}(\alpha_0) \quad \mathbb{P}_{\alpha_0}\text{a.s. as } \varepsilon\to 0.$$

It remains to study $\sup_{t\in[0,1]} |\varepsilon^2\nabla^2_{\alpha_j\alpha_i}\bar{U}_\varepsilon(\alpha_0 + t(\bar{\alpha}_\varepsilon - \alpha_0)) - \varepsilon^2\nabla^2_{\alpha_j\alpha_j}\bar{U}_\varepsilon(\alpha_0)|$.
We have $\varepsilon^2\nabla^2_{\alpha_j\alpha_i}\bar{U}_\varepsilon(\alpha) = \frac{1}{\Delta}(A_1^{ij}(\alpha) + A_2^{ij}(\alpha))$, where

$$A_1^{ij}(\alpha) = 2\sum_{k=1}^n \nabla_{\alpha_i} B_k^*(\alpha) S_k^{-1}(\alpha)\nabla_{\alpha_j} B_k(\alpha), \quad A_2^{ij}(\alpha) = \sum_{k=1}^n Z_k^*(\alpha) B_k(\alpha)$$

with

$$\begin{aligned} Z_k^*(\alpha) = {} & 2\nabla_{\alpha_j} B_k^*(\alpha)\nabla_{\alpha_i} S_k^{-1}(\alpha) + B_k^*(\alpha)\nabla^2_{\alpha_i\alpha_j} S_k^{-1}(\alpha) + 2\nabla_{\alpha_i} B_k^*(\alpha)\nabla_{\alpha_j} S_k^{-1}(\alpha) \\ & + 2\nabla^2_{\alpha_i\alpha_j} B_k^*(\alpha) S_k^{-1}(\alpha). \end{aligned}$$

Similarly to the previous section, we need that, under $\mathbb{P}_{\alpha_0}$, the properties stated below hold.

$$\|B_k(\alpha) - B_k(\alpha_0)\| \le \|\alpha - \alpha_0\| (C_1 + C_2 O_P(1)) \text{ uniformly with respect to } k,\alpha; \tag{3.5.13}$$

$$\left\| \frac{1}{\varepsilon} B_k(\alpha_0) \right\| \text{ are uniformly bounded random variables;} \tag{3.5.14}$$

$$\sup_{k\le n, \alpha\in\Theta} \|\nabla_{\alpha_i} B_k(\alpha)\| = O_P(1); \quad \text{and } \|\nabla_{\alpha_i} B_k(\alpha) - \nabla_{\alpha_i} B_k(\alpha_0)\| \le C_1 \|\alpha - \alpha_0\|. \tag{3.5.15}$$

The proofs of these properties are similar to the previous section and omitted.

Therefore,

$$A_2^{ij}(\alpha) - A_2^{ij}(\alpha_0) = \sum_{k=1}^n (Z_k^*(\alpha) - Z_k^*(\alpha_0)) B_k(\alpha_0) + \sum_{k=1}^n Z_k^*(\alpha)(B_k(\alpha) - B_k(\alpha_0)).$$

Using (3.5.13), (3.5.14) we get

$$|A_2^{ij}(\alpha) - A_2^{ij}(\alpha_0)| \le \sup \|Z_k(\alpha)\| \left(2n\varepsilon \sup \left\| \frac{B_k(\alpha_0)}{\varepsilon} \right\| + \|\alpha - \alpha_0\| (C_1 + C_2 O_P(1)).$$

Consider now $A_1^{ij}(\alpha) - A_1^{ij}(\alpha_0)$. It reads as

$$\begin{aligned} A_1^{ij}(\alpha) - A_1^{ij}(\alpha_0) = {} & 2\sum_{k=1}^n [\nabla_{\alpha_i} B_k^*(\alpha) S_k^{-1}(\alpha)(\nabla_{\alpha_j} B_k(\alpha) - \nabla_{\alpha_j} B_k(\alpha_0))] \\ & + [\nabla_{\alpha_j} B_k^*(\alpha) S_k^{-1}(\alpha)(\nabla_{\alpha_i} B_k(\alpha) - \nabla_{\alpha_i} B_k(\alpha_0))] \end{aligned}$$

$$+ [\nabla_{\alpha_i} B_k^*(\alpha)(S_k^{-1}(\alpha) - S_k^{-1}(\alpha_0))\nabla_{\alpha_j} B_k(\alpha_0)].$$

Hence, $\left\| A_1 ij(\alpha) - A_1^{ij}(\alpha_0) \right\| \leq 2nC \|\alpha - \alpha_0\|$.

Using the consistency $\bar{\alpha}_\varepsilon$, we get that

$$\sup_{t\in[0,1]} |\varepsilon^2 \nabla^2_{\alpha_j\alpha_i)} \bar{U}_\varepsilon(\alpha_0 + t(\bar{\alpha}_\varepsilon - \alpha_0)) - \varepsilon^2 \nabla^2_{\alpha_j\alpha_j} \bar{U}_\varepsilon(\alpha_0)| \to 0.$$

This achieves the proof of (ii) and of Theorem 3.5.1. □

3.5.2.1 Comments

(1) The term $\sum_{k=1}^n \log\det S_k(\alpha)$ could have been omitted in the definition of $\bar{U}_\varepsilon(\alpha)$. It has no influence on the asymptotic properties of $\bar{\alpha}_\varepsilon$. However, we have observed in the simulation results that it yields better estimators (less biased). An explanation lies in the fact that in practice ε is small, but probably not enough to compensate this first term. the observations of less biased estimators non-asymptotically.

(2) In [61], we considered the case of an unknown parameter β in the diffusion coefficient and therefore used a Conditional Least Square estimator based on $U_\varepsilon(\alpha) = \sum_{k=1}^n B_k^*(\alpha) B_k(\alpha)$. The CLS estimator obtained is consistent. It converges at the same rate, but with a larger covariance matrix $J_\Delta^{-1}(\alpha) I_\Delta(\alpha) J_\Delta^{-1}(\alpha)$ with $J_\Delta^{ij} = \sum_{k=1}^n (G_k^i(\alpha))^* G_k^j(\alpha)$ and $I_\Delta(\alpha) = \sum_{k=1}^n (G_k^i(\alpha))^* S_k(\alpha) G_k^j(\alpha)$.

(3) We can compare the result of Theorem 3.5.1 to the inference of an unknown parameter in the drift coefficient for a continuously observed diffusion on $[0,T]$ in the asymptotics $\varepsilon \to 0$. According to [92], assuming a known diffusion coefficient $\varepsilon\sigma(x)$, the Maximum Likelihood Estimator is consistent and the Fisher information matrix is

$$(I_b(\alpha_0,\beta_0))_{ij} = \int_0^T (\nabla_{\alpha_i} b(\alpha_0, z(\alpha_0,s)))^* \Sigma^{-1}(z(\alpha_0,s)) \nabla_{\alpha_j} b(\alpha_0, z(\alpha_0,s)) ds. \tag{3.5.16}$$

To compare the estimator $\bar{\alpha}_{\varepsilon,\Delta}$ with the CLS estimator, we can study the limits of the two Information matrices when Δ goes to zero. Using that $z(\alpha,\cdot)$ satisfies the ODE (3.2.8), we have,

$$G_k(\alpha_0) = -\nabla_\alpha b(\alpha_0, z(\alpha_0, t_{k-1})) + o_\Delta(1), \text{ as } \Delta \text{ goes to zero.} \tag{3.5.17}$$

This result together with Lemma 3.3.5 implies that $I_\Delta(\alpha_0,\beta_0) \to I_b(\alpha_0,\beta_0)$ as $\Delta \to 0$. Since $I_b(\alpha_0,\beta_0)$ is the optimal information matrix for continuous time observation, this convergence provides some kind of optimality result for fixed Δ.

Consider now the covariance matrix of the CLS estimator. We have, $\varepsilon \to 0$,

$$(J_\Delta(\alpha))_{ij} \to \int_0^T \nabla_{\alpha_i} b(\alpha_0, z(\alpha_0,t))^* \nabla_{\alpha_j} b(\alpha_0, z(\alpha_0,t)) dt, \text{ and}$$
$$(I_\Delta(\alpha))_{ij} \to \int_0^T \nabla_{\alpha_i} b(\alpha_0, z(\alpha_0,t))^* \Sigma(\beta_0, z(\alpha_0,t)) \nabla_{\alpha_j} b(\alpha_0, z(\alpha_0,t)) dt.$$

This clearly differs from the optimal asymptotic variance and confirms that the CLS estimator is not efficient. However, it might be easier to minimize the CLS function $\sum_{k=1}^n G_k(\alpha)^* G_k(\alpha)$ than the actual contrast function $\sum_{k=1}^n G_k(\alpha)^* S_{k-1}^{-1}(\alpha) G_k(\alpha)$. Therefore this CLS estimator can be useful to serve as an initialization for other computations or algorithms.

3.6 Assessment of estimators on simulated data sets

We consider two examples of epidemic dynamics, the *SIR* and the *SIRS* presented in the first part of these notes and recalled in Section 3.2.2 for the diffusion approximation. We used the Gillespie algorithm (see Part I of these notes) to simulate the *SIR* epidemic dynamics $(\mathscr{Z}^N(t), 0 \leq t \leq T)$ and, for the *SIRS* model, the τ-leaping method

([22]), which is more efficient for large populations.

As pointed in the introduction, diffusion approximations are relevant in case of a major outbreak in a large community. Therefore, we keep only in the analysis what we called "non-extinct trajectories", chosen according to a frequently used empirical criterion: we keep epidemic trajectories such that the final epidemic size is larger than the observed empirical size minus the standard empirical error of the final epidemic size.

The inference is based only on non-extinct trajectories . Since we possess, for each simulation, the whole sample path of the epidemic process, we can compute the maximum likelihood estimator (see Chapter 4 of this part) which depends on the whole path of the jump process. For instance, for the *SIR* case, the MLE is

$$\hat{\lambda}_N = \frac{1}{N}\frac{\#\text{ Infections}}{\int_0^T S^N(t)I^N(t)dt}; \quad \hat{\gamma}_N = \frac{1}{N}\frac{\#\text{ Recoveries}}{\int_0^T I^N(t)dt}. \tag{3.6.1}$$

We call this MLE based on complete epidemic data **the reference estimator**. This is the best result that can be achieved from these epidemic data.

In order to investigate the influence of various parameters, we consider various scenarios. Each scenario corresponds to the choice of the model, the parameters θ, the population size N, the time interval of observation $[0,T]$ and the sampling interval Δ. We proceeded to 1000 repetitions for each scenario.
Hence, we varied the total size of the population N, the parameters ruling the *SIR*, *SIRS* epidemics, the time interval for observations $[0,T]$. Then, we sampled with sampling Δ each path of the Markov jump process. This sampling interval also varies. Therefore the observations coming from the simulations are

$$\frac{\mathscr{Z}^N(k\Delta)}{N} = Z^N(k\Delta)\ k = 1,\dots,n \text{ with } T = n\Delta.$$

Each scenario corresponds to the choice of the model, the parameters θ, the population size N, the time interval of observation $[0,T]$ and the sampling interval Δ.

We compare the estimators obtained with the method described in the two previous sections with the MLE (3.6.1). The properties of our minimum contrast estimators are assessed and compared to reference estimators.
For parameters with dimension greater than two, confidence ellipsoids are projected on planes, by considering all pairs of parameters. Theoretical confidence ellipsoids are built as follows. Let $V(\theta_0)$ denote the covariance matrix of the asymptotic normal distribution of parameters estimation in drift term (i.e. $I_b^{-1}(\theta_0)$ defined in (3.4.3) and $M^{-1}(\theta_0)$ defined in (3.5.9). Since $\varepsilon^{-1}V(\theta_0)^{-1/2}(\hat{\theta}_{\varepsilon,\Delta}-\theta_0) \to_{\mathscr{L}} \mathscr{N}(0,I_k)$ (where $\hat{\theta}_{\varepsilon,\Delta}$ represents $\check{\alpha}_{\varepsilon,\Delta}$ obtained minimizing (3.4.1) or $\bar{\eta}_{\varepsilon,\Delta}$ in (3.5.8) Then, for $k=a$ (dimension of α), we have,

$$\frac{1}{\varepsilon^2}\,(\hat{\theta}_{\varepsilon,\Delta}-\theta_0)^* V(\theta_0)^{-1}(\hat{\theta}_{\varepsilon,\Delta}-\theta_0) \to_{\mathscr{L}} \chi_2(k). \tag{3.6.2}$$

The matrix $V(\theta_0)^{-1}$ being positive, the quantity $(\hat{\theta}_{\varepsilon,\Delta}-\theta_0)^* V(\theta_0)^{-1}(\hat{\theta}_{\varepsilon,\Delta}-\theta_0)$ is the squared norm of vector $\hat{\theta}_{\varepsilon,\Delta}-\theta_0$ for the scalar product associated to $V(\theta_0)^{-1}$. If we denote by $\chi_k^2(0.95)$ the 95% quantile of the χ_k^2 distribution, the relation (3.6.2) could be rewritten as $||(\hat{\theta}_{\varepsilon,\Delta}-\theta_0)^2 M(\theta_0)^{-1}|| \leq \varepsilon^2\chi_k^2(0.95)$ and define an ellipsoid in $\mathbb{R}^k$.
Empirical confidence ellipsoids are based on the variance-covariance matrix of centered estimators (based on 1000 independent estimations), whose eigenvalues define the axes of ellipsoids.
In the two epidemic models detailed below, we assume both components of $Z^N(t) = S^N(t), I^N(t)$ are observed with sampling interval Δ, $((S^N(k\Delta), I^N(k\Delta)), k = 1,\dots n)$ with $T = n\Delta$.

3.6.1 The SIR model

The parameters of interest for epidemics are considered following a reparameterization: the basic reproduction number, $R_0 = \frac{\lambda}{\gamma}$, which represents the average number of secondary cases generated by one infectious in a completely susceptible population, and the average infectious duration, $d = \frac{1}{\gamma}$. Two values were tested for $R_0 = \{1.5, 5\}$

and d was set to 3 (in days, an average value consistent with influenza infection). Three values for the population size $N = \{400, 1000, 10000\}$ and of the number of observations $n = \{5, 10, 40\}$ were considered, along with two values for the final time of observation, $T = \{20, 40\}$ (in days). For each scenario defined by a combination of parameters, the analytical maximum likelihood estimator (*MLE*), calculated from the observation of all the jumps of the Markov process (see 4), was taken as reference.

Effect of the parameter values $\{R_0, d\}$ *and of the number of observations n*
The accuracy of the two estimators $\check{\alpha}_{\varepsilon,\Delta}$ and $\bar{\alpha}_{\varepsilon}$, for $N = 1000$ and from trajectories with weak ($R_0 = 5$) and strong ($R_0 = 1.5$) stochasticity is illustrated in Figure 3.6.1. We observe that R_0 and d are moderately correlated (ellipsoids are deviated with respect to the x-axis and y- axis). The shape of confidence ellipsoids depends on parameter values: for $R_0 = 5$, the 95% confidence interval is larger for R_0 than for d, whereas the opposite occurs for $R_0 = 1.5$. For $R_0 = 5$, all these confidence intervals are almost superimposed, which suggests that the estimation accuracy is not altered by the fact that all the jumps are not observed. However, for $R_0 = 1.5$ the shape of ellipsoids varies with n. Point estimates for *MLE* derived for complete observation of $(\mathscr{Z}^N(t)$ of the original jump process and the estimators $\check{\alpha}_{\varepsilon,\Delta}$, $\bar{\alpha}_{\varepsilon}$ are very similar for different values of n, which confirms the interest of using these estimators when small number of observations is available.

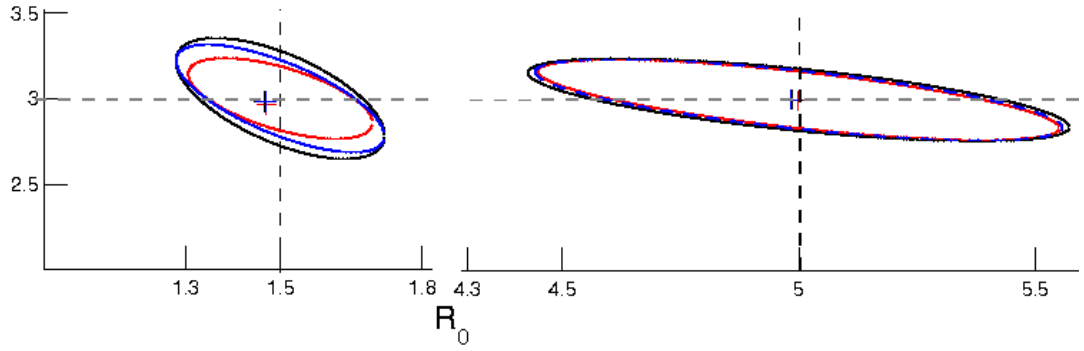


Figure 3.6.1: Point estimators (+) are computed by averaging over 1000 independent simulated trajectories of the *SIR* stochastic model (completely observed) together with their associated theoretical confidence ellipses centered on the true value: *MLE* with complete observations (red), *CE* for one observation/day, $n = 40$ (blue) and *CE* for $n = 10$ (black). Two scenarios are illustrated: $(R_0, d, T) = \{(1.5, 3, 40); (5, 3, 20)\}$, with $N = 1000$. For both scenarios $(S(0), I(0)) = (0.99, 0.01)$.The value of d is reported on the y-axis. Horizontal and vertical dotted lines cross at the true value

.

Effect of the parameter values $\{R_0, d\}$ *and of the population size N* From Figure 3.6.2, we can notice that $\sqrt{N}$ has an impact on estimation accuracy (the width of the confidence intervals decreases with $\sqrt{N}$). The case of very few observations ($n = 5$) leads to the largest confidence intervals. The *MLE* appears biased for $N = 400$. This could be due to the fact that the *MLE* is optimal when data represent a 'typical' realization (i.e. a trajectory that emerges leading to a non-negligible number of infected individuals) of the Markov process, but could yield a bias when observations are far from the average behaviour. Our *CE*s seem robust to the departure from the 'typical' behaviour (i.e. for noisy trajectories obtained either for small N or small R_0).

3.6.2 The SIRS model

For the *SIRS* model introduced in Section 3.2.2.2, four parameters were estimated: R, d, λ_1 and δ. Concerning the remaining parameters,μ was set to $1/50$ years^{-1} (a value usually considered in epidemic models), T_{per} was set to 365 days (corresponding to annual epidemics) and η was taken equal to 10^{-6} (which corresponds to 10 individuals

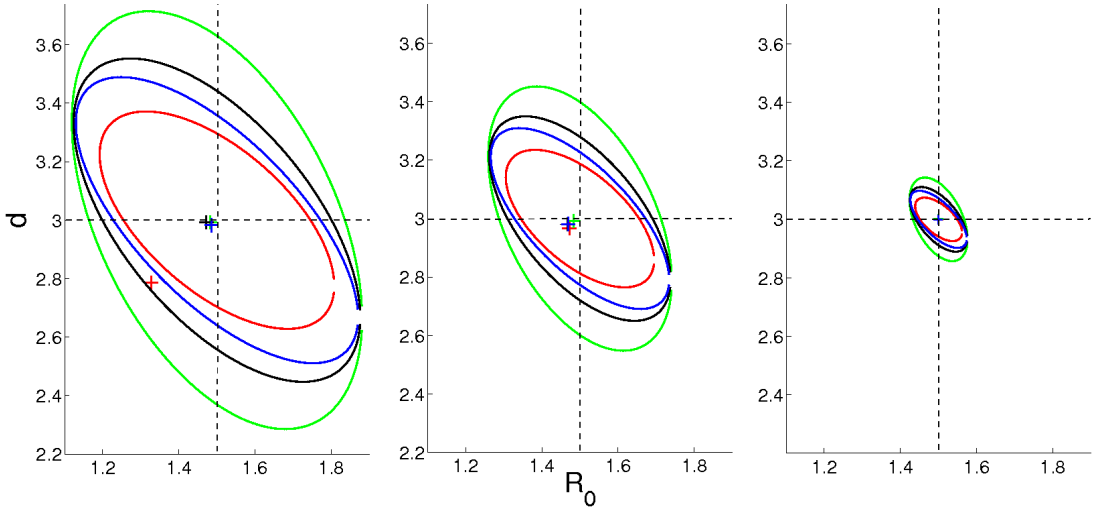


Figure 3.6.2: Point estimators (+) computed by averaging over 1000 independent simulated trajectories of the *SIR* stochastic model completely observed and their associated theoretical confidence ellipses centered on the true value: *MLE* with complete observations (red), *CE* for one observation/day, $n = 40$ (blue), *CE* for $n = 10$ (black) and *CE* for $n = 5$ (green) for $(S(0), I(0)) = (0.99, 0.01)$, $(R_0, d) = (1.5, 3)$ and $N = \{400, 1000, 10000\}$ (from left to right). Horizontal and vertical dotted lines cross at the true value.

in a population size of $N = 10^7$). We should notice that instead of estimating the real R_0 (more complicated to calculate for periodical dynamics), we prefer to estimate a parameter combination similar to the R_0 for *SIR* model, λ_0/γ, which was called here R.The performances of *CE*s were assessed on parameter combinations: $(R, d, \lambda_1, \delta) = \{(1.5, 3, 0.05, 2) \text{ and } (1.5, 3, 0.15, 2)\}$ and $T = 20$ years, with $\lambda_1 = 0.05$ leading to annual cycles and $\lambda_1 = 0.15$ to biennial dynamics (Figure 3.2.2). Numerically, the scenarios considered are consistent with influenza seasonal outbreaks. The accuracy of estimation is relatively high, as illustrated in Figure 3.6.3, regardless of the parameter. For one observation per day (which can be assimilated to a limit of data availability), the accuracy is very similar to the one based on a complete observation of the epidemic process (blue and red ellipsoids respectively). Estimations based on one observation per week are less but still reasonably accurate.

3.7 Inference for partially observed epidemic dynamics

In the case of epidemics, numbers of susceptible and infected individuals over time are generally not observed. In practice, (sometimes noisy) observations are often assumed to correspond to cumulated numbers, over the sampling interval Δ, of newly infected individuals (i.e. $\int_{t_{k-1}}^{t_k} \lambda S(s) I(s) ds$). In the *SIR* diffusion model, this corresponds to the recovered individuals $\{(R(t_k) - R(t_{k-1})), k = 1, \dots n\}$ for diseases with short duration of the infected period. Hence, this situation can be assimilated, as a first attempt, to the case where only one coordinate can be observed.

In this section, we consider the case of a two-dimensional diffusion process $X(t) = (X_1(t), X_2(t))^*$

$$dX(t) = b(\alpha, X(t))dt + \varepsilon\sigma(\beta, X(t)dB(t))dt; \quad X(0) = x, \tag{3.7.1}$$

where $B(t)$ is a Brownian motion on $\mathbb{R}^2$ and x non-random.

We assume that only the first coordinate $X_1(t)$ is observed on a fixed time interval $[0, T]$ with sampling Δ. We consider the diffusion on $\mathbb{R}^2$ satisfying the stochastic differential equation Therefore, the observations are now

$$X_1(t_k),\ k = 1, \dots n, \quad \text{with } t_k = k\Delta, \quad T = n\Delta. \tag{3.7.2}$$

For continuous observations of $(X_1(t))$ on a finite time interval $[0, T]$, two studies [80], [93] are concerned with parametric inference in this statistical framework. Both studied the maximum likelihood estimator of parameters

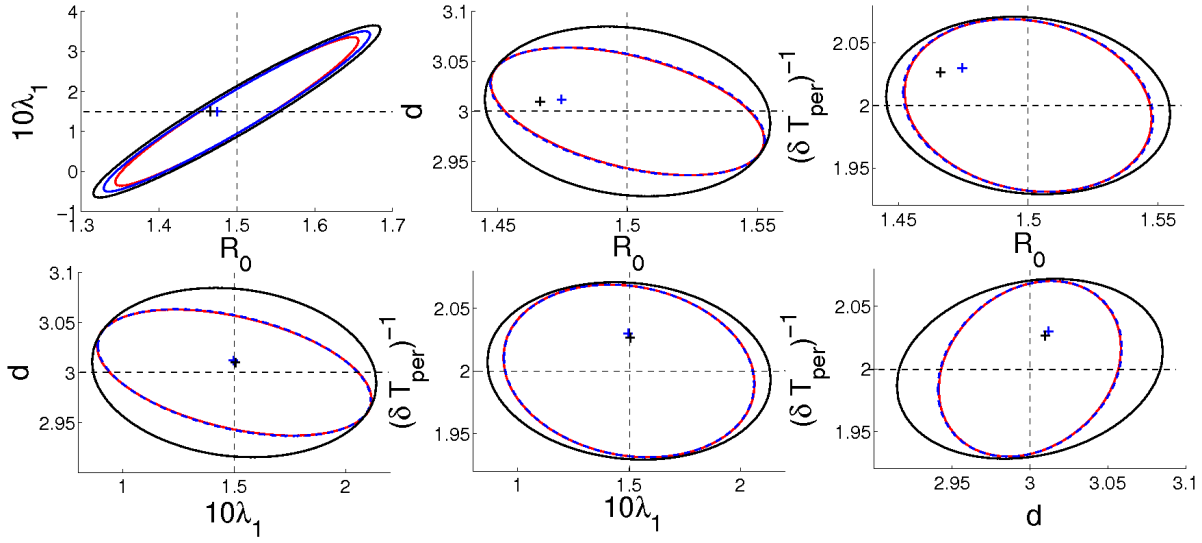


Figure 3.6.3: Point estimators (+) computed by averaging over 1000 independent simulated trajectories of the *SIRS* stochastic model with demography and seasonal forcing in transmission, completely observed (red), and their associated planar projections of theoretical confidence ellipsoids centered on the true value: *CE* for one observation/day (blue) and for one observation/week (black) for $(R,d,\lambda_1,\delta) = (1.5,3,0.15,2)$, $T = 20$ years and $N = 10^7$. Asymptotic confidence ellipsoids ($n \to \infty$) are also represented (red,blue,black). Horizontal and vertical dotted lines cross at the true value.

in the drift function for a diffusion matrix equal to $\varepsilon^2 I_p$. This likelihood is difficult to compute since it relies on integration on the unobserved coordinate. [80], [93] proposed filtering approaches to compute this likelihood, as it is done for general Hidden Markov Models (see e.g. [23] , [38]). Here, we can take advantage of the presence of ε and extend to partial observations the method by contrast processes and M- estimators that had been developed for complete observations ([47], [57], [61]), [119]).

We study the case of small (or high frequency) sampling interval, $\Delta = \Delta_n \to 0$, on a fixed time interval $[0,T]$ with $T = n\Delta$, which yields explicit results. This allows us to disentangle problems coming from discrete observations and those coming from the missing observation of one coordinate and hence provides a better understanding of the problems rising in this context. The case of Δ fixed could be studied similarly, with more cumbersome notations and no such insights .

First, the notations required are introduced, results are then stated, and finally, to illustrate this approach, the example of a two-dimensional Ornstein–Uhlenbeck process, where all the computations are explicit is developed. The consequences on diffusion approximations of Epidemic models where computations are no longer explicit are detailed later.

3.7.1 Inference for high frequency sampling of partial observations

Some specific notations need to be introduced.
For $x \in \mathbb{R}^2$, $X^\varepsilon(t)$, the diffusion process, $B(t)$ the Brownian motion, and M a 2×2 matrix, we write

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix}; B(t) = \begin{pmatrix} B_1(t) \\ B_2(t) \end{pmatrix}; M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \tag{3.7.3}$$

For functions $f(\theta,x)$ defined for $x \in \mathbb{R}^2$, we use (3.3.3) for differentiating with respect to x and (3.3.3), (3.3.4) for differentiation with respect to θ.

The observations are $(X_1(k\Delta), k = 0,\dots n)$. Since x_2 is not observed and unknown, we add it to the parameters. Therefore, setting $x_2 = \xi$, define using (S4),

$$\eta = (\alpha,\xi) \in \mathbb{R}^{a+1}; \quad \theta = (\alpha,\xi,\beta) = (\eta,\beta) \in \mathbb{R}^{a+b+1}. \tag{3.7.4}$$

The quantities introduced in (3.2.15) depend on α, η or θ and can be written, using (3.7.3), The expansion of $X(t)$ stated in (3.2.15) yields that $X_1(t)$ satisfies, using notations (3.7.3),

$$X_1(t) = z_1(\eta,t) + \varepsilon g_1(\theta,t) + \varepsilon^2 R_1^\varepsilon(\theta,t) \text{ with} \tag{3.7.5}$$

$$\begin{aligned} g_1(\theta,t) = &\int_0^t (\Phi(\eta,t,u)\sigma(\beta,z(\eta,u)))_{11}\, dB_1(u) \\ &+ (\Phi(\eta,t,u)\sigma(\beta,z(\eta,u)))_{12}\, dB_2(u). \end{aligned} \tag{3.7.6}$$

Using that $\Phi(t,u) = \Phi(t,s)\Phi(s,u)$ yields another expression for $g_1(\theta,t_k)$,

$$\begin{aligned} g_1(\theta,t_k) = &(\Phi(\eta,t_k,t_{k-1})g(\theta,t_{k-1}))_1 \\ &+ \int_{t_{k-1}}^{t_k} (\Phi(\eta,t,u)\sigma(\beta,z(\eta,u)))_{11}\, dB_1(u) \\ &+ (\Phi(\eta,t,u)\sigma(\beta,z(\eta,u)))_{12}\, dB_2(u). \end{aligned} \tag{3.7.7}$$

For estimating the unknown parameters, we use, instead of a filtering approach, the stochastic expansion of $X(t)$, where the unobserved component $X_2(t)$ is substituted by its deterministic counterpart $z_2(\eta,t)$. For building a tractable estimation function, we also simplify the expression of $B_k(\alpha,X)$ (see (3.3.7)) by replacing $\Phi(\eta;t_k,t_{k-1})$ by its first-order approximation $I_2 + \Delta\nabla_x b(\alpha,z(\eta,t_{k-1}))$, so that $\Phi_{11}(\eta,t_k,t_{k-1}) \simeq 1 + \Delta\nabla_{x_1} b_1(\alpha,z(\eta,t_{k-1}))$.

The path used in (3.3.7) is $\begin{pmatrix} X_1(t) \\ z_2(\eta,t) \end{pmatrix}$ leading, instead of $B_k(\alpha,X)$ to $\begin{pmatrix} A_k(\eta,X_1) \\ 0 \end{pmatrix}$, with

$$A_k(\eta,X_1) = X_1(t_k) - z_1(\eta,t_k) - (1 + \Delta\nabla_{x_1} b_1(\alpha,z(\eta,t_{k-1})))\,(X_1(t_{k-1}) - z_1(\eta,t_{k-1})). \tag{3.7.8}$$

For a first approach, we consider an estimation method based on the Conditional Least Squares built on the $A_k(\eta,X_1)$'s defined in (3.7.8).

$$\bar{U}_{\varepsilon,\Delta}(\eta,X_1) = \frac{1}{\varepsilon^2\Delta}\sum_{k=1}^n A_k(\eta,X_1)^2. \tag{3.7.9}$$

This CLS functional does not depend on β, and therefore β cannot be estimated using $\bar{U}_{\varepsilon,\Delta}$. estimated. The associated estimators are then defined as

$$\bar{\eta}_{\varepsilon,\Delta} = \underset{\eta\in K_a\times K_z}{\operatorname{argmin}}\ \bar{U}_{\varepsilon,\Delta}(\eta,X_1). \tag{3.7.10}$$

Note that this process could also be used for estimating η for fixed Δ and low frequency data, using $\Phi_{11}(t_k,t_{k-1})$ instead of its approximation.

Assume that $\eta = (\alpha,\xi) \in \Theta$, with Θ compact set of $\mathbb{R}^a \times \mathbb{R}$. Denote by $\eta_0 = (\alpha_0,\xi_0)$ the true parameter value and consider the estimation of η. The distribution of $(X(t))$ satisfying (3.7.1) depends on $\theta = (\eta,\beta)$. Set $\theta_0 = (\eta_0,\beta_0)$ and $\mathbb{P}_{\theta_0}$ the distribution of $(X(t))$ on $(C([0,T],\mathbb{R}^2),\mathscr{C}_T)$.

Let us first study $\bar{U}_{\varepsilon,\Delta}(\eta,X_1)$.

Lemma 3.7.1. *Assume (S1)–(S5). Then, the process $\bar{U}_{\varepsilon,\Delta}(\eta,X_1)$ defined in (3.7.9) satisfies that, under $\mathbb{P}_{\theta_0}$, as $\varepsilon,\Delta \to 0$,*

$$\varepsilon^2\bar{U}_{\varepsilon,\Delta}(\eta,X_1) \to J_T(\eta_0,\eta) = \int_0^T (\Gamma_1(\eta_0,\eta;t))^2 dt \quad \textit{a.s. where} \tag{3.7.11}$$

$$\begin{aligned} \Gamma_1(\eta_0,\eta;t) = &b_1(\alpha_0,z(\eta_0,t)) - b_1(\alpha,z(\eta,t)) \\ &- \nabla_{x_1} b_1(\alpha,z(\eta,t))(z_1(\eta_0,t) - z_1(\eta,t)). \end{aligned} \tag{3.7.12}$$

So, to get that $\bar{U}_{\varepsilon,\Delta}(\eta, Y)$ is a contrast function for estimating $\eta = (\alpha, \xi)$, we need an assumption that ensures that $\{\eta \neq \eta_0 \Rightarrow J_T(\eta_0, \eta) > 0\}$. This leads to the additional identifiability assumption,

(S8): $\eta \neq \eta_0 \Rightarrow \{t \to \Gamma_1(\eta_0, \eta; t) \not\equiv 0\}$.

For deterministic systems, the notion of observability is used in the case of partial observations (see e.g. [108], [113]), which sums up to $\{\eta \neq \eta_0 \Rightarrow z(\eta, \cdot) \not\equiv z(\eta_0, \cdot)\}$. If the underlying deterministic system is not observable, Assumption (S8) which makes reference to the identifiability of the model with respect to the parameters is not satisfied. But the converse is not true, Assumption (S8) being a bit stronger.

Proof. The proof of Lemma 3.7.1 is a repetition of the proof of Lemma 3.3.7. First, an application of the stochastic Taylor expansion yields that, as $\varepsilon \to 0$, $(X_1(t), 0 \leq t \leq T) \to (z_1(\eta_0, t), 0 \leq t \leq T)$ almost surely under $\mathbb{P}_{\theta_0}$. Second, letting $\Delta \to 0$, we get that, there exists a constant $C > 0$ such that

$$\frac{1}{\Delta} A_k(\alpha, z_1(\eta_0, \cdot)) = \Gamma_1(\eta_0, \eta, t_{k-1}) + \Delta \|\eta - \eta_0\| \, r_k(\eta_0, \eta), \tag{3.7.13}$$

with $\sup_k \sup_{\eta \in \Theta} \|r_k(\eta_0, \eta)\| \leq C$. □

To study the asymptotic behaviour of $\bar{\eta}_{\varepsilon,\Delta}$, we have to introduce additional quantities. First, we define the vector $D(\eta, t) \in \mathbb{R}^{a+1}$, using the notations defined in (3.3.3),

$$\begin{aligned} D_i(\eta, t) &= -(\nabla_{\alpha_i} b_1)(\alpha, z(\eta, t)) - \nabla_{x_2} b_1(\alpha, z(\eta, t)) \nabla_{\alpha_i} z_2(\eta, t) \quad \text{for } i = 1, \ldots, a, \\ D_i(t) &= -\nabla_{x_2} b_1(\alpha, z(\eta, t)) \nabla_{\xi} z_2(\eta, t) \quad \text{if } i = a+1, \end{aligned} \tag{3.7.14}$$

Then, built on the D_i's, define the matrix $\Lambda(\eta) = (\Lambda_{ij}(\eta))$ by

$$\Lambda_{ij}(\eta) = 2 \int_0^T D_i(\eta, t) D_j(\eta, t) \, dt. \tag{3.7.15}$$

Finally, define the three functions for $\theta = (\alpha, \xi, \beta)$,

$$\begin{aligned} v_1(\theta; t) &= \sigma_{11}^2(\beta, z(\eta, t)) + \sigma_{12}^2(\beta, z(\eta, t)) \\ &= \Sigma_{11}(\beta, z(\eta, t)), \\ v_2(\theta; t, s) &= \sigma_{11}(\beta, z(\eta, s)) \left(\Phi(\eta, t, s) \sigma(\beta, z(\eta, s))\right)_{21} \\ &\quad + \sigma_{12}(\beta, z(\eta, s)) \left(\Phi(\eta, t, s) \sigma(\beta, z(\eta, s))\right)_{22} \\ &= \left(\Phi(\eta; t, s) \Sigma(\beta, x(\eta, s))\right)_{21}, \\ v_3(\theta, t, s) &= \int_0^{t \wedge s} \left(\Phi(\eta, t, u) \sigma(\beta, z(\eta, u))\right)_{11} \left(\Phi(\eta, s, u) \sigma(\beta, z(\eta, u))\right)_{11} du \\ &\quad + \int_0^{t \wedge s} \left(\Phi(\eta, t, u) \sigma(\beta, z(\eta, u))\right)_{22} \left(\Phi(\eta, s, u) \sigma(\beta, z(\beta, u))\right)_{22} du. \end{aligned} \tag{3.7.16}$$

We can now state the main result of this section.

Theorem 3.7.2. *Assume (S1)–(S8). Then under $\mathbb{P}_{\theta_0}$, as $\varepsilon, \Delta \to 0$,*

(i) $\bar{\eta}_{\varepsilon,\Delta} \to \eta_0$ in probability .

(ii) If moreover $\varepsilon^2 \Delta^{-1} = n\varepsilon^2 \to 0$ and $\Lambda(\eta_0)$ defined in (3.7.15) is invertible, then

$$\varepsilon^{-1}(\bar{\eta}_{\varepsilon,\Delta} - \eta_0) \to \mathcal{N}(0, \Lambda(\eta_0)^{-1} V(\theta_0) \Lambda(\eta_0)^{-1}) \quad \text{in distribution}, \tag{3.7.17}$$

where $V(\theta) = V^{(1)}(\theta) + V^{(2)}(\theta) + V^{(3)}(\theta)$ *with, using (3.7.14), (3.7.16),*

$$V_{ij}^{(1)}(\theta) = \int_0^T D_i(\eta,t) D_j(\eta,t) v_1(\theta,t)\, dt, \tag{3.7.18}$$

$$V_{ij}^{(2)}(\theta) = \int \int_{0\le s\le t\le T} D_i(\eta,s) D_j(\eta,t) \nabla_{x_2} b_1(\alpha, z(\eta,s)) v_2(\theta,t,s)) ds\, dt, \tag{3.7.19}$$

$$V_{ij}^{(3)}(\theta) = \tag{3.7.20}$$

$$\int_0^T \int_0^T D_i(\eta,s) D_j(\eta,t) \nabla_{x_2} b_1(\alpha, z(\eta,s)) \nabla_{x_2} b_1(\alpha, z(\eta,t)) v_3(\theta,t,s) ds\, dt.$$

Based on (3.7.11) and Assumption (S8), the proof of the consistency of $\bar{\eta}_{\varepsilon,\Delta}$ is obtained by standard tools and omitted.

For the proof of (ii), the main difficulty lies in a precise study of $\varepsilon \nabla_i \bar{U}_{\varepsilon,\Delta}(\eta_0, X_1)$, which is the sum of n terms that are no longer conditionally independent. The three terms in the matrix $V(\theta_0)$ come from this expansion. Indeed,

$$\varepsilon (\nabla_i \bar{U}(\eta_0, Y))_i, \to \mathcal{N}_{a+1}\big(0, V(\theta_0)\big) \quad \text{in distribution under } \mathbb{P}_{\theta_0}. \tag{3.7.21}$$

Then, studying $\varepsilon^2 \nabla_{ij} \bar{U}(\eta, Y)$ yields, using (3.7.9), (3.7.14), as $\varepsilon, \Delta \to 0$,

$$\varepsilon^2 \nabla_{ij} \bar{U}(\eta_0, Y) \to \Lambda_{ij}(\eta_0) = 2 \int_0^T D_i(\eta_0, t) D_j(\eta_0, t) dt \quad \text{a.s. under } \mathbb{P}_{\theta_0}. \tag{3.7.22}$$

The proof is quite technical and is omitted.

Let us describe our method on a partially observed two-dimensional Ornstein–Uhlenbeck diffusion process $X(t) = (X_1(t), X_2(t))^*$ where all the computations are explicit. Let

$$dX(t) = AX(t)dt + \varepsilon \varsigma dB(t), \quad X(0) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{3.7.23}$$

with $A = \begin{pmatrix} a & b \\ 0 & a+h \end{pmatrix}, \varsigma = \sigma \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

We assume that $h \neq 0$, $\sigma > 0$. The parameter in the drift is $\alpha = (a, b, h)$. For partial observations, we also need introducing $\eta = (a, b, h, \xi)$ and $\theta = (a, b, h, \xi, \sigma)$. The observations are $(X_1(t_k), k = 1, \ldots, n)$ with $t_k = k\Delta$, $T = n\Delta$ and $\Delta = \Delta_n \to 0$.

The solution of the ODE (3.2.8) applied to the drift of diffusion process (3.7.23) is

$$z_1(\eta,t) = (z_1 - \frac{\xi b}{h}) e^{at} + \frac{\xi b}{h} e^{(a+h)t}; \quad z_2(\eta,t) = \xi e^{(a+h)t}. \tag{3.7.24}$$

Let us compute the matrix $\Phi(\alpha,t,u) = e^{(t-u)A}$, we have $A = PDP^{-1}$, with

$$P = \begin{pmatrix} 1 & b/h \\ 0 & 1 \end{pmatrix}, \; D = \begin{pmatrix} a & 0 \\ 0 & a+h \end{pmatrix}, \text{ so that}$$

$$\Phi(\alpha,t,s) = \begin{pmatrix} e^{a(t-s)} & \frac{b}{h}\big(e^{(a+h)(t-s)} - e^{a(t-s)}\big) \\ 0 & e^{(a+h)(t-s)} \end{pmatrix}.$$

The solution of (3.7.23) is therefore $X(t) = Pe^{tD}P^{-1}X(0) + \varepsilon\sigma \int_0^t Pe^{(t-s)D}P^{-1} dB(s)$. Hence,

$$X_1(t) = z_1(\eta,t) + \varepsilon\sigma \left(\int_0^t e^{a(t-s)} dB_1(s) + \frac{b}{h} \int_0^t \left(e^{(a+h)(t-s)} - e^{a(t-s)}\right) dB_2(s) \right). \tag{3.7.25}$$

Using that $\nabla_{x_1} b_1(\alpha, z(\eta,t)) = a$ and (3.7.24) yields that

$$A_k(\eta, X_1) = X_1(t_k) - z_1(\eta, t_k) - (1 + a\Delta)\left(X_1(t_{k-1}) - z_1(\eta, t_{k-1})\right). \tag{3.7.26}$$

$$\Gamma_1(\eta_0,\eta,t) = (a_0-a)z_1(\eta_0,t) + b_0\xi_0 e^{(a_0+h_0)t} - b\xi e^{(a+h)t}.$$

Assumptions (S1)–(S7) are satisfied. Looking at the analytical expression of $z_1(\eta,t)$, we have that $b\xi = \tilde{b}\tilde{\xi}$ leads to identical solutions $z_1(\eta,t)$. Therefore, Assumption (S8) is not satisfied and b, ξ cannot be estimated separately when observing one coordinate only. This is also true for the deterministic ODE and the non-identifiability is here an intrinsic problem to this partial observation example.

Therefore, we define a new parameter $b' = b\xi$ and consider that the parameter to estimate is now $\eta = (a,b',h)$. Then, checking (S8) is straightforward.
The various quantities introduced in the previous section have a closed expression. Indeed, the functions $D_i(\eta,t)$ defined in (3.7.14) write, using (3.7.22), (3.7.24) with $\eta = (a,b',h)$,

$$\begin{aligned}
D_1(\eta,t) &= -(z_1 - \frac{b'}{h})e^{at} - (\frac{b'}{h} + b't)e^{(a+h)t},\\
D_2(\eta,t) &= -e^{(a+h)t},\\
D_3(\eta,t) &= -b'te^{(a+h)t}.
\end{aligned}$$

The matrix $\Lambda(\eta)$ is defined as $\Lambda(\eta) = (\Lambda_{ij}(\eta))$ with $\Lambda_{ij}(\eta) = \int_0^T D_i(\eta,t)D_j(\eta,t)dt$ $(= \int_0^T D(\eta,t)D^*(\eta,t)dt$. The functions defined in (3.7.16) are, with $\theta = (a,b,h,\sigma)$,

$$v_1(\theta,t) = \sigma^2; \quad v_2(\theta,t,s) = 0; \quad v_3(\theta,t,s) = \sigma^2\left(\frac{e^{a|t-s|}}{2a} + \frac{e^{(a+h)|t-s|}}{2(a+h)}\right).$$

Therefore,

$$\begin{aligned}
V_{ij}(\theta) = \sigma^2 &\int_0^T D_i(\eta,t)D_j(\eta,t)dt\\
&+ \frac{\sigma^2b^2}{2}\int_0^T\int_0^T D_i(\eta,s)D_j(\eta,t)\left(\frac{e^{a|t-s|}}{a} + \frac{e^{(a+h)|t-s|}}{(a+h)}\right)dsdt.
\end{aligned}$$

The estimator $\bar{\eta}_{\varepsilon,\Delta}$ defined by (3.7.10) is a consistent estimator of $\eta_0 = (a_0,b'_0,h_0)$ and satisfies (3.7.17) with the matrices $\Lambda(\eta_0)$ and $V(\theta_0)$ obtained above. The asymptotic covariance matrix is therefore

$$\begin{aligned}
&\sigma^2\Lambda^{-1}(\eta) +\\
&\frac{\sigma^2b^2}{2}\Lambda^{-1}(\eta)\left(\int_0^T\int_0^T D_i(\eta,t)D_j(\eta,s)\left(\frac{e^{a|t-s|}}{a} + \frac{e^{(a+h)|t-s|}}{a+h}\right)dsdt\right)_{ij}\Lambda^{-1}(\eta).
\end{aligned} \tag{3.7.27}$$

In the case of complete discrete observations, the first term of (3.7.27) is the asymptotic variance obtained with conditional least squares. Therefore, the loss of information coming from partial observations is measured by the second term of (3.7.27) (added to the fact that only bz_0 is identifiable).

3.7.2 Assessment of estimators on simulated and real data sets

We first present the results on the *SIR* studied in the previous section but assuming partial observations. Then we investigate the inference on the real data set of Influenza dynamics modeled with the *SIRS* studied in the previous section.

3.7.2.1 Inference for partial observation of the *SIR* model with sampling interval Δ

In this section, we consider the case where one component of the epidemic process $X^N(t) = (S^N(t),I^N(t))$ is observed on $[0,T]$. The observations are the successive numbers of infected individuals

$$(I^N(k\Delta), k = 1,\dots n) \text{ with sampling } \Delta; T = n\Delta.$$

According to the notations of Section 3.7, we have to interchange the coordinates of S, I and set $X(t) = (I(t), S(t))^*$; the drift term can be written as

$$X(t) = \begin{pmatrix} I(t) \\ S(t) \end{pmatrix}; \quad b((\lambda,\gamma),(i,s)) = \begin{pmatrix} \lambda si - \gamma i \\ -\lambda si \end{pmatrix}; \quad \Sigma(i,s) = \begin{pmatrix} \lambda si + \gamma i & -\lambda si \\ -\lambda si & \lambda si \end{pmatrix}.$$

We assume that $I(0) = i_0, S(0) = s_0$. Setting $\xi = s_0$, then the parameter defined in the previous section is $\eta = (\lambda, \gamma, \xi)$. Denote by $z(\eta,t) = (i(\eta,t), s(\eta,t))$ the solution of the ODE

$$di/dt = \lambda si - \gamma i; i(0) = i_0, \quad ds/dt = \lambda si; s(0) = \xi.$$

Then, the conditional least square method now reads as

$$\bar{U}_{\varepsilon,\delta}(\eta, I) = \frac{1}{\varepsilon^2 \Delta} \sum_{k=1}^{n} (I(t_k) - i(\eta, t_k) - (1 + \lambda s(\eta, t_{k-1}) - \gamma)(I(t_{k-1}) - i(\eta, t_{k-1}))^2.$$

Using definition 3.7.12, the function $\Gamma_1(\eta_0, \eta, t)$ reads as

$$\Gamma_1(\eta_0, \eta, t) = i(\eta_0, t)(\lambda_0 s(\eta_0, t) - \lambda s(\eta, t) - \gamma_0 + \gamma).$$

To investigate the identifiability assumption, let us check **(S8)**. It reads as $\eta \neq \eta_0 \Rightarrow \{t \to \Gamma(\eta_0, \eta, t)\} \not\equiv 0$.

Assume that we have observed that the epidemic spreads, so that we have $\forall t \in [0,T], i(\eta_0, t) > 0$. Therefore, we have to prove that

$$\{t \to (\lambda_0 s(\eta_0, t) - \lambda s(\eta, t) - \gamma_0 + \gamma) \equiv 0\} \Rightarrow \{\eta = \eta_0\}. \tag{3.7.28}$$

Differentiating this relation with respect to t yields

$$\forall t, \lambda_0^2 s(\eta_0, t) i(\eta_0, t) - \lambda^2 s(\eta, t) i(\eta, t) = 0. \tag{3.7.29}$$

Using (3.7.28), we get the second relation

$$\forall t, \frac{s(\eta, t)}{i(\eta_0, t)} (\lambda i(\eta, t) - \lambda_0 i(\eta_0, t)) = \frac{\lambda_0(\gamma_0 - \gamma)}{\lambda}.$$

Differentiating this relation with respect to t yields that

$$\lambda \frac{s(\eta, t) i(\eta, t)}{i(\eta_0, t)} (\lambda_0 i(\eta_0, t) - \lambda i(\eta, t)) \equiv 0.$$

Since at time 0, $i(\eta, 0) = i(\eta_0, 0) = i_0$, we get that $\lambda = \lambda_0$. Using now (3.7.29) yields that, at time 0, $s(\eta, 0) = s(\eta_0, 0)$ so that $\xi = \xi_0$. Finally, by relation (3.7.28), we get $\gamma = \gamma_0$.

We conclude that the two parameters λ, γ as well as the initial state s_0 are identifiable when observing $(I(t_k), k = 0, \ldots n)$. The same holds true for $R_0 = \lambda/\gamma$, $d = 1/\gamma$ and s_0.

Performances of estimators in the case of partially observed *SIR* model are assessed on simulations obtained with the following parameters: $N = 10000$, $R_0 = 1.5$, $d = 3$, $s_0 = 0.97$, $T = 40$. Observations are represented by vector $I^N(k\Delta)$. Estimations of parameters (R_0, d, s_0) are performed on 1000 simulated trajectories. Theoretical and empirical confidence ellipses are built as detailed in the introduction of Section 3.6.
As shown in Figure 3.7.1, confidence ellipsoids are quite large in the case of partial data. However, they do not include unreasonable values from the epidemiological point of view. Quantile based empirical 95% confidence intervals are still quite large.

The relatively unexpected large volume of confidence ellipsoids, obtained despite theoretical identifiability of model parameters when observing only one component of the system (here $I_N(k\Delta)$) is probably due to the fact that the numerical variance-covariance matrix is ill-conditioned (the order of magnitude of the third eigenvalue is 100 times smaller than that of the first two eigenvalues. It probably corresponds to the notion of "Numerical Identifiability", which does not necessarily coincide with "Theoretical Identifiability".

Concerning point estimators, we successively considered the mean and the median of the estimators obtained for the 1000 simulation experiments. Assuming the complete observation of both coordinates of the *SIR* jump process yields, as expected, accurate values for R_0, d. Assuming that only $I(k\Delta), k = 1, \ldots n$ with $n = 40$, we obtain for a true parameter value $(1.5, 3, 0.97)$ that the mean point estimator is $(1.89, 3.43, 0.88)$ and for the median estimator $(1.54, 3.24, 0.99)$.

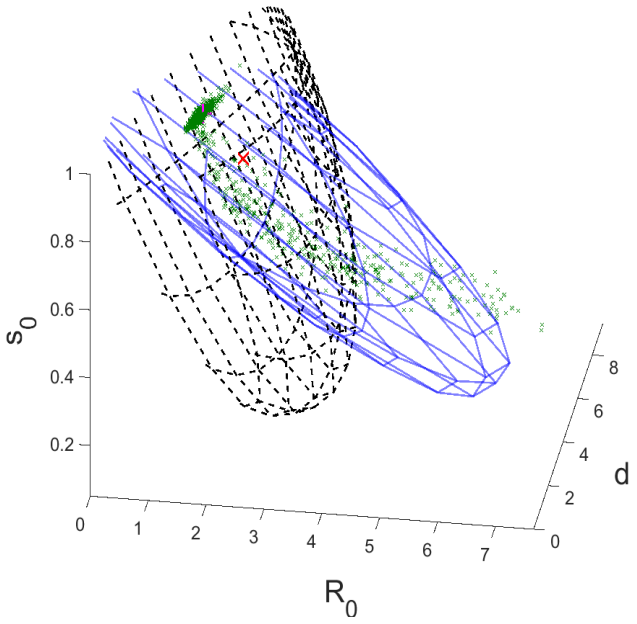


Figure 3.7.1: Point estimators (green) computed by averaging over 1000 independent simulated trajectories of the *SIR* stochastic model, partially observed ($I(k\Delta)$ only) for $(R_0, d, s_0) = (1.5, 3, 0.15, 0.97)$, $T = 40$ days and $N = 10000$. Theoretical confidence ellipsoid (black), centered on the true value and empirical confidence ellipsoid (blue), centered on mean estimated value are provided. Both ellipsoids are truncated at plausible limits on each direction. Mean and median point estimator are $(R_0, d, s_0) = (1.89, 3.43, 0.88)$ (red cross) and $(1.54, 3.24, 0.99)$ (purple cross), respectively.

3.7.2.2 Partial observations: *SIRS* model, real data on influenza epidemics

The performances of the contrast estimators for the case where only one coordinate of a diffusion process is observed are evaluated on data related to influenza outbreaks in France, collected by the French Sentinel Network (FSN), providing surveillance for several health indicators (www.sentiweb.org). These data are represented by numbers of individuals seeing a doctor during a given time interval, for symptoms related to influenza infection and are reported by a group of general practitioners (GP) voluntarily enrolled into the FSN. Several levels of errors of observation are associated to these data: (i) the state of individuals consulting a GP from the FSN is not exactly known: it can be assimilated to a new infection or to a new recovery, given that symptoms and infectiousness are not necessarily simultaneous and that a certain delay occurs between symptoms onset and consultation time (more correctly, the observed state is probably "infected" but not "newly infected"); (ii) not all infected individuals go and see a GP; (iii) the GP's supplying the FSN database represent only a proportion of all French GP's; (iv) the exact dates of consultations are not known, data are aggregated over two-week time periods; (v) data are preprocessed by the FSN to produce observations with a daily time step.
Here, we account partly for (i) on one hand and jointly for (ii) and (iii) on the other hand and assume that observations $Y(t_k)$ represent a proportion of daily (observation times $t_k = k\Delta$, with $\Delta = 1$ day) numbers of newly recovered individuals: $Y(t_k) = \rho\gamma I(t_k)$, where ρ can be interpreted as the reporting rate. Since data are available

over several seasons of influenza outbreaks (data from 1990 to 2011, hence $[0,T]=[0,21.5]$ years), an appropriate model allowing to reproduce periodic dynamics is the *SIRS* model described in Section 3.2.2.2.

$$(S,I) \overset{\frac{\lambda(t)}{N}S(I+N\eta)}{\longrightarrow} (S-1,I+1) \quad ; \qquad (S,I) \overset{\mu S}{\longrightarrow} (S-1,I);$$
$$(S,I) \overset{(\gamma+\mu)I}{\longrightarrow} (S,I-1) \quad ; \qquad (S,I) \overset{\mu N+\delta(N-S-I)}{\longrightarrow} (S+1,I).$$

The seasonality in transmission is modeled via $\lambda(t)=\lambda_0(1+\lambda_1 sin(2\pi t/T_{per}))$.

The parameter is $\theta=(\lambda_0,\lambda_1,\gamma,\delta,\eta,\mu)$, the associated drift function $b(\theta,t,(s,i))$ and diffusion matrix $\Sigma(\theta,t,(s,i))$ are

$$b(\theta,t,(s,i)) = \begin{pmatrix} -\lambda(t)s(i+\eta)+\delta(1-s-i)+\mu(1-s) \\ \lambda(t)s(i+\eta)-(\gamma+\mu)i \end{pmatrix}, \tag{3.7.30}$$

$$\Sigma(\theta,t,(s,i)) = \begin{pmatrix} \lambda(t)s(i+\eta)+\delta(1-s-i)+\mu(1+s) & -\lambda(t)s(i+\eta) \\ -\lambda(t)s(i+\eta) & \lambda(t)s(i+\eta)+(\gamma+\mu)i \end{pmatrix}. \tag{3.7.31}$$

In summary, the data used are assumed to be discrete high frequency observations of one coordinate of the following two-dimensional diffusion with small variance:

$$\begin{cases} dS(t) &= -\lambda(t)S(t)(I(t)+\eta)+\delta(1-S(t)-I(t)+\mu(1-S(t)))dt \\ &\quad +\frac{1}{\sqrt{N}}(\sigma_{11}dB_1(t)+\sigma_{12}dB_2(t)) \\ dI(t) &= (\lambda(t)S(t)(I(t)+\eta)-(\gamma+\mu)I(t)dt+\frac{1}{\sqrt{N}}(\sigma_{21}dB_1(t)+\sigma_{22}dB_2(t)). \end{cases}$$

The vector of parameters to be estimated is $\alpha=(R=\lambda_0/\gamma,10\lambda_1,d=1/\gamma,\delta_{per}=1/\delta T_{per},10\rho)$, where parameters are defined in equation (3.2.19) and more generally in the entire Section 3.2.2.2. Parameters η, μ and T_{per} are fixed at plausible values: $\eta=10^{-6}$, $\mu=\frac{1}{50}$ (years^{-1}) and $T_{per}=365$ days. The starting point of the ODE system is unknown, but since we are interested in the stationary behaviour of this process, we fix ($r_{-20T_{per}}=0.27, i_{-20T_{per}}=0.0001$, see [27] for example) and let the system evolve until $t=0$ for the tested set of parameter α to obtain our initial starting point.

Estimation results are summarized in Figure 3.7.2, which represents multi-annual dynamics of influenza cases: observed dynamics (blue curve) and simulated ones (using the ODE version of the *SIRS* model based on estimated parameter values; red curve). Estimators are associated to contrast process defined in (3.7.9). Point estimates of parameters are: $(R,10\lambda_1,d,\delta_{per},10\rho)=(1.47,1.94,2.20,5.66,0.87)$. These values are in agreement with independent estimation based on data from the same database but using a different inference method, the maximum iterating filtering proposed by [18] (personal communication S. Ballesteros). As shown in Figure 3.7.1 for the *SIR* model, widths of theoretical confidence intervals for each parameter should be larger than those corresponding to complete observations of the *SIRS* model (drawn in Figure 3.6.3). In particular, for λ_1, the width of the confidence interval for partial observations will be larger than $0.35*\sqrt{(10^7/6*10^7)}=0.14$ (after correction for the population size, which is $N=10^7$ in Figure 3.6.3 and $N=6*10^7$ in Figure 3.7.2).
We can notice from Figure 3.7.2 that predicted trajectories correspond to a regime with bi-annual cycles, composed of two different peaks (red curve). The bifurcation diagram with respect to λ_1 (similar to Figure 3.2.2), when the remaining parameters are either set to fixed values (defined in this section) or to estimated values, exhibits the bifurcation from one annual cycle to bi-annual cycle at $\lambda_1=0.035$. This value is likely to belong to the confidence interval of estimated $\lambda_1=0.19$, since the width of this interval should be greater than 0.14. Hence, this can have some influence on estimation, influence which is not well characterized in the literature for models exhibiting bifurcation profiles, especially for trajectories corresponding to parameter values close to the bifurcation point. We also observe that the smaller peak in the bi-annual cycles is underestimated, leading to almost no epidemic burst every other year. The presence of a bifurcation in the *SIRS* ODE model probably requires a better approximation of the original jump point process.

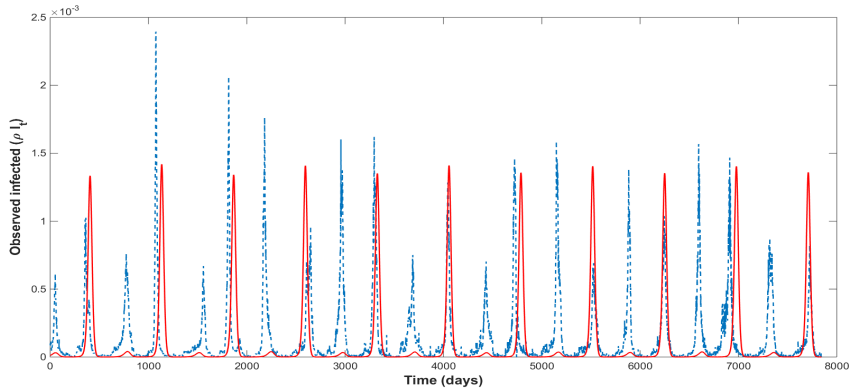


Figure 3.7.2: Time series of reported cases (expressed as a fraction of the total population in France) of influenza-like illness provided by the FSN (www.sentiweb.org) (blue curve) and deterministic trajectories (mean behaviour) predicted by the *SIRS* model based on estimated parameters using contrast (3.7.9) (red curve).

3.7.2.3 Discussion and concluding remarks

Several extensions of this study are possible for partial observations. First, we have chosen to detail the case of high sampling interval. The study in the case of a fixed sampling interval Δ should be obtained with similar tools, leading to similar results. Another extension concerns our choice of a Conditional Least squares for $\bar{U}_{\varepsilon,\Delta}$. An estimation criterium similar to the one used in Section 3.4 could be studied, using $S_k(\alpha,\beta)$ (see (3.3.8)) or substituting $\Sigma(\beta,X(t_k))$ by $\Sigma(\beta,x(\eta,t_k))$ for small sampling. This yields the new process, using (3.7.8),

$$\bar{U}_{\varepsilon,n}(\eta,(Y(t_k))) = \sum_{k=1}^{n} log\,\Sigma(\beta,x(\eta,t_k)) + \frac{1}{\varepsilon^2\Delta}\Sigma(\beta,x(\eta,t_k))^{-1}(A_k(\eta,Y))^2. \tag{3.7.32}$$

The study of this process should yield estimators in the diffusion coefficient β with probably additional assumptions linking ε and Δ. Finally for fixed Δ, $S_k(\alpha,\beta)$ defined in (3.3.8) could be substituted by $(S_k(\alpha,\beta))_{11}$ in the case of two distinct parameters in the drift and diffusion coefficient, and $(S_k(\alpha)_{11}$ in the case corresponding to epidemics where the same parameters are present in the drift and diffusion coefficients. Another extension of the method described in Section 3.7 is the case of a p-dimensional diffusion process where only the first l-coordinates are observed (for instance the *SEIR* model with only Infected observed).

Chapter 4

Inference for Continuous Time SIR models

by Catherine Larédo and Viet Chi Tran

4.1 Introduction

Consider the *SIR* epidemic model with exponential times in a finite population of size N where $S(t), I(t), R(t)$ denote the number of Susceptible, infected/infectious and Removed individuals at time t with infection rate λ and recovery rate γ ($S(t)+I(t)+R(t)=N$ for all t). There are various ways of describing this process using pure jump Markov processes. We refer to Chapter **??** of Part I of these notes and to Section A.5 of the Appendix for a recap on these processes.

This description now belongs to the domain of event time data, which are conveniently studied by the use of counting processes. We refer to Section A.5 of the Appendix for a short introduction to counting processes in continuous time.

At this point, we need an asymptotic framework to study the properties of these estimators. Two frameworks have been proposed.
Case (1): Assume that the number of initially infected $I(0)=a$ remains fixed and that the number of initial Susceptible is $S(0)=n:=N-a$. We also assume for the sake of simplicity that $R(0)=0$. This leads to a total population size $N=n+a$ that goes to infinity.
Case (2): Assume that the population size $N\to\infty$ and that both $S(0), I(0)$ tend to infinity with N such that $S(0)/N\to s_0>0; I(0)/N\to i_0>0$ as $N\to\infty$.

Case (1) has been studied by Rida [109], to which we refer for a detailed presentation. We focus here mainly on Case (2).

4.2 Maximum likelihood in the SIR case

To ease notation, we work here on a simplification of the SEIR process studied in Part I of these notes. We omit the state E and consider an SIR model (corresponding to the limiting case when $\nu\to+\infty$). Recall that the population size is N, that the infection rate is λ and the removal rate γ. We assume that we observe the whole trajectory on a time window $[0,T]$ with $T>0$: $(S_t^N, I_t^N, R_t^N)_{t\in[0,T]}$. The successive times of events are $(T_i)_{1\le i\le K_N(T)}$, where $K_N(T)=\sum_{i\ge 0}\mathbf{1}_{T_i\le T}$ is the number of events. At each event, $J_i=0$ if we have an infection and $J_i=1$ if we have a recovery. Notice that we are here in the case where we have knowledge of all recovery and infection events, i.e. that we have *complete epidemic data*. The case where some data are missing is treated in the next subsections.

Writing the likelihood of our data is important to calibrate the parameters of the model, $\theta=(\lambda,\gamma)\in\mathbb{R}_+^2$ in the case of the SIR model, but also because this is also useful for designing EM or MCMC procedures.

Definition 4.2.1. *We define the likelihood $\mathscr{L}_T^N(\theta)$ of the observations as the density, in $\mathbb{D}([0,T],[0,1]^3)$ of the process $(S_t^N, I_t^N, R_t^N)_{t\in[0,T]}$ with respect to the SIR process where intervals between events follow independent exponential distributions of parameter $2N$ and where each event is an infection with probability $1/2$ and a recovery with probability $1/2$. The likelihood is of course a function of $\theta \in \mathbb{R}_+^2$ and of the observations $(S_t^N, I_t^N, R_t^N)_{t\in[0,T]}$ which are omitted in the notation for the sake of notation.*

This definition has been proposed in [31] for example. The dominating measure with respect to which the distribution of $(S_t^N, I_t^N, R_t^N)_{t\in[0,T]}$ is written is here the distribution of the process corresponding to the sequence (J_i, T_i)'s where the J_i's are i.i.d. Bernoulli random variables with parameter $1/2$, and where the intervals $\Delta T_i = T_i - T_{i-1}$ are i.i.d. exponential random variables with expectation $1/(2N)$. With the notation above:

$$\begin{aligned}\mathscr{L}_T^N(\theta) &= \mathscr{L}_T^N\big((S_t^N, I_t^N, R_t^N)_{t\in[0,T]}; \lambda, \gamma\big) \\ &= \exp\big(NT - \int_0^T (\lambda S_s^N I_s^N - \gamma I_s^N) ds\big) \prod_{i=1}^{K_N(T)} (\lambda S_{T_i-}^N I_{T_i-}^N)^{1-J_i} (\gamma I_{T_i-}^N)^{J_i}. \end{aligned} \tag{4.2.1}$$

Taking the log, and using the formulation of the processes (S_t, I_t, R_t) by means of Poisson point processes Q^1 and Q^2 as in Part I, Chapter 2 of these notes,

$$\begin{aligned}\log \mathscr{L}_T^N(\theta) &= NT - \int_0^T (\lambda S_s^N I_s^N - \gamma I_s^N) ds \\ &\quad + \sum_{i=1}^{K_N(T)} \Big[(1-J_i)\log\big(\lambda S_{s_-}^N I_{s_-}^N\big) + J_i \log\big(\gamma I_{s_-}^N\big)\Big] \\ &= NT - \int_0^T (\lambda S_s^N I_s^N - \gamma I_s^N) ds + \int_0^T \log\big(\lambda S_{s_-}^N I_{s_-}^N\big) \mathbf{1}_{u\le \lambda N S_{s_-}^N I_{s_-}^N} Q^1(ds, du) \\ &\quad + \int_0^T \log\big(\gamma I_{s_-}^N\big) \mathbf{1}_{u\le \gamma N I_{s_-}^N} Q^2(ds, du).\end{aligned}$$

The above function is concave in λ and γ, for a given observations $(S_t^N, I_t^N)_{t\in[0,T]}$, and maximizing it, we obtain:

Proposition 4.2.2. *The maximum likelihood estimator $\widehat{\theta}_N = (\widehat{\lambda}_N, \widehat{\gamma}_N)$ of θ (MLE) is then given by:*

$$\widehat{\lambda}_N = \frac{1}{N} \frac{\sum_{i=1}^{K_N(T)} (1-J_i)}{\int_0^T S_s^N I_s^N ds}, \qquad \widehat{\gamma}_N = \frac{1}{N} \frac{\sum_{i=1}^{K_N(T)} J_i}{\int_0^T I_s^N ds}. \tag{4.2.2}$$

These estimators have already been mentioned in (3.6.1) and it had been noticed that the numerators of $\widehat{\lambda}_N$ and $\widehat{\gamma}_N$ are respectively the numbers of infections and recoveries on the period $[0,T]$. Remark also that the estimators (4.2.2) are the same for the Cases (1) and (2) presented in Section 4.1. In what follows, we concentrate on the Case (2).

Using the Law of Large Numbers and the Central Limit Theorem stated in Part I, Section 2 of these notes we obtain that

Proposition 4.2.3. *The estimator $\widehat{\theta}_N$ is convergent and asymptotically Gaussian when $N \to +\infty$:*

$$\sqrt{N}\big(\widehat{\theta}_N - \theta\big) = \sqrt{N} \begin{pmatrix} \widehat{\lambda}_N - \lambda \\ \widehat{\gamma}_N - \gamma \end{pmatrix} \Rightarrow \mathscr{N}\big(0_{\mathbb{R}^2}, I^{-1}(\lambda, \gamma)\big),$$

where the Fisher information matrix is:

$$I(\lambda, \gamma) = \begin{pmatrix} V_{11}(t) & 0 \\ 0 & V_{22}(t) \end{pmatrix}$$

with $(s(t),i(t))_{t\in[0,T]}$ the solution of the limiting ODE that approximates $(S_t^N,I_t^N)_{t\in[0,T]}$ when $N\to+\infty$ (see Example 2.2.10 in Part I) and with

$$V_{11}(t)=\frac{\int_0^T s(t)i(t)dt}{\lambda}=\frac{1-s(T)}{\lambda^2};\ V_{22}(t)=\frac{\int_0^T i(t)dt}{\gamma}=\frac{1+\mu-s(T)-i(T)}{\gamma^2}. \tag{4.2.3}$$

Proof. Notice that the estimator $\widehat{\lambda}$ given in Proposition 4.2.2 can be rewritten, with the notations of Example 2.2.1 of Part I of these notes, as

$$\widehat{\lambda}_N=\frac{1}{N}\frac{P_1\Big(\lambda N\int_0^T S_s^N I_s^N ds\Big)}{\int_0^T S_s^N I_s^N ds}.$$

Using the Law of Large Numbers given in Part I, Section 2.2, the process $(S_t^N,I_t^N)_{t\in[0,T]}$ converges uniformly when $N\to+\infty$ to the unique solution of the ODE

$$\begin{aligned} s'(t)&=-\lambda s(t)i(t),\\ i'(t)&=\lambda s(t)i(t)-\gamma i(t).\end{aligned}$$

Moreover,

$$\lim_{N\to+\infty}\widehat{\lambda}_N=\lambda\frac{\int_0^T s(t)i(t)dt}{\int_0^T s(t)i(t)dt}=\lambda.$$

Now,

$$\sqrt{N}\big(\widehat{\lambda}_N-\lambda\big)=\frac{1}{\int_0^T S_s^N I_s^N ds}\left[\frac{1}{\sqrt{N}}P_1\Big(\lambda N\int_0^T S_s^N I_s^N ds\Big)-\sqrt{N}\lambda\int_0^T S_s^N I_s^N ds\right].$$

From Part I, Section 2.3, we have the following convergence in distribution

$$\frac{1}{\sqrt{N}}P_1\Big(\lambda N\int_0^T s(t)i(t)dt\Big)-\sqrt{N}\lambda\int_0^T s(t)i(t)dt\Rightarrow B_1\Big(\lambda\int_0^T s(t)i(t)dt\Big)$$

where B_1 is a standard real Brownian motion. As in the proof of Proposition 2.3.1, the bracket in the right term is then shown to converge to the same limit $B_1(\lambda\int_0^T s(t)i(t)dt)$. Since the denominator of the right-hand side converges in probability to $\int_0^T s(t)i(t)dt$, we obtain the asymptotic normality of $\widehat{\lambda}_N$ with asymptotic variance

$$\frac{\lambda}{\int_0^T s(t)i(t)dt}.$$

Proceeding similarly for $\widehat{\gamma}_N$ and using the asymptotic independence between the two estimators provides the result. Notice that the Fisher information matrix can also be computed from the log-likelihood, and that all regularity assumptions of generic asymptotic normality results are satisfied (see e.g. Chapter 4 of [96]). □

Corollary 4.2.4. *An estimator of $R_0=\lambda/\gamma$ is $\widehat{R}_0^{(t)}=\frac{\hat{\lambda}_t}{\hat{\gamma}_t}$. Applying the functional delta-theorem (e.g. [124]), it converges in distribution to*

$$\sqrt{n}(\widehat{R}_0^{(t)}-R_0)\to\mathcal{N}(0,\sigma^2(t))\quad\text{ with }\quad\sigma^2(t)=\frac{V_{11}^{-1}(t)+R_0^2V_{22}^{-1}(t)}{\gamma^2}. \tag{4.2.4}$$

Remark 4.2.5 (Maximum likelihood estimators in the Case (1))**.** *Let us denote by $(N_t)_{t\in\mathbb{R}_+}$ the counting processes associated to the infection process:*

$$N_t=P_1\Big(\int_0^t\lambda NS_s^N I_s^N ds\Big),$$

and by τ_N the extinction time, when there is no infective individual left. Because the population is finite, $\tau_N<+\infty$ almost surely and $N(\tau_N)\le N$. Let

$$A=\{\omega;N(\tau_N,\omega)\to\infty\ as\ N\to\infty\}$$

be the event on which a major outbreak occurs. Ball [8] proved that $\mathbb{P}(A) = 1 - \min\{1, (\gamma/\lambda)^a\}$. *Moreover if* $R_0 = \lambda/\gamma > 1$, *then* $\mathbb{P}(A) > 0$ *and as* $n \to \infty$,

$$\frac{N(\tau_N)}{N} \to \pi 1_A \text{ where } \pi \text{ is such that } \frac{\lambda}{\gamma} = -\frac{\log(1-\pi)}{\pi}.$$

Asymptotic results for the estimators are obtained on A and A^c. *The maximum likelihood estimator satisfies that*

$$\hat{\lambda}_N \to \lambda 1_A + Z 1_{A^c}$$

in distribution where Z is a positive explicit random variable such that $\mathbb{E}(Z) < 1/\lambda$ *if* $\lambda/\gamma > 1$. *Note that in this case,* $\hat{\lambda}_N$ *is not a consistent estimator. We refer to [109] for a detailed presentation of the results.*

These methods can be extended to other epidemic models. We will detail later for the SEIR and SIRS epidemic models. The main drawback of this approach is that the epidemic process is rarely observed in such details, which prevents this kind of statistical approach. However, this study sums up the best statistical results that can be obtained when complete observations are available. When incomplete observations are available, the loss of information will be measured with respect to this general reference.

4.2.1 MCMC estimation

The preceding subsection treated the case of complete observation. In practice, parameter estimation for SIR models is usually a difficult task because of missing observations, which is a recurrent issue in epidemiology. O'Neill Roberts [107] developed a Markov chain Monte Carlo method (MCMC) to make inferences about the missing data and the unknown parameters in a Bayesian framework.

We consider an SIR model as in Section 4.2. Instead of observing the sequence $(J_i, T_i)_{i\in\{1\dots K_T^N\}}$ (type – infection or recovery – and time of occurrence of the successive events, as described in the beginning of Section 4.2), we observe only the T_i's such that $J_i = 1$ (recovery events, that can also be detection events in some applications) and the total number of events K_T^N is unknown. In this section, we adopt the following notation. Let us assume that there are m infections at times $\sigma = (\sigma_1 < 0, \dots \sigma_m)$ that are unobserved and n removals at times $\tau = (\tau_1 = 0, \dots \tau_n)$ which constitute our observations. For later purposes, we will denote by $\sigma_{-1} = (\sigma_2, \dots \sigma_m)$ the vector of infection times starting from the second infection. We observe the total size of the population N, the number n of removal times and the vector τ of these removal times. The parameter of interest is $(\lambda, \gamma, \sigma_1)$ and the vector σ_{-1} is the vector of nuisance parameters.

The MCMC algorithm proposed by O'Neill and Roberts [107] take place in a Bayesian framework. Given λ, γ and the first infection time σ_1, the likelihood of $(\sigma_{-1}, \tau) = (\sigma_2 \dots \sigma_m, \tau_1, \dots \tau_m)$ is obtained from adapting (4.2.1):

$$\mathscr{L}_T^N(\sigma_{-1}, \tau | \lambda, \gamma, \sigma_1) = \exp\big(NT - \int_{\sigma_1}^T (\lambda S_s^N I_s^N - \gamma I_s^N) ds\big) \prod_{i=1}^n (\lambda S_{\sigma_i -}^N I_{\sigma_i -}^N) \prod_{i=1}^m (\gamma I_{\tau_i -}^N). \tag{4.2.5}$$

4.2.1.1 A priori distributions

We suppose that λ and γ have *a priori* Gamma distribution with parameters $(\alpha_\lambda, \beta_\lambda)$ and $(\alpha_\gamma, \beta_\gamma)$ respectively, where we recall that the density of a Gamma distribution with parameter (α, β) is:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbf{1}_{(0,+\infty)}(x)$$

where $\Gamma(x)$ is the gamma function such that for any positive integer k, $\Gamma(k) = (k-1)!$. Following [107], we also chose for the *a priori* distribution of σ_1 the 'exponential' distribution with density (on $\mathbb{R}_-$) with $\rho > 0$:

$$\rho e^{\rho \sigma_1} \mathbf{1}_{(-\infty, 0)}(\sigma_1).$$

4.2.1.2 A posteriori distributions

The purpose is now to generate a sample from the *a posteriori* distribution $\pi(\sigma,\lambda,\beta|\tau)$. For this, O'Neill and Roberts propose a Metropolis–Hastings algorithm.

Recall the principle of the Metropolis–Hastings algorithm used to obtain a sample $\mathbf{x}$ in a distribution with a density $\pi(x)$ that is proportional to some $f(x)$. Consider a transition kernel with a density $q(y|x)$ from which it is easy to simulate. Starting from a first point x_0, construct a sequence of points $(x_k)_{k\in N}$ with f and q as follows. Assume that x_k has been constructed, then:

- draw y from $q(y|x_k)$.
- With probability
$$\phi(x_k,y)=\min\Big(\frac{f(y)q(x_k|y)}{f(x_k)q(y|x_k)},1\Big)$$
define $x_{k+1}=y$.
With probability $1-\phi(x_k,y)$, define $x_{k+1}=x_k$.

This defines a reversible Markov chain whose stationary distribution is π.

We apply the above idea to sample σ,λ,β from the *a posteriori* distribution. To choose the transition kernels, notice first that with direct computation, we obtain:

$$\begin{aligned}\pi\big(\sigma_1|\tau,\sigma_{-1},\lambda,\gamma\big)\sim&(\rho+\lambda N+\gamma)e^{-(\theta+\lambda N+\gamma)(\sigma_2-y)}\mathbf{1}_{y<\sigma_2}\\ \pi\big(\lambda|\tau,\sigma,\gamma\big)\sim&\Gamma\big(\alpha_\lambda+\int_{\sigma_1}^T S_s^N I_s^N ds, m-1+\beta_\lambda\big)\\ \pi\big(\gamma|\tau,\sigma,\lambda\big)\sim&\Gamma\big(\alpha_\gamma+\int_{\sigma_1}^T I_s^N ds, n+\beta_\gamma\big).\end{aligned}$$

Hence, it is natural to choose the above distributions for the proposals of σ_1, λ and β. It remains to propose a transition kernel for σ_{-1}. O'Neill and Roberts propose a Hasting algorithm with the three following moves:

- Move an infection time chosen at random by sampling the candidate uniformly in $[0,T]$. If the infection time chosen at random was at time s and the proposal time drawn uniformly in $[0,T]$ is t, the move is accepted with probability
$$\begin{aligned}\phi(\sigma,\sigma\cup\{t\}\setminus\{s\})=&\frac{\mathscr{L}_T^N\big(\sigma\cup\{t\}\setminus\{s\},\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{|\sigma|-1}\frac{1}{T}}{\mathscr{L}_T^N\big(\sigma,\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{|\sigma|-1}\frac{1}{T}}\wedge 1\\ =&\frac{\mathscr{L}_T^N\big(\sigma\cup\{t\}\setminus\{s\},\tau|\lambda,\gamma,\sigma_1\big)}{\mathscr{L}_T^N\big(\sigma,\tau|\lambda,\gamma,\sigma_1\big)}\vee 1.\end{aligned}$$
- Remove an infection time chosen at random. If the chosen infection time was at time s, the acceptation probability is then:
$$\frac{\mathscr{L}_T^N\big(\sigma\setminus\{s\},\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{T-\sigma_1}}{\mathscr{L}_T^N\big(\sigma,\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{|\sigma|-1}}\wedge 1=\frac{\mathscr{L}_T^N\big(\sigma\setminus\{s\},\tau|\lambda,\gamma,\sigma_1\big)(|\sigma|-1)}{\mathscr{L}_T^N\big(\sigma,\tau|\lambda,\gamma,\sigma_1\big)(T-\sigma_1)}\wedge 1.$$
- Add a new infection at a time t drawn uniformly on $[0,T]$:
$$\frac{\mathscr{L}_T^N\big(\sigma\cup\{t\},\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{|\sigma|}}{\mathscr{L}_T^N\big(\sigma,\tau|\lambda,\gamma,\sigma_1\big)\frac{1}{(T-\sigma_1)}}\wedge 1=\frac{\mathscr{L}\big(\sigma+\{t\}\big)(T-\sigma_1)}{\mathscr{L}\big(\sigma\big)|\sigma|}\wedge 1.$$

A numerical application is performed in [107] for small epidemics. This algorithm is simulated and compared with other ones in Section 4.3.2.

4.2.2 EM algorithm for discretely observed Markov jump processes

We consider now the situation where the Markov jump process is only observed at discrete time points. This has been considered by Bladt and Sorensen [14]. We study the maximum likelihood estimation of the Q-matrix based on a discretely sampled Markov jump process. The problem of identifiability and of existence and uniqueness of the MLE is related to the following problem in probability: can a given discrete time Markov chain be obtained as a discrete time sampling of a continuous time Markov jump process?

4.2.2.1 Likelihood function

Let $X = (X(s), s \geq 0)$ be a Markov jump process with finite state space $E = \{1, \ldots, N\}$ and Q-matrix $\mathbf{Q} = (q_{kl})$. If X is continuously observed on the time interval $[0,T]$, the likelihood function is given by,

$$L_T(\mathbf{Q}) = \prod_{k=1}^{N} \prod_{l \neq k} q_{kl}^{N_{kl}(T)} \exp(-q_{kl} R_k(T)), \text{ where} \tag{4.2.6}$$

the process $N_{kl}(t)$ is the number of transitions from state k to state l in the time interval $[0,t]$ and $R_k(t)$ is the time spent in state k before time t.

$$R_k(t) = \int_0^t \delta_{\{X(s)=k\}} \, ds. \tag{4.2.7}$$

For details see e.g. [73] .
Therefore, if the process is continuously observed on $[0,T]$, the maximum likelihood estimator of its $\mathbf{Q}$- matrix is easily obtained:

$$\hat{\mathbf{Q}}_{kl} = \frac{N_{kl}(T)}{R_k(T)}. \tag{4.2.8}$$

Assume now that the process is observed with a sampling interval Δ with $T = n\Delta$. Then, setting $X_i = X(t_i)$ is a discrete time Markov chain with transition matrix

$$P^{\Delta}(\mathbf{Q}) \quad \text{where } P^t(\mathbf{Q}) = \exp(t\mathbf{Q}), \quad t > 0,$$

with $\exp(\cdot)$ denoting the matrix exponential function.

Hence the likelihood for the discrete observations $(x_0, \ldots, x_n)$ is

$$L_{n,\Delta}(\mathbf{Q}) = \prod_{i=1}^{n} P^{\Delta}(\mathbf{Q})_{x_{i-1} x_i},$$

with the notation that the ij entry of a matrix A is denoted A_{ij}. Since it is a discrete time Markov chain, it satisfies,

$$\begin{aligned} L_{n,\Delta}(\mathbf{Q}) &= \prod_{k=1}^{N} \prod_{l=1}^{N} (P^{\Delta}(\mathbf{Q})_{kl}^{N^{kl}(n)}, \\ N^{kl}(n) &= \sum_{i=1}^{n} \delta_{\{X_{i-1}=k, X_i=l\}}. \end{aligned}$$

The random variables $(N^{kl}(n))$ are the number of transitions from state k to state l before n. We have proved in Section 2.1) that the associated MLE of the transition matrix $\hat{\mathbf{P}}$ is explicit. But building an estimator of Q from $\hat{\mathbf{P}}$ is not straightforward.

Indeed, let $\mathscr{P}_0 = \{\exp \mathbf{Q} \mid \mathbf{Q} \in \mathscr{Q}\}$ denote the set of transition matrices that correspond to discrete time observation of a continuous time Markov jump process. If $\hat{\mathbf{P}} \in \mathscr{P}_0$, there exists a $\hat{\mathbf{Q}} \in \mathscr{Q}$ such that $P^{\Delta}(\hat{\mathbf{Q}}) = \hat{\mathbf{P}}$. This raises two distinct problems. First the set $\mathscr{P}_0$ is quite complex, and second the matrix exponential function is not an injection on its domain, so $\hat{\mathbf{Q}}$ may not be unique leading to identifiability questions for the statistical model. Additional assumptions are thus required in order to ensure the convergence of stochastic algorithms such as *EM*, *MCMC*. We refer to Bladt and Sorensen [14] for details.

4.2.2.2 The Expectation-Maximization (EM) algorithm

This is a broadly used method for optimizing the likelihood function in cases where only partial information is available (see e.g. [34, 35, 123, 127]). A discretely observed Markov jump process is such an example where only data $Y_i = X(t_i); i = 1,\dots,n$ are available. Let $X = \{X(t); 0 \le t \le T\}$ and $Y = \{Y_i; i = 1\dots,n\}$. The EM-algorithm aimed at estimating the Q-matrix $Q = (q_{ij},; i, j \in E)$ iterating the two steps:
E-step: replace the unobserved parts by their conditional expected values given the data $Y = y$
M-step: perform maximum likelihood on the complete data.

The difficult part in the EM algorithm here is the **E-step**:
i.e. compute $\mathbb{E}_{Q_0}[\log L_T(\mathbf{Q})|Y = y]$ where Q_0 is an arbitrary Q-matrix.
Indeed, consider the **M-step**. From equation (4.2.6), we have

$$\mathbb{E}_{\mathbf{Q}_0}(\log L_T(\mathbf{Q})|Y = y) = \sum_{k=1}^{N}\sum_{l\neq k}\log(q_{kl})\mathbb{E}_{\mathbf{Q}_0}(N_{kl}(T)|Y = y) - \sum_{k=1}^{N}\sum_{l\neq k} q_{kl}\mathbb{E}_{\mathbf{Q}_0}(N_k(T)|Y = y).$$

This is the likelihood of a continuous time process with observed statistics
$\mathbb{E}_{\mathbf{Q}_0}(N_{kl}(T)|Y = y), \mathbb{E}_{\mathbf{Q}_0}(N_k(T)|Y = y)$. It is maximized, as a function of Q, according to (4.2.8) by

$$\hat{Q}_{kl} = \frac{\mathbb{E}_{\mathbf{Q}_0}(N_{kl}(T)|Y = y)}{\mathbb{E}_{\mathbf{Q}_0}(N_k(T)|Y = y)}. \tag{4.2.9}$$

Therefore, to perform the algorithm, we have to compute the two quantities $\mathbb{E}_{\mathbf{Q}_0}(N_{kl}(T)|Y = y)$ and $\mathbb{E}_{\mathbf{Q}_0}(N_k(T)|Y = y)$.
For this, let us consider a fixed intensity matrix $\mathbf{Q}$ and omit the index $\mathbf{Q}$. Denote by e_i the unit vector with i^{th} coordinate equal to 1, and for U a vector or a matrix, let U^* the transpose of U.

Noting that $N^k(T) = \sum_{p=1}^{n}(N^k(t_p) - N^k(t_{p-1}))$, we get by the Markov property and the time homogeneity of $X = X(t)$,

$$\begin{aligned}\mathbb{E}(N^k(t_p) - N^k(t_{p-1})/Y = y) &= \mathbb{E}(N^k(t_p) - N^k(t_{p-1})|X(t_p) = y_p, X(t_{p-1}) = y_{p-1})\\ &= \mathbb{E}(N^k(t_p - t_{p-1})|X(t_p - t_{p-1}) = y_p, X(0) = y_{p-1}).\end{aligned}$$

Similarly $N^{kl}(T) = \sum_{p=1}^{n}(N^{kl}(t_p) - N^{kl}(t_{p-1}))$, and

$$\mathbb{E}(N^{kl}(t_p) - N^{kl}(t_{p-1})/Y = y) = \mathbb{E}(N^{kl}(t_p - t_{p-1})|X(t_p - t_{p-1}) = y_p, X(0) = y_{p-1}).$$

Hence,

$$\mathbb{E}(N^k(T)|Y = y) = \sum_{p=1}^{n} E^k_{y_{p-1}y_p}(t_p - t_{p-1}); \quad \mathbb{E}^{kl}(T)|Y = y) = \sum_{p=1}^{n} F^{kl}_{y_{p-1}y_p}(t_p - t_{p-1}); \tag{4.2.10}$$

where if (i, j) and $(k, l) \in E$, and $t > 0$,

$$\begin{aligned}E^k_{ij}(t) &= \mathbb{E}_{\mathbf{Q}_0}(N^k(t)|X(t) = j, X(0) = i),\\ F^{kl}_{ij}(t) &= \mathbb{E}_{\mathbf{Q}_0}(N^{kl}(t)|X(t) = j, X(0) = i).\end{aligned}$$

Fix $k \in E$ and define the matrix $M^k(t)$ by

$$M^k_{ij}(t) = \mathbb{E}(N_k(t)1_{X(t)=j}|X(0) = i). \tag{4.2.11}$$

Then, according to [13],

$$\frac{d}{dt}M^k_{ij}(t) = \sum_{l=1}^{N} M^k_{il}(t)q_{lj} + \exp(t\mathbf{Q})_{ij}\delta_{jk}; \quad M^k_{ij}(t_0) = 0.$$

This equation has an explicit solution which reads as $\mathbf{M}^k(t) = (M^k_{ij}(t), i, j \in E)$,

$$\mathbf{M}^k(t) = \int_0^t \exp(s\mathbf{Q})(e_k e_k^*) \exp((t-s)\mathbf{Q})\, ds. \tag{4.2.12}$$

Fix now $k, l \in E$ and define the matrix $\mathbf{f}^{kl}_{ij}(t) = \mathbb{E}(N^{kl}(t) 1_{X(t)=j} | X(0) = i)$. Similarly

$$\mathbf{f}^{kl}(t) = q_{kl} \int_0^t \exp(s\mathbf{Q})(e_k e_l^*) \exp((t-s)\mathbf{Q}) ds. \tag{4.2.13}$$

Hence, using that $\mathbb{P}(X(t) = j | X(0) = i) = e_i^* \exp(\mathbf{Q}t) e_j$ yields that

$$E^k_{ij}(t) = \frac{M^k_{ij}(t)}{e_i^* \exp(t\mathbf{Q}) e_j}; \quad F^{kl}_{ij}(t) = \frac{\mathbf{f}^{kl}_{ij}(t)}{e_i^* \exp(t\mathbf{Q}) e_j}. \tag{4.2.14}$$

So the EM-algorithm works along the successive iterations. Start from an initial Q-matrix $\mathbf{Q}_0$. Let $\mathbf{Q}_m$ denote the Q-matrix of iteration m. Then

- For all $k, l \in E$, compute using (4.2.12), (4.2.13), (4.2.14) the matrices $E_{y_i y_{i+1}}(t_{i+1} - t_i)$, and $F^{kl}_{y_i y_{i+1}}(t_{i+1} - t_i)$ associated to $Q = \mathbf{Q}_m$
- Compute the two quantities $\mathbb{E}(N^k(T)|Y = y)$, $\mathbb{E}(N^{kl}(T)|Y = y)$ using (4.2.10)
- Define $\mathbf{Q}_{m+1}$ by (4.2.9).

Let $\mathbf{Q}_0, \mathbf{Q}_1, \ldots, \mathbf{Q}_p, \ldots$ a sequence a Q- matrices obtained by the EM algorithm. Then $L_{n,\Delta}(\mathbf{Q}_{p+1}) \geq L_{n,\Delta}(\mathbf{Q}_p)$ for $p = 0, 1, 2, \ldots$ (see e.g. [35]). Under additional regularity conditions, one can prove (cf [14], Theorem 4) that, If $\mathbf{Q}_0$ satisfies that, for all $k, l \in E$, $(\mathbf{Q}_0)_{kl} > 0$, then the sequence $(\mathbf{Q}_p)$ converge to a stationary point of the likelihood function $L_{n,\Delta}$ or $det\{exp(\mathbf{Q}_p)\} \to 0$.

4.3 ABC estimation

Markov Chain Monte Carlo (MCMC) methods that treat the missing data as extra parameters, have become increasingly popular for calibrating stochastic epidemiological models with missing data [26, 105, 107]. However, MCMC may be computationally prohibitive for high-dimensional missing observations [27, 120] and fine tuning of the proposal distribution is required for efficient algorithms [53]. The computation of the likelihood can sometimes be numerically infeasible because it involves integration over the unobserved events. In discrete time, or when the total population size is known and small as in [107], this is possible. But in (4.2.1) for example, because we are in continuous time, the likelihood of removal times, when the infection times and K_t^N are unknown, involves a summation over all possibilities which is impossible: the sum is over all the possible numbers of infections between each successive removal times, plus on the possible times of these infections. An alternative is given by Approximate Bayesian Computation (ABC), which was originally proposed for making inference in population genetics [10]. This approach is not based on the likelihood function but relies on numerical simulations and comparisons between simulated and observed summary statistics. We detail here the ABC procedure and its application to epidemiology. For more information on ABC methods, the interested reader is referred to [100, 114]. In particular, there have been many refinements of the ABC method presented here, for instance using simulations to modify the sampling distributions (e.g. [9, 116, 121]).

In [17], the development of ABC estimation techniques for SIR models is motivated by the study of the Cuban HIV-AIDS database. In this case, the population is separated into the following compartments: 1) susceptible individuals who can be infected by HIV, 2) non-detected HIV positive infectious individuals who can propagate the disease, and 3) detected HIV positive individuals. When an individual is detected as HIV positive, we assume that the transmission of the disease ceases. So detection corresponds here to 'recovery' events in the classical SIR model presented in Part I of this book. The Cuban database contains the dates of detection of the 8,662 individuals

that have been found to be HIV positive in Cuba between 1986 and 2007 [4]. The database contains additional covariates including the manner by which an individual has been found to be HIV positive. The individuals can be detected either by *random screening* (individuals 'spontaneously' take a detection test) or *contact-tracing*. The total number of infectious individuals as well as the infection times are unknown. Blum and Tran [17] proposed an ABC estimation procedure when all detection times are known, which they then extend to noisy or binned detection times. They also propose an extension of ABC to path-valued summary statistics consisting of the cumulated number of detections through time. They introduce a finite-dimensional vector of summary statistics and compare the statistical properties of point estimates and credibility intervals obtained with full and binned detection times. We present here these methods for a simple SIR model and compare numerically the posterior distributions obtained with ABC and MCMC. We refer the reader to [17] for more details and treatment of Cuban HIV data. Other use of ABC estimation techniques in public health can be found in [39, 103] for example.

4.3.1 Main principles of ABC

For simplicity, we deal here with densities and not general probability measures. Let $\mathbf{x}$ be the available data and $\pi(\theta)$ be the prior where θ is the parameter. Two approximations are at the core of ABC.

Replacing observations with summary statistics Instead of focusing on the posterior density $p(\theta \,|\, \mathbf{x})$, ABC aims at a possibly less informative *target* density $p(\theta \,|\, S(\mathbf{x}) = s_{obs}) \propto \Pr(s_{obs}|\theta)\pi(\theta)$ where S is a summary statistic that takes its values in a normed space, and s_{obs} denotes the observed summary statistic. The summary statistic S can be a d-dimensional vector or an infinite-dimensional variable such as a L^1 function. Of course, if S is sufficient, then the two conditional densities are the same. The target distribution will also be coined as the *partial posterior distribution*.

Simulation-based approximations of the posterior Once the summary statistics have been chosen, the second approximation arises when estimating the partial posterior density $p(\theta \,|\, S(\mathbf{x}) = s_{obs})$ and sampling from this distribution. This step involves nonparametric kernel estimation and possibly correction refinements.

4.3.1.1 Sampling from the posterior

The ABC method with smooth rejection generates random draws from the target distribution as follows (see e.g. [10])

1. Generate N random draws (θ_i, s_i), $i = 1, \ldots, N$. The parameter θ_i is generated from the prior distribution π and the vector of summary statistics s_i is calculated for the i^{th} data set that is simulated from the generative model with parameter θ_i.

2. Associate to the i^{th} simulation the weight $W_i = K_\delta(s_i - s_{obs})$, where δ is a tolerance threshold and K_δ a (possibly multivariate) smoothing kernel.

3. The distribution $(\sum_{i=1}^N W_i \delta_{\theta_i})/(\sum_{i=1}^N W_i)$, in which δ_θ denotes the Dirac mass at θ, approximates the target distribution.

4.3.1.2 Point estimation and credibility intervals

Assume here that $\theta = (\theta_1, \ldots \theta_d)$ is a d-dimensional vector. We denote by $\theta_i = (\theta_{1,i}, \ldots \theta_{d,i})$ the simulated vectors of parameters in the previous paragraph. Once a sample from the target distribution has been obtained, several estimators may be considered for point estimation of each one-dimensional component θ_j, $j \in \{1, \ldots d\}$. Using the weighted sample $(\theta_{j,i}, W_i)$, $i = 1, \ldots, N$, the *mean* of the target distribution $p(\theta_j | s_{obs})$ is estimated by

$$\hat{\theta_j} = \frac{\sum_{i=1}^N \theta_{j,i} W_i}{\sum_{i=1}^N W_i} = \frac{\sum_{i=1}^N \theta_{j,i} K_\delta(s_i - s_{obs})}{\sum_{i=1}^N K_\delta(s_i - s_{obs})}, \quad j = 1,2,3 \tag{4.3.1}$$

which is the well-known Nadaraya–Watson regression estimator of the conditional expectation $\mathbb{E}(\theta_j \,|\, s_{obs})$ (see e.g. [122, Chapter 1]). We also compute the *medians*, *modes*, and 95% credibility intervals (CI) of the marginal posterior distribution (see Section 3 of the supplementary material).

4.3.1.3 Summary statistics

We are here interested in estimating the parameter $\theta = (\lambda, \gamma)$ of a SIR model (see Part I of this book). Two different sets of summary statistics are considered.

First, we consider the (infinite-dimensional) statistics $(R_t, t \in [0,T])$ consisting of the cumulated number of recoveries at time t since the beginning of the epidemic. Because the data consist of the recovery times this curve $(R_t, t \in [0,T])$ can simply be viewed as a particular coding of the whole dataset. It is thus a sufficient statistic implying that the partial posterior distribution $p(\theta \,|\, R^1, R^2)$ is equal to the posterior distribution $p(\theta \,|\, \mathbf{x})$.
The L^1-norm between the ith simulated path R_i and the observed one R_{obs} is

$$\|R_{obs} - R_i\|_1 = \int_0^T |R_{obs,s} - R_{i,s}|\, ds \quad , i = 1, \ldots, N. \tag{4.3.2}$$

The weights W_i are then computed as $W_i = K_\delta(\|R_{obs} - R_i\|_1)$ where δ is a tolerance threshold found by accepting a given percentage P_δ of the simulations and where an Epanechnikov kernel is chosen for K.

Second, when there is noise or when the recovery times have been binned, the full observations $(R_t, t \in [0,T])$ are unavailable. Then, we replace these summary statistics by a vector of summary statistics such as the numbers of recoveries per year during the observation period. We consider a d-dimensional vector of summary statistics of three different types: 1) number R_T of individuals detected by the end of the observation period, 2) for each year j, numbers of removed individuals $R_{j+1} - R_j$, 3) numbers of new infectious in the first years (assuming for instance that all of them have been detected since) $I_{j+1} - I_j$ for $j = 0, \ldots, J_0$, where J_0 is a small number of years where the information is supposed to be known, 4) mean time during which an individual is infected but has not been detected in the J_0 first years. This mean time corresponds to the mean sojourn time in the class I for the J_0 first years. Since these new summary statistics are not sufficient anymore, the new partial posterior distribution may be different from the posterior $p(\theta \,|\, \mathbf{x})$.
In order to compute the weights W_i, we consider the following spherical kernel $K_\delta(x) \propto K(\|\mathbf{H}^{-1}x\|/\delta)$. Here K denotes the one-dimensional Epanechnikov kernel, $\|\cdot\|$ is the Euclidean norm of $\mathbb{R}^d$ and $\mathbf{H}^{-1}$ a matrix. Because the summary statistics may span different scales, $\mathbf{H}$ is taken equal to the diagonal matrix with the standard deviation of each one-dimensional summary statistic on the diagonal.

4.3.2 Comparisons between ABC and MCMC methods for a standard SIR model

Following [10] a performance indicator for ABC techniques consists in their ability to replicate likelihood-based results given by MCMC. Here the situation is particularly favourable for comparing the two methods since the partial and the full posterior are the same. In the following examples, we choose samples of small sizes ($n = 3$ and $n = 29$) so that the dimension of the missing data is reasonable and MCMC achieves fast convergence. For large sample sizes with high-dimensional missing data, MCMC convergence might indeed be a serious issue and more thorough updating scheme shall be implemented [27, 120].

We consider the standard SIR model with infection rate λ and recovery rate γ. The data consist of the recovery times and we assume that the infection times are not observed. We implement the MCMC algorithm of [107]. A total of $10,000$ steps are considered for MCMC with an initial burn-in of $5,000$ steps. For ABC, the summary statistic consists of the cumulative number of recoveries as a function of time. A total of $100,000$ simulations are performed for ABC.

The first example was previously considered by [107]. They simulated recovery times by considering one initial infectious individual and by setting $S_0 = 9$, $\lambda = 0.12$, and $\gamma = 1$. We choose gamma distributions for the

priors of λ and γ with a shape parameters of 0.1 and rate parameters of 1 and 0.1. As displayed by Figure 4.3.1, the posterior distributions obtained with ABC are extremely close to the ones obtained with MCMC provided that the tolerance rate is sufficiently small. We see that the tolerance rate changes importantly the posterior distribution obtained with ABC (see the posterior distributions for λ).

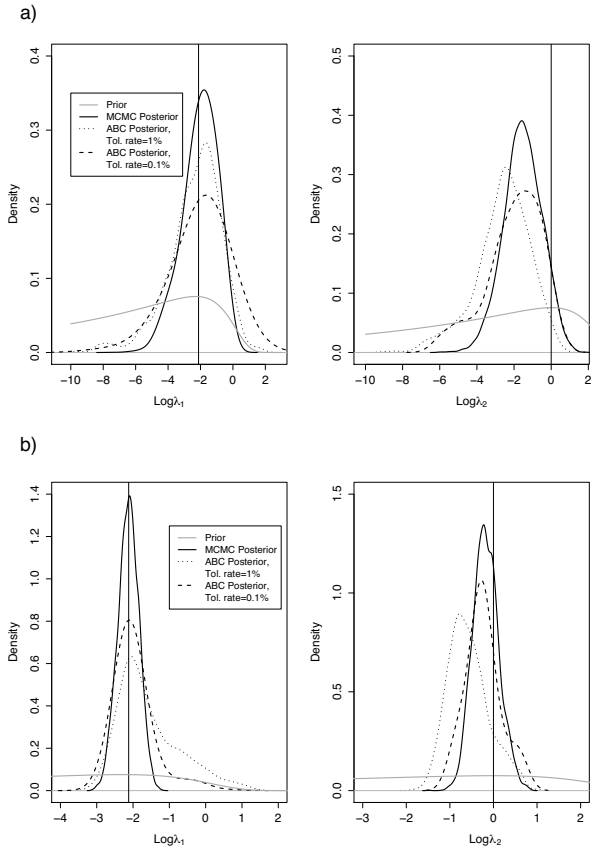


Figure 4.3.1: *Comparison of the posterior densities obtained with MCMC and ABC. The vertical lines correspond to the values of the parameters used for generating the synthetic data. Left: the data consist of 3 recovery times that have been simulated by [107]. Right: The data consist of 29 recovery times that we simulated by setting* $\lambda = 0.12$, $\gamma = 1$, $S_0 = 30$, $I_0 = 1$, *and* $T = 5$ *(see the supplementary material of [17] for the 29 recovery times).*

In a second example, we simulate a standard SIR trajectory with $\lambda = 0.12$, $\gamma = 1$, $S_0 = 30$ and $I_0 = 1$. The data now consist of 29 recovery times (and are given in the supplementary material of [17]). Once again, Figure 4.3.1 shows that the ABC and MCMC posteriors are close provided that the tolerance rate is small enough. ABC produces posterior distributions with larger tails compared to MCMC, even with the lowest tolerance rate of 0.1%. This can be explained by considering the extreme scenario in which the tolerance threshold δ goes to infinity: every simulation has a weight of 1 so that ABC targets the prior instead of the posterior. As the prior has typically larger tails than the posterior, ABC inflates the posterior tails.

4.3.3 Comparison between ABC with full and binned recovery times

4.3.3.1 The curse of dimensionality and regression adjustments

In this case, the first set of summary statistics presented in Section 4.3.1 can not be used any more and we have to use the second set of summary statistics, which constitute a vector of descriptive statistics as is much often encountered in the literature. In the case of a d-dimensional vector of summary statistics, the estimator of the conditional mean (4.3.1) is convergent if the tolerance rate satisfies $\lim_{N\to+\infty}\delta_N = 0$, so that its bias converges to 0, and $\lim_{N\to+\infty} N\delta_N^d = +\infty$, so that its variance converges to 0 [41]. As d increases, a larger tolerance threshold shall be chosen to keep the variance small. As a consequence, the bias may increase with the number of summary statistics. This phenomenon known as the *curse of dimensionality* may be an issue for the ABC-rejection approach. The following paragraph presents regression-based adjustments that cope with the curse of dimensionality.

The adjustment principle is presented in a general setting within which the corrections of [10] and [16] can be derived. Correction adjustments aim at obtaining from a random couple (θ_i, s_i) a random variable distributed according to $p(\theta \mid s_{obs})$. The idea is to construct a coupling between the distributions $p(\theta|s_i)$ and $p(\theta|s_{obs})$, through which we can shrink the θ_i's to a sample of i.i.d. draws from $p(\theta|s_{obs})$. In the remaining of this subsection, we describe how to perform the corrections for each of the one-dimensional components separately. For $\theta \in \mathbb{R}$, correction adjustments are obtained by assuming a relationship $\theta = G(s,\varepsilon) =: G_s(\varepsilon)$ between the parameter and the summary statistics. Here G is a (possibly complicated) function and ε is a random variable with a distribution that does not depend on s. A possibility is to choose $G_s = F_s^{-1}$, the (generalized) inverse of the cumulative distribution function of $p(\theta|s)$. In this case, $\varepsilon = F_s(\theta)$ is a uniform random variable on $[0,1]$. The formula for adjustment is given by

$$\theta_i^* = G_{s_{obs}}^{-1}(G_{s_i}(\theta_i)) \quad i = 1,\dots,N. \tag{4.3.3}$$

For $G_s = F_s^{-1}$, the fact that the θ_i^*'s are i.i.d. with density $p(\theta|s_{obs})$ arises from the standard inversion algorithm. Of course, the function G shall be approximated in practice. As a consequence, the adjusted simulations θ_i^*, $i = 1,\dots,N$, constitute an approximate sample of $p(\theta \mid s_{obs})$. The ABC algorithm with regression adjustment can be described as follows

1. Simulate, as in the rejection algorithm, a sample (θ_i, s_i), $i = 1,\dots,N$.
2. By making use of the sample of the (θ_i, s_i)'s weighted by the W_i's, approximate the function G such that $\theta_i = G(s_i, \varepsilon_i)$ in the vicinity of s_{obs}.
3. Replace the θ_i's by the adjusted θ_i^*'s. The resulting weighted sample (θ_i^*, W_i), $i = 1,\dots,N$, form a sample from the target distribution.

Local linear regression (LOCL) The case where G is approximated by a linear model $G(s,\varepsilon) = \alpha + s^t\beta + \varepsilon$, was considered by [10]. The parameters α and β are inferred by minimizing the weighted squared error

$$\sum_{i=1}^{N} K_\delta(s_i - s_{obs})(\theta_i - (\alpha + (s_i - s_{obs})^T\beta))^2.$$

Using (4.3.3), the correction of [10] is derived as

$$\theta_i^* = \theta_i - (s_i - s_{obs})^T\hat{\beta},\ i = 1,\dots,N. \tag{4.3.4}$$

Asymptotic consistency of the estimators of the partial posterior distribution with the correction (4.3.4) is obtained by [15].

Nonlinear conditional heteroscedastic regressions (NCH) To relax the assumptions of homoscedasticity and linearity inherent to local linear regression, Blum and Francois [16] approximated G by $G(s,\varepsilon) = m(s) + \sigma(s)\times\varepsilon$ where $m(s)$ denotes the conditional expectation, and $\sigma^2(s)$ the conditional variance. The estimators $\hat{m}$ and $\log\hat{\sigma}^2$

are found by adjusting two feed-forward neural networks using a regularized weighted squared error. For the NCH model, parameter adjustment is performed as follows

$$\theta_i^* = \hat{m}(s_{obs}) + (\theta_i - \hat{m}(s_i)) \times \frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s_i)},\ i = 1,\dots,N.$$

In practical applications of the NCH model, we train $L = 10$ neural networks for each conditional regression (expectation and variance) and we average the results of the L neural networks to provide the estimates $\hat{m}$ and $\log \hat{\sigma}^2$.

Reparameterization In both regression adjustment approaches, the regressions can be performed on transformations of the responses θ_i rather that on the responses themselves. Parameters whose prior distributions have finite supports are transformed via the logit function and non-negative parameters are transformed via the logarithm function. These transformations guarantee that the θ_i^*'s lie in the support of the prior distribution and have the additional advantage of stabilizing the variance.

Comparison between the first and second set of summary statistics A simulation study is carried to compare the ABC methods based on the two different sets of summary statistics presented in Section 4.3.1 has been carried in [17] using a slightly more elaborate SIR model with contact-tracing introduced in [31]. Blum and Tran simulated $M = 200$ synthetic data sets epidemic. When using the finite-dimensional vector of summary statistics, they perform the smooth rejection approach as well as the LOCL and NCH corrections with a total of 21 summary statistics. Each of the $M = 200$ estimations of the partial posterior distributions are performed using a total of $N = 5000$ simulations.

Figure 4.3.2 displays the boxplots of the 200 estimated modes, medians, 2.5% and 97.5% quantiles of the posterior distribution for λ as a function of the tolerance rate P_δ. First, the medians and modes are found to be equivalent except for the rejection method with 21 summary statistics for which the mode is less biased. For the lowest tolerance rates, the point estimates obtained with the four possible methods are close to the value λ used in the simulations, with smaller CI for the LOCL and NCH variants. When increasing the tolerance rate, the bias of the point estimates obtained with the rejection method with 21 summary statistics slightly increases. By contrast, up to tolerance rates smaller than 50%, the biases of the point estimates obtained with the three other methods remain small. As can be expected, the widths of the CI obtained with the rejection methods increase with the tolerance rate while they remain considerably less variable for the methods with regression adjustment.

For further comparison of the different methods, we can compute the rescaled mean square errors (ReMSEs):

$$\text{ReMSE}(\lambda) = \frac{1}{M}\sum_{k=1}^{M} \frac{(\log(\hat{\lambda^k}) - \log(\lambda))^2}{\text{Range}(\text{prior}(\lambda))^2}, \tag{4.3.5}$$

where $\hat{\lambda}^k$ is a point estimate obtained with the kth synthetic data set.

To compare the whole posterior distributions obtained with the four different methods, we can also compute the different CIs. The rescaled mean CI (RMCI) is defined as follows

$$\text{RMCI} = \frac{1}{M}\sum_{k=1}^{M} \frac{|IC^k|}{\text{Range}(\text{prior}(\lambda))}, \tag{4.3.6}$$

where $|IC^k|$ is the length of the kth estimated 95% CI for the parameter λ. As displayed by Figure 4.3.2, the CIs obtained with smooth rejection increase importantly with the tolerance rate whereas such an important increase is not observed with regression adjustment.

4.4 Sensitivity analysis

Epidemiological models designed in order to test public health scenarios by simulations or disentangle various factors for a better understanding of the disease propagation are often over-parameterized. Input parameters are

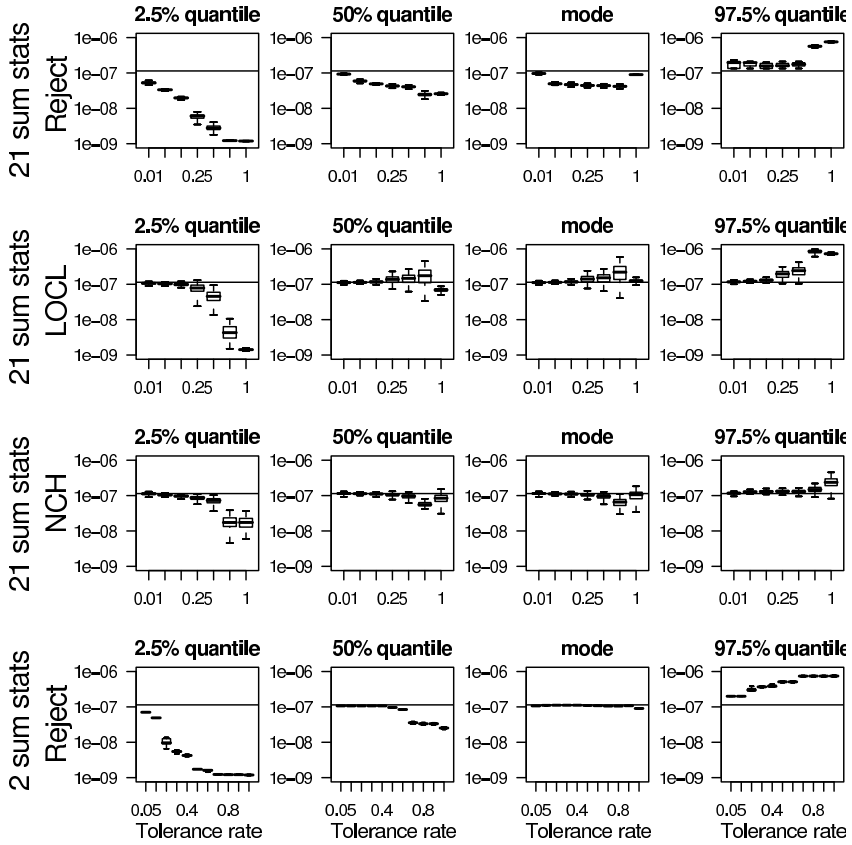


Figure 4.3.2: *Boxplots of the* $M = 200$ *estimated modes and quantiles (*2.5%, 50%, *and* 97.5%*) of the partial posterior distributions of* λ *in a model presented in Blum and Tran [17]. For each ABC method and each value of the tolerance rate,* 200 *posterior distributions are computed for each of the 200 synthetic data sets. The horizontal lines correspond to the true value* $\lambda = 1.14 \times 10^{-7}$ *used when simulating the 200 synthetic data sets. The different tolerance rates are 0.01, 0.05, 0.10, 0.25, 0.50, 0.50, 0.75, and 1 for all the ABC methods except the rejection scheme with the two summary statistics. For the latter method, the tolerance rates are* 0.007, 0.02, 0.06, 0.13, 0.27, 0.37, 0.45, 0.53, 0.66, 0.80, 1.

the rates describing the times that individuals stay in each compartment, for example. The sources that are used to calibrate the model can also be numerous: some parameters are for example obtained from epidemiological studies or clinical trials, but there can be uncertainty on their values due to various reasons. The restricted size of the sample in these studies brings uncertainty on the estimates, which are given with uncertainty intervals (classically, a 95% confidence interval). Different studies can provide different estimates for the same parameters. The study populations can be subject to selection biases. In the case of clinical trials where the efficacy of a treatment is estimated, the estimates can be optimistic compared with what will be the effectiveness in real-life, due to the protocol of the trials. It is important to quantify how theses uncertainties on the input parameters can impact the results and the conclusion of an epidemiological modelling study. To check the robustness of some output with respect to the parameters, sensitivity analyses are often performed.

In a mathematical model where the output $y \in \mathbb{R}$ depends on a set of $p \in \mathbb{N}$ input parameters $x = (x_1, ...x_p) \in \mathbb{R}^p$ through the relation $y = f(x)$, there are various ways to measure the influence of the input x_ℓ, for $\ell \in \{1, \ldots, p\}$, on y. In this article, we are interested in Sobol indices [117], which are based on an ANOVA decomposition (see [112, 77, 78] for a review). These indices have been proposed to take into account the uncertainty on the input parameters that are here considered as a realisation of a set of independent random variables $X = (X_1, ...X_p)$, with

a known distribution and with possibly correlated components. Denoting by $Y = f(X)$ the random response, the first-order Sobol indices can be defined for $\ell \in \{1,\dots,p\}$ by

$$S_\ell = \frac{\mathrm{Var}\big(\mathbb{E}[Y \mid X_\ell]\big)}{\mathrm{Var}(Y)}. \tag{4.4.1}$$

This first-order index S_ℓ corresponds to the sensitivity of the model to X_ℓ alone. Higher order indices can also be defined using ANOVA decomposition: considering $(\ell,\ell') \in \{1,\dots,p\}$, we can define the second order sensitivity, corresponding to the sensitivity of the model to the interaction between X_ℓ and $X_{\ell'}$ index by

$$S_{\ell\ell'} = \frac{\mathrm{Var}\big(\mathbb{E}[Y \mid X_\ell, X_{\ell'}]\big)}{\mathrm{Var}(Y)} - S_\ell - S_{\ell'} \tag{4.4.2}$$

We can also define the total sensitivity indices by

$$S_{T_\ell} = \sum_{L \subset \{1,\dots,p\} \,|\, \ell \in L} S_L. \tag{4.4.3}$$

As the estimation of the Sobol indices can be computer time consuming, a usual practice consists in estimating the first-order and total indices, to assess 1) the sensitivity of the model to each parameter taken separately and 2) the possible interactions, which are quantified by the difference between the total order and the first-order index for each parameter. Several numerical procedures to estimate the Sobol indices have been proposed, in particular by Jansen [82] (see also [111, 112]). These estimators, that we recall in the sequel, are based on Monte Carlo simulations of $(Y, X_1 \dots X_p)$.

The literature focuses on deterministic relations between the input and output parameters. In a stochastic framework where the model response Y is not unique for given input parameters, few works have been done, randomness being usually limited to input variables. Assume that:

$$Y = f(X, \varepsilon), \tag{4.4.4}$$

where $X = (X_1, \dots X_p)$ still denotes the random variables modelling the uncertainty of the input parameters and where ε is a noise variable. When noise is added in the model, the classical estimators do not always work: Y can be very sensitive to the addition of ε. Moreover, this variable is not always controllable by the user.

When the function f is linear, we can refer to [44]. In the literature, meta-models are used: approximating the mean and the dispersion of the response by deterministic functions allows us to come back to the classical deterministic framework (e.g. Janon et al. [81], Marrel et al. [101]). We study here another point of view, which is based on the non-parametric statistical estimation of the term $\mathrm{Var}\big(\mathbb{E}[Y \mid X_\ell]\big)$ appearing in the numerator of (4.4.1). Approaches based on the Nadaraya–Watson kernel estimator have been proposed by Da Veiga and Gamboa [126] or Solís [118] while an approach based on warped wavelet decompositions is proposed by Castellan et al. [25]. An advantage of these non-parametric estimators is that their computation requires less simulations of the model. For Jansen estimators, the number of calls of f required to compute the sensitivity indices is $n(p+1)$, where n is the number of independent random vectors $(Y^i, X_1^i, \dots X_p^i)$ $(i \in \{1,\dots n\})$ that are sampled for the Monte Carlo procedure, making the estimation of the sensitivity indices time-consuming for sophisticated models with many parameters. In addition, for the non-parametric estimators, the convergence of the mean square error to zero may be faster than for Monte Carlo estimators, depending on the regularity of the model.

4.4.1 A non-parametric estimator of the Sobol indices of order 1

Denoting by $V_\ell = \mathbb{E}\big(\mathbb{E}^2(Y \mid X_\ell)\big)$ the expectation of the square conditional expectation of Y knowing X_ℓ, we have:

$$S_\ell = \frac{V_\ell - \mathbb{E}(Y)^2}{\mathrm{Var}(Y)}, \tag{4.4.5}$$

which can be approximated by

$$\widehat{S}_\ell = \frac{\widehat{V}_\ell - \bar{Y}^2}{\widehat{\sigma}_Y^2} \tag{4.4.6}$$

where

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j \text{ and } \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

are the empirical mean and variance of Y. We consider here two approximations $\widehat{V}_\ell$ of V_ℓ, based on Nadaraya–Watson and on warped wavelet estimators.

Assume that we have n independent couples $(Y^i, X_1^i, \dots X_p^i)$ in $\mathbb{R}\times\mathbb{R}^p$, for $i \in \{1,\dots,n\}$, generated by (4.4.4). Let us start with the kernel-based estimator:

Definition 4.4.1. *Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a kernel such that $\int_{\mathbb{R}} K(u)du = 1$. Let $h > 0$ be a window and let us denote $K_h(x) = K(x/h)/h$. An estimator of S_ℓ for $\ell \in \{1,\dots p\}$ is:*

$$\widehat{S}_\ell^{(NW)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n} Y_j K_h(X_\ell^j - X_\ell^i)}{\sum_{j=1}^{n} K_h(X_\ell^j - X_\ell^i)}\right)^2 - \bar{Y}^2}{\widehat{\sigma}_Y^2}. \tag{4.4.7}$$

This estimator is based on the Nadaraya–Watson estimator of $\mathbb{E}(Y \,|\, X_\ell = x)$ given by (e.g. [122])

$$\frac{\sum_{j=1}^{n} Y_j K_h(X_\ell^j - x)}{\sum_{j=1}^{n} K_h(X_\ell^j - x)}.$$

Replacing this expression in (4.4.6) provides $\widehat{S}_\ell^{(NW)}$. This estimator and the rates of convergence have been studied by Solís [118]. If we instead use a warped wavelet decomposition of $\mathbb{E}(Y \,|\, X_\ell = x)$ (see e.g. [29, 85]), this provides an estimator studied by Castellan et al. [25]. Let us present this second estimator.

Let us denote by G_ℓ the cumulative distribution function of X_ℓ. Let $(\psi_{jk})_{j\geq -1, k\in\mathbb{Z}}$ be a Hilbert wavelet basis of L^2, the space of real functions that are square integrable with respect to the Lebesgue measure on $\mathbb{R}$. In the sequel, we denote by $\langle f, g\rangle = \int_{\mathbb{R}} f(u)g(u)du$, for $f, g \in L^2$, the usual scalar product of L^2. The wavelet ψ_{-10} is the father wavelet, and for $k \in \mathbb{Z}$, $\psi_{-1k}(x) = \psi_{-10}(x-k)$. The wavelet ψ_{00} is the mother wavelet, and for $j \geq 0$, $k \in \mathbb{Z}$, $\psi_{jk}(x) = 2^{j/2}\psi_{00}(2^j x - k)$.

Definition 4.4.2. *Let us define for $j \geq -1$, $k \in \mathbb{Z}$,*

$$\widehat{\beta}_{jk}^{\ell} = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{jk}(G_\ell(X_\ell^i)). \tag{4.4.8}$$

Then, we define the (block thresholding) estimator of S_ℓ as

$$\widehat{S}_\ell^{(WW)} = \frac{\widehat{V}_\ell - \bar{Y}^2}{\widehat{\sigma}_Y^2}, \tag{4.4.9}$$

where $\widehat{V}_\ell$ is an estimator of the variance V_ℓ given by:

$$\widehat{V}_\ell = \sum_{j=-1}^{J_n}\Big[\sum_{k\in\mathbb{Z}}\left(\widehat{\beta}_{jk}^{\ell}\right)^2 - w(j)\Big]\mathbf{1}_{\sum_{k\in\mathbb{Z}}\left(\widehat{\beta}_{jk}^{\ell}\right)^2 \geq w(j)} \tag{4.4.10}$$

with $w(j) = K\left(\frac{2^j + \log 2}{n}\right)$ and $J_n := \left[\log_2\left(\sqrt{n}\right)\right]$ (where $[\cdot]$ denotes the integer part) and K a positive constant.

Let us present the idea explaining the estimator proposed in Definition 4.4.2. Let us introduce centered random variables η_ℓ such that

$$Y = f(X,\varepsilon) = \mathbb{E}(Y\,|\,X_\ell) + \eta_\ell. \tag{4.4.11}$$

Let $g_\ell(x) = \mathbb{E}(Y\,|\,X_\ell = x)$ and $h_\ell(u) = g_\ell \circ G_\ell^{-1}(u)$. h_ℓ is a function from $[0,1] \mapsto \mathbb{R}$ that belong to L^2 since $Y \in L^2$. Then

$$h_\ell(u) = \sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u), \quad \text{with} \tag{4.4.12}$$

$$\beta_{jk}^\ell = \int_0^1 h_\ell(u)\psi_{jk}(u)du = \int_{\mathbb{R}} g_\ell(x)\psi_{jk}(G_\ell(x))G_\ell(dx). \tag{4.4.13}$$

Notice that the sum in k is finite because the function h_ℓ has compact support in $[0,1]$. It is then natural to estimate $h_\ell(u)$ by

$$\widehat{h}_\ell = \sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \widehat{\beta}_{jk}^\ell \psi_{jk}(u), \tag{4.4.14}$$

and we then have:

$$\begin{aligned} V_\ell &= \mathbb{E}\big(\mathbb{E}^2(Y\,|\,X_\ell)\big) \\ &= \int_{\mathbb{R}} G_\ell(dx)\Big(\sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \beta_{jk}^\ell \psi_{jk}\big(G_\ell(x)\big)\Big)^2 \\ &= \int_0^1 \Big(\sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u)\Big)^2 du \\ &= \sum_{j\geq -1}\sum_{k\in\mathbb{Z}} \big(\beta_{jk}^\ell\big)^2 = \|h_\ell\|_2^2. \end{aligned} \tag{4.4.15}$$

Adaptive estimation of $\|h_\ell\|_2^2$ has been studied in [95], which provides the block thresholding estimator $\widehat{V}_\ell$ in Definition 4.4.2. The idea is: 1) to sum the terms $\big(\beta_{jk}^\ell\big)^2$, for $j \geq 0$, by blocks $\{(j,k),\ k\in\mathbb{Z}\}$ for $j \in \{-1,\dots,J_n\}$ with a penalty $w(j)$ for each block to avoid choosing too large j's, 2) to cut the blocks that do not sufficiently contribute to the sum, in order to obtain statistical adaptation.

Notice that $\widehat{V}_\ell$ can be seen as an estimator of V_ℓ resulting from a model selection on the choice of the blocks $\{(j,k),\ k\in\mathbb{Z}\},\ j\in\{-1,\dots,J_n\}$ that are kept, with the penalty function $\mathrm{pen}(\mathscr{J}) = \sum_{j\in\mathscr{J}} w(j)$, for $\mathscr{J} \subset \{-1,\dots,J_n\}$. Indeed:

$$\begin{aligned} \widehat{V}_\ell &= \sup_{\mathscr{J}\subset\{-1,0,\dots,J_n\}} \sum_{j\in\mathscr{J}} \Big[\sum_{k\in\mathbb{N}} \big(\widehat{\beta}_{jk}^\ell\big)^2 - w(j)\Big] \\ &= \sup_{\mathscr{J}\subset\{-1,0,\dots,J_n\}} \sum_{j\in\mathscr{J}}\sum_{k\in\mathbb{N}} \big(\widehat{\beta}_{jk}^\ell\big)^2 - \mathrm{pen}(\mathscr{J}). \end{aligned} \tag{4.4.16}$$

Note that the definition of the estimator and the penalization depend on a constant K through the definition of $w(j)$. The value of this constant is chosen in order to obtain oracle inequalities. In practice, this constant is hard to compute, and can be chosen by a slope heuristic approach (see e.g. [5]).

4.4.2 Statistical properties

In this Section, we are interested in the rate of convergence to zero of the mean square error (MSE) $\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big)$. Let us consider the generic estimator $\widehat{S}_\ell$ defined in (4.4.6), where $\widehat{V}_\ell$ is an estimator of $V_\ell = \mathbb{E}(\mathbb{E}^2(Y\mid X_\ell))$ (not necessarily (4.4.10)). We first start with a Lemma stating that the MSE can be obtained from the rate of convergence of $\widehat{V}_\ell$ to V_ℓ.

Lemma 4.4.3. *Consider the generic estimator $\widehat{S}_\ell$ defined in* (4.4.6) *and $\widehat{V}_\ell$ an estimator of V_ℓ (not necessarily* (4.4.10)*). Then there is a constant C such that:*

$$\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) \leq \frac{C}{n} + \frac{4}{\mathrm{Var}(Y)^2}\mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell\big)^2\Big]. \tag{4.4.17}$$

Proof. From (4.4.5) and (4.4.6),

$$\begin{aligned}\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) =& \mathbb{E}\Big[\Big(\frac{V_\ell - \mathbb{E}(Y)^2}{\mathrm{Var}(Y)} - \frac{\widehat{V}_\ell - \bar{Y}^2}{\widehat{\sigma}_Y^2}\Big)^2\Big] \\ \leq& 2\mathbb{E}\Big[\Big(\frac{\mathbb{E}(Y)^2}{\mathrm{Var}(Y)} - \frac{\bar{Y}^2}{\widehat{\sigma}_Y^2}\Big)^2\Big] + 2\mathbb{E}\Big[\Big(\frac{V_\ell}{\mathrm{Var}(Y)} - \frac{\widehat{V}_\ell}{\widehat{\sigma}_Y^2}\Big)^2\Big]. \end{aligned} \tag{4.4.18}$$

The first term in the right-hand side (r.h.s.) is in C/n. For the second term in the right-hand side of (4.4.18):

$$\mathbb{E}\Big[\Big(\frac{V_\ell}{\mathrm{Var}(Y)} - \frac{\widehat{V}_\ell}{\widehat{\sigma}_Y^2}\Big)^2\Big] \leq 2\mathbb{E}\Big[\widehat{V}_\ell^2\Big(\frac{1}{\mathrm{Var}(Y)} - \frac{1}{\widehat{\sigma}_Y^2}\Big)^2\Big] + \frac{2}{\mathrm{Var}(Y)^2}\mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell\big)^2\Big]. \tag{4.4.19}$$

The first term in the r.h.s. is also in C/n, which concludes the proof. □

The preceding lemma implies that the rate of convergence of $\widehat{V}_\ell$ to V_ℓ is determinant for the rate of convergence of $\widehat{S}_\ell$. We recall the result of Solís [118], where an elbow effect for the MSE is shown when the regularity of the density of (X_ℓ, Y) varies. The case of the warped wavelet estimator introduced by Castellan et al [25] is studied at the end of the section and the rate of convergence is stated in Corollary 4.4.8.

4.4.2.1 MSE for the Nadaraya–Watson estimator

Using the preceding Lemma, Loubes Marteau and Solís prove an elbow effect for the estimator $\widehat{S}_\ell^{(NW)}$. Let us introduce $\mathscr{H}(\alpha, L)$, for $\alpha, L > 0$, the set of functions ϕ of class $[\alpha]$, whose derivative $\phi^{([\alpha])}$ is $\alpha - [\alpha]$ Hölder continuous with constant L.

Proposition 4.4.4 (Loubes Marteau and Solís [99, 118])**.** *Assume that $\mathbb{E}(X_\ell^4) < +\infty$, that the joint density $\phi(x,y)$ of (X_ℓ, Y) belongs to $\mathscr{H}(\alpha, L)$, for $\alpha, L > 0$ and that the marginal density of X_ℓ, ϕ_ℓ belongs to $\mathscr{H}(\alpha', L')$ for $\alpha' > \alpha$ and $L' > 0$. Then:*
If $\alpha \geq 2$, there exists a constant $C > 0$ such that

$$\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) \leq \frac{C}{n}.$$

If $\alpha < 2$, there exists a constant $C > 0$ such that

$$\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) \leq C\big(\frac{\log^2 n}{n}\big)^{\frac{2\alpha}{\alpha+2}}.$$

For smooth functions ($\alpha \geq 2$), Loubes et al. recover a parametric rate, while they still have a nonparametric one when $\alpha < 2$. Their result is based on (4.4.17) and a bound for $\mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell\big)^2\Big]$ given by [99, Th. 1], whose proof is technical. Since their result is not adaptive, they require the knowledge of the window h for numerical implementation. Our purpose is to provide a similar result for the warped wavelet adaptive estimator, with a shorter proof.

4.4.2.2 MSE for the warped wavelet estimator

Let us introduce first some additional notation. We define, for $\mathscr{J} \subset \{-1, \ldots, J_n\}$, the projection $h_{\mathscr{J},\ell}$ of h on the subspace spanned by $\{\psi_{jk},$ with $j \in \mathscr{J}, k \in \mathbb{Z}\}$ and its estimator $\widehat{h}_{\mathscr{J},\ell}$:

$$h_{\mathscr{J},\ell}(u) = \sum_{j \in \mathscr{J}} \sum_{k \in \mathbb{Z}} \beta_{jk}^\ell \psi_{jk}(u) \tag{4.4.20}$$

$$\widehat{h}_{\mathscr{J},\ell}(u) = \sum_{j\in\mathscr{J}}\sum_{k\in\mathbb{Z}} \widehat{\beta}^{\ell}_{jk}\psi_{jk}(u). \tag{4.4.21}$$

We also introduce the estimator of V_ℓ for a fixed subset of resolutions $\mathscr{J}$:

$$\widehat{V}_{\mathscr{J},\ell} = \|\widehat{h}_{\mathscr{J},\ell}\|_2^2 = \sum_{j\in\mathscr{J}}\sum_{k\in\mathbb{Z}} \big(\widehat{\beta}^{\ell}_{jk}\big)^2. \tag{4.4.22}$$

Note that $\widehat{V}_{\mathscr{J},\ell}$ is one possible estimator $\widehat{V}_\ell$ in Lemma 4.4.3.

The estimators $\widehat{\beta}_{jk}$ and $\widehat{V}_{\mathscr{J},\ell}$ have natural expressions in term of the empirical process $\gamma_n(dx)$ defined as follows:

Definition 4.4.5. *The empirical measure associated with our problem is:*

$$\gamma_n(dx) = \frac{1}{n}\sum_{i=1}^{n} Y_i\delta_{G_\ell(X^i_\ell)}(dx) \tag{4.4.23}$$

where $\delta_a(dx)$ denotes the Dirac mass in a.
For a measurable function f, $\gamma_n(f) = \frac{1}{n}\sum_{i=1}^n Y_i f\big(G_\ell(X^i_\ell)\big)$. We also define the centered integral of f with respect to $\gamma_n(dx)$ as:

$$\bar{\gamma}_n(f) = \gamma_n(f) - \mathbb{E}\big(\gamma_n(f)\big) \tag{4.4.24}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i f\big(G_\ell(X^i_\ell)\big) - \mathbb{E}\big[Y_i f\big(G_\ell(X^i_\ell)\big)\big]\Big). \tag{4.4.25}$$

Using the empirical measure $\gamma_n(dx)$, we have:

$$\widehat{\beta}^{\ell}_{jk} = \gamma_n\big(\psi_{jk}\big) = \beta^{\ell}_{jk} + \bar{\gamma}_n\big(\psi_{jk}\big).$$

Let us introduce the correction term

$$\zeta_n = 2\bar{\gamma}_n\big(h_\ell\big) \tag{4.4.26}$$

$$= 2\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i h_\ell\big(G_\ell(X^i_\ell)\big) - \mathbb{E}\Big(Y_1 h_\ell\big(G_\ell(X^1_\ell)\big)\Big)\Big]$$

$$= 2\Big[\frac{1}{n}\sum_{i=1}^{n} h^2_\ell\big(G_\ell(X^i_\ell)\big) - \|h_\ell\|_2^2\Big] + \frac{2}{n}\sum_{i=1}^{n}\eta^i_\ell h_\ell\big(G_\ell(X^i_\ell)\big). \tag{4.4.27}$$

The rate of convergence of the estimator (4.4.10) is obtained in [25] based on the estimate presented in the next theorem. This result is derived using ideas due to Laurent and Massart [95] who considered estimation of quadratic functionals in a Gaussian setting. Because we are not necessarily in a Gaussian setting here, we rely on empirical processes and use sophisticated technology developed by Castellan [24].

Theorem 4.4.6 (Castellan, Cousien, Tran [25])**.** *Let us assume that the random variables Y are bounded by a constant M, and let us choose a father and a mother wavelets ψ_{-10} and ψ_{00} that are continuous with compact support (and thus bounded). The estimator $\widehat{V}_\ell$ defined in* (4.4.10) *is almost surely finite, and:*

$$\mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell - \zeta_n\big)^2\Big] \le C \inf_{\mathscr{J}\subset\{-1,\dots,J_n\}}\Big(\|h_\ell - h_{\mathscr{J},\ell}\|_2^4 + \frac{\mathrm{Card}^2(\mathscr{J})}{n^2}\Big) + \frac{C'\log_2^2(n)}{n^{3/2}}, \tag{4.4.28}$$

for constants C and $C' > 0$.

We deduce the following corollary from the estimate obtained above. Let us consider the Besov space $\mathscr{B}(\alpha,2,\infty)$ of functions $h = \sum_{j\ge -1}\sum_{k\in\mathbb{Z}}\beta_{jk}\psi_{jk}$ of L^2 such that

$$|h|_{\alpha,2,\infty} := \sum_{j\ge 0} 2^{j\alpha}\sqrt{\sup_{0<v\le 2^{-j}}\int_0^{1-v}|h(u+v)-h(u)|^2du} < +\infty.$$

For a $h \in \mathscr{B}(\alpha,2,\infty)$ and $h_{\mathscr{J}}$ its projection on

$$\mathrm{Vect}\{\psi_{jk},\ j \in \mathscr{J} = \{-1,\dots J_{\max}\},\ k \in \mathbb{Z}\},$$

we have the following approximation result from [66, Th. 9.4].

Proposition 4.4.7 (Härdle, Kerkyacharian, Picard and Tsybakov)**.** *Assume that the wavelet function ψ_{-10} has compact support and is of class $\mathscr{C}^N$ for an integer $N > 0$. Then, if $h \in \mathscr{B}(\alpha,2,\infty)$ with $\alpha < N+1$,*

$$\sup_{\mathscr{J}\subset\mathbb{N}\cup\{-1\}} 2^{\alpha J_{max}} \|h - h_{\mathscr{J}}\|_2 = \sup_{\mathscr{J}\subset\mathbb{N}\cup\{-1\}} 2^{\alpha J_{max}} \Big(\sum_{j\geq J_{max}} \sum_{k\in\mathbb{Z}} \beta_{jk}^2 \Big)^{1/2} < +\infty. \tag{4.4.29}$$

Notice that Theorem 9.4 of [66] requires assumptions that are fulfilled when ψ_{-10} has compact support and is smooth enough (see the comment after the Corol. 8.2 of [66]).

Corollary 4.4.8. *If ψ_{-10} has compact support and is of class $\mathscr{C}^N$ for an integer $N > 0$ and if h_ℓ belongs to a ball of radius $R > 0$ of $\mathscr{B}(\alpha,2,\infty)$ for $0 < \alpha < N+1$, then*

$$\sup_{h\in\mathscr{B}(\alpha,2,\infty)} \mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell\big)^2\Big] \leq C\Big(n^{-\frac{8\alpha}{4\alpha+1}} + \frac{1}{n}\Big). \tag{4.4.30}$$

As a consequence, we obtain the following elbow effect:
If $\alpha \geq \frac{1}{4}$, there exists a constant $C > 0$ such that

$$\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) \leq \frac{C}{n}.$$

If $\alpha < \frac{1}{4}$, there exists a constant $C > 0$ such that

$$\mathbb{E}\big((S_\ell - \widehat{S}_\ell)^2\big) \leq C n^{-\frac{8\alpha}{4\alpha+1}}.$$

Proof. Using (4.4.28) and the fact that

$$\mathbb{E}\big(\zeta_n^2\big) = \frac{4}{n}\mathrm{Var}\Big(Y_1 h_\ell\big(G_\ell(X_\ell^1)\big)\Big) \leq \frac{2M^2\|h_\ell\|_2^2}{n}, \tag{4.4.31}$$

we obtain:

$$\mathbb{E}\Big[\big(\widehat{V}_\ell - V_\ell\big)^2\Big] \leq C\Big[\inf_{\mathscr{J}\subset\{-1,\dots,J_n\}} \Big(\|h_\ell - h_{\mathscr{J},\ell}\|_2^4 + \frac{\mathrm{Card}^2(\mathscr{J})}{n^2}\Big) + \frac{1+\|h_\ell\|_2^2}{n}\Big]. \tag{4.4.32}$$

If $h_\ell \in \mathscr{B}(\alpha,2,\infty)$, then from Proposition 4.4.7, we have for $\mathscr{J} = \{-1,\dots,J_{\max}\}$ that $\|h_\ell - h_{\mathscr{J},\ell}\|_2^4 \leq 2^{-4\alpha\, J_{\max}}$. Thus, for subsets $\mathscr{J}$ of the form considered, the infimum is attained when choosing $J_{\max} = \frac{2}{4\alpha+1}\log_2(n)$, which yield an upper bound in $n^{8\alpha/(4\alpha+1)}$.

For h_ℓ in a ball of radius R, $\|h_\ell\|_2^2 \leq R^2$, and we can find an upper bound that does not depend on h. Because the last term in (4.4.32) is in $1/n$, the elbow effect is obtained by comparing the order of the first term in the r.h.s. ($n^{8\alpha/(4\alpha+1)}$) with $1/n$ when α varies. □
□

Let us remark that in comparison with the result of Loubes et al. [99], the regularity assumption here is on the function h_ℓ rather than on the joint density $\phi(x,y)$ of (X_ℓ, Y). The adaptivity of the estimator is then welcomed since the function h_ℓ is *a priori* unknown. Note that in applications, the joint density $\phi(x,y)$ also has to be estimated and hence has an unknown regularity.

When $\alpha < 1/4$ and $\alpha \to 1/4$, the exponent $8\alpha/(4\alpha+1) \to 1$. In the case when $\alpha > 1/4$, we can show from the estimate of Th. 4.4.6 that:

$$\lim_{n\to+\infty} n\mathbb{E}\left[\left(\widehat{V}_\ell - V_\ell - \zeta_n\right)^2\right] = 0, \tag{4.4.33}$$

which yields that $\sqrt{n}\left(\widehat{V}_\ell - V_\ell - \zeta_n\right)$ converges to 0 in L^2. Since $\sqrt{n}\zeta_n$ converges in distribution to $\mathcal{N}\left(0, 4\mathrm{Var}\left(Y_1 h_\ell(G_\ell(X_\ell^1))\right)\right)$ by the central limit theorem, we obtain that:

$$\lim_{n\to+\infty} \sqrt{n}\left(\widehat{V}_\ell - V_\ell\right) = \mathcal{N}\left(0, 4\mathrm{Var}\left(Y_1 h_\ell(G_\ell(X_\ell^1))\right)\right), \tag{4.4.34}$$

in distribution.

4.4.2.3 Numerical illustration on an SIR model

Let us consider an SIR model. The input parameters are the rates λ and γ. The output parameter is the final size of the epidemic, i.e. at a time $T > 0$ where $I_T^N = 0$, $Y = R_T^N$.

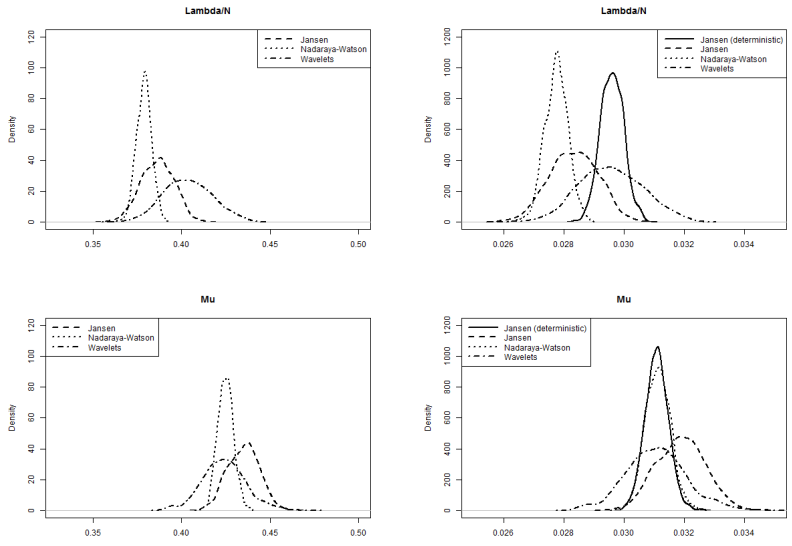
Recall from Chapter 2 that the fractions $(S_t^N/N, I_t^N/N, R_t^N/N)_{t\in[0,T]}$ can be approximated by the unique solution $(s(t), i(t), r(t))_{t\in[0,T]}$ of a system of ODE (see Example **??** of Chapter **??** in Part I of this volume). These limiting equations provide a natural deterministic approximating meta-model (recall [101]) for which sensitivity indices can be computed.

For the numerical experiment, we consider a close population of 1200 individuals, starting with $S_0 = 1190$, $I_0 = 10$ and $R_0 = 0$. The parameters distributions are uniformly distributed with $\lambda/N \in [1/15000, 3/15000]$ and $\gamma \in [1/15, 3/15]$. Here the randomness associated with the Poisson point measures is treated as the nuisance random factor in (4.4.4).
We compute the Jansen estimators of S_λ and S_γ for the deterministic meta-model constituted by the Kermack–McKendrick ODEs of Chapter 2 in Part I of this volume, with $n = 30,000$ simulations. For the estimators of S_λ and S_γ in the SDE, we compute the Jansen estimators with $n = 10,000$ (i.e. $n(p+1) = 30,000$ calls to the function f), and the estimators based on Nadaraya–Watson and on wavelet regressions with $n = 30,000$ simulations.

Let us comment on the results. First, the comparison of the different estimation methods is presented in Fig. 4.4.1. Since the variances in the meta-model and in the stochastic model differ, we start with comparing the distributions of $\mathbb{E}(Y \mid \lambda)$ and $\mathbb{E}(Y \mid \gamma)$ that are centered around the same value, independently of whether the meta-model or the stochastic model is used. These distributions are obtained from 1,000 Monte-carlo simulations. In Fig. 4.4.1(b), taking the meta-model as a benchmark, we see that the wavelet estimator performs well for both λ and γ while Nadaraya–Watson regression estimator performs well only for γ and exhibit biases for λ. Jansen estimator on the stochastic model exhibit biases for both λ and γ.

In a second time, we focus on the estimation of the Sobol indices for the stochastic model. The smoothed distributions of the estimators of S_λ and S_γ, for 1,000 Monte Carlo replications, are presented in Fig. 4.4.1 (a); the means and standard deviations of these distributions are given in Table 4.4.1. Although there is no theoretical values for S_λ and S_γ, we can see (Table 4.4.1) that the estimators of the Sobol indices with non-parametric regressions all give similar estimates in expectation for γ. For λ, the estimators are relatively different, with the Nadaraya–Watson showing the lower estimate. This is linked with the bias seen on Fig. 4.4.1 (b) and discussed below. In term of variance, the Nadaraya–Watson estimator gives the tightest distribution, while the wavelet estimator gives the highest variance.

The advantage of using the estimators with wavelets lies in their robustness to the inclusion of high frequencies and in the fact that they can overcome some smoothing biases that the Nadaraya–Watson regressions exhibit (Fig. 4.4.1 (b)). This can be understood when looking at Fig. 4.4.2: the simulations can give very noisy Y's. For example, extinctions of the epidemics can be seen in very short time in simulations, due to the initial randomness of the trajectories. This produces distributions for Y's that are not unimodal or with peaks at 0, which makes the



(a) (b)

Figure 4.4.1: *Estimations of the first-order Sobol indices, using Jansen estimators on the meta-model with* $n = 10,000$ *and the non-parametric estimations based on Nadaraya–Watson and wavelet regressions. (a): the distributions of the estimators of* S_λ *and* S_γ *is approximated by Monte-carlo simulations. (b): the distributions of* $\mathbb{E}(Y \mid \lambda)$ *and* $\mathbb{E}(Y \mid \gamma)$ *are approximated by Monte Carlo simulations.*

	Jansen	Nadaraya–Watson	Wavelet
$\widehat{S_\lambda}$	0.39	0.38	0.40
s.d.	(9.2e-3)	(4.3e-3)	(1.4e-2)
$\widehat{S_\gamma}$	0.44	0.42	0.42
s.d.	(9.0e-3)	(4.4e-3)	(1.2e-2)

Table 4.4.1: *Estimators of the Sobol indices for* λ *and* γ *and their standard deviations using n =10,000 Monte Carlo replications of the stochastic SIR model.*

estimation of $\mathbb{E}(Y \mid \lambda)$ or $\mathbb{E}(Y \mid \gamma)$ more difficult. The variance of the estimator with wavelets is however the widest and in practice, finding the thresholding constants for the wavelet coefficients can be somewhat tricky when the number of input parameters is large.

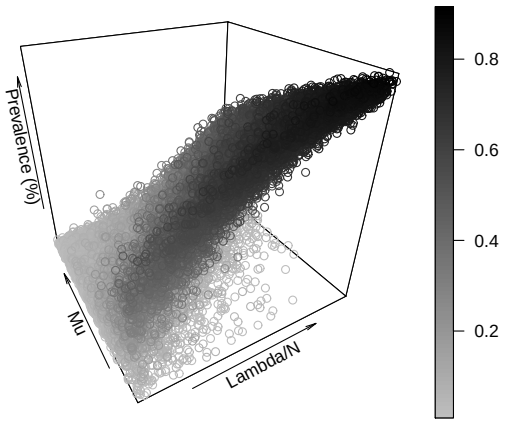


Figure 4.4.2: *Prevalence (Y) simulated from the* $n(p+1) = 30,000$ *simulations of* λ *and* γ*, for the SIR model.*

Appendix

A.1 Some classical results in statistical inference

In this section, we have gathered results on inference useful for this part of these notes.

A.1.1 Heuristics on Maximum Likelihood Methods

As a guide for statistical inference for epidemic dynamics, we first describe the heuristics for getting properties of Maximum likelihood Estimators, each family of statistical models having to be studied specifically (see [23] for more details).

Definitions and properties are given for general discrete time stochastic processes. Consider a sequence $(X_1,\dots,X_n)$ of random variables with values in E, and let P_θ^n denote the distribution of $(X_1,\dots,X_n)$ on $(E^n,\mathscr{E}^n)$. Assume that the parameter set Θ is included in $\mathbb{R}^q$ and that θ_0 the true value of the parameter belongs to $Int(\Theta)$.

The properties on the MLE relies on three basic results that hold as $n\to\infty$ under $\mathbb{P}_{\theta_0}^n$:

(i) a law of large numbers for the log-likelihood $\ell_n(\theta)$,

(ii) a central limit theorem for the score function $\nabla_\theta \ell_n(\theta_0)$

(iii) a law of large numbers for the observed information $\nabla_\theta^2 \ell_n(\theta_0)$ under $\mathbb{P}_{\theta_0}^n$.

For a regular statistical model with a standard rate of convergence $\sqrt{n}$,

(i) For all $\theta\in\Theta$, $n^{-1}\ell_n(\theta)\to J(\theta_0,\theta)$ in $P_{\theta_0}^n$-probability. uniformly w.r.t. θ, $\theta\to J(\theta_0,\theta)$ is a continuous function with a global unique maximum at θ_0.

(ii) $n^{-1/2}\nabla_\theta\ell_n(\theta_0)\to\mathscr{N}(0,\mathscr{I}(\theta_0))$ in distribution under $P_{\theta_0}^n$,

(iii) $-\frac{1}{n}\nabla_\theta^2\ell_n(\theta_0)\to\mathscr{I}(\theta_0)$ in $P_{\theta_0}^n$-probability.

Condition (i) ensures consistency of the MLE $\hat\theta_n$.

Assuming that $\mathscr{I}(\theta_0)$ is non-singular, a Taylor expansion of the score function $\nabla_\theta\ell_n$ at point θ_0 leads, using that $\nabla_\theta\ell_n(\hat\theta_n)=0$,

$$0=\nabla_\theta\ell_n(\hat\theta_n)=\nabla_\theta\ell_n(\theta_0)+\Big(\int_0^1\nabla_\theta^2\ell_n(\theta_0+t(\hat\theta_n-\theta_0))dt\Big)\,(\hat\theta_n-\theta_0). \tag{A.1.1}$$

From this expansion, we get, using that $\mathscr{I}(\theta_0)$ is non-singular,

$$\sqrt{n}(\hat\theta_n-\theta_0)=\Big(-\frac{1}{n}\int_0^1\nabla_\theta^2\ell_n(\theta_0+t(\hat\theta_n-\theta_0))dt\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\nabla_\theta\ell_n(\theta_0)\Big). \tag{A.1.2}$$

Since $\hat\theta_n\to\theta_0$ in $P_{\theta_0}^n$-probability we get, using (iii), that

- the first factor of the r.h.s. of the equation above converges to $\mathscr{I}(\theta_0)^{-1}$ P_{θ_0} a.s.

- the second factor converges in distribution under P_{θ_0} to $\mathcal{N}(0, \mathcal{I}(\theta_0))$.

Finally, Slutsky's Lemma yields that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_{\mathcal{L}} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ under P_{θ_0}.

A.1.2 Miscellaneous results

We first state a theorem concerning the properties of the $\phi(\theta)$.

Theorem A.1.1. *Let (X_n) be a sequence of random variables with values in $\mathbb{R}^p$ and $a_n > 0$ such that $a_n \to \infty$ as $n \to \infty$. Assume that $a_n(X_n - m)$ converges in distribution to a random variable Z. Let $\phi : \mathbb{R}^p \to \mathbb{R}^q$ a continuously differentiable application. Then $a_n(\phi(X_n) - \phi(m))$ converges in distribution to the random variable $\nabla_x \phi(m)Z$, where $\nabla_x \phi$ is the Jacobian matrix of ϕ: $\nabla_x \phi = (\frac{\partial \phi_k}{\partial x_l})_{1 \le k \le q, 1 \le l \le p}$.*

We refer to [124] for the proof.

For sake of clarity, we also give a recap on Exponential families of distributions (see e.g. [11] or [124]). Indeed, among parametric families of distributions, exponential families of distributions, widely used in statistics, provide here a nice framework to study the likelihood.

Let X be a random variable in $\mathbb{R}^k$ (or $\mathbb{Z}^k$) with distribution P_θ and density $p(\theta, x)$, with $\theta \in \Theta$, subset of $\mathbb{R}^q$.

Definition A.1.2. *The family $\{P_\theta, \theta \in \Theta\}$ is an exponential family if there exist q functions $(\eta_1, \dots, \eta_q)$ and ϕ defined on Θ, q real functions $T_1, \dots, T_q$ and a function $h(\cdot)$ defined on $\mathbb{R}^k$ such that*

$$p(\theta, x) = h(x) \exp\{\sum_{j=1}^{q} \eta_j(\theta) T_j(x) - \phi(\theta)\} \; ; x \in \mathbb{R}^k \tag{A.1.3}$$

Then $T(X) = (T_1(X), \dots, T_q(X))$ is a sufficient statistic in the i.i.d. case. The random variable X satisfies

$$m(\theta) := \mathbb{E}_\theta(X) = \nabla_\theta \phi(\theta); \quad \sigma^2(\theta) := \mathrm{Var}_\theta(X) = \nabla^2_\theta \phi(\theta). \tag{A.1.4}$$

A.2 Inference for Markov chains

In order to present a good overview of the statistical problems, we detail the statistical inference for Markov chains. We have rather focus here on parametric inference since epidemic models always include in their dynamics parameters that need to be estimated in order to derive predictions.

A.2.1 Recap on Markov chains

We first begin setting the notations used throughout this chapter and introducing the basic definitions.

Let $(X_n, n \ge 0)$ a Markov chain on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$ with state space $(E, \mathscr{E})$, transition kernel Q and initial distribution μ on $(E, \mathscr{E})$.

The space of observations: $(E^{\mathbb{N}}, \mathscr{E}^{\otimes \mathbb{N}})$. Based on a classical theorem of probability, there exists a unique probability measure on $(E^{\mathbb{N}}, \mathscr{E}^{\otimes \mathbb{N}})$, denoted $P_{\mu,Q}$ such that the coordinate process $(X_n, n \ge 0)$ is a Markov chain (with respect to its natural filtration) with initial distribution μ and transition kernel Q. Then, based on a classical theorem in probability, there exists a unique probability measure on $(E^{\mathbb{N}}, \mathscr{E}^{\otimes \mathbb{N}})$, denoted $P_{\mu,Q}$ such that the coordinate process $(X_n, n \ge 0)$ is a Markov chain (with respect to its natural filtration) with initial distribution μ and transition kernel Q.

The probability $P_{\mu,Q}$ has the property:

- if $A_0, A_1, \dots, A_n$ are measurable sets in E, then

$$P_{\mu,Q}(X_i \in A_i; i = 0, \dots, n) = \int_{A_0} \mu(dx_0) \int_{A_1} Q(x_0, dx_1) \dots \int_{A_n} Q(x_{n-1}, dx_n).$$

Let Θ denote some subset of "probability measures $\times$ transition kernels on $(E,\mathscr{E})$". The canonical statistical model is $(E^{\mathbb{N}},\mathscr{E}^{\mathbb{N}},(P_{\mu,Q},(\mu,Q)\in\Theta))$. Let us denote by $P^n_{\mu,Q}$ the distribution of $(X_0,\dots,X_n)$ on E^{n+1}. The successive observations of (X_i) allow to estimate μ,Q.

Let α be a σ-finite positive measure on $(E,\mathscr{E})$ dominating all the distributions $\{\mu(dy),(Q(x,dy),x\in E)\}$ and assume that $\mu(dy)=\mu(y)\alpha(dy),Q(x,dy)=Q(x,y)\alpha(dy)$. Then, the likelihood of the observations $(x_0,\dots,x_n)$ is the probability density function of $(X_0,\dots,X_n)$, $P^n_{\mu,Q}$, with respect to the measure $\alpha_n=\otimes_{i=0}^n\alpha^i(dy)$ on E^{n+1}, with $\alpha^i(\cdot)$ copies of $\alpha(\cdot)$.

$$\frac{dP^n_{\mu,Q}}{d\alpha_n}(x_i,i=0,\dots,n)=\mu(x_0)Q(x_0,x_1)\dots Q(x_{n-1},x_n).$$

Then, the likelihood function at time n is

$$L_n(\mu,Q)=\frac{dP^n_{\mu,Q}}{d\alpha_n}(X_0,\dots,X_n)=\mu(X_0)Q(X_0,X_1)\dots Q(X_{n-1},X_n). \tag{A.2.1}$$

The associated Loglikelihood is

$$\ell_n(\mu,Q)=\log L_n(\mu,Q). \tag{A.2.2}$$

A.2.1.1 Maximum likelihood method for Markov chains

Let us consider the case of positive recurrent Markov chains. We follow the sketch detailed above to study the properties of MLE estimators.

Assume that the parameter set Θ is is a compact subset of $\mathbb{R}^q$.

Definition A.2.1. *A family $(Q_\theta(x,dy),\theta\in\Theta)$ of transition probability kernels on $(E,\mathscr{E})\to[0,1]$ is dominated by the transition kernel $Q(x,dy)$ if*
$\forall x\in E, Q_\theta(x,dy)=f_\theta(x,y)Q(x,dy)$, with $f_\theta:(E\times E,\mathscr{E}\times\mathscr{E})\to\mathbb{R}^+$ measurable.

Assume that the initial distribution μ is known and let $\mathbb{P}_\theta$ (resp. $\mathbb{Q}$ denote the distribution of the Markov chain (X_n) with initial distribution μ and transition kernel Q_θ (resp. $Q(x,dy)$. Then the likelihood function and loglikelihood write

$$L_n(\theta)=\frac{d\mathbb{P}_\theta}{d\mathbb{Q}}(X_0,\dots,X_n)=\Pi_{i=1}^n f_\theta(X_{i-1},X_i),\quad \ell_n(\theta)=\sum_{i=1}^n\log f_\theta(X_{i-1},X_i). \tag{A.2.3}$$

The maximum likelihood estimator is defined as: $\hat\theta_n=\text{argsup}_{\theta\in\Theta}L_n(\theta)$.

A.2.1.2 Consistency

Denote by θ_0 the true value of the parameter. In order to study the properties of t$\hat\theta_n$ as $n\to\infty$, we introduce some assumptions.

(H0): The family $(Q_\theta(x,dy),\theta\in\Theta)$ is dominated by the transition kernel $Q(x,dy)$.

(H1): The Markov chain (X_n) with transition kernel Q_{θ_0} is irreducible, positive recurrent and aperiodic, with stationary measure $\lambda_{\theta_0}(dx)$ on E.

(H2): $\lambda_{\theta_0}(\{x,Q_\theta(x,\cdot)\neq Q_{\theta_0}(x,\cdot)\})>0$.

(H3): $\forall\theta,\log f_\theta(x,y)$ is integrable with respect to $\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy):=\lambda_{\theta_0}\otimes Q_{\theta_0}$.

(H4): $\forall(x,y)\in E^2,\ \theta\to f_\theta(x,y)$ is continuous w.r.t. θ.

(H5): There exists a function $h(x,y)$ integrable w.r.t. $\lambda_{\theta_0}\otimes Q_{\theta_0}$ and such that

$$\forall\theta\in\Theta,|\log f_\theta(x,y)|\le h(x,y).$$

Assumption (H0) ensures the existence of the likelihood, (H1) is analogous for Markov chains to repetitions in a n sample of i.i.d. random variables, (H2) corresponds to an identifiability assumption, which ensures that different parameter values lead to distinct distributions for the observations. Assumptions (H3)–(H5) are regularity assumptions.

Theorem A.2.2. *Assume (H0)–(H5) and that Θ is a compact subset of $\mathbb{R}^q$. Then the MLE $\hat{\theta}_n$ is consistent: it converges in $\mathbb{P}_{\theta_0}$-probability to θ_0 as $n \to \infty$.*

Proof. Using that, under (H0),(H1), the sequence $(Y_n = (X_{n-1}, X_n), n \geq 1)$ is a positive recurrent Markov chain on $(E \times E, \mathscr{E} \times \mathscr{E})$ with stationary distribution $\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy)$, the ergodic theorem applies to (Y_n) and yields that, under (H3),

$$\frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_{i-1}, X_i) \to J(\theta_0, \theta) := \int\int_{E\times E} \log f_\theta(x,y)\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy) \quad P_{\theta_0}\text{-a.s.} \tag{A.2.4}$$

Rewriting this equation yields that $J(\theta_0, \theta)$ defined in (A.2.4),

$$J(\theta_0, \theta) = \int\int \log \frac{f_\theta(x,y)}{f_{\theta_0}(x,y)} \lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy) + A(\theta_0),$$

with $A(\theta_0) = \int\int \log f_{\theta_0}(x,y)\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy)$. Under (H0),

$$Q_\theta(x,dy) = f_\theta(x,dy)Q(x,dy),$$

so that

$$\begin{aligned} J(\theta_0, \theta) &= \int \lambda_{\theta_0}(dx) \int \log \frac{Q_\theta(x,dy)}{Q_{\theta_0}(x,dy)} Q_{\theta_0}(x,dy) + A(\theta_0) \\ &= -\int K(Q_{\theta_0}(x,\cdot), Q_\theta(x,\cdot))\, \lambda_{\theta_0}(dx) + A(\theta_0), \end{aligned}$$

where $K(P,Q)$ denotes the Kullback–Leibler divergence between two probabilities. Recall that it satisfies

- if $P << Q$, then $K(P,Q) = \mathbb{E}_P(\log \frac{dP}{dQ}) = \int \log \frac{dP}{dQ} dP = E_Q(\phi(\frac{dP}{dQ}))$ with $\phi(x) = x\log(x) + 1 - x$.
- $K(P,Q) = +\infty$ otherwise.

A well-known property is that $K(P,Q) \geq 0$ and $K(P,Q) = 0$ if and only if $P = Q$ a.s. Assumption (H2) ensures that $\theta \to J(\theta_0, \theta)$ possesses a global unique maximum at $\theta = \theta_0$.

The MLE $\hat{\theta}_n$ satisfies that $\hat{\theta}_n = \text{Argsup}_\theta(\frac{1}{n}\ell_n(\theta))$. The maximum of the right-hand side of (A.2.4) is θ_0. Hence to get consistency, we have to prove that "lim Argsup $\frac{1}{n}\ell_n(\theta)$" is equal to "Argsup lim $\frac{1}{n}\ell_n(\theta)$", which is θ_0. Note that, for all $\theta \in \Theta$, $\ell_n(\hat{\theta}_n) \geq \ell_n(\theta)$ and $J(\theta_0, \theta_0) \geq J(\theta_0, \hat{\theta}_n)$. Combining these two inequalities we get,

$$\begin{aligned} 0 \leq J(\theta_0, \theta_0) - J(\theta_0, \hat{\theta}_n) &\leq J(\theta_0, \theta_0) - \frac{1}{n}\ell_n(\theta_0) + \frac{1}{n}\ell_n(\theta_0) - \frac{1}{n}\ell_n(\hat{\theta}_n) \\ &\quad + \frac{1}{n}\ell_n(\hat{\theta}_n) - J(\theta_0, \hat{\theta}_n) \\ &\leq 2 \sup_{\theta\in\Theta} |J(\theta_0, \theta) - \frac{1}{n}\ell_n(\theta)|. \end{aligned}$$

Therefore, by taking Θ a compact subset of $\mathbb{R}^q$, we get that $J(\theta_0, \hat{\theta}_n) \to J(\theta_0, \theta_0)$ $\mathbb{P}_{\theta_0}$- a.s. as $n \to \infty$. Assumptions (H4),(H5) ensure that $J(\theta_0, \cdot)$ is continuous with a unique global maximum at θ_0 so that the MLE converges to θ_0 in $\mathbb{P}_{\theta_0}$-probability. □

A.2.1.3 Limit distribution

This section is based on general results presented in [64]. For V a vector or a matrix, let V^* denote its transposition. Define the $q \times q$ matrix

$$\mathscr{I}(\theta_0) = \int\int \frac{\nabla_\theta f_{\theta_0}(x,y)\, \nabla_\theta^* f_{\theta_0}(x,y)}{f_{\theta_0}(x,y)^2} \lambda_{\theta_0}(dx) Q_{\theta_0}(x,dy). \tag{A.2.5}$$

Let us introduce the additional assumptions.

(H6) $\theta \to \ell_n(\theta)$ is $C^2(\Theta)$ $\mathbb{P}_{\theta_0}$ -a.s.

(H7) $\mathscr{I}(\theta_0)$ defined in (A.2.5) is non-singular.

(H8) $\int \phi_{\theta_0}(r,x,y)\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy) \to 0$ as $r \to 0$ where

$$\phi_{\theta_0}(r,x,y) = \sup\{\| \nabla_\theta^2 \log f_\theta(x,y) - \nabla_\theta^2 \log f_{\theta_0}(x,y) \| \cdot \| \theta - \theta_0 \| \leq r\}.$$

We can state the result on the asymptotic normality of the MLE

Theorem A.2.3. *Assume (H0)–(H8). Then the MLE $\hat{\theta}_n$ is asymptotically Gaussian: under P_{θ_0},*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_{\mathscr{L}} \mathscr{N}_q(0, \mathscr{I}(\theta_0)^{-1}).$$

Proof. Under (H6), the score function is well defined and reads as

$$\nabla_\theta \ell_n(\theta) = \sum_{i=1}^n \nabla_\theta \log f_\theta(X_{i-1}, X_i) = \sum_{i=1}^n v_i(\theta). \tag{A.2.6}$$

The score function satisfies

Proposition A.2.4. *Under assumptions (H0)–(H5), $\nabla_\theta \ell_n(\theta_0)$ is a q-dimensional $\mathbb{P}_{\theta_0}$-martingale w.r.t. $(\mathscr{F}_n)_{n\geq 0}$, which is centered and square integrable.*

Proof: By (A.2.6), we have, $\nabla_\theta \ell_n(\theta_0) = \nabla_\theta \ell_{n-1}(\theta_0) + v_n(\theta_0)$. We get, using that, under (H5), $\int \nabla_\theta f = \nabla_\theta(\int f)$ holds true,

$$\begin{aligned}
\mathbb{E}_{\theta_0}(v_i(\theta_0)|\mathscr{F}_{i-1}) &= \mathbb{E}_{\mathbb{Q}}(\nabla_\theta \log f_{\theta_0}(X_{i-1},X_i) f_{\theta_0}(X_{i-1},X_i)|\mathscr{F}_{i-1}) \\
&= \mathbb{E}_{\mathbb{Q}}(\nabla_\theta f_{\theta_0}(X_{i-1},X_i)|\mathscr{F}_{i-1}) \\
&= \nabla_\theta\, \mathbb{E}_{\mathbb{Q}}(f_{\theta_0}(X_{i-1},X_i)|\mathscr{F}_{i-1}) = \nabla_\theta 1 = 0.
\end{aligned}$$

Noting that $E_{\theta_0}(\nabla_\theta \ell_1(\theta_0)) = \nabla_\theta(E_{\theta_0} 1) = 0$, $\nabla_\theta \ell_n(\theta_0)$ is a centered martingale.

Consider now the increasing process associated with this martingale. We have
$\langle \nabla_\theta \ell_n(\theta_0) \rangle = \sum_{i=1}^n E_{\theta_0}(v_i(\theta_0)\, v_i^*(\theta_0)|\mathscr{F}_{i-1})$.
An application of the ergodic theorem yields $\frac{1}{n}\sum v_i(\theta_0)\, v_i^*(\theta_0) \to \mathscr{I}(\theta_0)$ $\mathbb{P}_{\theta_0}$ a.s.
Therefore for $j = 1, \ldots q$, $E_{\theta_0}\langle \nabla_\theta \ell_n(\theta_0) \rangle_{jj} \to \infty$ as $n \to \infty$. Applying a central limit theorem, we get that

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta_0) \to \mathscr{N}_q(0, \mathscr{I}(\theta_0)).$$

The matrix $\mathscr{I}(\theta_0)$ is the Fisher information matrix .

A Taylor expansion of the score function $\nabla_\theta \ell_n$ at point θ_0 leads, using that $\nabla_\theta \ell_n(\hat{\theta}_n) = 0$, to

$$0 = \frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\hat{\theta}_n) = \frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta_0) + \frac{1}{n}\Big(\int_0^1 \nabla_\theta^2 \ell_n(\theta_0 + t(\hat{\theta}_n - \theta_0)) dt \Big) \frac{\hat{\theta}_n - \theta_0}{\sqrt{n}}. \tag{A.2.7}$$

Now, (A.2.4) yields, using (A.2.1),

$$\frac{1}{n}\nabla^2_\theta \ell_n(\theta_0) \to \int \lambda_{\theta_0}(dx) \int \nabla^2_\theta (\log f_{\theta_0}(x,y)) Q_{\theta_0}(x,dy) = -\mathscr{I}(\theta_0),$$

Indeed, the last equality is obtained using Assumptions (H3)–(H6) and

$$\int \frac{\nabla^2_\theta f_{\theta_0}(x,y)}{f_{\theta_0}(x,y)} Q_{\theta_0}(x,dy) = \nabla^2_\theta \Big(\int f_{\theta_0}(x,y) Q(x,dy) \Big) = 0.$$

Therefore, from expansion (A.2.7), we get,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Big(-\frac{1}{n} \int_0^1 \nabla^2_\theta \ell_n(\theta_0 + t(\hat{\theta}_n - \theta_0)) dt \Big)^{-1} \Big(\frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta_0) \Big). \tag{A.2.8}$$

Since $\hat{\theta}_n \to \theta_0$ in $P^n_{\theta_0}$-probability we get that the first factor of the r.h.s. of (A.2.8) converges to $\mathscr{I}(\theta_0)^{-1}$ under $\mathbb{P}_{\theta_0}$ a.s., and that the second factor converges in distribution under P_{θ_0} to $\mathscr{N}(0, \mathscr{I}(\theta_0))$. Finally, Slutsky's Lemma yields that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to $\mathscr{N}(0, \mathscr{I}(\theta_0)^{-1} \mathscr{I}(\theta_0) \mathscr{I}(\theta_0)^{-1}) = \mathscr{N}(0, \mathscr{I}(\theta_0)^{-1})$ in distribution. □

A.2.2 Other approaches than the likelihood

It often occurs in practice that the likelihood is difficult to compute. One way to overcome this problem relies on stochastic algorithms. However, another way round is to build other processes than the likelihood to derive estimators. These methods include for the i.i.d. case the M-estimators ([124]) and, for stochastic processes, Estimating equations, approximate likelihoods, pseudolikelihoods. ([88]), Generalized Moment Methods ([65]), Contrast functions ([32]).

A.2.2.1 Minimum contrast approaches

What if, instead of the likelihood, another process (contrast process) $U_n(\theta)$ is used as for instance the C.L.S. method (in essence think of $U_n \simeq -\ell_n$))

Let us assume that $U_n(\theta) = U_n(\theta, X_0, \ldots, X_n)$ satisfies

(H1b) For all $\theta \in \Theta$, $U_n(\theta)$ is $\mathscr{F}_n$-measurable and $\theta \to U_n(\theta)$ is under $\mathbb{P}_{\theta_0}$ a.s. continuous and twice continuously differentiable on a subset $V(\theta_0)$.

(H2b) For all θ, $n^{-1}U_n(\theta) \to K(\theta_0, \theta)$ in P_{θ_0}-probability uniformly over compacts subsets of Θ, where $\theta \to K(\theta_0, \theta)$ is continuous with a unique global minimum at θ_0.

(H3b) $n^{-1/2}\nabla_\theta U_n(\theta_0) \to \mathscr{N}_q(0, I_U(\theta_0))$ in distribution under $\mathbb{P}_{\theta_0}$.

(H4b) There exists a symmetric positive matrix $J_U(\theta_0)$ such that

$$\lim_{n\to\infty} \sup_{|\theta - \theta_0| \le \delta} \| \frac{1}{n} \nabla^2_\theta U_n(\theta) - J_U(\theta_0) \| \to 0 \text{ as } \delta \to 0 \quad P_{\theta_0}\text{-a.s.}$$

Define the MCE estimator $\tilde{\theta}_n$ associated with $U_n(\theta)$ as any solution of

$$U_n(\tilde{\theta}_n) = \inf_{\theta \in \Theta} U_n(\theta). \tag{A.2.9}$$

Then, using similar proofs than in Section A.2.1.1 yields that

Theorem A.2.5. *Assume that (H1b)–(H4b) hold. Then, the MCE defined in* (A.2.9)

(1) $\tilde{\theta}_n \to \theta_0$ *in* $P_{\theta_0}-$ *probability.*

(2) $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_{\mathscr{L}} \mathscr{N}_q(0, J_U(\theta_0)^{-1} I_U(\theta_0) J_U^{-1}(\theta_0))$ *under* P_{θ_0}.

Note that contrary to the MLE where $J_U(\theta_0) = I_U(\theta_0)$, the asymptotic covariance matrix of $\tilde{\theta}_n$ is no longer $I_U(\theta_0)^{-1}$. Analytic properties of matrices yield that $J_U(\theta_0)^{-1} I_U(\theta_0) J_U^{-1}(\theta_0))$ is always greater (as a linear form) than $I_U(\theta_0)^{-1}$.

A.2.2.2 Conditional Least Squares

A classical approach associated to this method is the Conditional Least Squares method.
Let (X_n) be an Markov chain on $\mathbb{R}^p$ with transition kernel $Q_\theta(x,dy)$ on $\mathbb{R}^p$ and initial distribution μ. Assume that it is positive recurrent with stationary distribution $\lambda_\theta(dx)$.

Define the two functions

$$g(\theta,x) = \int y Q_\theta(x,dy) \text{ and}$$
$$V(\theta,x) = \int {}^t(y - g(\theta,x))\,(y - g(\theta,x)) Q_\theta(x,dy).$$

Clearly, $E_\theta(X_i|X_{i-1}) = g(\theta,X_{i-1})$ and $\mathrm{Var}_\theta(X_i|X_{i-1}) = V(\theta,X_{i-1})$. We assume
The CLS method is associated with the process

$$U_n(\theta) = \frac{1}{2}\sum_{i=1}^{n} (X_i - E_\theta(X_i|X_{i-1}))^* \,(X_i - E_\theta(X_i|X_{i-1})). \tag{A.2.10}$$

Applying the ergodic theorem to $((X_{i-1},X_i), i \geq 1)$ yields that, under $\mathbb{P}_{\theta_0}$

$$\frac{1}{n} U_n(\theta) \to K(\theta_0,\theta) = \frac{1}{2}\int\int (y - g(\theta,x))^* (y - g(\theta,x)) \lambda_{\theta_0}(dx) Q_{\theta_0}(x,dy) \text{ a.s.}$$

Rewriting this limit yields that

$$K(\theta_0,\theta) = \frac{1}{2}\int\int (g(\theta,x) - g(\theta_0,x))^* (g(\theta,x) - g(\theta_0,x)) \lambda_{\theta_0}(dx) Q_{\theta_0}(x,dy) + A(\theta_0)$$

with

$$A(\theta_0) = \frac{1}{2}\int\int (y - g(\theta,x))^* (y - g(\theta,x)) \lambda_{\theta_0}(dx) Q_{\theta_0}(x,dy).$$

To study the MCE $\tilde{\theta}_n = \mathrm{Argmin}\{U_n(\theta), \theta \in \Theta\}$, we assume

(A1) For all $x \in \mathbb{R}^p$, $g(\theta,x)$ and $V(\theta,x)$ are finite and C^2 with respect to θ.

(A2) $\theta \to K(\theta_0,\theta)$ continuous and $\lambda_{\theta_0}(\{x, g(\theta,x) \neq g(\theta_0,x)\}) > 0$.

(A3) The matrix $J_U(\theta) = \int (\nabla_\theta g(\theta,x)\, \nabla_\theta^* g(\theta,x) \lambda_\theta(dx)$ is non-singular at θ_0.

(A4) The function $\phi(\delta,x) = \sup_{||\theta-\theta_0||\leq\delta} ||\nabla_\theta^2 g(\theta,x) - \nabla_\theta^2 g(\theta_0,x)||$ satisfies

$$\int \phi(\delta,x) \lambda_{\theta_0}(dx) \to 0 \text{ as } \delta \to 0.$$

Assumption (A1) ensures that U_n is well defined, (A2) that $\theta \to K(\theta_0,\theta)$ has a global unique minimum at θ_0. Assumption (A3),(A4) ensure that (H3b), (H4b) hold.

Let us study $\nabla_\theta U_n(\theta)$. We have that

$$0 = \nabla_\theta U_n(\tilde{\theta}_n) = \nabla_\theta U_n(\theta_0) + \Big(\int_0^1 \nabla^2_\theta U_n(\theta_0 + t(\hat{\theta}_n - \theta_0))dt \Big) \, (\hat{\theta}_n - \theta_0). \tag{A.2.11}$$

The first term of the r.h.s. of (A.2.11) reads as

$$\nabla_\theta U_n(\theta_0) = -\sum_{i=1}^n \left(\nabla_\theta g(\theta_0, X_{i-1})\right)^* \, (X_i - g(\theta_0, X_{i-1})).$$

Hence, under (A1), $\nabla_\theta U_n(\theta_0)$ is a centered L^2-martingale under P_{θ_0} with

$$\langle \nabla_\theta U_n(\theta_0) \rangle = \sum_{i=1}^n E_{\theta_0}\Big(\left(\nabla_\theta g(\theta_0, X_{i-1})\right)^* \, V(\theta_0, X_{i-1}) \nabla_\theta g(\theta_0, X_{i-1}) \Big).$$

Applying the ergodic theorem yields

$$\frac{1}{n} \langle \nabla_\theta U_n(\theta_0) \rangle_n \to \int \left(\nabla_\theta g_{\theta_0}(x)^*\right) V(\theta_0, x) \, \nabla_\theta g_{\theta_0}(x) \lambda_{\theta_0}(dx) := I_U(\theta_0) \text{ a.s.}$$

Therefore, we can apply the central limit theorem for martingales (see Theorem A.4.2) and obtain,

$$\frac{1}{\sqrt{n}} \nabla_\theta U_n(\theta_0) \to_{\mathscr{L}} \mathscr{N}_q(0, I_U(\theta_0)) \text{ under } P_{\theta_0}.$$

For the second term, we get $\nabla^2_\theta U_n(\theta_0) = \sum_{i=1}^n \nabla_\theta g(\theta_0, X_{i-1})^* \nabla_\theta g(\theta_0, X_{i-1})$ which satisfies

$$\frac{1}{n} \nabla^2_\theta U_n(\theta_0) \to J_U(\theta_0) := \int \nabla_\theta g(\theta_0, x)^* \, \nabla_\theta g(\theta_0, x) \, \lambda_{\theta_0}(dx) \quad P_{\theta_0} \text{ a.s.}$$

Therefore under (A3), (A4), $J_U(\theta_0)$ is invertible. Therefore, $\tilde{\theta}_n$ is consistent and $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to \mathscr{N}(0, \Sigma(\theta_0))$ with $\Sigma(\theta_0) = J_U^{-1}(\theta_0) I_U(\theta_0) J_U^{-1}(\theta_0)$.

A.2.3 Hidden Markov Models

A Hidden Markov Model is, roughly speaking, a Markov chain observed with noise. This raises new problems for the statistical inference of parameters ruling the Markov chain model (X_n).
Consider a Markov chain $(X_n, n \geq 0))$ with state space E. The term "hidden " corresponds to the situation where the Markov chain cannot be directly observable, Instead of (X_n) , the observations consists in another stochastic process (Y_n) whose distribution is ruled by (X_n). The simplest case is for instance the case of measurements errors $Y_n = X_n + \varepsilon_n$, with (ε_i) i.i.d. random variables. All the statistical inference for (X_n) has to be done in terms of (Y_n) only, since (X_n) cannot be observed.

For epidemic data, this situation occurs when the exact status of individuals cannot be observed or when there is a systematic error in the reporting rate of Infected individuals.

The precise definition of a Hidden Markov Model (HMM) is:

Definition A.2.6. *A Hidden Markov Model (HMM) is a bivariate discrete time process $((X_n, Y_n), n \geq 0)$ with state space $\mathscr{X} \times \mathscr{Y}$ such that*
(i) (X_n) is a Markov chain with state space $\mathscr{X}$.
(ii) For all $i \leq n$, the conditional distribution of Y_i given $(X_0, \ldots, X_n)$ only depends on X_i.

A classical example of Hidden Markov models is obtained as follows:
Let (ε_n) is a sequence of i.i.d. random variables on E and $F(.,.) : \mathscr{X} \times E \to \mathscr{Y}$ a given measurable function. Then, if $Y_n = F(X_n, \varepsilon_n)$, the bivariate sequence (X_n, Y_n) is a Hidden Markov Model.

It follows from this definition that (X_n, Y_n) is a Markov chain on $\mathscr{X} \times \mathscr{Y}$, while the sequence (Y_n) is no longer Markov:
$\mathscr{L}(Y_n|Y_0, \ldots, Y_{n-1})$ effectively depends on all the past observations.

This is why the inference for parameters ruling (X_n) is difficult and rely on specific tools (see e.g. [23], [125]).

A.3 Results for statistics of diffusions processes

Inference for diffusion processes observed on a finite time-interval presents some specific properties. For sake of comprehensiveness, a short recap of classical results for diffusion processes inference is then given.We first present the general framework required for time-dependent diffusions and then detail these results. (see [88] for a presentation of available results).

On a probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t, t \geq 0), \mathbb{P})$, consider the stochastic differential equation

$$d\xi_t = b(t, \xi_t)dt + \sigma(t, \xi_t)dB_t, \xi_0 = \eta. \tag{A.3.1}$$

We assume that (B_t) is a p-dimensional Brownian motion, that b and σ satisfy regularity assumptions which ensure the existence and uniqueness of solutions of (A.3.1) and that η is $\mathscr{F}_0$-measurable and that

We detail results on the inference on parameters in the drift and diffusion coefficient depending on various kinds of observations of $(\xi_t, t \in [0,T])$. For this, let us recall some basic definitions concerning these processes. The state space of $(\xi_t, t \leq T)$ is $C_T = \{x = (x(t)) : [0,T] \to \mathbb{R}^p \text{ continuous}, \mathscr{C}_T\}$, where $\mathscr{C}_T$ denote the Borel filtration associated with the uniform topology. Denote by $X_t : C_T \to \mathbb{R}^p$, $X_t(x) = x(t)$. the coordinate functions defined for $0 \leq t \leq T$. The distribution of $\xi^T : (\xi_t, t \in [0,T])$ on $(C_T, \mathscr{C}_T)$ is denoted by $P^T_{b,\sigma}$.

A.3.1 Continuously observed diffusions on [0,T]

The distributions $P_{b,\sigma} P_{b',\sigma'}$ of two diffusion processes having distinct diffusion coefficients are singular. Therefore, we assume that $\sigma(\cdot) = \sigma'(\cdot)$. From a statistical point of view, this means that $\sigma(\cdot)$ can be identified from the continuous observation of (ξ_t). Consider the parametric model associated to the diffusion (ξ_t) in $\mathbb{R}^p$:

$$d\xi_t = b(\theta, t, \xi_t)dt + \sigma(t, \xi_t)dB_t, \xi_0 = x_0. \tag{A.3.2}$$

Define the diffusion matrix $\Sigma(t,x) = \sigma(t,x)\sigma^*(t,x)$.

Consider the estimation of a q-dimensional parameter $\theta \in \Theta$, with Θ a subset of $\mathbb{R}^q$. Then, under conditions ensuring existence and uniqueness of solutions (see e.g. [83]) and additional assumptions for the Girsanov formula (cf. [68], [97]) on $C([0,T], \mathbb{R}^p), \mathscr{C}_T)$,

$$\begin{aligned} L_T(\theta) &= \frac{dP^T_\theta}{dP^T_0}(X) \\ &= \exp\left[\int_0^T \Sigma^{-1}(t, X_t) b(\theta; t, X_t) dX_t - \frac{1}{2}\int_0^T b^*(\theta; t, X_t)\Sigma^{-1}(t, X_t) b(\theta, t, X_t) dt\right]. \end{aligned} \tag{A.3.3}$$

The statistical model is $(C_T, \mathscr{C}_T, (P^T_{\theta,\sigma}, \theta \in \Theta))$. The loglikelihood is $\ell_T(\theta) = \log L_T(\theta)$. The Maximum Likelihood Estimator is $\hat{\theta}_T$ s.t.

$$\ell_T(\hat{\theta}_T) = \sup\{\ell_T(\theta), \theta \in \Theta\}. \tag{A.3.4}$$

There is no general theory for the properties of the MLE as $T \to \infty$, except in the case of ergodic diffusions.

Consider the case of an autonomous diffusion ξ_t satisfying the stochastic differential equation on $\mathbb{R}^p$:

$$d\xi_t = b(\theta, \xi_t)dt + \sigma(\xi_t)dB_t;\ \xi_0 \simeq \eta.$$

Assume that, for $\theta \in \Theta \in \mathbb{R}^q$, (ξ_t) positive recurrent diffusion process with stationary distribution $\lambda(\theta, x)dx$ on $\mathbb{R}^p$. Then, under assumptions ensuring that the statistical model is regular (see [69] for general results and [94] for ergodic diffusions), then, as $T \to \infty$, the MLE $\hat{\theta}_T$ is consistent and

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \quad \stackrel{\mathscr{L}}{\to} \quad \mathscr{N}_k(0, I^{-1}(\theta_0)) \text{ under } \mathbb{P}_{\theta_0}, \text{ with}$$

$$I(\theta) \quad = \quad I(\theta) = \int_{\mathbb{R}^p} \nabla_\theta b^*(\theta,x)\Sigma^{-1}(x)\nabla_\theta b(\theta,x)\lambda(\theta,x)dx.$$

A.3.2 Discrete observations with sampling Δ on a time interval [0,T]

Consider the stochastic differential equation (A.3.2), where parameters in the drift are α and in the diffusion coefficient β.

$$d\xi_t = b(\alpha,t,\xi_t)dt + \sigma(\beta,t,\xi_t)dB_t, \xi_0 = x_0. \tag{A.3.5}$$

Let $T = n\Delta$ and assume that the observations are obtained at times $(t_i^n = i\Delta; i = 0,\dots n)$.
The space of observations is $((\mathbb{R}^p)^n, (\mathscr{B}(\mathbb{R}^p))^n$. Let $\mathbb{P}^n_{\alpha,\beta}$ denote the distribution of the n-tuple. Contrary to continuous observations, the probabilities $\mathbb{P}^n_{\alpha,\beta}, \mathbb{P}^n_{\alpha',\beta'}$ are absolutely continuous, leading to a likelihood $L_n(\alpha,\beta)$ for the n-tuple. However, it depends on the transition probabilities $\mathbb{P}_\theta(X(t_{i+1}) \in A|X(t_i) = x)$ of the underlying Markov chain. The main difficulty here lies in the intractable likelihood. This is a well known problem for discrete observations of diffusion processes. Alternative approaches based on M-estimators or contrast processes (see [124] for i.i.d. observations, [88] for SDE) have to be investigated.

Several cases can be considered according to T and Δ with $T = n\Delta$.

(a) $T \to \infty$. Results are obtained for ergodic diffusions.

1- Δ fixed: Both parameters in the drift coefficient α and in the diffusion coefficient β can be consistently estimated and ([86]),

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}_n - \alpha_0 \\ \hat{\beta}_n - \beta_0\end{pmatrix} \to \mathscr{N}(0, I_\Delta^{-1}(\alpha_0,\beta_0). \tag{A.3.6}$$

2- $\Delta = \Delta_n \to 0$ and $T = n\Delta_n \to \infty$ as $n \to \infty$. As $n \to \infty$, there is a double asymptotics $\Delta_n \to 0$ and $T = n\Delta_n \to \infty$. Both parameters in the drift coefficient α and in the diffusion coefficient β can be consistently estimated and the following holds (see [86] and [87]

- Parameters in the drift coefficient α are estimated at rate $\sqrt{n\Delta_n}$.
- Parameters in the diffusion coefficient β are estimated at rate $\sqrt{n}$.

$$\begin{pmatrix}\sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\beta}_n - \beta_0)\end{pmatrix} \to \mathscr{N}(0, I^{-1}(\alpha_0,\beta_0).) \tag{A.3.7}$$

(b) $T = n\Delta_n$ fixed and $\Delta = \Delta_n \to 0$ as $n \to \infty$.
It presents the following properties.

- Except for specific models, there is no consistent estimators for parameters in the drift.
- Parameters in the diffusion coefficient can be consistently estimated and satisfy

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathscr{L}} Z = \eta\, U, \text{ with } \eta, U \text{ independent, } U \sim \mathscr{N}(0,I).$$

The random variable Z is not normally distributed but Gaussian but has Mixed variance Gaussian law. It corresponds to a Local Asymptotic Mixed Normal statistical model (see [124], [68] for general references on LAMN; [37], [49] and [58] for diffusion processes).

A.3.3 Inference for diffusions with small diffusion matrix on $[0,T]$

The asymptotic properties of estimators are now studied with respect to the asymptotic framework "$\varepsilon \to 0$". Consider the SDE

$$d\xi_t = b(\alpha,\xi_t)dt + \varepsilon\sigma(\xi_t)dB_t, \xi_0 = x_0.$$

Contrary to the previous section, it is possible to estimate parameters in the drift α.
For continuous observations on $[0,T]$, Kutoyants ([92]) has studied the estimation of α using the likelihood and proved that the MLE is consistent and satisfies

$$\varepsilon^{-1}(\hat{\alpha}_\varepsilon - \alpha_0) \to \mathcal{N}(0, I_b^{-1}(\alpha_0)) \text{ with} \tag{A.3.8}$$
$$I_b(\alpha) = \int_0^T (\nabla_\alpha b)^*(\alpha, z(\alpha,t))\Sigma^{-1}(z(\alpha,t))\nabla_\alpha b(\alpha, z(\alpha,t))dt.$$

The Fisher information of this statistical model is $I_b(\alpha)$.

The statistical inference based on discrete observations of the sample path with sampling interval $\Delta = \Delta_n \to 0$ has first been studied for one-dimensional diffusions with $\sigma \equiv 1$ ([47]), and [119], [57] assuming a parameter β in the diffusion coefficient $\sigma(\beta,x)$. Under assumptions linking the two asymptotics ε and n, [57] proved the existence of consistent and asymptotically Gaussian estimators $(\tilde{\alpha}_{\varepsilon,n}, \tilde{\beta}_{\varepsilon,n})$ of (α_0,β_0), which converge at different rates, parameters in the drift function being estimated at rate ε^{-1} and parameters in the diffusion coefficient at rate $\sqrt{n} = \Delta_n^{-1/2}$.

$$\begin{pmatrix} \varepsilon^{-1}(\hat{\alpha}_{\varepsilon,n} - \alpha_0) \\ \sqrt{n}(\hat{\beta}_{\varepsilon,n} - \beta_0) \end{pmatrix} \underset{n\to\infty,\varepsilon\to 0}{\longrightarrow} \mathcal{N}\left(0, \begin{pmatrix} I_b^{-1}(\alpha_0,\beta_0) & 0 \\ 0 & I_\sigma^{-1}(\alpha_0,\beta_0) \end{pmatrix}\right). \tag{A.3.9}$$

The matrix I_b is the matrix (A.3.8) and the matrix I_σ is

$$I_\sigma(\alpha,\beta)_{ij} = \tag{A.3.10}$$
$$\left(\frac{1}{2T}\int_0^T \mathrm{Tr}(\nabla_{\beta_i}\Sigma(\beta,s,z(\alpha,s))\Sigma^{-1}(\beta,s,z(\alpha,s))\nabla_{\beta_j}\Sigma(\beta,s,z(\alpha,s))ds\right),$$

where $I_b(\alpha_0,\beta_0)$ and $I_\sigma(\alpha_0,\beta_0)$ are assumed invertible.

A.4 Some limit theorems for martingales and triangular arrays

A.4.1 Central limit theorems for martingales

This Central Limit Theorem for martingales in $\mathbb{R}$ is stated in [64].

Let $M_n = \sum_{i=1}^n X_i$ and $\langle M\rangle_n = \sum_{i=1}^n E(X_i^2/\mathcal{F}_{i-1})$. Set $s_n^2 = EM_n^2 = E\langle M\rangle_n$.

Theorem A.4.1. *Assume that the sequence (M_n) of L^2 centered martingales satisfy that, as $n\to\infty$, $s_n^2 \to \infty$ and*

***(H1)**: $\forall \varepsilon > 0$, $\frac{1}{s_n^2}\sum_{i=1}^n E(X_i^2 \mathbf{1}_{|X_i|\geq s_n\varepsilon}|\mathcal{F}_{i-1}) \to 0$ in probability.*

***(H2)**: $\frac{1}{s_n^2}\langle M\rangle_n \to \eta^2$ in probability (η is an r.v. such that, if $\eta^2 < \infty$, $E\eta^2 = 1$).*

Then $(\frac{M_n}{s_n}, \frac{\langle M\rangle_n}{s_n^2}) \to_{\mathcal{L}} (\eta\, N, \eta^2)$ with η, N independent r.v.s, $N \sim \mathcal{N}(0,1)$.

Note that $Z = \eta N$ satisfies $E(\exp(iuZ)) = E(\exp(-u^2\eta^2/2))$.

The Lindberg condition (H1) is often replaced by the stronger assumption:

(H1b): $\exists\, \delta > 0$, $\frac{1}{s_n^{2+\delta}}\sum_{i=1}^n E(|X_i|^{2+\delta}|\mathcal{F}_{i-1}) \to 0$ in probability.

If the dimension of the parameter is q, the score function $\nabla_\theta \ell_n(\theta_0)$ is a $\mathbb{P}_{\theta_0}$-martingale in $\mathbb{R}^q$. So we need theorems for multidimensional martingales in $\mathbb{R}^q$.

Let (M_n) be a sequence of random variables in $\mathbb{R}^q$ with $M_n^* = (M_n^1, \dots, M_n^q)$. Then (M_n) is a $\mathscr{F}_n$-martingale if (M_n^p) is a $\mathscr{F}_n$-martingale for $p = 1, \dots q$.

Assume that (M_n) is a centered L^2-martingale in $\mathbb{R}^q$ and set $X_i = M_i - M_{i-1}$ with $X_i^* = (X_i^1, \dots, X_i^q)$.
Then the increasing process $\langle M \rangle_n$ is the $q \times q$ random matrix defined by $\langle M \rangle_0 = 0$ and $\langle M \rangle_n - \langle M \rangle_{n-1} = E(X_n X_n^* | \mathscr{F}_{n-1}) = \left(E(X_n^p X_n^l | \mathscr{F}_{n-1})\right)_{1 \le p,l \le q}$.
Hence, for $1 \le p, l \le q$, $\langle M \rangle_n^{pl} = \sum_{i=1}^n E(X_i^p \, X_i^l | \mathscr{F}_{i-1})$.

This theorem is derived from a convergence theorem for triangular arrays stated in [74].
For each p, assume that $\mathbb{E}(\langle M_n^p \rangle) = (s_n^p)^2 \to \infty$ and define

$$\zeta_i^{n,p} = \frac{X_i^p}{s_n^p} \quad \text{and } (\zeta_i^n)^* = (\zeta_i^{n,1}, \dots \zeta_i^{n,q}).$$

Theorem A.4.2. *Assume that there exists a positive random matrix Γ such that, as $n \to \infty$,*

(H1): $\sum_{i=1}^n \mathbb{E}(\zeta_i^n (\zeta_i^n)^* | \mathscr{F}_{i-1}) \to \Gamma$ *in probability.*

(H2): *There exists $\delta > 0$, $\sum_{i=1}^n E(\|\zeta_i^n\|^{2+\delta} | \mathscr{F}_{i-1}) \to 0$ in probability.*

Then the following holds

$$\left(\sum_{i=1}^n \zeta_i^n, \sum_{i=1}^n \mathbb{E}(\zeta_i^n (\zeta_i^n)^* | \mathscr{F}_{i-1} \right) \xrightarrow{\mathscr{L}} \left(\Gamma^{1/2} N_q \, , \Gamma \right)$$

with $N_q \sim \mathscr{N}_q(0, I)$ and Γ, N_q independent.

Here again, if $Z = \Gamma^{1/2} N_q$, then, for $u \in \mathbb{R}^q$, $E(\exp(iuZ)) = E(\exp(-\frac{u^* \Gamma u}{2}))$.

A.4.2 Limit theorems for triangular arrays

When dealing with discrete observations with small sampling interval, classical limit theorems for martingales can no longer be used since the σ-algebras $\mathscr{G}_k^n = \sigma(Z(s), s \le k/n)$ do not satisfy the nesting property. We need general theorems for triangular arrays as stated in [74].

A.4.2.1 Recap on triangular arrays

Let $(\Omega, \mathscr{F}, (\mathscr{F}_t, t \ge 0), \mathbb{P})$ be a filtered probability space satisfying the usual conditions. Assume that for each n, there is a strictly increasing sequence $(T(n,k), k \ge 0)$ of finite $(\mathscr{F}_t)$-stopping times with limit $+\infty$ and $T(n,0) = 0$. The stopping rule is defined as

$$N_n(t) = \sup\{k, T(n,k) \le t\} = \sum_{k \ge 1} 1_{T(n,k) \le t}.$$

A q-dimensional triangular array is a double sequence $(\zeta_k^n), n, k \ge 1)$ of q-dimensional variables $\zeta_k^n = (\zeta_k^{n,j})_{1 \le j \le q}$. such that each ζ_k^n is $\mathscr{F}_{T(n,k)}$-measurable.

We consider the behavior of the sums

$$S_t^n = \sum_{k=1}^{N_n(t)} \zeta_k^n.$$

The triangular array is asymptotically negligible (A.N.) if

$$\sum_{k=1}^{N_n(t)} \zeta_k^n \overset{u.c.p.}{\to} 0 \quad \text{i.e. } \sup_{s\le t} \Big| \sum_{k=1}^{N_n(s)} \zeta_k^n \Big| \overset{\mathbb{P}}{\to} 0.$$

In the sequel, we assume that the $T(n,k)$ are non-random and set $\mathscr{G}_k^n = \mathscr{F}_{T(n,k)}$.
The example we have in mind consists in the deterministic times

$$T(n,k) = \inf\{t, [nt] \ge k\Delta\} \Rightarrow N_n(t) = \sup\{k, \ \frac{k\Delta}{n} \le t\}. \tag{A.4.1}$$

Triangular arrays often occur as follows: ζ_k^n may be a function of the increment $Y_{T(n,k)} - Y_{T(n,k-1)}$ for some underlying adapted càdlàg process Y. For discretely observed diffusion processs, we have $\zeta_k^n = X(k\Delta/n) - X((k-1)\Delta/n)$. We first state a lemma proved in [49].

Lemma A.4.3. *Let ζ_k^n, U be random variables with ζ_k^n being $\mathscr{G}_k^n$-measurable. Assume that*

(i) $\sum_{k=1}^n \mathbb{E}(\zeta_k^n | \mathscr{G}_{k-1}^n) \to U$ *in $\mathbb{P}$-probability,*

(ii) $\sum_{k=1}^n \mathbb{E}[(\zeta_k^n)^2 | \mathscr{G}_{k-1}^n)] \to 0$ *in $\mathbb{P}$-probability,*

Then

$$\sum_{k=1}^n \zeta_k^n \to U \quad \text{in } \mathbb{P}\text{-probability}.$$

Corollary A.4.4. *Let ζ_k^n, U be d-dimensional random variables with ζ_k^n being $\mathscr{G}_k^n$-measurable. Assume*

(i) $\sum_{k=1}^n \mathbb{E}(\zeta_k^n | \mathscr{G}_{k-1}^n) \to U$ *in $\mathbb{P}$-probability,*

(ii) $\sum_{k=1}^n \mathbb{E}[\big\|\zeta_k^n\big\|^2 | \mathscr{G}_{k-1}^n)] \to 0$ *in $\mathbb{P}$-probability,*

Then

$$\sum_{k=1}^n \zeta_k^n \to U \quad \text{in } \mathbb{P}\text{-probability}.$$

A.4.2.2 Convergence in law of triangular arrays

Let (ζ_k^n) be a triangular array of d-dimensional random variables such that ζ_k^n is $\mathscr{G}_k^n$-measurable.

Theorem A.4.5. *Assume that (ζ_k^n) satisfy for $N_n(t)$ defined in* (A.4.1)

(i) $\sum_{k=1}^{N_n(t)} \mathbb{E}(\zeta_k^n | \mathscr{G}_{k-1}^n) \overset{u.c.p.}{\to} A_t$ *with A an $\mathbb{R}^d$-valued deterministic function.*

(ii) $\sum_{k=1}^{N_n(t)} \mathbb{E}(\zeta_k^{n,i} \zeta_k^{n,j} | \mathscr{G}_{k-1}^n) - \mathbb{E}(\zeta_k^{n,i} | \mathscr{G}_{k-1}^n)\mathbb{E}(\zeta_k^{n,j} | \mathscr{G}_{k-1}^n) \overset{\mathbb{P}}{\to} C_t^{ij}$ *for $1 \le i,j \le d$ and for all $t \ge 0$, where $C = (C^{ij})$ is a deterministic continuous $\mathscr{M}_{d\times d}^+$-valued function.*

(iii) For some $p > 2$, $\sum_{k=1}^{N_n(t)} \mathbb{E}(\big\|\zeta_k^n\big\|^p | \mathscr{G}_{k-1}^n) \overset{\mathbb{P}}{\to} 0$.

Then, we have

$$\sum_{k=1}^{N_n(t)} \zeta_k^n \overset{\mathscr{L}}{\to} A + Y, \ \textit{w.r.t. the Skorokhod topology,} \tag{A.4.2}$$

where Y is a continuous centered Gaussian process on $\mathbb{R}^d$ with independent increments s.t. $\mathbb{E}(Y_t^i Y_t^j) = C_t^{ij}$.

Remark: If (ii) holds for a single time t, the convergence $\sum_{k=1}^{N_n(t)} \zeta_k^n \overset{\mathscr{L}}{\to} A_t + Y_t$ for this particular t fails in general. There is an exception detailed below (Theorem VII-2-36 of [76]).

Theorem A.4.6. *Assume that for each n, the variables $(\zeta_k^n, k \geq 1)$ are independent and let l_n be integers, or ∞. Assume that, for all $i, j = 1, \ldots, d$ and for some $p > 2$,*

$$\begin{aligned}
&\sum_{k=1}^{l_n} \mathbb{E}(\zeta_k^{n,i}) \xrightarrow{\mathbb{P}} A_i, \\
&\sum_{k=1}^{l_n} \left(\mathbb{E}(\zeta_k^{n,i} \zeta_k^{n,j}) - \mathbb{E}(\zeta_k^{n,i}) \mathbb{E}(\zeta_k^{n,j}) \right) \xrightarrow{\mathbb{P}} C^{ij}, \\
&\sum_{k=1}^{l_n} \mathbb{E}(\|\zeta_k^n\|^p) \xrightarrow{\mathbb{P}} 0,
\end{aligned}$$

where C^{ij} and A_i are deterministic numbers. Then the variables $\sum_{k=1}^{l_n} \zeta_k^n$ converge in distribution to a Gaussian vector with mean $A = (A^i)$ and covariance matrix $C = (C^{ij})$.

A.5 Inference for pure jump processes

In statistical applications, we study likelihood ratios formed by taking Radon–Nikodym derivatives of members of the family of probability measures $(P_\theta, \theta \in \Theta \subset \mathbb{R}^q)$ with respect to one fixed reference distribution.

A.5.1 Girsanov type formula for counting processes

Rather than giving the general expression of the Girsanov formula for semi-martingales (see [76]), we state it first for the case of a counting process on $\mathbb{N}$ and then for multivariate counting processes.

Let X be a stochastic process such that the predictable compensator Λ of X satisfies $\Lambda(t) = \int_0^t \lambda(s)ds)$. assume that, under $\mathbb{P}_\theta$, it is a counting process with intensity $\lambda^\theta(t)$ where $\lambda^\theta(t) > 0$ for all $t > 0$. Denote by $T_1, T_2, \ldots$ the sequence of jump times of X and let $N(t)$ denotes the number of jumps up to time t. Then

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta_0}} | \mathscr{F}_t = \exp\{ \sum_{i=1}^{N(t)} [\log(\lambda^\theta(T_i)) - \log(\lambda^{\theta_0}(T_i))] - \int_0^t [\lambda^\theta(s) - \lambda^{\theta_0}(s)] ds. \} \tag{A.5.1}$$

Consider now multivariate counting processes $N(t) = (N_1(t), \ldots, N_k(t))$. We refer to Jacod's formula (see e.g. Andersen [1, II.7]) for a general expression of two probability measures $\mathbb{P}, \tilde{\mathbb{P}}$ on a filtered probability space under which $\mathbf{N}$ has compensators $\mathbf{\Lambda}$, $\tilde{\mathbf{\Lambda}}$ respectively. Usually, we will have continuous or absolutely continuous compensators with intensities $\lambda_l(t), \tilde{\lambda}_l(t)$. Since no jumps can occur simultaneously, the sequence of jump times T_i is well defined, together with the mark $J_i \in \{1, \ldots, k\}$ ($J_i = l$ if the jump T_i occurs in N_l ($\Delta N_l(T_i) = 1$). The process $N_\cdot(t) = \sum_{l=1}^k N_l(t)$ is a counting process with compensator $\Lambda_\cdot(t) = \sum_{l=1}^k \Lambda_l(t)$. Assume $\tilde{\mathbb{P}}$ is absolutely continuous with respect to $\mathbb{P}$ (written $\tilde{\mathbb{P}} << \mathbb{P}$).

Theorem A.5.1. *Assume that $\tilde{\mathbb{P}} << \mathbb{P}$. Then*

$$\tilde{\Lambda}_l << \Lambda_l \text{ for all } l = 1, \ldots, k, \quad P\text{- a.s.}$$

$$\Delta\Lambda_\cdot(t) = 1 \text{ for any time } t \text{ implies } \Delta\tilde{\Lambda}_\cdot(t) = 1, \quad P\text{-a.s.}$$

$$\begin{aligned}
\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} | \mathscr{F}_t &= \frac{d\tilde{P}}{dP} | \mathscr{F}_0 \, \frac{\prod_{l=1}^k \prod_{s \leq t} \tilde{\lambda}_l(t)^{\Delta N_l(t)} \exp(-\int_0^t \tilde{\lambda}_\cdot(s)ds)}{\prod_{l=1}^k \prod_{s \leq t} \lambda_l(t)^{\Delta N_l(t)} \exp(-\int_0^t \lambda_\cdot(s)ds)} \\
&= \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} | \mathscr{F}_0 \, \exp\left\{ \sum_{l=1}^k \sum_{i=1}^{N(t)} [\log \tilde{\lambda}_l(T_i) - \log \lambda_l(T_i)] \Delta N_l(T_i) - \sum_{l=1}^k \int_0^t [\tilde{\lambda}_l(s) - \lambda_l(s)] ds \right\}.
\end{aligned}$$

Note that the products in the above formula are just $\prod_n \tilde{\lambda}_{J_n}(T_n)$, $\prod_n \lambda_{J_n}(T_n)$.

A.5.2 Likelihood for Markov pure jump processes

Let us consider a pure jump process with state space $E = \{0,\dots,N\}$ and Q-matrix $\mathbf{Q} = (q_{ij})$ observed up to time T. The likelihood is

$$L_T(\mathbf{Q}) = \prod_{i=0}^{N}\prod_{j\neq i} q_{ij}^{N_{ij}(T)} \exp(-q_{ij}N_i(T)), \tag{A.5.2}$$

where the process $N_{ij}(t)$ counts the number of transitions from state i to state j on the time interval $[0,t]$ and $N_i(t)$ is the time spent in state i before time t:

$$N_i(t) = \int_0^t \delta_{\{X(s)=i\}} ds.$$

We refer to [73] for a complete study of Marked point processes.
This yields that the maximum likelihood estimator of $\mathbf{Q}$ is

$$\hat{q}_{ij}(T) = \frac{N_{ij}(T)}{N_i(T)}, \quad \text{for } j \neq i \quad \text{and } N_i(T) > 0. \tag{A.5.3}$$

If $N_T(i) = 0$, the process has not been in state i: there is no information about q_{ij} in the observations and the MLE of q_{ij} does not exist. As for Markov chains with countable state space, $\hat{q}_{ij}(T)$ is the empirical estimate of q_{ij}.

A.5.3 Martingale properties of likelihood processes

In statistical applications, we want to consider a whole family of probability measures $\mathbb{P}$, not necessarily mutually absolutely continuous and therefore cannot apply the above theorem to obtain $\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}|_{\mathscr{F}_t}$ for each $\tilde{\mathbb{P}},\mathbb{P}$ considered. However, for any two probability measures $\tilde{\mathbb{P}},\mathbb{P}$, the measure $\mathbf{Q} = \frac{1}{2}(\tilde{\mathbb{P}}+\mathbb{P})$ dominates both $\tilde{\mathbb{P}}$ and $\mathbb{P}$. We can therefore calculate $d\mathbb{P}/d\mathbf{Q}$ and $d\tilde{\mathbb{P}}/d\mathbf{Q}$ and finally set,

$$\begin{aligned}
\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} &= \frac{d\tilde{\mathbb{P}}}{d\mathbf{Q}}\Big/\frac{d\mathbb{P}}{d\mathbf{Q}} \text{ where } \frac{d\mathbb{P}}{d\mathbf{Q}} > 0,\\
\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} &= \infty \text{ where } \frac{d\mathbb{P}}{d\mathbf{Q}} = 0.
\end{aligned}$$

Suppose now that we have a statistical model $(\mathbb{P}_\theta, \theta\in\Theta)$ for some subset $\Theta \in \mathbb{R}^q$. Suppose that all $\mathbb{P}_\theta$ are dominated by a fixed probability measure $\mathbf{Q}$. For simplicity, we assume that all the $\mathbb{P}_\theta$'s coincide on $\mathscr{F}_0$ and consider only the absolute continuous case:
under $\mathbb{P}_\theta$, $\mathbf{N} = (N_1,\dots,N_k)$ has compensator $\mathbf{\Lambda}^\theta = (\int \lambda_l^\theta), l = 1,\dots,k)$ for certain intensity process λ^θ. We consider the likelihood function as depending on both $t\in\mathbb{R}^+$ and $\theta\in\Theta$. Dropping the denominator in Theorem A.5.1 (which does not depend on θ), we have that the likelihood at time t as a function of θ is proportional to

$$\begin{aligned}
L(\theta,t) &= \exp(-\sum_{l=1}^{k}\int_0^t \lambda_l^\theta(s)ds)\prod_{T_n\le t}\lambda_{J_n}^\theta(T_n),\\
&= \exp\{\sum_{l=1}^{k}\int_0^t [\log\lambda_l^\theta(s)dN_l(s) - \lambda_l^\theta(s)ds]\}.
\end{aligned}$$

Remark A.5.2. *This is another expression of the general Girsanov formula given in the appendix of Part 1 of these notes.*

The likelihood process $L(\theta,t)$ is a $(\mathbf{Q},(\mathscr{F}_t))$-martingale. Indeed, let Y a $\mathscr{F}_s$ measurable random variable. We have $E_\mathbf{Q}(YL(\theta,t)) = E_\mathbf{Q}(Y\frac{d\mathbb{P}_\theta}{dQ}) = \mathbb{E}_\theta(Y) = E_\mathbf{Q}(YL(\theta,s))$ since $Y\in\mathscr{F}_s$.
Hence $E_\mathbf{Q}(L(\theta,t)|\mathscr{F}_s) = L(\theta,s)$.

Consider now the log-likelihood

$$\log L(\theta,t) = \sum_{l=1}^{k}\int_0^t (\log\lambda_l^\theta(s)dN_l(s) - \lambda_l^\theta(s)ds). \tag{A.5.4}$$

The score process is defined as $\nabla_\theta \log L(\theta,t)$. Assuming that differentiation may be taken under the integral sign, we get

$$\begin{aligned}\nabla_{\theta_j} \log L(\theta,t) &= \frac{\partial}{\partial \theta_j} \log L(\theta,t) \\ &= \sum_{l=1}^{k} \int_0^t \nabla_{\theta_j} \log \lambda_l^\theta(s)\ (dN_l(s) - \lambda_l^\theta(s)ds),\ j = 1,\ldots,q.\end{aligned} \tag{A.5.5}$$

Hence the score process is a $(P_\theta, (\mathscr{F}_t))$- local martingale in $\mathbb{R}^q$. It is a centered L^2-martingale with associated predictable $q \times q$ matrix variation process

$$\langle \nabla_\theta \log L(\theta;\cdot)\rangle_{r,j} = \sum_{l=1}^{k} \int_0^t \nabla_{\theta_r} \log \lambda_l^\theta(s) \nabla_{\theta_j} \log \lambda_l^\theta(s)\ \lambda_l^\theta(s) ds. \tag{A.5.6}$$

The "observed information" at θ is obtained by differentiating again with respect to θ. If differentiation can be taken under the integral sign, we get

$$\begin{aligned}\nabla^2_{\theta_r\theta_j} \log L(\theta;t) = &\sum_{l=1}^{k} \int_0^t \nabla^2_{\theta_r\theta_j j} \lambda_l^\theta(s)(dN_l(s) - \lambda_l(s)ds) \\ &- \int_0^t \nabla_{\theta_r} \log \lambda_l^\theta(s) \nabla_{\theta_j} \log \lambda_l^\theta(s) \lambda_l^\theta(s) ds.\end{aligned} \tag{A.5.7}$$

Using (A.5.6) yields that the compensator of the process $-\nabla^2 \log L(\theta;\cdot)$ is $\langle \nabla_\theta \log L(\theta;\cdot)\rangle$. This is a version of a well-known result: the variance matrix of the score coincides with the expected information matrix.

References for Part IV

[1] P.K. Andersen, O. Borgan, R.D. Gill and N. Keiding, *Statistical Models Based on Counting Processes*, Springer Series in Statistics. Springer, New York, 1993.

[2] H. Andersson and T. Britton, *Stochastic Epidemic Models and their Statistical Analysis*, Lecture Notes in Statistics Series. Springer, 2000.

[3] C. Andrieu, A. Doucet and R. Holenstein, Particle Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society: Series B*, 72:269–342, 2010.

[4] H. De Arazoza, J. Joanes, R. Lounes, C. Legeai, S. Clémencon, J. Perez and B. Auvert, The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low prevalence, *BMC Infectious Disease*, 7:130, 2007.

[5] S. Arlot and P. Massart, Data-driven calibration of penalties for least-squares regression, *Journal of Machine Learning Research*, 10:245–279, 2009.

[6] K.B. Athreya and P. Ney, *Branching Processes*, Springer Series in Probability. Springer, 1972.

[7] R. Azencott, Formule de Taylor stochastique et développement asymptotique integrales de Feynmann, *Séminaire de Probabilités XVI*, pages 237–285, 1982.

[8] F. Ball, The threshold behaviour of epidemic models, *Journal of Applied Probability*, 20(2):227–241, 1983.

[9] M.A. Beaumont, J.M. Marin, J.M. Cornuet and C.P. Roberts, Adaptive Approximate Bayesian Computation, *Biometrika*, 96(4):983–990, 2009.

[10] M.A. Beaumont, W. Zhang and D.J. Balding, Approximate Bayesian Computation in population genetics, *Genetics*, 162:2025–2035, 2002.

[11] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I*, Chapman and Hall/CRC, 2nd edition, 2015.

[12] O.N. Bjørnstad, B.F. Finkenstadt and B.T. Grenfell, Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model, *Ecological Monographs*, 72(2):169–184, 2002.

[13] M. Bladt, B. Meini, M.F. Neuts and B. Sericola, Distributions of reward functions on continuous-time Markov chains, *Matrix-Analytic Methods: Theory and Applications*, pages 39–62, 2002.

[14] M. Bladt and M. Sørensen, Statistical inference for discretely observed Markov jump processes, *Journal of the Royal Statistical Society, Series B*, 67(3):395–410, 2005.

[15] M.G. Blum, Approximate Bayesian Computation: a non-parametric perspective, *Journal of the American Statistical Association*, 105:1178–1187, 2010.

[16] M.G. Blum and O. Francois, Non-linear regression models for Approximate Bayesian Computation, *Statistics and Computing*, 20:63–73, 2010.

[17] M.G. Blum and V.C. Tran, HIV with contact-tracing: a case study in Approximate Bayesian Computation, *Biostatistics*, 11(4):644–660, 2010.

[18] C. Bretó, E.L. He, D. Ionides and A.A. King, Time series analysis via mechanistic models, *Annals of Applied Statistics*, 3(1):319–348, 2009.

[19] T. Britton and F. Giardina, Introduction to statistical inference for infectious diseases. *Journal de la Société Francaise de Statistiques*, 157(1):53–70, 2016.

[20] F. Ball, T. Britton, C. Larédo, E. Pardoux, D. Sirl and V.C. Tran, *Stochastic Epidemic Models with Inference*, T. Britton and E. Pardoux eds., Lecture Notes in Mathematics, Mathematical Biosciences, Vol. 2255, 2019.

[21] A. Camacho, A. Kucharski, Y. Aki-Sawyerr, M.A. White, S. Flasche, M. Baguelin, T. Pollington, J.R. Carney, R. Glover, E. Smout, A. Tiffany, W.J. Edmunds, and S. Funk, Temporal changes in ebola transmission in sierra leone and implications for control requirements: a real-time modelling study, *PLOS Currents Outbreaks*, Edition 1(10.1371), 2015.

[22] Y. Cao, D.T. Gillespie and L.R. Petzold, Avoiding negative populations in explicit Poisson tau-leaping. *Journal of Chemical Physics*, 123:054–104, 2005.

[23] O. Cappé, E. Moulines and T. Ryden, *Inference in Hidden Markov Models*, Springer Series in Statistics. Springer, 2005.

[24] G. Castellan, *Sélection d'histogrammes ou de modèles exponentiels de polynômes par morceaux à l'aide d'un critère de type Akaike*, PhD thesis, Université d'Orsay, 2000.

[25] G. Castellan, A. Cousien and V.C. Tran, Nonparametric adaptive estimation of order 1 Sobol indices in stochastic models, with an application to epidemiology, Submitted, 2017.

[26] S. Cauchemez, F. Carrat, C. Viboud, A.J. Valleron and P.Y. Boelle, A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data, *Statistics in Medicine*, 23(22):3469–3487, 2004.

[27] S. Cauchemez and N.M. Ferguson, Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London, *Journal of The Royal Society Interface*, 5(25):885–897, 2008.

[28] S. Cauchemez and N.M. Ferguson, Methods to infer risk factors in complex outbreak data, *Journal of The Royal Society Interface*, 9(68):456–469, 2012.

[29] G. Chagny, Penalization versus Goldenshluger-lepski strategies in warped bases regression, *ESAIM: P&S*, 17:328–358, 2013.

[30] A. Chatzilena, E. van Leeuwen, O. Ratmann, M. Baguelin and N. Demiris, Contemporary statistical inference for infectious disease models using Stan, *arXiv: 1903.00423*, 2019.

[31] S. Clémencon, V.C. Tran and H. De Arazoza, A stochastic SIR model with contact-tracing: large population limits and statistical inference, *Journal of Biological Dynamics*, 2(4):391–414, 2008.

[32] D. Dacunha-Castelle and M. Duflo, *Probabilités et Statistiques: 2. Problèmes à Temps Mobile*, Masson, Paris, 1993.

[33] D.J. Daley and J. Gani, *Epidemic Modelling: an Introduction*, Cambridge University Press, 2001.

[34] B. Delyon, M. Lavielle and E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, *The Annals of Statistics*, 27(1):94–128, 1999.

[35] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[36] O. Diekmann, H. Heesterbeek and T. Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics*, Princeton University Press, 2013.

[37] G. Dohnal, On estimating the diffusion coefficient, *Journal of Applied Probability*, 24:105–114, 1987.

[38] A. Doucet, N. De Freitas and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[39] C.C. Drovandi and A.N. Pettitt, Using approximate bayesian computation to estimate transmission rates of nosocomial pathogens, *Statistical Communications in Infectious Diseases*, 3(1):online, 2011.

[40] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, 2nd edition, 2005.

[41] J. Fan, Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, 87(420):998–1004, 1992.

[42] P. Fearnhead, O. Papaspiliopoulos and G.O. Roberts, Particle filters for partially observed diffusions, *Journal of the Royal Statistical Society: Series B*, 70(4):755–777, 2008.

[43] P. Fearnhead and D. Prangle, Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC, *Journal of the Royal Statistical Society*, 74(3):419–474, 2012.

[44] J.C. Fort, T. Klein, A. Lagnoux and B. Laurent, Estimation of the Sobol indices in a linear functional multidimensional model, *Journal of Statistical Planning and Inference*, 143(9):1590–1605, 2013.

[45] M.I. Freidlin and A.D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer, 1978.

[46] C. Fuchs, *Inference for diffusion processes*, Springer, 2013.

[47] V. Genon-Catalot, Maximum contrast estimation for diffusion processes from discrete observations, *Statistics*, 21(1):99–116, 1990.

[48] V. Genon-Catalot, *Cours de Statistique des diffusions*, Preprint, 2018.

[49] V. Genon-Catalot and J. Jacod, On the estimation of the diffusion coefficient for multi-dimensional diffusion processes, *Annales de l'I.H.P. Probabilités et statistiques*, 29(1):119–151, 1993.

[50] V. Genon-Catalot, T. Jeantheau and C. Larédo, Stochastic volatility models as hidden markov models and statistical applications, *Bernoulli*, 6(6):105, 2000.

[51] V. Genon-Catalot and C. Larédo, Leroux's method for general Hidden Markov Models, *Stochastic Processes and their Applications*, 116(2):222–243, 2006.

[52] W.R. Gilks, S. Richardson and D.J. Spieghalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[53] W.R. Gilks and G.O. Roberts, Strategies for improving mcmc, In W.R. Gilks, S. Richardson and D.J. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pages 89–114, Boca Raton, Chapman and Hall/CRC, 1996.

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

[55] A. Gloter. Parameter estimation for a discrete sampling of an integrated Ornstein–Uhlenbeck process, *Statistics*, 35(3):225–243, 2001.

[56] A. Gloter, Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient, *ESAIM: Probability and Statistics*, 4:205–227, 2010.

[57] A. Gloter and M. Sørensen, Estimation for stochastic differential equations with a small diffusion coefficient, *Stochastic Processes and their Applications*, 119(3):679–699, 2009.

[58] E. Gobet, Local asymptotic normality property for elliptic diffusion: a malliavin calculus approach, *Bernoulli*, 7(6):899–912, 2001.

[59] M. Greenwood, On the statistical measure of infectiousness, *J. Hyg.*, 31:336–351, 1931.

[60] P. Guttorp, *Statistical Inference for Branching Processes*, Wiley Series in Probability and Mathematical Statistics. Wiley, 1991.

[61] R. Guy, C. Larédo and E. Vergu, Parametric inference for discretely observed multidimensional diffusions with small diffusion coefficient, *Stochastic Processes and their Applications*, 124:51–80, 2014.

[62] R. Guy, C. Larédo and E. Vergu, Approximation of epidemic models by diffusion processes and their statistical inference, *Journal of Mathematical Biology*, 70:621–646, 2015.

[63] R. Guy, C. Larédo and E. Vergu, Approximation and inference of epidemic dynamics by diffusion processes, *Journal de la Société Francaise de Statistiques*, 157(1):71–100, 2016.

[64] P. Hall and C.C. Heyde, *Martingale Limit Theory and its Application*, Probability and Mathematical Statistics. Academic Press, 1980.

[65] L.P. Hansen and J.A. Scheinkman, Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes, *Econometrica*, 63:767–804, 1995.

[66] W. Härdle, G. Kerkyacharian, D. Picard and A. Tsybakov, *Wavelets, Approximation and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*, Springer, New York, 1987.

[67] M. Høhle, E. Jørgensen and P.D. O'Neill, Inference in disease transmission experiments by using stochastic epidemic models, *Journal of the Royal Statistical Society: Series C*, 54(2):349–366, 2005.

[68] R. Hopfner, *Asymptotic Statistics with a View to Stochastic Processes*, Graduate. Walter de Gruyter GmbH, Berlin/Boston, 2014.

[69] I.A. Ibragimov and R.Z. Has'minskii, *Statistical Estimation. Asymptotic Theory*, Applications of Mathematics. Springer, 1981.

[70] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, volume 24, North-Holland Publishing Company, 1989, Second Edition.

[71] E.L. Ionides, A. Bhadra and A. King, Iterated filtering, *Annals of Statistics*, 39:1776–1802, 2011.

[72] E.L. Ionides, C. Bretó and A.A. King, Inference for nonlinear dynamical systems, *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

[73] M. Jacobsen, *Marked Point and Piecewise Deterministic Processes*, Applied Mathematics. Probability and Its Applications. Birkhäuser, Berlin, 2006.

[74] J. Jacod and P. Protter, *Discretization of Processes*, volume 67 of *Stochastic Modelling and Applied Probability*, Springer, 2012.

[75] J. Jacod and A.N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, 1987.

[76] J. Jacod and A.N. Shiryaev, *Limit Theorems for Stochastic Processes*, volume 288 of *A series of Comprehensive Studies in Mathematics*, Springer, 2003.

[77] J. Jacques, *Contributions à l'analyse de sensibilité et à l'analyse discriminante*, PhD thesis, Université Joseph Fourier, Grenoble, 12, 2005.

[78] J. Jacques, Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique, *IRMA Lille*, 71(III), 2011.

[79] P. Jagers, *Branching Processes with Biological Applications*, Wiley, 1975.

[80] M.R. James and F. Le Gland, Consistent parameter estimation for partially observed diffusions with small noise, *Applied Mathematics and Optimization*, 32(1):47–72, 1995.

[81] A. Janon, M. Nodet and C. Prieur, Uncertainties assessment in global sensitivity indices estimation from metamodels, *International Journal for Uncertainty Quantification*, 4(1):21–36, 2014.

[82] M.J.W. Jansen, Analysis of variance designs for model output, *Computer Physics Communications*, 117:35–43, 1999.

[83] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, volume Second Edition of *Graduate Texts in Mathematics*, Springer, 2000.

[84] M.J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press, 2011.

[85] G. Kerkyacharian and D. Picard, Regression in random design and warped wavelets, *Bernoulli*, 10(6):1053–1105, 2004.

[86] M. Kessler, Estimation of an ergodic diffusion from discrete observations, *Scandinavian Journal of Statistics*, 24:221–229, 1997.

[87] M. Kessler, Simple and explicit estimating functions for a discretely observed diffusion process, *Scandinavian Journal of Statistics*, 27(1):65–82, 2000.

[88] M. Kessler, A. Lindner and M. Sørensen, *Statistical Methods for Stochastic Differential Equations*, volume 124 of *Monographs on Statistics and Applied Probability*, CRC Press, 2012.

[89] A.A. King, E.L. Ionides, M. Pascual and M.J. Bouma, Inapparent infections and cholera dynamics, *Nature*, 454(7206):877–880, 2008.

[90] A.A. King, D. Nguyen and E.L. Ionides, Statistical inference for partially observed markov processes via the r package pomp, *Journal of Statistical Software*, 69(12):1–43, 2016.

[91] E. Kuhn and M. Lavielle, Coupling a stochastic approximation version of em with an mcmc procedure, *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[92] Y.A. Kutoyants, *Parameter Estimation for Stochastic Processes*, volume 6 of *Research and Exposition in Mathematics*, Heldermann, 1984.

[93] Y.A. Kutoyants, *Identification of Dynamical Systems with Small Noise*, Springer, 1994.

[94] Y.A. Kutoyants, *Statistical Inference for Ergodic Diffusion Processes*, Springer, 2004.

[95] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection, *The Annals of Statistics*, 28(5):1302–1338, 2000.

[96] Y.N. Linkov, *Asymptotic Statistical Methods for Stochastic Processes*, volume 196 of *Translations of Mathematical Monographs*, American Mathematical Society, Providence, 2001, translated from the 1993 Russian original by V. Kotov.

[97] R.S. Liptser and A.N. Shiryaev, *Statistics of Random Processes. I. General Theory*, volume Second Edition, Springer, 2001.

[98] R.S. Liptser and A.N. Shiryaev, *Statistics of Random Processes .II. Applications*, volume Second Edition, Springer, 2001.

[99] J.-M. Loubes, C. Marteau and M. Solís, Rates of convergence in conditional covariance matrix estimation, `ArXiv:1310.8244`, 2014.

[100] J.-M. Marin, P. Pudlo, C.P. Robert and R. Ryder, Approximate Bayesian computation methods, *Statistics and Computing*, 22(6):1167–1180, 2012.

[101] A. Marrel, B. Iooss, S. Da Veiga and M. Ribatet, Global sensitivity analysis of stochastic computer models with joint metamodels, *Statistics and Computing*, 22(3):833–847, 2012.

[102] T.J. McKinley, A.R. Cook and R. Deardon, Inference in epidemic models without likelihoods, *International Journal of Biostatistics*, 5(1), 2009.

[103] T.J. McKinley, J.V. Ross, R. Deardon and A.R. Cook, Simulation-based Bayesian inference for epidemic models, *Computational Statistics & Data Analysis*, 71:434–447, 2014.

[104] S. Méléard, *Modèles aléatoires en Ecologie et Evolution*, Springer, 2016.

[105] P.D. O'Neill, A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov Chain Monte Carlo methods , *Mathematical Biosciences*, 180:103–114, 2002.

[106] P.D. O'Neill, Introduction and snapshot review: Relating infectious disease transmission models to data, *Statistics in medicine*, 29(20):2069–2077, 2010.

[107] P.D. O'Neill and G.O. Roberts, Bayesian inference for partially observed stochastic epidemics, *Journal of the Royal Statistical Society: Series A* 162:121–129, 1999.

[108] H. Pohjanpalo, System identifiability based on the power series expansion of the solution, *Mathematical Biosciences*, 41(1-2):21–33, September 1978.

[109] W.N. Rida, Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model, *Journal of the Royal Statistical Society. Series B*, 53(1):269–283, 1991.

[110] J.V. Ross, D.E. Pagendam and P.K. Polett, On parameter estimation in population models II: Multi-dimensional processes and transient dynamics, *Theoretical Population Biology*, 75(2-3):123–132, 2009.

[111] A. Saltelli, K. Chan and E.M. Scott, *Sensitivity analysis*, Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2000.

[112] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola, *Global sensitivity analysis*, John Wiley & Sons, Chichester, 2008.

[113] A. Sedoglavic, A probabilistic algorithm to test local algebraic observability in polynomial time, *Journal of Symbolic Computation*, 33(5):735–755, 2002.

[114] S.A. Sisson, M.A. Beaumont and Y. Fan, editors, *Handbook fo Approximate Bayesian Computation*, Handbooks of Modern Statistical Methods. Chapman & Hall/CRC Press, 2018.

[115] S.A. Sisson, Y. Fan and M. Beaumont, *Handbook of Approximate Bayesian Computation*, Handbooks of Modern Statistical Methods. Chapman and Hall /CRC, 2018.

[116] S.A. Sisson, Y. Fan and M. Tanaka, Sequential Monte Carlo without likelihoods, *Proc. Nat. Acad. Sci. USA*, 104:1760–1765, 2007.

[117] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models, *Math. Modeling Comput. Experiment*, 1(4):407–414, 1993.

[118] M. Solís, *Conditional covariance estimation for dimension reduction and sensivity analysis*, Phd thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, Toulouse, France, 2014.

[119] M. Sørensen and M. Uchida, Small-diffusion asymptotics for discretely sampled stochastic differential equations, *Bernoulli*, 9(6):1051–1069, 2003.

[120] I. Chis Ster, B.K. Singh and N.M. Ferguson, Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain, *Epidemics*, 1:21–34, 2009.

[121] T. Toni, D. Welch, N. Strelkowa, A. Ipsen and M.P. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of The Royal Society Interface*, 6:187–202, 2009.

[122] A.B. Tsybakov, *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications*, Springer, 2004.

[123] F. Vaida, Parameter convergence for EM and MM algorithms, *Statistica Sinica*, 15:831–840, 2005.

[124] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[125] R. van Handel, *Lecture Notes on Hidden Markov Models*, Princeton University, 2008.

[126] S. Da Veiga and F. Gamboa, Efficient estimation of sensitivity indices, *Journal of Nonparametric Statistics*, 25(3):573–595, 2013.

[127] C.F.J. Wu, On the convergence properties of the EM algorithm, *Annals of Statistics*, 11(1):95–103, 1983.