Approximation of epidemic models by diffusion processes and their statistical inference

Catherine Larédo [*]

Laboratoire MIA, I.N.R.A. and LPMA, Université Denis Diderot, CNRS-UMR 7599.

Chapter 1: Introduction, first examples and recap on parametric inference for Markov chains

[*] *Email address*: catherine.laredo@jouy.inra.fr (Catherine Larédo)

1 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.

1.1 General set-up

Multidimensional continuous-time Markov jump processes $(Z(t))$ on $\mathbb{Z}^p$ form a usual set-up for modeling SIR-like epidemics. However, when facing incomplete epidemic data, inference based on $(Z(t))$ is not easy to be achieved.

There are different type of situations where missing data are present. One situation concerns Hidden Markov Models, that is, loosely speaking, a Markov process observed in noise. It corresponds for Epidemics to the fact that the exact status of all the individuals within a population is not observed, or that detecting the status has some noise. 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, concerning multidimensional processes , is that some coordinates cannot be observed in practice. While the statistical inference for stochastic processes has a longstanding theory when sample paths are thoroughly observed (complete data) , this is no longer true for many cases that occur in practice. The aim of this course is to provide some tools to estimate these parameters on the basis of available data. Discrete time Markov chains are simple models for modeling stochastic epidemics. It is also interesting to study their inference because all the questions that can arise in more complex models can be illustrated in this set-up. Hence, classical results for the parametric inference for Markov chains are detailed here.

Density-dependent epidemic processes can be approximated by diffusion processes. This leads to new tools for studying inference for incomplete eptdemic data. We present these diffusion approximations for classical epidemic models such as SIR, $SIRS$, corresponding to single or recurrent outbreaks, or $SEIR$ corresponding to a simplified model of Ebola Dynamics. Then, we develop a framework for estimating the key parameters of epidemic models based on statistics of diffusion processes approximating $(Z(t))$. For this, we first introduce some classical tools used for statistics of diffusion processes. When necessary, recap on this topic are given along these notes. Various methods are assessed on simulated data sets modeling SIR and $SIRS$ epidemics and on real $SIRS$ data.

1.2 Parametric inference

There are various methods that can be used to estimate parameters in statistical models, that are summed up below.

- Maximum Likelihood Estimation
This entails that one can compute the likelihood of the observations. For a continuously and completely observed process, this is generally possible, but for a discrete time observation of a continuous-time process or for other kinds of incomplete observation, it is often not possible. This opens the whole domain of stochastic algorithms which aim at completing the data in order to estimate parameters with Maximum Likelihood methods. In regular statistical models, maximum likelihood estimators (MLE) are consistent and efficient (best variance).

- Minimum Contrast Estimation or Estimating Functions
When it is difficult to use or to compute the exact likelihood, pseudo-likelihoods (contrast functions; approximate likelihoods,..), or pseudo -score functions (approximations of the score function (obtained by differentiating the likelihood with respect to parameters), estimating functions,..) are often used. When they are well designed, these methods lead to estimators that are consistent and converge at the right rate. They might loose the efficiency property of MLE in regular statistical models

- Empirical Methods
This comprises all the methods that rely on limit theorems (such as the ergodic theorem) associated with various functionals built on the observations. Moments methods and Generalized Moment Methods belong to these empirical methods. Here estimation can be parametric or non parametric.

2 First classical examples of Markov chains in Epidemics

These two examples are taken from Andersson and Britton (Stochastic Epidemic Models and their Statistical Analysis, Lecture Notes in Statistics 151, 2000).

2.1 Greenwood model

Consider a population of size N composed of S_0 Susceptible individuals and I_0 infected at time 0 ($S_0 + I_0 = N$). Assume that the latent period equals 1. After it, susceptible individuals can be infected with probability p ($0 < p < 1$). Infected individuals are removed from the ifection chain.

Denote by S_n the number of Susceptible and I_n Infected at time n. Let I_n denote the number of Newly infected at time n. Then, the distribution of S_1 given (S_0, I_0) is ,

$S_1 = S_0 - I_1$ where $\mathcal{L}(I_1|(S_0, I_0) \sim Bin(S_0, p)$ and $S_1 = S_0 - I_1$.

Similarly, assume that at time n there are S_n susceptible and I_n infected individuals. Then,

$I_{n+1} \sim Bin(S_n, p)$ and $S_{n+1} = S_n - I_{n+1}$.

The process keeps going on up to the time where there is no longer Susceptible in the population. Let $\mathcal{F}_n = \sigma((S_i, I_i), i = 0, \dots n)$. Hence,

$$\begin{aligned}
\mathbb{P}((S_{n+1}, I_{n+1}) = (s_{n+1}, i_{n+1})/\mathcal{F}_n) &= 0 \text{ if } s_{n+1} > s_n \text{ or if } i_{n+1} + s_{n+1} \neq s_n, \\
&= \mathbb{P}(I_{n+1} = s_n - s_{n+1}/S_n = s_n) \\
&= C_{s_n}^{s_n - s_{n+1}} \, p^{s_n - s_{n+1}}(1-p)^{s_{n+1}}
\end{aligned}$$

Note that in this model $\mathcal{F}_n = \sigma(S_i, i = 0, \dots n)$ and that (S_n) is a Markov chain with state space $\{0, N\}$ and that the conditional distribution of S_{n+1} given $\mathcal{F}_n$ is a $Bin(S_n, (1-p))$

Some relevant questions: Let us asssume that the successive numbers of susceptibles $(s_0, s_1, \dots, s_K)$ have been observed. Is it possible to estimate p on the basis of these observations?

What is the duration of the epidemics (which depends on p)?

Likelihood approach

The underlying idea is that a good estimator of p is a value that yields for the successive observations $s_0, s_1, \dots, s_K$ the highest probability. Denote by P_p the probability associated with the Markov chain with parameter p. Assume that, at time 0, there is s_0 susceptible individuals. Then the likelihood associated with parameter p and observations $(s_0, \dots, s_K)$

is, if $s_0 \geq s_1 \cdots \geq s_N$

$$\begin{aligned} L_K(p; s_0, s_1, \dots, s_K) &= \mathbb{P}_p(S_0 = s_0, \dots, S_K = s_K) \\ &= \mathbb{P}_p(S_0 = s_0)\Pi_{n=1}^K \mathbb{P}_p(S_n = s_n / S_{n-1} = s_{n-1}) \\ &= \mathbb{P}_p(S_0 = s_0)\Pi_{n=1}^K C_{s_{n-1}}^{s_n}(1-p)^{s_n} p^{s_{n-1}-s_n} \\ &= C(s_0, \dots, s_K) p^{s_0 - s_K}(1-p)^{\sum_{n=1}^K s_n}. \end{aligned}$$

The likelihood $L_K(p; s_0, s_1, \dots, s_K))$ is equal to 0 otherwise.
In the term $C(s_0, , \dots, s_K)$, all the quantities independent of the parameter have been gathered. They only depend on the model and the observations, and therefore have no influence on the estimation of p.
Setting $l_K(p; s_0, \dots, s_K) = \log L_K(p; s_0, \dots, s_K) = l_K(p)$, an elementary computation yields that the value of p that maximizes the likelihood is

$$\hat{p}_K = \frac{s_0 - s_K}{\sum_{n=1}^K s_n}.$$

Now, we could have considered another well-known estimator : a Conditional Least Squares estimator. Indeed, $E(S_n|\mathcal{F}_{n-1}) = (1-p)S_{n-1}$. Therefore the CLS contrast process is:
$U_K(p; s_0, s_1, \dots, s_K)) = \sum_{n=1}^K (S_n - E(S_n|\mathcal{F}_{n-1})^2 = \sum_{n=1}^K (S_n - (1-p)S_{n-1})^2$. The CLS estimator is defined as a value p minimizing $U_K(p; s_0, s_1, \dots, s_K))$. This yields another estimator

$$\tilde{p}_K = 1 - \frac{\sum_{n=1}^K S_{n-1} S_n}{\sum_{n=1}^K S_{n-1}^2}.$$

A concern in Statistics is to answer the question: how do such estimators (or other ones) behave when the observation time increases (and thus the quantity of information)?
Remarks
1. If, instead of observing (S_n), the successive number of Infecteds (I_n) had been available, the inference is different since (I_n) is not a Markov chain.
2. In this model, the number of infected individuals at time n has no unfluence on the number of Infected at time $n+1$, which might be unrealistic.

2.2 Reed-Frost model

This is also a chain binomial model, which can be used to model the evolution of an ordinary influenza in a small group of individuals.

Assume that the latent period is long with respect to a short infectious period and that new infections occur at successive generations separated by the latent periods. Then, the epidemics dynamics is a SIR model (Susceptible, Infected, Removed) built as follows.

Denote by (S_n, I_n) the number of Susceptible and Infected individuals at time n. Assume that p is now the probability of contact between a Susceptible and Infected individual and let $q = 1 - p$ (probability of no contact between a Susceptible and an Infected). Assume moreover that contacts between susceptible and infecteds are independent.

Then, if the number of Susceptibles and Infected at time n is (s_n, i_n), the probability that a Susceptible remains Susceptible at time $n+1$ is q^{i_n} (probability of no contact with the i_n infected). Therefore, the probability of infection for a Susceptible is $p_n = (1 - q^{i_n})$. As before, the distribution of I_{n+1} given $\mathcal{F}_n$ is $Bin(S_n, p_n)$ 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,

$$\begin{aligned} \mathbb{P}(S_{n+1} = s_{n+1}, I_{n+1} = i_{n+1}/\mathcal{F}_n) &= 0 \ \text{ if } s_{n+1} + i_{n+1} \neq s_n, \\ &= C_{s_n}^{s_{n+1}} (q^{i_n})^{s_{n+1}} (1 - q^{i_n})^{s_n - s_{n+1}}. \end{aligned}$$

Let us stress that, contrary to the Greenwood model, the sequence of r.v. (S_n) is no longer Markov.

As before, the questions are how long before the end of the epidemics, what is the total number of infected individuals during the epidemics... The related random variables are: $T = inf\{n, S_n = 0\}$, and $Z = \sum_{n \geq 1} I_n$. their distributions both depend on p.

Likelihood approach

Assume now that the successive numbers of Susceptibles and Infected individuals have been observed up to time K: $(s_0, i_0), \ldots, (s_K, i_K)$, and that at time 0, there is (s_0, i_0) susceptibles and infecteds. Consider the estimation of $q = 1 - p$ and denote P_q the associated probability. Then, if $s_n + i_n = s_{n+1}$ for $n = 0, \ldots, K-1$,

$$\begin{aligned} L_K(q; (s_0, i_0), \ldots, (s_K, i_K)) &= P_q((S_0, I_0) = (s_0, i_0), \ldots (S_K, I_K) = (s_K, i_K)) \\ &= \Pi_{n=0}^{K-1} P_q((S_{n+1}, I_{n+1}) = (s_{n+1}, i_{n+1}) | (s_n, i_n)) \\ &= \Pi_{n=1}^{K} C_{s_n}^{s_{n+1}} (q^{i_n})^{s_{n+1}} (1 - q^{i_n})^{s_n - s_{n+1}}. \end{aligned}$$

Therefore, $\log L_K(q) = C((s_n, i_n)_n) + \sum_{n=0}^{K-1}(s_{n+1}i_n \log q + (s_n - s_{n+1})(1 - q^{i_n}))$.
Differentiating with respect to q yields

$$\frac{d \log L_N}{dq} = \sum_{n=0}^{N-1}(\frac{s_{n+1}i_n}{q} - \frac{i_{n+1}i_n q^{i_n-1}}{1-q^{i_n}}) = \frac{1}{q}\sum_{n=0}^{N-1}\frac{i_n}{1-q^{i_n}}(s_{n+1} - s_n q^{i_n}) \tag{1}$$

Hence a MLE estimator $\hat{q}_N$ of q is a solution of the equation
$\sum_{n=0}^{N-1}\frac{i_n}{1-q^{i_n}}(s_{n+1} - s_n q^{i_n}) = 0$.

Its properties can be studied as the number of observations increase.
Here, a problem which often occurs in practice already appears in this simple model:The number of both Susceptibles and Infecteds cannot be observed, but only, the number of Infected individuals are recorded. Therefore observations consist of the sequence $\{(I_{n+1}, n = 0, \dots, K-1\}$.
This corresponds to the statistical problem of inference for partially observed Markov models, in the special case of a deterministic relation between S_n and I_n ($S_n + I_n = S_{n-1}$).

3 Parametric inference for Markov chains

Discrete time Markov chains models are interesting here because all the 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 often sums up to a Markov chain model. Therefore, this allows to illustrate some classical statistical methods for stochastic models used in epidemics.

3.1 Canonical Statistical model and Likelihood

Consider n independent identically distributed random variables $(X_i, i = 1, \dots, n)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with density $f_\theta(x)dx$ on $\mathbb{R}$, with θ is an unknown parameter. Assume that $\theta \in \Theta$ with Θ a subset of $\mathbb{R}^k$ and let θ_0 denote the true value of the parameter that we want to estimate on the basis of the observations. The canonical statistical model associated with the observations $(X_i, i = 1, \dots, n)$ is defined as
- the observations space i.e. $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$,
- the probability distribution of the observations, i.e. the probability distribution $P^n_{\theta_0,}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ of the vector $(X_i, i = 1, \dots, n)$.
This distribution depends on the unknown value θ_0, and since it is unknown, we have

to consider the family of distributions P_θ^n. The canonical statistical experiment is here $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_\theta^n, \theta \in \Theta)$. Let us first recall some classical definitions.

Definition 1. *A likelihood function is a function such that*

$$\theta \to L(\theta) = \frac{dP_\theta^n}{d\mu}(X_i, i = 1, \ldots, n),$$

where μ is a positive σ-finite measure dominating all the distributions P_θ^n and $(X_i, i = 1, \ldots, n)$ is the observed sample.

The corresponding Loglikelihood is

$$l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_{i-1}, X_i). \tag{2}$$

Definition 2. *A Maximum Likelihood Estimator is any solution $\hat{\theta}$ of the equation*

$$L(\hat{\theta}) = sup_{\theta \in \Theta}\, L(\theta).$$

Since the r.v. (X_i) are independent, the density of P_θ^n with respect to the Lebesgue measure λ on $\mathbb{R}^n$ is, $\frac{dP_\theta^n}{d\lambda}(x_i, i = 1, \ldots, n) = \Pi_{i=1}^n f_\theta(x_i).$, and a likelihood function $L(\theta) = \Pi_{i=1}^n f_\theta(X_i)$.

3.2 Canonical Markov chain

Let us first introduce some notations and concepts that we use hereafter for Markov chains. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space. A E-valued (discrete time) stochastic process (X_n) is a collection of E- valued random variables. A filtration of $(\Omega, \mathcal{F})$ is a non-decreasing sequence $(\mathcal{F}_n), n \geq 0$ of sub-σ fields of $\mathcal{F}$. A filtered space is a triple $(\Omega, \mathcal{F}, \mathbb{F})$, where $\mathbb{F}$ is a filtration. For any filtration $\mathbb{F}$, we denote $\mathcal{F}_\infty = \vee_{n=0}^\infty \mathcal{F}_n$ the σ-field generated by $\mathbb{F}$.

Definition 3. *Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be two measurable spaces. An unnormalized transition kernel from $(E, \mathcal{E})$ to $(G, \mathcal{G})$ is a function $Q : E \times \mathcal{G} \to [0, \infty]$ that satisfies*
(i) For all $x \in E, Q(x, .)$ is a positive measure on $(G, \mathcal{G})$.
(ii) For all $A \in \mathcal{G},,$ the function $x \to Q(x, A)$ is measurable.

If $Q(x, G) = 1$ for all $x \in E$, then Q is called a transition kernel.

Consider now an E-valued stochastic process $(X_n, n \geq 0)$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

Definition 4. *The stochastic process $(X_n, n \geq 0)$ is a Markov chain under $\mathbb{P}$, with respect to filtration $\mathbb{F}$ and with transition kernel Q if for all $n \geq 0$,*
- X_n is $\mathcal{F}_n$- measurable, and
- for all $A \in \mathcal{E}$, $\mathbb{P}(X_{n+1} \in A | \mathcal{F}_n) = Q(X_n, A)$.
The distribution of X_0 is the initial distribution and E the state space.

Some elementary exemples
Example 1: i.i.d random variables
Let $(X_n), n \geq 0$ be a sequence of independent identically distributed random variables with distribution μ, then $Q(x, dy) = \mu(dy)$ or$\forall B \in \mathcal{E}, Q(x, B) = \mu(B)$.
Example 2: Random walk on $\mathbb{R}$.
Let $((Y_n))$ i.i.d sequence of random variables on $\mathbb{R}$ with density $\nu(dy) = h(y)dy$, $\mathcal{F}_n = \sigma(Y_0, \ldots, Y_n)$ and define the Markov chain (X_n),
$X_0 = Y_0$,
$X_1 = X_0 + Y_1$, and $\ldots$ $X_{n+1} = X_n + Y_{n+1}$.
Then $\mathbb{E}(f(X_{n+1}/\mathcal{F}_n)) = \mathbb{E}(f(X_{n+1}/X_n)) = \int \ f(X_n + y)h(y)dy = \int f(y)h(y - X_n)dy := Qf(X_n)$.
Hence the transition probability of (X_n) is $Q(x, dy) = h(y - x)dy$.
Example 3: Markov chain $X_{n+1} = \phi(X_n, Y_{n+1})$
Consider a mesurable known function $\phi(x, y)$, $\phi : (E \times E, \mathcal{E} \times \mathcal{E}) \rightarrow (E, \mathcal{E})$ and let (Y_n) a sequence of i.i.d. random variables on E with distribution $\nu(dy)$. Assume that $X_0 \sim \mu$ is independent of $(Y_n, n \geq 1)$.
Then, the sequence of r.v. defined as: $X_{n+1} = \phi(X_n, Y_{n+1})$ is a Markov chain on $(E, \mathcal{E})$.
For $x \in E$ and $A \in \mathcal{E}$, define $\nu_x(A) := \nu(\{y, \phi(x, y) \in A\})$ ((the image of measure ν by $\phi(x, .)$).
Then $\mathbb{E}(f(X_{n+1})/\mathcal{F}_\backslash) = \mathbb{E}(f(\phi(X_n, Y_{n+1}))/X_n) = \int f(y)\nu_{X_n}(dy)$.
Therefore the transition probability of (X_n) is $Q(x, dy) = \nu_x(dy)$.
In Example 2, $\phi(x, y) = x + y$ and $\nu_x(A) = \nu(A - x)$.

In order to do statistics, we need to define the canonical version of (X_n). The canonical space is the space of observations $(E^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$. The coordinate process is the E- valued stochastic process defined on the canonical space by $X_n(\omega) = \omega(n)$. The canononical space is endowed with the natural filtration $\mathbb{F}^X$ of the coordinate process.

Theorem 1. *Let μ be a probability measure on $(E, \mathcal{E})$ and Q a transition kernel on $(E, \mathcal{E})$. Then, there exists a unique probability measure on $(E^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$, denoted $P_{\mu,Q}$ such that the coordinate process $(X_n, n \geq 0)$ is a Markov chain (with respect to its natural filtration) with initial distribution μ and transition kernel Q.*

Properties

• Let $A_0, A_1, \dots, A_n$ belong $\mathcal{E}$, then
$P^n_{\mu,Q}(X_0 \in A_0, \dots, X_n \in A_n) = \int_{A_0} \mu(dx_0) \int_{A_1} Q(x_0, dx_1) \dots \int_{A_n} Q(x_{n-1}, dx_n)$.
• If f a bounded measurable function on E^{n+1},
$P^n_{\mu,Q}(f) = \int_{E^{n+1}} f(x_0, x_1, \dots, x_n) \mu(dx_0) Q(x_0, dx_1) \dots Q(x_{n-1}, dx_n)$.

Definition 5. *The canonical statistical model associated with the observation $(X_i, i = 0, \dots n)$ is defined by*
- the observations space : $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}})$,
- the probability distribution of the observations, that is the probability $\mathbb{P}_{\mu_0,Q_0}$ (defined in Theorem 1) on $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}})$ of the vector $(X_i, i = 0, \dots n)$.

As before, this distribution depends on the unknown initial distribution μ_0, and unknown density kernel $Q_0(x, dy)$. Since there are unknown, we have to consider the family of distributions $P_{\mu,Q}$. The canonical statistical model is $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}}, (P_{\mu,Q}, (\mu, Q) \in \Theta))$, with Θ some subset of "probability measures x density kernels" on $(E, \mathcal{E})$.

3.3 Likelihood process

The successive observations of (X_i) allow to estimate μ, Q. One expects that longer observations lead to better estimators of (μ, Q).
Let λ be a positive measure on $(E, \mathcal{E})$ dominating all the distributions
$\{\mu(dy), (Q(x, dy), x \in E)\}$. Assume that $\mu(dy) = \mu(y)\lambda(dy), Q(x, dy) = Q(x, y)\lambda(dy)$.
Then, if $P^n_{\mu,Q} = P_{\mu,Q}|\mathcal{F}_n$, the density of $P^n_{\mu,Q}$ with respect to the measure λ^{n+1} on E^{n+1} is

$$\frac{d\mathbb{P}^n_{\mu,Q}}{d\lambda^{n+1}}(x_i, i = 0, \dots, n) = \mu(x_0) Q(x_0, x_1) \dots Q(x_{n-1}, x_n).$$

Therefore, a likelihood function is

$$L_n(\mu, Q) = \mu(X_0) Q(X_0, X_1) \dots Q(X_{n-1}, X_n).$$

This is called the likelihood at time n. The sequence $(L_n(\theta), n \in N)$ is the likelihood process.
Example: Markov chain with finite state space $E = \{1, \dots, K\}$.
Let λ be the uniform measure on E $()\lambda(k) = 1$ for all $k \in E)$. Then,

$$\frac{dP^n_{\mu,Q}}{d\lambda^{n+1}}(x_i, i = 0, \dots, n) = \mu(x_0)Q(x_0, x_1) \dots Q(x_{n-1}, x_n).$$

Therefore, a likelihood function at time n is

$$\mu(X_0)Q(X_0, X_1) \dots Q(X_{n-1}, X_n).$$

3.4 Maximum likelihood estimator for Markov chains

Assume that the parameter set Θ is is a subset of $\mathbb{R}^l$ and that θ_0 is the true value of the parameter.

Definition 6. *A family* $(Q_\theta(x, dy), \theta \in \Theta)$ *of transition probability kernels on* $(E, \mathcal{E}) \to [0, 1]$ *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* $f_\theta : (E \times E, \mathcal{E} \times \mathcal{E}) \to \mathbb{R}^+$ *measurable.*

Assume that the initial distribution μ is known and let $P_\theta = P_{\mu,Q_\theta}$ the distribution of the Markov chain (X_n) with initial distribution μ and transition kernel Q_θ. Then a likelihood function is

$$L_n(\theta) = \Pi_{i=1}^n f_\theta(X_{i-1}, X_i).$$

A maximum likelihood estimator is defined as any solution $\hat{\theta}_n$ of

$$L_n(\hat{\theta}_n) = sup\{L_n(\theta), \theta \in \Theta\}.$$

In order to study the properties of this estimator 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): $\lambda_{\theta_0}(\{x, Q_\theta(x, .) \neq Q_{\theta_0}(x, .)\}) > 0$.
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.

Studying the properties of the MLE is two-fold: first prove the consistency of $\hat{\theta}_n$, then study the limit distribution. We detail euristically how these properties are obtained.

Consistency: It relies on several steps.

Step (1) : the convergence of the loglikelikelihood $\ell_n(\theta)$ suitably normalized to a deterministic limit $J(\theta_0, \theta)$ under P_{θ_0}.

Step (2) : the property that this limit has a unique global maximum at θ_0.

Step (3) Since $\hat{\theta}_n = Argsupl_n(\theta)$ and $\theta_0 = ArgsupJ(\theta_0, \theta)$, assumptions ensuring that "lim Argsup= Argsup lim".

Step (1): Define for $n \geq 1$, $Y_n = (X_{n-1}, X_n)'$. Under (H0), (H1), $(Y_n, n \geq 1)$ is a positive recurrent Markov chain on $(E \times E, \mathcal{E} \times \mathcal{E})$ with stationary distribution $\lambda_{\theta_0}(dx)Q_{\theta_0}(x, dy)$. Applying the ergodic theorem yields that, under P_{θ_0},

$$\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_{i-1}, X_i) \to \int\int_{E \times E} \log f_\theta(x, y) \lambda_{\theta_0}(dx) Q_{\theta_0}(x, dy) \ a.s. \tag{3}$$

Step (2): Let us set

$$J(\theta_0, \theta) = \int\int_{E \times E} \log f_\theta(x, y) \lambda_{\theta_0}(dx) Q_{\theta_0}(x, dy). \tag{4}$$

Recall the definition of the Kullback-Leibler divergence between two probabilities.

Definition 7. *Let P, Q be two probability distributions defined on a probability space $(\Omega, \mathcal{A})$. Then the Kullback-Leibler divergence $K(P, Q)$ of Q with respect to P is:*
-if $P << Q$, $K(P, Q) = \mathbb{E}_P(\log \frac{dP}{dQ}) = \int \log \frac{dP}{dQ} dP$.
- $K(P, Q) = +\infty$ otherwise.

The Kullback-Leibler divergence measures the "difference"" between two probabilities. It satisfies
$K(P, Q) \geq 0$ and $K(P, Q) = 0$ if and only if $Q = P$, P-a.s.
Rewriting equation (4) yields

$$J(\theta_0, \theta) = \int\int \log \frac{f_\theta(x, y)}{f_{\theta_0}(x, y)} \lambda_{\theta_0}(dx) Q_{\theta_0}(x, dy) + A(\theta_0),$$

with $A(\theta_0) = \int\int \log f_{\theta_0}(x, y) \lambda_{\theta_0}(dx) Q_{\theta_0}(x, dy)$.

Hence, using that, under (H0), $Q_\theta(x,dy) = f_\theta(x,dy)Q(x,dy)$ yields

$$\begin{aligned} J(\theta_0,\theta) &= \int \lambda_{\theta_0}(dx) \int \log \frac{Q_\theta(x,dy)}{Q_{\theta_0}(x,dy)} Q_{\theta_0}(x,dy) + A(\theta_0) \quad (5) \\ &= -\int K(Q_{\theta_0}(x,.), Q_\theta(x,.))\, \lambda_{\theta_0}(dx) + A(\theta_0). \quad (6) \end{aligned}$$

Assumption (H2) ensures that $\theta \to J(\theta_0,\theta)$ possesses a unique global maximum at $\theta = \theta_0$.

Step (3) relies on additional uniform integrability assumptions.

Theorem 2. *Assume (H0),(H1) and (H2) and that* Θ *is a compact subset of* $\mathbb{R}^l$. *Assume moreover,*
(i) $\forall \theta, \log f_\theta(x,y)$ *is integrable with respect to* $\lambda_{\theta_0}(dx)Q_{\theta_0}(x,dy) := \lambda_{\theta_0} \otimes Q_{\theta_0}$
(ii)$\forall (x,y) \in E^2,\ \theta \to f\theta(x,y)$ *is continuous w.r.t.* θ,
(iii) There exists a function $h(x,y)$ *integrable w.r.t.* $\lambda_{\theta_0} \otimes Q_{\theta_0}$ *and such that*

$$\forall \theta, |\log f_\theta(x,y)| \leq h(x,y).$$

Then the maximum likelihood estimator $\hat{\theta}_n$ *is consistent, i.e. it converges in probability under* $\mathbb{P}_{\theta_0}$ *to* θ_0 *as* $n \to \infty$.

3.5 A primer on MLE asymptotics

We sum up Section 12.1 of the book of Cappé, Moulines, Ryden (2005), Inference in Hidden Markov Models.
For standard models, asymptotic properties of the MLE rely on tree basic results: a law of large numbers for the log-likelihood $\ell_n(\theta)$, a central limit theorem for the score function and a law of large numbers for the observed information. this sums up,
(i) For all $\theta \in \Theta$, $n^{-1}\ell_n(\theta) \to J(\theta_0,\theta)$ $P_{\theta_0}-$ a.s. uniformly over compacts subsets of Θ, where $\ell_n(\theta)$ is the log-likelihood of the parameter θ given the first n observations and $\theta \to J(\theta_0,\theta)$ is a continuous function with a global unique maximum at θ_0.
(ii) $n^{-1/2}\nabla_\theta \ell_n(\theta_0) \to \mathcal{N}(0, \mathcal{I}(\theta_0))$ in distribution under P_{θ_0} , where $\mathcal{I}(\theta))$ is the Fisher information matrix at θ.
(iii) $lim_{n\to\infty} sup_{|\theta-\theta_0|\leq\delta} \parallel -\frac{1}{n}\nabla^2_\theta \ell_n(\theta) - \mathcal{I}(\theta_0) \parallel \to 0$ as $\delta \to 0$ $P_{\theta_0}-$ a.s.

Condition (i) ensures strong consistency of the MLE. Euristically, the argument is the following. The MLE $\hat{\theta}_n$ satisfies $\ell(\hat{\theta}_n) \geq \ell(\theta)$ for all $\theta \in \Theta$. Because $J(\theta_0, \theta)$ has a global unique maximum at θ_0, $J(\theta_0, \theta_0) - J(\theta_0, \hat{\theta}_n) \geq 0$. Combining the two inequalities yields

$$\begin{aligned} 0 &\leq J(\theta_0, \theta_0) - J(\theta_0, \hat{\theta}_n) \\ &\leq J(\theta_0, \theta_0) - \frac{1}{n}\ell_n(\theta_0) + \frac{1}{n}\ell_n(\theta_0) - \frac{1}{n}\ell_n(\hat{\theta}_n) + \frac{1}{n}\ell_n(\hat{\theta}_n) - J(\theta_0, \hat{\theta}_n) \\ &\leq 2\sup_{\theta \in \Theta} |J(\theta_0, \theta) - \frac{1}{n}\ell_n(\theta)|. \end{aligned}$$

Therefore, if Θ is a compact set, $\ell(\hat{\theta}_n) \to J(\theta_0, \theta_0)$ P_{θ_0}- a.s.as $n \to \infty$, which in turn implies, as $J(\theta_0, .)$ is continuous with a unique global maximum at θ_0, that the MLE converges to θ_0 P_{θ_0}- a.s. Therefore, the MLE is strongly consistent.
If consistency holds , then properties (ii) and (iii) yield asymptotic normality of the MLE. Assuming that θ_0 is an interior point of Θ and that the Fisher information $\mathcal{I}(\theta_0)$ is non -singular, a Taylor expansion of the score function $\nabla_\theta \ell_n$ at point θ_0 leads, using that $\nabla_\theta \ell_n(\hat{\theta}_n) = 0$

$$0 = \nabla_\theta \ell_n(\hat{\theta}_n) = \nabla_\theta \ell_n(\theta_0) + \Big(\int_0^1 \nabla_\theta^2(\theta_0 + t(\hat{\theta}_n - \theta_0))dt\Big)\,(\hat{\theta}_n - \theta_0). \tag{7}$$

Form this expansion, we get, using that $\mathcal{I}(\theta_0)$ is non-singular,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Big(-\frac{1}{n}\int_0^1 \nabla_\theta^2(\theta_0 + t(\hat{\theta}_n - \theta_0))dt\Big)^{-1}\Big(\frac{1}{\sqrt{n}}\nabla_\theta \ell_n(\theta_0)\Big). \tag{8}$$

Using that $\hat{\theta}_n \to \theta_0$ P_{θ_0}- a.s and (iii) yields that the first factor of the right hand side of (8) converges to $\mathcal{I}(\theta_0)^{-1}$ P_{θ_0} a.s. The second factor converges in distribution under P_{θ_0} to $\mathcal{N}(0, \mathcal{I}(\theta_0))$. An application of Slutsky's theorem yields that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to $\mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ in distribution under P_{θ_0}. This is the standard result on asymptotic normality of the MLE.

Let us specialize these results to the Markov Chain Statistical model.
Point (i) has been checked: $J(\theta_0, \theta) = \int K(Q_{\theta_0}(x, .), Q_\theta(x, .))\, \lambda_{\theta_0}(dx) + A(\theta_0)$.
Point (ii): $\nabla_\theta \ell_n(\theta_0) = \sum_{i=1}^n \frac{1}{f_{\theta_0}(X_{i-1}, X_i)} \nabla_\theta f_{\theta_0}(X_{i-1}, X_i)$.
Now, under P_{θ_0}, $\frac{dP_{\theta_0}}{d\mathbb{P}}(X_1, \ldots, X_n) = \Pi_{i=1}^n f_{\theta_0}(X_{i-1}, X_i)$.
Hence, $E_{\theta_0}(\nabla_\theta \ell_n(\theta_0)) = E_{\theta_0}(L_n(\theta_0)^{-1}\nabla_\theta L_n(\theta_0)) = E_{\mathbb{P}}(\nabla_\theta L_n(\theta_0))$.
With appropriate assumptions, $E_{\mathbb{P}}(\nabla_\theta L_n(\theta_0)) = \nabla_\theta E_{\mathbb{P}}(L_n(\theta_0)) = 0$.

Hence, setting $v_i(\theta) = \nabla_\theta \ell_i(\theta) - \nabla_\theta \ell_{i-1}(\theta) = f_\theta(X_{i-1}, X_i))^{-1}\nabla_\theta f_\theta(X_{i-1}, X_i)$, we get $\nabla_\theta \ell_n(\theta_0) = \sum_{i=1}^n v_i(\theta_0)$ with $E_{\theta_0}(v_i(\theta_0)|\mathcal{F}_{i-1}) = 0$. Hence $\nabla_\theta \ell_n(\theta_0))$ is a centered P_{θ_0}-martingale, say $M_n(\theta_0)$. with associated increasing process,
$< M(\theta_0) >_n = \sum_{i=1}^n E_{\theta_0}(v_i(\theta_0)^2|\mathcal{F}_{i-1})$. Applying the ergodic theorem yields that:

$$\frac{1}{n} < M(\theta_0)_n \to \int\int \frac{1}{f_{\theta_0}(x,y)^2} \nabla_\theta f_{\theta_0}(x,y))\ {}^t\nabla_\theta f_{\theta_0}(x,y) \lambda_{\theta_0}(x) Q_{\theta_0}(x,y) dxdy := \mathcal{I}(\theta_0).$$

Therefore the CLT for martingales yields that $\frac{1}{\sqrt{n}} M_n(\theta_0) \to \mathcal{N}(0, \mathcal{I}(\theta_0))$ in distribution under P_{θ_0}.
Similarly, another application of the ergodic theorem together with integability assumptions yields that
$\frac{1}{n}\nabla_\theta^2 \ell_n(\theta_0)$ converges P_{θ_0} a.s. to a deteministic limit equal to $\mathcal{I}(\theta_0)$.

3.6 Going a step further

What if, instead of the likelihood, nother process is used , e.g. the conditional least squares method or a nother contrast process$xU_n(\theta)$? (in essence think of $U_n = -\ell_n$).
Define $\tilde{\theta}_n$ such that

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

(i) For all $\theta \in \Theta$, $n^{-1}\ell_n(\theta) \to K(\theta_0, \theta)$ in $P_{\theta_0}-$ probability. uniformly over compacts subsets of Θ , where $\theta \to K(\theta_0, \theta)$ is a continuous function with a global unique minimum at θ_0.
(ii) $n^{-1/2}\nabla_\theta U_n(\theta_0) \to \mathcal{N}(0, I_U(\theta_0))$ in distribution under P_{θ_0}.
(iii) There exists a symmetric positive matrix $J_U(\theta_0)$ such that $lim_{n\to\infty} sup_{|\theta-\theta_0|\leq\delta} \parallel -\frac{1}{n}\nabla_\theta^2 U_n(\theta) - J_U(\theta_0) \parallel \to 0$ as $\delta \to P_{\theta_0}-$ a.s.

Then, if $J_U(\theta_0)$ is non singular, $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges to $\mathcal{N}(0, J_U(\theta_0)^{-1} I_U(\theta_0) J_U(\theta_0))$ in distribution under P_{θ_0}.
Indeed, (ii) holds in the case of the MLE because $\nabla_\theta \ell(\theta_0)$ is a martingale under P_{θ_0} which is centered and belongs to L^2. For other functionals of the observations, a CLT is obtained building a martingale (M_n) based on the observations. Then, if $< M >_n \to \infty$, under appropriate assumptions, we get (ii) and (iii) using limit theorems for martingales.

Coming back to the Conditional Least Squares method for ergodic Markov chains.

Assume that (X_n) is a Markov chain with state space $\mathbb{R}$ and define

$$U_n(\theta) = \frac{1}{2}\sum_{i=1}^{n}(X_i - E_\theta(X_i|X_{i-1})^2).$$

Let us check successively points (i),(ii) and (iii) for $U_n(\theta)$.
Point (i): We have that $E_\theta(X_i|X_{i-1}) = \int yQ_\theta(X_{i-1}, y)dy := g_\theta(X_{i-1})$.
Hence, $\frac{1}{n}U_n(\theta) \to \frac{1}{2}\int\int(y - g_\theta(x))^2\lambda_{\theta_0}(x)Q_{\theta_0}(x, y)dxdy := K(\theta_0, \theta)$.
Rewriting this limit yields,
$K(\theta_0, \theta) = \frac{1}{2}\int(g_\theta(x) - g_{\theta_0}(x))^2\lambda_{\theta_0}(x)Q_{\theta_0}(x, y)dxdy + A(\theta_0)$,
with $A(\theta_0) = \frac{1}{2}\int\int(y - g_{\theta_0}(x))^2\lambda_{\theta_0}(x)Q_{\theta_0}(x, y)dxdy$. So point (1) holds under an assumption that $\theta \neq \theta_0 \Rightarrow g_\theta(.), g_{\theta_0}(.)$ are non identical functionsâĂę
Point (2): Let us study $\nabla_\theta U_n(\theta_0)$.
$M_n(\theta_0) := \nabla_\theta U_n(\theta_0) = -\sum_{i=1}^{n}(X_i - g_{\theta_0}(X_{i-1}))\nabla_\theta g_{\theta_0}(X_{i-1})$.
Hence, using the definition of g_θ, we get that $M_n(\theta_0)$ is a centered L^2 P_{θ_0}- martingale. Its increasing process (crochet) $< M(\theta_0) >_n$ is
$< M(\theta_0) >_n = \sum_{i=1}^{n} E_{\theta_0}\Big((X_i - g_{\theta_0}(X_{i-1}))^2|X_{i-1}\Big)\nabla_\theta g(X_{i-1})\ {}^t\nabla_\theta g(X_{i-1})$.
Let $V_\theta(x)$ denote the conditional variance of X_i given $\{X_{i-1} = x\}$.
Then, $< M(\theta_0) >_n = \sum_{i=1}^{n}\ V_{\theta_0}(X_{i-1})\nabla_\theta g_{\theta_0}(X_{i-1})\ {}^t\nabla_\theta g_{\theta_0}(X_{i-1})$.
An application of the ergodic theorem yields
$\frac{1}{n} < M(\theta_0) >_n \to \int V_{\theta_0}(x)\nabla_\theta\ g_{\theta_0}(x)\ {}^t\nabla_\theta\ g_{\theta_0}(x)\lambda_{\theta_0}(x)dx := \Sigma(\theta_0)$. P_{θ_0} a.s.
Finally, under appropriate assumptions, we can apply the central limit theorem for martingales and get that, in distribution under P_{θ_0},

$$\frac{1}{\sqrt{n}}M_n \to \mathcal{N}(0, \Sigma(\theta_0).$$

Hence the matrix $\Sigma(\theta) = I_U(\theta)$.
Point (iii): We have to study now $\frac{1}{n}\nabla_\theta^2 U_n(\theta)$. We have

$$\nabla_\theta^2 U_n(\theta) = \sum_{i=1}^{n}\nabla_\theta\ g_\theta(X_{i-1})\ {}^t\nabla_\theta\ g_\theta(X_{i-1}) + (X_i - g_\theta(X_{i-1}))\nabla_\theta^2 g_\theta(X_{i-1}).$$

Hence $\frac{1}{n}\nabla_\theta^2 U_n(\theta_0) \to \int\nabla_\theta\ g_{\theta_0}(x)\ {}^t\nabla_\theta\ g_{\theta_0}(x)\lambda_{\theta_0}(x)dx := J_U(\theta_0)$.

Let us stress that, contrary to the MLE approach, we have no longer that $I_U(\theta) = J_U(\theta)$. Joining these results, we get that the minimum least squares estiamtor $\tilde{\theta}_n$ is consistent asymptotically Gaussian at rate $\sqrt{n}$. We have just increased the asymptotic variance which is no longer the optimal one (in other words, $\tilde{\theta}_n$ is not efficient).

4 Coming back to examples

4.1 Birth and death chain

Assume that in a large infinite population, an epidemic model is described by a birth and death chain on $\mathbb{N}$ describing the number of Infected individuals I_n at time n.
The birth and death chain associated with parameters $p, q, r \in (0,1)$ such that $p+q+r = 1$ is
- if $k \geq 1$, $\mathbb{P}(I_{n+1} = k+1 | I_n = k) = p$,
$\mathbb{P}(I_{n+1} = k-1 | I_n = k) = q$, and $\mathbb{P}(I_{n+1} = k | I_n = k) = r$.
- if $k = 0$, $\mathbb{P}(I_{n+1} = 1 | I_n = 0) = p$ and $\mathbb{P}(I_{n+1} = 0 | I_n = 0) = 1 - p$.
(there is still some possible infection (coming for instance from the environment) even if there is no longer infecteds).
This description corresponds for instance to an epidemic where individuals (e.g. animal farms are infected by the environment only) and recovery is obtained by vaccination, assuming that only one animal can be vaccinated by time unit.

Set $\theta = (p, q)$ with $(0 < p, q < 1; p + q < 1)$ and $\Theta = (0,1)^2$. Let $\theta_0 = (p_0, q_0)$ br e the true parameter value and assume that The initial number of infected $I_0 = i_0$ is known. Then (I_n) is a Markov chain on $\mathbb{N}$ with transition kernel,
$Q_\theta(i,j) = p\delta_{i+1}(j) + q\delta_{i-1}(j) + r\delta_i(j)$ if $i \neq 0$,
$Q_\theta(0,j) = p\delta_1(j) + (1-p)\delta_0(j)$,
where $\delta_i(.)$ denotes the Dirac measure at point i: $\delta_i(j) = 1$ if $j = i$, 0 if $j \neq i$.

This is an irreductible aperiodic Markov chain on $\mathbb{N}$ and if $p < q$, (I_n) is positive recurrent with stationary distribution the geometric distribution

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

Choosing the positive measure on $\mathbb{N}$ $\gamma(i) = 1$ for all i, we get that the distribution P_θ of (I_n) satisfies,
$\frac{dP_\theta^n}{d\gamma^n}(I_k, k = 1, \ldots, n) = \prod_1^n Q_\theta(I_{k-1}, I_k)$.

Let us define the random variables $N_n^{i,j}$, which count the number of transitions from state i to state j up to time n:

$$N_n^{i,j} = \sum_{k=1}^{n} \delta_{i,j}(I_{k-1}, I_k),$$

where $\delta_{i,j}(x,y) = 1$ if $(x,y) = (i,j)$ and 0 otherwise. Then, we get another expression for $L_n(\theta)$,

$$L_n(\theta) = \frac{dP_\theta^n}{d\gamma^{n+1}}(I_k, k = 1, \ldots, n) = \mu(I_0) \prod_{i,j\geq 0} Q_\theta(i,j)^{N_n^{i,j}}.$$

Clearly, if $j \neq \{i-1, i, i+1\}$, $N_n^{i,j} = 0$. Hence, the loglikelihood $l_n(\theta)$ satisfies

$$l_n(\theta) = (\sum_{i\geq 0} N_n^{i,i+1}) \log p + (\sum_{i\geq 1} N_n^{i,i-1}) \log q + (\sum_{i\geq 1} N_n^{i,i}) \log r + N_n^{0,0} \log(1-p).$$

The stationary distribution of (I_n, I_{n+1}) is $(1 - \frac{p_0}{q_0})(\frac{p_0}{q_0})^i Q_{\theta_0}(i,j)$. Therefore, applying the ergodic theorem yields that, under $\mathbb{P}_{\theta_0}$:
$\frac{1}{n} N_n^{i,i+1} \to p_0 \lambda_{\theta_0}(i) \Rightarrow \frac{1}{n}\sum_{i\geq 0} N_n^{i,i+1} \to p_0$ a.s.
$\frac{1}{n} N_n^{i,i-1} \to q_0 \lambda_{\theta_0}(i) \Rightarrow \frac{1}{n}\sum_{i\geq 1} N_n^{i,i-1} \to q_0 \times \frac{p_0}{q_0} = p_0$ a.s.
$\frac{1}{n} N_n^{i,i} \to r_0 \lambda_{\theta_0}(i) \Rightarrow \frac{1}{n}\sum_{i\geq 1} N_n^{i,i} \to \frac{r_0 p_0}{q_0}$ and $\frac{1}{n} N_n^{0,0} \to (1 - \frac{p_0}{q_0})(1-p_0)$.
Therefore, $\frac{1}{n} l_n(\theta)$ converges under P_{θ_0} to $J(\theta_0, \theta)$ which is
$J(\theta_0,\theta) = p_0 \log p + p_0 \log q + \frac{r_0 p_0}{q_0} \log r + (1 - \frac{p_0}{q_0})(1-p_0)\log(1-p)$.
Using that, for $i \neq 0$, $K(Q_{\theta_0}(i,.), Q_\theta(i,.)) = p_0 \log \frac{p_0}{p} + q_0 \log \frac{q_0}{q} + r_0 \log \frac{r_0}{r}$ and
$K(Q_{\theta_0}(0,.), Q_\theta(0,.)) = p_0 \log \frac{p_0}{p} + (1-p_0) \log \frac{1-p_0}{1-p}$, we can check that $\theta \to J(\theta_0,\theta)$ possesses a unique global minimum at $\theta = \theta_0$.
The Maximum likelihood estimator is $(\hat{p}_n, \hat{q}_n)$ writes
$\hat{p}_n = \frac{1}{n}\sum_{i\geq 0} N_n^{i,i+1})$ and $\hat{q}_n = \frac{\sum_{i\geq 1} N_n^{i,i-1}}{\sum_{i\geq 1} N_n^{i,i-1}) + \sum_{i\geq 1} N_n^{i,i}}(1 - \hat{p}_n)$.

4.2 Model of infection in the Intensive Care Unit (ICU)

This example is taken from Chapter 4 of Diekmann, Heesterbeck, Britton (2013).
This concerns a finite population of size N (N small) but high turnover (most patients stay only a couple of days).There are two routes for infection (colonization): endogenous route (α mechanism) and exogenous route (β transmission). New admitted individuals are susceptible.
Concentrating on long time intervals leads to a Markov chain description. Let us consider the probabilities of the various compositions of the ICU in terms of Infected and susceptible individuals. Assume that discharge and admission take place evey day at noon. The

bookkeeping scheme concerns the state of the ICU immediately after discharge (12h05). Each patient has probability $1/\delta$ of being discharged by unit of time. For sake of clarity, assume that the probability that both infection events occur in the same time interval is zero.

Consider the simplest example : an ICU with two beds. It corresponds to three possible states: state 0 (both patients are Susceptible), state 1 (one patient is sucsceptible, one is colonized) and 2 (both are colonized). Assume that both patients are susceptible on day i at 12h05. The probability that both patients are susceptible at 11h55 the next day is $(e^{-\alpha})^2$, in state 1 with probability $2e^{-\alpha}(1-e^{-\alpha})$ and in state 2 with probability $(1-e^{-\alpha})^2$. After discharge (12h05), the probability that they are
in state 2 is $Q(0,2) = (1-\frac{1}{\delta})^2(1-e^{-\alpha})^2$;
in state 1 is $Q(0,1) = 2\frac{1}{\delta}(1-\frac{1}{\delta})^2 + 2(1-\frac{1}{\delta})e^{-\alpha}(1-e^{-\alpha})$,
in state 0 $Q(0,0) = (e^{-\alpha})^2 + \frac{2}{\delta}(1-\frac{1}{\delta})e^{-\alpha} + (\frac{1}{\delta})^2(1-e^{-\alpha})^2$.
Consider now the case that we are in state 1at 12h05 (one patient is susceptible, and the other is colonized). At 11h55 the next day the state is 1 with probability $\exp(-(\alpha+\beta))$ and 2 with probability $(1-exp(-(\alpha+\beta)))$. Hence at 12h05 on that same day,it will be
- 2 with probability $Q(1,2) = (1-\frac{1}{\delta})^2(1-exp(-(\alpha+\beta)))$;
- 1 with probability $Q(1,1) = (1-\frac{1}{\delta})\exp(-(\alpha+\beta)) + \frac{2}{\delta}(1-\frac{1}{\delta})(1-\exp(-(\alpha+\beta)))$,
- 0 with probability : $Q(1,0) = \frac{1}{\delta}\exp(-(\alpha+\beta)) + \frac{1}{\delta})^2(1-e^{-(\alpha+\beta)})$.
Finally, if we start with two colonized patients, the situation can only change by discharge and admission. Therefore, $Q(2,2) = (1-\frac{1}{\delta})^2$, $Q(2,1) = \frac{2}{\delta}(1-\frac{1}{\delta})$ and $Q(2,0) = (\frac{1}{\delta})^2$.
This is a positive recurrent Markov chain with transition kernel $Q = Q_\theta$ depending on $\theta = (\alpha, \beta, \delta)$. Observing after discharge the state $(X_i, i = 0, \dots, n)$ of the ICU allows to estimate θ. The loglikelihood is

$$l_n(\theta) = \sum_{k=1}^{n} \log Q_\theta(X_{k-1}, X_k) = \sum_{i,j=0,1,2} N_n^{ij} \log Q_\theta(i,j),$$

where $N_n^{i,j} = \sum_{k=1}^n \mathbf{1}_{(X_{k-1}=i, X_k=j)}$. The MLE $(\hat\alpha_n, \hat\beta_n, \hat\delta_n)$ can be obtained by maximizing this loglikelihood. According to the previous theorem, there are consistent.

The next problem lies in the fact that the exact status of the patients is not always known (no systematic control). Assume for instance that each patient is tested with probability p. Then, the observations are no longer X_n, but (Y_n) which are obtained as follows:
-if $X_n = 0$, then $Y_n = 0$ with probability $P(0,0) = (1-p)^2$, $Y_n = 1$ with probability

$P(0,1) = 2p(1-p)$, 2 with probability $P(0,2) = p^2$;
-if $X_n = 1$, then $Y_n = 0$ with probability $P(1,0) = p(1-p)$; 1 with probability $P(1,1) = p^2 + (1-p)^2$; 2 with probabilility $P(1,2) = p(1-p)$
- if $X_n = 2$, $Y_n = 0$ with probability $P(2,0) = p^2$, 1 with probability $P(2,1) = 2p(1-p)$, and 2 with probabilility $P(2,2) = (1-p)^2$.
Therefore the distribution of Y_n conditionally on X_n is an explicit distribution depending on the parameter p. Denote $f_p(x,y)$ the conditional distribution of Y_n given $X_n = x$. Only (Y_n) is observed. Can we estimate θ and p from thes observations?

We have now to deal with a Hidden Markov Model (X_n, Y_n), i.e.
(i) (X_n) is a Markov chain
(ii) The conditional distribution of Y_n given $\{(X_i, Y_i), i = 1, \ldots, n-1, X_n\}$ only depends on X_n.

It follows from this definition that $(Y_n, n \geq 0)$ does not verify the Markov property.
Note that (X_n, Y_n) is a Markov chain with transition kernel,

$$\begin{aligned} \mathbb{P}(X_{n+1} = x', Y_{n+1} = y' | X_n = x, Y_n = y) &= \mathbb{P}(X_{n+1} = x', Y_{n+1} = y' | X_n = x) \\ &= Q_\theta(x, x') f_p(x', y'). \end{aligned}$$