UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXII

COMPOSITE LIKELIHOOD INFERENCE IN STATE SPACE MODELS

Direttore della Scuola: Ch.ma Prof. ssa ALESSANDRA SALVAN

Supervisore: Ch.mo Prof. MARCO FERRANTE

Dottoranda: NADIA FRIGO

1 Febbraio 2010

Acknowledgements

This thesis is the result of three years of hard work, with alternation of periods of enthusiasm and frustration. I would here like to express my thanks to the people who have been very helpful to me during this period.

I am deeply and sincerely grateful to my supervisor, Professor Marco Ferrante for encouraging me in moments of difficulty and for teaching me to seize opportunities and deal calmly the deadlines and commitments. His wide knowledge and his logical way of thinking have been of great value for me. His understanding and personal guidance have provided a good basis for the present thesis.

My warm thanks are due to Professor Christophe Andrieu, that I visited in Bristol, UK, who introduced me to the fields of filtering theory and composite likelihood. His detailed and constructive comments have been a remarkable support throughout this work.

I also wish to thank Professor Stuart Coles for his valuable advices and friendly help in the initial steps of my research.

I would like to gratefully acknowledge Professor Paolo Vidoni and Doctor Cristiano Varin who gave me interesting feedback and useful suggestions.

I owe my most sincere gratitude to Professor Alessandra Salvan, Chair of Ph.D. School in Statistics, for the perfect organization of the doctoral program and for the readiness shown in these three years.

I cannot forget in these acknowledgements my colleagues and friends of the Ph.D. program, Vanna, Nicola, Daniele, James, Laura, Susanna and Francesco for having shared with me the joy and pain of the doctoral school.

Special thanks go to my good friends, without whom life would be bleak. In particular, I thank Michele and Maria (listening to little Vera's songs was sometimes exactly what I needed after a day spent fighting with the computer), Laura, Ida, Andrea, Anna, Corrado, Stefano, Andrea, Genni and the list is very long, for the movies, happy hours, dinners, parties and concerts we enjoy together and for distracting me completely and joyously. Thanks to all of you, for making me laugh and for reminding me that there are more important things in life than a Ph.D. thesis.

I owe my loving thanks to my honey Antonio for his endless patience and ever-present support. He has lost a lot due to my research abroad. Without his encouragement and understanding it would have been impossible for me to finish this work. My special gratitude is due to my parents and my sister Michela for their loving support. Thanks to all of them, still apologizing for those periods in which I was intractable.

Contents

List of Figures

List of Tables

Abstract

In general state space models, where the computational effort required in the evaluation of the full likelihood function is infeasible, we analyze the problem of static parameter estimation based on composite likelihood functions, in particular pairwise and split data likelihood functions. We discuss consistency and efficiency properties of these estimators (related to the characteristics of the model) and the bias in stationary models where the invariant distribution is unknown.

We focus on numerical methods to compute estimates of the parameter describing a general state space model. We develop an on line Expectation- Maximization algorithm in order to obtain the maximum pairwise likelihood estimate in a general state space framework. We illustrate this method for a linear gaussian model and we extend it to make inference also in jump Markov linear systems. In this framework, some sampling procedures need to be developed to estimate the parameters of the model. In particular, we present an algorithm to sample from the latent discrete state Markov chain given the pairs of observations.

Riassunto

Nell'ambito di modelli state space, per i quali ricavare la funzione di verosimiglianza completa non è computazionalmente possibile, si è analizzato il problema della stima di parametri statici mediante funzioni di verosimiglianza composita, in particolare funzioni di verosimiglianza a coppie e a blocchi. L'interesse si è concentrato sullo studio delle proprietà di consistenza e di efficienza di tali stimatori (in relazione alle caratteristiche del processo stazionario sottostante il modello) nonchè su problemi di distorsione in modelli stazionari per i quali la distribuzione invariante non è nota.

Sono stati presi in esame metodi numerici per il calcolo delle stime dei parametri che descrivono un modello state space generale. Si è sviluppato un algoritmo Expectation- Maximization sequenziale per ottenere stime di massima verosimiglianza a coppie nel contesto di modelli state space generali. Tale metodo è illustrato per modelli lineari gaussiani e viene esteso per l'inferenza in sistemi lineari markoviani con salti. In questo contesto, è stato necessario sviluppare adeguate procedure di campionamento. In particolare, viene presentato un algoritmo per campionare dalla catena markoviana a stati discreti date le coppie di osservazioni.

Chapter 1

Introduction

1.1 Overview

State space models are a general class of time series capable of modeling dependent observations in a natural and interpretable way. They consist of a Markov process (called hidden/latent state process) not observed directly, but only through another process. When the parameter describing the model is known, sequential inference on the latent process is typically based on the sequence of joint posterior distributions, where each summarizes all the information collected about the latent process up to the current time. Sequential estimation of these distributions is achieved by *optimal filtering* recursions. Such recursions rarely admit a closed form expression, but it is possible to resort to efficient numerical approximations, e.g. Sequential Monte Carlo methods (SMC).

In most real-world scenarios, the parameter is unknow and needs to be estimated. Although apparently simpler than optimal filtering, the static parameter estimation problem has proved to be much more difficult: no closed form solutions are, in general, available, even for linear gaussian and finite state space hidden Markov models. A possible way to address this problem is based on SMC methods. There have been many attempts to develop elaborate sequential algorithms, but all of them suffer from a common intrinsic problem, namely *path degeneracy*. This phenomenon reflects a fundamental weakness of SMC methods: with limited resources, it is not possible to consistently estimate the sequence of posterior distributions at every instant time [Del Moral, 2004]. Direct application of SMC techniques is hence inappropriate for static parameter inference.

A different approach consists on developing an inferential procedure based on full likelihood function to compute point estimates from the data. Recently, some results on the consistency and asymptotic normality of the maximum likelihood estimator in state space models have been proved [Douc et al., 2004]. Anyway, when the latent process is continuous, the computational effort required in the evaluation of the full likelihood function is infeasible. Approximated solutions, based on Monte Carlo or numerical methods, have been considered, but none of the proposed solutions are completely satisfactory.

A possible way to overcome this problem is to replace the likelihood function by another function, easier to determine. In this direction, composite likelihood approaches have been suggested. The term composite likelihood indicates a likelihood type object formed by taking the product of individual component likelihoods, each of which corresponds to a marginal or conditional event. This is useful when the joint density is difficult to evaluate but computing likelihoods for some subsets of the data is possible, as in general state space models framework. This idea dates back probably to Besag [1974] even though the term composite likelihood was stated by Lindsay [1988].

All these topics are revised in Chapter 2.

1.2 Main contributions of the Thesis

In the present thesis we analyze the problem of static parameter estimation based on composite likelihood functions, in particular *pairwise* and *split data likelihood* functions. We discuss the asymptotic properties of the parameter estimators obtained by maximizing these functions in state space scenario in connection with stationary and ergodic properties of the processes involved. We develop also numerical methods to compute such estimates.

In Chapter 3, we take into account the *pairwise likelihood* function and we study its asymptotic properties. We discuss which kind of pairwise likelihood function is better to use among all possible choices for the weights. We analyze motivations that justify the preference of pairwise likelihood of order L, L being the maximum distance between the pairs, instead of pairwise likelihood with all the pairs. We study the asymptotic properties of the maximum pairwise likelihood estimator, related to the characteristic of the state space model. We prove the consistency of the maximum pairwise likelihood estimator of order L. In particular, we need that the joint process is an uniformly ergodic Markov chain. Moreover, we present an expression for a central limit theorem and we quantify the bias of the estimate in the case where the invariant distribution is unknown and it is substituted by a generic distribution. Our result confirms the intuition that the bias, introduced when using a generic distribution instead of the stationary distribution in the pairwise likelihood function, depends on how close the two distributions are and on the ergodic properties of the latent process. We suggest a possible way to choose a suitable approximation for the invariant distribution. In the case in which the invariant distribution is unknown, but transitions for the latent process are simple, the idea is to approximate the invariant distribution sampling from this transition kernel and to take advantage of the geometric ergodicity of the process.

If L is fixed, the use of pairwise likelihood of order L suggests that information about the parameter can be extracted from the dependence structure of the pairs of observations with a lag distance not greater than L. Usually it happens that the maximum pairwise likelihood estimators tend to lose efficiency, with respect to those based on full likelihood. Until now, no general results about evaluation of this gap are available. In Chapter 4, we empirically compare the efficiency between maximum pairwise likelihood and maximum full likelihood estimators as well as the efficiency between maximum split data likelihood (when blocks of observations are allowed to overlap) and maximum pairwise likelihood estimators. We prove that the loss of efficiency of the maximum split data likelihood estimator vanishes as L increases, while the variance of the maximum pairwise likelihood estimator decreases until a certain L^* and then it tends to increase. We suggest the existence of a ‘best lag’ L^*, in terms of variance of the maximum pairwise likelihood estimator.

We focus on numerical methods to compute estimates of the parameter describing a general state space model. We develop an on line Expectation- Maximization algorithm in order to obtain the maximum pairwise likelihood estimate in a general state space framework (Chapter 5). This algorithm increases the pairwise likelihood at each iteration step. We illustrate this method for a linear gaussian model, deriving the update equations in fairly explicit details. We modify standard Kalman filter recursions in order to take into account conditioning on pairs of observations instead of all observations. This simple example, where the invariant distribution is known, as well as the conditional distribution of the latent states given the pairs of observations, allows us to apply the idea of approximating the stationary distribution by sampling from the transition kernel. We give an empirical evidence of our bias theorem, i.e. starting from a generic distribution and sampling from the transition kernel reduces the bias in the estimates for each parameter in the model.

Chapter 6, is devoted to inference issues in jump Markov linear systems. We present an algorithm that generalizes what derived for a linear gaussian model. In this framework, some sampling procedures need to be developed to estimate the parameters of the model. In particular, we present an algorithm to sample from the latent discrete state Markov chain given a pair of observations.

Chapter 2

State space models

State space models are a general class of time series capable of modeling dependent observations in a natural and interpretable way. These models can be defined in the following form. For any parameter $\theta \in \Theta$, the hidden/latent state process $\{X_k; k \geq 1\} \subset \mathcal{X}^{\mathbb{N}}$ is a Markov process, characterized by its Markov transition probability distribution $f_\theta(x'|x)$, i.e. $X_1 \sim \nu$ and for $n \geq 1$,

$$X_{n+1}|(X_n = x) \sim f_\theta(\cdot|x). \tag{2.0.1}$$

The process $\{X_k; k \geq 1\}$ is observed, not directly, but through another process $\{Y_k; k \geq 1\} \subset \mathcal{Y}^{\mathbb{N}}$. The observations are assumed to be conditionally independent given $\{X_k; k \geq 1\}$, and their common marginal probability distribution is of the form $g_\theta(y|x)$, i.e. for $1 \leq n \leq m$,

$$Y_n|(X_1, \ldots, X_n = x, \ldots, X_m) \sim g_\theta(\cdot|x). \tag{2.0.2}$$

From now on, we will assume that the process $\{Z_k; k \geq 1\} = \{(X_k, Y_k); k \geq 1\}$ is stationary (in the strict sense) with joint distribution given by

$$p_\theta(x_{1:n}, y_{1:n}) = \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i), \tag{2.0.3}$$

where we denote by π_θ the marginal for $\{X_k; k \geq 1\}$ of this invariant distribution.

When the static parameter θ is known, sequential inference on the process

$\{X_k; k \geq 1\}$ is typically based on the sequence of joint posterior distributions $\{p_\theta(x_{1:n}|y_{1:n}); n \geq 1\}$, where each summarizes all the information collected about $X_{1:n}$ up to time n. *Optimal filtering* is concerned with the sequential estimation of these distributions which can be -at least conceptually- easily achieved using the following updating formula for $n \geq 2$

$$p_\theta(x_{1:n}|y_{1:n}) = \frac{g_\theta(y_n|x_n) f_\theta(x_n|x_{n-1})}{p_\theta(y_n|y_{1:n-1})} p_\theta(x_{1:n-1}|y_{1:n-1}), \qquad (2.0.4)$$

and $p_\theta(x_1|y_1) \propto g_\theta(y_1|x_1)\pi_\theta(x_1)$.

Although simple, the recursion formula (2.0.4) rarely admits a closed form expression: this is typically the case as soon as f_θ or g_θ are non- gaussian, or $\mathcal{X}$ is not a finite set. In such scenarios it is possible to resort to numerical approximations. Sequential Monte Carlo (SMC) methods (aka particle filters) are a class of numerical algorithms available to approximate $p_\theta(x_{1:n}|y_{1:n})$ sequentially in time. They have been recently proved to be efficient tools to propagate in time sample approximations of these distributions' marginals $p_\theta(x_{n-L+1:n}|y_{1:n})$ for a given integer $L > 0$ [Doucet et al., 2001]. This methodology is now well developed and the theory supporting this approach is also well established [Del Moral, 2004].

Example 2.0.1. *AR(1) model with additive observation noise*

$$\begin{aligned} X_{n+1} &= \phi X_n + W_n, \quad W_n \sim N(0, \tau^2) \\ Y_n &= X_n + V_n, \quad V_n \sim N(0, \sigma^2). \end{aligned}$$

In this case

$$f_\theta(x'|x) = N(\phi x, \tau^2) \quad \textit{and} \quad g_\theta(y|x) = N(x, \sigma^2),$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. The parameter vector is $\theta = (\phi, \tau^2, \sigma^2)$. For stationarity, $\theta \in (-1, 1) \times \mathbb{R}^+ \times \mathbb{R}^+$ and $\pi_\theta \sim N\left(0, \frac{\tau^2}{1-\phi^2}\right)$. This is an example in which a close form expression for (2.0.4) is available. When the state space model rests on linear and gaussian assumptions, the updating procedure corresponds to the Kalman filter, which gives recursively the mean and the variance of the gaussian filtering and prediction distributions at *each time n. For these reasons this model is also known as linear gaussian model.*

In most real-world scenarios the parameter is unknown and needs to be estimated. We assume that there is a 'true' parameter value θ^* generating the data $\{Y_k; k \geq 1\}$ and that this value is unknown. We focus here on the estimation of this static parameter. Although apparently simpler than the optimal filtering, the static parameter estimation problem has proved to be much more difficult; no closed form solutions are, in general, available, even for linear gaussian and finite state-space hidden Markov models (see Example 2.2.1).

2.1 SMC methods for static parameters

A possible way to address the static parameter estimation is based on SMC methods. We do not want here to review these methods in details, but simply to point out their intrinsic limitations which have fundamental practical consequences for the static parameter estimation problem. These limitations illustrate the complexity of static parameter estimation and motivate the approach developed in this thesis. Assuming that the static parameter θ is fixed for the time being, we describe the simplest SMC algorithm available to approximate $\{p_\theta(x_{1:n}, Y_{1:n}); n \geq 1\}$ sequentially. More elaborate algorithms are reviewed in Doucet et al. [2001], but crucially all such SMC algorithms suffer from a common problem, namely *path degeneracy*, as explained below.

2.1.1 Sampling Importance Resampling (SIR)

Assume that at time $n-1$, a collection of $N (N \gg 1)$ random samples named particles $\{\hat{X}^{(i)}_{1:n-1}, i = 1, \ldots, N\}$ distributed approximately according to $p_\theta(x_{1:n-1}|Y_{1:n-1})$ is available. The empirical distribution

$$\hat{p}^N_\theta(dx_{1:n-1}|Y_{1:n-1}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{X}^{(i)}_{1:n-1}}(dx_{1:n-1}) \tag{2.1.1}$$

is an approximation of $p_\theta(x_{1:n-1}|Y_{1:n-1})$, where $\delta_{x_0}(dx)$ represents the Dirac delta mass function centered at x_0. Now, at time n, one wishes to produce N particles which will define an approximation $\hat{p}^N_\theta(dx_{1:n}|Y_{1:n})$ of $p^N_\theta(dx_{1:n}|Y_{1:n})$. A simple method to achieve this consists of setting $\tilde{X}^{(i)}_{1:n-1} = \hat{X}^{(i)}_{1:n-1}$ and then sampling, for example, $\tilde{X}^{(i)}_n \sim f_\theta\left(\cdot|\tilde{X}^{(i)}_{n-1}\right)$. The resulting empirical distribution of the particles $\{\tilde{X}^{(i)}_{1:n}; i = 1, \ldots, N\}$ is an approximation of the joint density

$$p_\theta(x_{1:n-1}|Y_{1:n-1})\, f_\theta(x_n|x_{n-1}).$$

We correct for the discrepancy between this density and the target $p_\theta(x_{1:n}|Y_{1:n})$ using importance sampling. This yields the following approximation of $p_\theta(x_{1:n}|Y_{1:n})$

$$\hat{p}^N_\theta(dx_{1:n}|Y_{1:n}) = \sum_{i=1}^{N} \omega^{(i)}_n \delta_{\tilde{X}^{(i)}_{1:n}}(dx_{1:n}),$$

where each particle $\tilde{X}^{(i)}_{1:n}$ has now a weight $\omega^{(i)}_n$ given by

$$\omega^{(i)}_n \propto g_\theta\left(Y_n|\tilde{X}^{(i)}_{1:n}\right) \quad \text{and} \quad \sum_{i=1}^{N} \omega^{(i)}_n = 1.$$

To obtain an unweighted approximation of $p_\theta(x_{1:n}|Y_{1:n})$ of the form (2.1.1), we resample particles $\{\tilde{X}^{(i)}_{1:n}; i = 1, \ldots, N\}$ according to probabilities proportional to their weights $\{\omega^{(i)}_n; i = 1, \ldots, N\}$. The underlying idea is to get rid of particles with small weights and multiply particles which are in the region of high probability masses (see Figure 2.1.1). Many such resampling schemes have been proposed in the literature [Doucet et al., 2001].

2.1.2 Limitation of SMC Methods

Under relatively weak assumptions on f_θ and g_θ, it can be proved that the resulting set of empirical posterior distributions $\{\hat{p}^N_\theta(dx_{1:n}|Y_{1:n})\}$ converges toward the true posterior as N goes to infinity. More precisely, it can be easily shown that for any $n \geq 1$ and any bounded test function $\varphi_n : \mathcal{X}^n \rightarrow \mathbb{R}$ there exists some constant

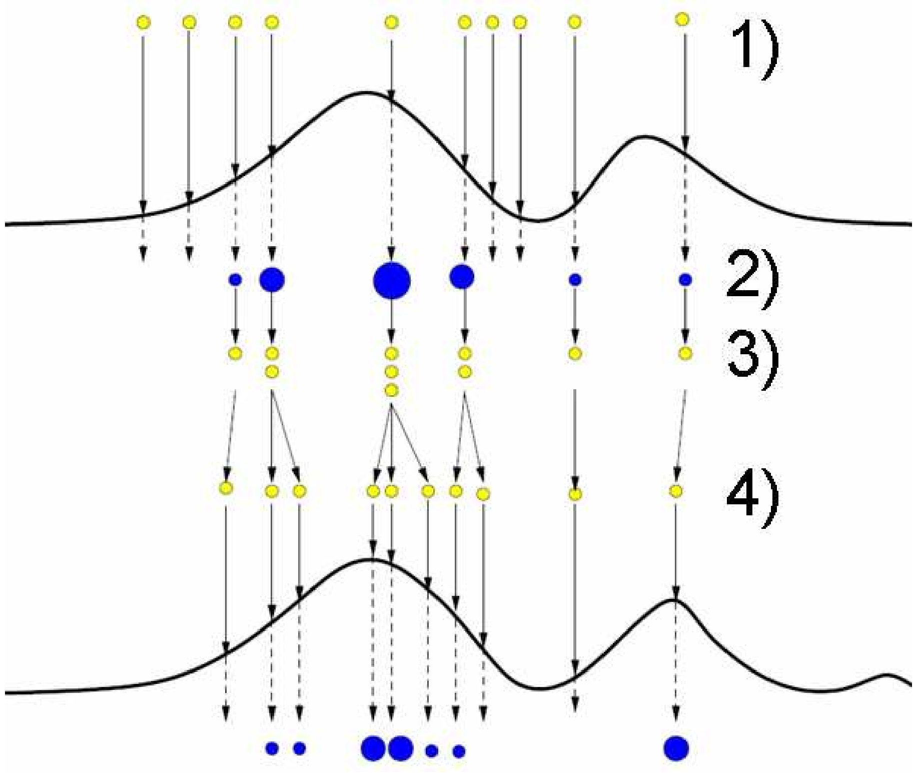

Figure 2.1.1: SMC method via SIR. Starting at time n with a sample distributed approximately according to $p_\theta(x_{1:n}|Y_{1:n-1})$ (1), for each particle we compute the weights using information at time n, getting a weighted approximation of $p_\theta(x_{1:n}|Y_{1:n})$ (2). We select those particles that are suitable to get the unweighted approximation (3). Then we propagate the particles according to the transition kernel to get an unweighted approximation of $p_\theta(x_{1:n+1}|Y_{1:n})$ (4).

$C_{\theta,n}(\varphi_n) < \infty$ such that for any $N \geq 1$

$$\mathbb{E}\left[\left(\int_{\mathcal{X}^n} \varphi_n(x_{1:n})(p_\theta(dx_{1:n}|Y_{1:n}) - \hat{p}_\theta^N(dx_{1:n}|Y_{1:n}))\right)^2\right] \leq \frac{C_{\theta,n}(\varphi_n)}{N}, \qquad (2.1.2)$$

where the expectation is with respect to the particles realizations. A much wider range of results is in fact available: L_p convergence, central limit theorems, large deviation. A complete treatment can be found in Del Moral [2004]. Although at first sight reassuring, Equation (2.1.2) is practically useless since the bound $C_{\theta,n}(\varphi_n)$ typically grows polynomially or exponentially with n, and reflects a fundamental weakness of SMC methods: with limited resources, i.e. N fixed and finite, it is not possible to consistently estimate the sequence of distributions $\{p_\theta(x_{1:n}|Y_{1:n})\}$.

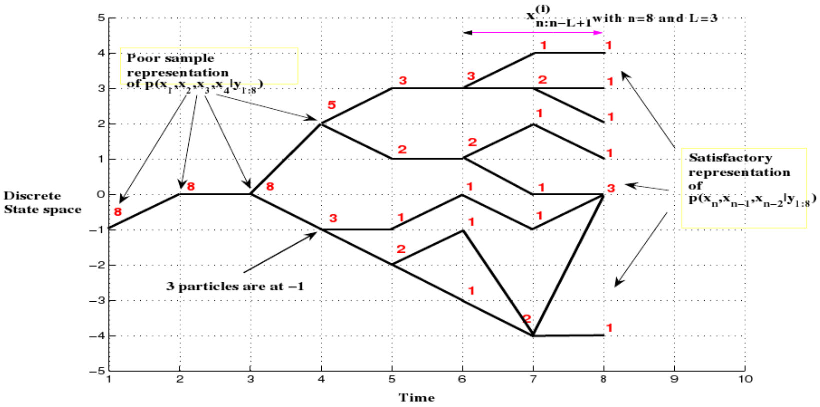

Figure 2.1.2: Realistic sequential methods suffer from path degeneracy.

We report here a simple example proposed in Andrieu et al. [2007] that illustrates the underlying phenomenon which explains the growth of $C_{\theta,n}(\varphi_n)$. The tree in Figure 2.1.2 represents a realization of the paths $\{\hat{X}^{(i)}_{1:n}; i = 1,\ldots,N\}$ of $N = 8$ particles up to time $n = 8$ for a system for which the state space is $\mathcal{X} = \{-5,-4,\ldots,0,1,5\}$. The numbers at each node represent the number of particles that effectively pass through it. This realization of the particle process is representative of what is generally observed in more complex scenarios: the paths tend to *coalesce* as we follow the paths backward in time. As a result, whereas $\{\hat{X}^{(i)}_8; i = 1,\ldots,N\}$ and $\{\hat{X}^{(i)}_7; i = 1,\ldots,N\}$ have a good coverage of $\mathcal{X}$, which will result in a good representation of $p(x_8|Y_{1:8})$ and $p(x_7|Y_{1:8})$, the sample representation deteriorates as we go back in time, resulting in poor approximation of $p(x_{1:4}|Y_{1:8})$, i.e. even if the true $p(x_{1:4}|Y_{1:8})$ is not degenerate, the sample representation is degenerate. This *coalescence* phenomenon is the result of the resampling stage and has long been observed [Gordon et al., 1993]. In fact, for a fixed time index k, there almost surely exists a finite random time n such that the particles $\{\hat{X}^{(i)}_{1:n}; i = 1,\ldots,N\}$ have all similar paths up to time k; i.e. $\{\hat{X}^{(i)}_{1:k}\} = \{\hat{X}^{(j)}_{1:k}\}$ for all $i, j \in \{1,\ldots,N\}$. In the light of this toy example, it should thus come as non surprise if it is impossible to bound $C_{\theta,n}(\varphi_n)$ uniformly over time.

A cure to the coalescence phenomenon could consist of rejuvenating the paths by sampling new paths $\{\hat{X}^{(i)}_{1:n-L}; i = 1, \ldots, N\}$ according to $p_\theta(x_{1:n-L}|Y_{1:n})$ for some $L > 0$. However, this would require a growing (in time) computational budget per iteration, which is unrealistic in a sequential framework or for large time series. Another potential fix could consist of stopping resampling of the past of the paths, say $\{\hat{X}^{(i)}_{1:n-L}; i = 1, \ldots, N\}$ for some integer L, from time n onwards, provided that the identity $p_\theta(x_{1:n-L}|Y_{1:n}) \approx p_\theta(x_{1:n-L}|Y_{1:n+1})$ holds [Kitagawa and Sato, 2001]. However it is difficult to quantify L practically. More crucially, this parameter might directly depend on the unknown static parameter that we are trying to estimate. As we shall see, this inability of SMC methods to approximate, with a constant computational budget per iteration, the joint distribution $p_\theta(x_{1:n}|Y_{1:n})$ makes SMC parameter estimation algorithms inappropriate. The success of SMC methods lies in the fact that results of the following form can be obtained under the relatively general assumption detailed in Del Moral [2004]. Let $L > 0$ be an integer and let $\varphi_L : \mathcal{X}^L \to \mathbb{R}$ be a bounded test function, then there exists some constant $D_{\theta,L}(\varphi_L) < \infty$ such that for any $n \geq 1$,

$$\mathbb{E}\left[\left(\int_{\mathcal{X}^L} \varphi_n(x_{n-L+1:n})(p_\theta(dx_{n-L+1:n}|Y_{1:n}) - \hat{p}^N_\theta(dx_{n-L+1:n}|Y_{1:n}))\right)^2\right] \leq \frac{D_{\theta,L}(\varphi_L)}{N}.$$

Clearly, in the light of the discussion above, this result can only hold when the particles $\{\hat{X}^{(i)}_{n-L+1:n}; i = 1, \ldots, N\}$ do not depend too heavily on their 'far past' $\{\hat{X}^{(i)}_{1:k}; i = 1, \ldots, N\}$ for $k \ll n - L + 1$, since these paths form a poor representation of $p_\theta(x_{1:k}|Y_{1:n})$. In other words, it is required that a property of the type $p_\theta(x_{n-L+1:n}|Y_{1:n}, x_{1:k}) \approx p_\theta(x_{n-L+1:n}|Y_{1:n})$ holds.

In summary, for a fixed computational budget per time instant, SMC methods can not properly approximate joint distribution sequences of the form $\{p_\theta(x_{1:n}|Y_{1:n}); n \geq 1\}$ sequentially in time because of the paths' coalescence phenomenon: as we shall see this is what makes the direct application of SMC techniques inappropriate for static parameter inference. However, under ergodic assumption, for a given lag $L > 0$, SMC methods can consistently approximate sequences of distributions $\{p_\theta(x_{n-L+1:n}|Y_{1:n}); n \geq 1\}$ for a fixed number N of particles.

2.1.3 Difficulties with Static Parameters

Here we briefly review recent SMC-based static parameter estimation techniques proposed in the literature. We can essentially classify most proposed approaches into three categories. The first approach consists of setting a prior on θ, incorporating θ in the state and applying standard SMC algorithms to the joint state $Z_n = (\theta_n, X_n)$, where the prior transition probability on $\{\theta_n; n \geq 1\}$ is $\theta_{n+1} \sim \delta_{\theta_n}(d\theta_{n+1})$, $\delta_{\theta_n}(\cdot)$ being the Dirac delta density function. However, in the light of the discussion above, the use of standard SMC methods to estimate the distributions $\{p(x_{1:n}, \theta|Y_{1:n}); n \geq 1\}$ (and thus $\{p(\theta|Y_{1:n}); n \geq 1\}$) is bound to fail, especially due to the lack of ergodicity of the process $\{Z_n; n \geq 1\}$. If we were to apply the generic SMC algorithm as described earlier, the parameter space would only be explored at the initialization of the algorithm as the transition probability of the Markov process $\{Z_n; n \geq 1\}$ includes a Dirac delta mass for the component θ. Consequently, after a few iterations, the marginal posterior distribution of the parameter is typically approximated by a single Dirac delta function, which corresponds to one of the initial values sampled from the prior distribution at time 1. In fact the particle deplation phenomenon has been quantified in the simpler scenario where no latent variable $x_{1:n}$ is present [Chopin, 2004]: it can be shown that the variance of functionals of θ grows polynomially fast in time and exponentially fast with the dimension of θ. This problem was historically identified very early on by SMC users [Gordon et al., 1993], and in order to limit it, various approximation strategies have been proposed.

The second, pragmatic, approach consist of modifying the state- space model, so that θ is not static anymore. In this scenario we consider the extended state $Z_n = (\theta_n, X_n)$ where, for example, $\theta_{n+1}|\theta_n \sim N(\theta_n, \sigma_\theta^2)$ for a typically small σ_θ^2 [Kitagawa, 1998]. However, the choice of σ_θ^2 is difficult, leading to a trade off between accuracy of the proxy model and speed of convergence of the particle filter. Historically it is related to a more general technique which was proposed in order to introduce diversity in the system of particles using a kernel approximation of the empirical distribution [Gordon et al., 1993]. Liu and West [2001] establish some form of duality between this approach and the modified system dynamic approach for some particular cases of interest. However, the selection of an appropriate kernel remains a difficult point.

In the third approach we also set a prior on θ and SMC is used to estimate the joint posterior $p(\theta, x_{1:n}|Y_{1:n})$. Contrary from the first approach, here diversity among particles in the parameter space is introduced using MCMC steps of invariant distribution $p(\theta|Y_{1:n}, x_{1:n})$. This is certainly more elegant than the second approach, as the model of interest is not artificially altered. This algorithm takes a simple form when $p(Y_{1:n}|x_{1:n}, \theta)$ can be summarized by a set of low dimensional sufficient statistics [Andrieu et al., 1999, Fernhead, 2002, Gilks and Berzuini, 2001, Storvik, 2002]. However, as noted for example in Andrieu et al. [1999] and in light of the limitations of SMC methods outlined previously, this approach might be unreliable. The problem is that the SMC estimates of the sufficient statistics, necessary to perform the MCMC updates, degrade as n increases because they are based on the approximation of the joint distribution $p_{\theta^*}(x_{1:n}|Y_{1:n})$. In general, it happens that initially the SMC estimate displays good performance but performs very poorly as n increases: this stems from the fact that the joint distributions $\{p_{\theta^*}(x_{1:n}|Y_{1:n}); n \geq 1\}$ can not be consistently estimated over time. What we usually observe using this approach is that, at first, the parameter seems to converge toward the correct region but then drifts away as the sufficient statistics used in the MCMC update are not properly estimated. The practical problem is that of assessing whether the algorithm is stable and has converged or not.

The development of an algorithm to approximate the sequence $\{p(\theta|Y_{1:n})\}$ with a fixed precision and fixed computational efforts at any time n seem to us to remain an open question.

2.2 Point estimation methods

A different approach consists on the estimation of the unknown parameter θ^* in a frequentist way, developing an inferential procedure based on likelihood quantities to compute point estimates of θ^* from $\{Y_k; k \geq 1\}$. The aim is to produce a point estimate of θ^* rather than a series of estimates of the posterior distributions $\{p(\theta, Y_{1:n}); n \geq 1\}$. As a result no particle method is required in the parameter space, and it should also be pointed out that SMC methods in the state- space $\mathcal{X}$ are, in general, also not necessary.

2.2.1 Full likelihood Inference

The most natural approach of point estimate consists of maximizing the series of likelihoods $\{p_\theta(Y_{1:n}); n \geq 1\}$. With our notation, the likelihood for a sequence of observations $y_1, \ldots, y_n$ is

$$L(\theta; y_{1:n}) = p_\theta(y_{1:n}) = \int_{\mathcal{X}^n} \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i) dx_{1:n}, \qquad (2.2.1)$$

which is simply obtained by taking into account the dependence structure characterizing the model. In many situations π_θ, i.e. the stationary distribution, is not known analytically. We denote with $p_\theta(y_{1:n}|\mu)$ the joint distribution of the observations when $X_1 \sim \mu$, obtained by substituting μ for the true invariant distribution π_θ in (2.2.1). With this notation, $p_\theta(y_{1:n}) := p_\theta(y_{1:n}|\pi_\theta)$. An alternative equivalent representation of the likelihood function is

$$L(\theta; y_{1:n}) = \prod_{i=1}^{n} p_\theta(y_i|y_{1:i-1}) = \prod_{i=1}^{n} \int_{\mathcal{X}} p_\theta(x_i|y_{1:i-1}) g_\theta(y_i|x_i) dx_i,$$

where the prediction densities are obtained recursively from

$$\begin{aligned} p_\theta(x_i|y_{1:i-1}) &= \int_{\mathcal{X}} f_\theta(x_i|x_{i-1}) p_\theta(x_{i-1}|y_{1:i-1}) dx_{i-1} & (2.2.2) \\ p_\theta(x_{i-1}|y_{1:i-1}) &= \frac{p_\theta(x_{i-1}|y_{1:i-2}) g_\theta(y_{i-1}|x_{i-1})}{\int_{\mathcal{X}} p_\theta(x_{i-1}|y_{1:i-2}) g_\theta(y_{i-1}|x_{i-1}) dx_{i-1}}. & (2.2.3) \end{aligned}$$

Formulae (2.2.2) and (2.2.3) specify the well-known recursion which enables the computation, at each time $n \geq 1$, of the filtering and the prediction densities.

Example 2.2.1. *Hidden Markov Model*
If the latent states $\{X_k; k \geq 1\}$ is a finite state Markov chain on $\{1, \ldots, K\}$, the state space model is usually called hidden Markov model. Denoting by $A_\theta = [\alpha_{ij}]$ the transition probability matrix, the likelihood function for the observations

$y_1, \ldots, y_n$ is

$$p_\theta(y_{1:n}) = \sum_{x_1=1}^{K} \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha^{(1)}_{x_1} g_\theta(y_1|x_1) \prod_{i=2}^{n} \alpha_{x_{i-1}x_i} g_\theta(y_i|x_i).$$

The initial probability distribution $\alpha^{(1)}$ used in the definition above is not necessarily the stationary probability distribution for the stochastic matrix A_θ, but any probability vector $\alpha^{(1)}$ with strictly positive elements. The consistency of the maximum likelihood estimators does not depend on the choice of $\alpha^{(1)}$ [Leroux, 1992]. In this case the likelihood function can be expressed as

$$p_\theta(y_{1:n}) = \alpha^{(1)} \left(\prod_{i=1}^{K} G_\theta(y_i) A_\theta \right) \mathbf{1}, \tag{2.2.4}$$

where $G_\theta(y) = diag\{g_\theta(y|x)\}$ and $\mathbf{1}$ is a $K \times 1$ vector of ones. It is clear that (2.2.4) is essentially a product of matrices and is hence easily evaluated. It can be maximized over θ using standard numerical optimization procedures or using the EM algorithm.

Recently, some results on the consistency and asymptotic normality of the maximum likelihood estimator (MLE) can be found in Douc et al. [2004] (see also the references therein). Their results allow one to consider the case where π_θ, and hence the true likelihood, is unknown. The technique relies primarily on the forgetting properties of the filter, uniformly in θ. Anyway, when $\{X_k; k \geq 1\}$ is continuous, evaluation of the full likelihood requires an integration over an n-dimensional space. This task is insurmountable for typical values of n and exact methods for computing and maximizing the likelihood function are usually not feasible. Approximated solutions, based on Monte Carlo or numerical methods, have been considered, but none of the proposed solutions are completely satisfactory. Markov Chain Monte Carlo (MCMC) methods are usually difficult to implement while Particle Filters (PF) are well suited but suffer from the well known degeneracy problem. A possible way to overcome this problem is to replace the likelihood by another function, easier to determine. Any function which (asymptotically) has its maximum at the true parameter point is a potential candidate. In this direction composite likelihood approaches have been suggested.

2.2.2 Composite likelihood Inference

Even if the full likelihood approach is the most natural and leads to an efficient estimation of the parameter, the computational effort required in the evaluation and maximization of the function suggests to develop new procedures in order to reduce the computational burden. In this way it is possible to fit highly structured statistical models, even when the use of standard likelihood methods is not practically possible. In the sequel we focus on composite likelihood.

The term composite likelihood indicates a likelihood type object formed by taking the product of individual component likelihoods, each of which corresponds to a marginal or conditional event. This is useful when the joint density is difficult to evaluate but computing likelihoods for some subsets of the data is possible, as in general state space models framework. This idea dates back probably to Besag [1974] even though the term composite likelihood was stated by Lindsay [1988].
Given the observations $y_{1:n}$, a composite likelihood is defined by specifying a set of K marginal or conditional events $A_k(y_{1:n}), k = 1, \ldots, K$, with likelihood given by $L_k(\theta; y_{1:n}) = L(\theta; A_k(y_{1:n}))$. Then, the composite likelihood is obtained by composing these likelihood objects and it corresponds to

$$L_C(\theta; y_{1:n}) = \prod_{k=1}^{K} L_k(\theta; y_{1:n})^{\omega_k},$$

with ω_k suitable non-negative weights. The composite loglikelihood is

$$l_C(\theta; y_{1:n}) = \sum_{k=1}^{K} \omega_k l_k(\theta; y_{1:n}),$$

with $l_k(\theta; y_{1:n}) = logL_k(\theta; y_{1:n})$.
This class contains, and thus generalizes, the usual ordinary likelihood, as well as many other interesting alternatives. We can group composite likelihoods into two general classes [Varin and Vidoni, 2005]: omission methods and composite marginal likelihoods.

Omission methods

The first group consists of ‘omission methods’, since composite likelihoods are obtained by removing complicated terms that are not very informative on the parameter of interest in such a way that the loss of efficiency may be tolerated. Examples include the Besag pseudolikelihood [Besag, 1974, 1977], introduced for making inference in spatial models before the advent of modern Monte Carlo methods. It consists on a composite likelihood constructed from conditional neighborhood densities. This is a quite natural suggestion since in the context of Markov Random Fields we might assume that the conditional distribution at site i depends only upon the values at those sites which are, in some sense, in the proximity of site i. Using this fact, he considered

$$L_C(\theta) = \prod_{i=1}^{n} p_\theta(y_i|y_j, j \in N(i)),$$

where p_θ is defined in (2.0.3) and $N(i)$ denotes some neighborhood of the i-th site. Another example is the m-th order likelihood for stationary processes [Azzalini, 1983], motivated by the fact that the exact likelihood in this framework can be written as a product whose i-th term is the probability density function of the corresponding sample element conditional on all previous observations. This suggests to replace the conditioning on all previous observations by only the m-th most recent ones, for some $m \geq 0$, leading to

$$L_C(\theta) = \prod_{i=1}^{n} p_\theta(y_i|y_{i-m:i-1}).$$

Composite marginal likelihoods

The other group contains composite likelihoods constructed from marginal densities [Cox and Reid, 2004]. Typical attention is paid to compositions of low-dimensional marginals, since their computation involves usually lower dimensional integrals. This is the case of the *pairwise likelihood* (PL) [Le Cessie and

Van Houwelingen, 1994],

$$L_{P,\omega}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_\theta(y_i, y_j)^{\omega_{ij}}, \tag{2.2.5}$$

where $\omega_{ij}, i = 1, \dots, n-1, j = i+1, \dots, n$ are suitable non-negative weights, or of the *split data likelihood* (SDL) proposed by Ryden [1994] as an alternative to maximum likelihood for inference in hidden Markov models. This is a composite likelihood constructed by splitting the $n = mL$ observations into m groups of fixed size L and assuming these groups are independent

$$L_{SD}(\theta; y_{1:n}) = \prod_{i=1}^{m} p_\theta(y_{L(i-1)+1:iL}).$$

In the SDL framework, it is also possible to consider overlapping blocks of the form $(Y_{1:L}, Y_{2:L+1}, \dots, Y_{n-L+1:n})$. In this case we define

$$\begin{aligned} L_{SD}^{(ov)}(\theta; y_{1:n}) &= \prod_{i=1}^{n-L+1} p_\theta(y_{i:L+i-1}) \\ &= \prod_{i=1}^{L(m-1)+1} p_\theta(y_{i:L+i-1}). \end{aligned} \tag{2.2.6}$$

In our study we take into account PL and SDL and we discuss the asymptotic properties of the parameter estimators obtained by maximizing these functions in the state space scenario.

Chapter 3

Pairwise likelihood inference in State space models

In this chapter we consider the pairwise likelihood function and the asymptotic properties of the parameter estimator obtained by maximizing this function in state space scenario. We discuss which kind of pairwise likelihood function is better to use among all possible choices for the weights and we prove the consistency of the pairwise likelihood estimator of order L. Moreover, we present an expression for a central limit theorem (if such theorem exists) and we quantify the bias of the estimate in the case where the invariant distribution is unknown and it is substituted by a generic distribution. Some comments about the efficiency problems are also suggested.

3.1 Different choices for the weights

Starting from (2.2.5), a suitable choice for the weights allows one to consider different types of PL. We shall concentrate on the PL that takes into account all the $n(n-1)/2$ pairs (obtained choosing $\omega_{ij} = 1, \quad \forall i = 1, \dots, n-1, \, j = i+1, \dots, n$), that is

$$L_P(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_\theta(y_i, y_j) \tag{3.1.1}$$

and on the so called L-th order PL, which is based on all the pairs of observations with a lag distance not greater than $L \in \{1, \dots, n-1\}$, that is

$$L_P^{(L)}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i, y_j).$$

Note that $L_P^{(n-1)}(\theta; y_{1:n})$ corresponds to (3.1.1). Given the dependence structure of the model (2.0.1, 2.0.2), for every $i = 1, \dots, n-1$, $j = i+1 \dots, n$

$$\begin{aligned} p_\theta(y_i, y_j) &= \int_{\mathcal{X}^{j-i+1}} p_\theta(y_i, y_j, x_{i:j}) dx_{i:j} \\ &= \int_{\mathcal{X}^{j-i+1}} \pi_\theta(x_i) g_\theta(y_i|x_i) \left[\prod_{k=i+1}^{j} f_\theta(x_k|x_{k-1}) \right] g_\theta(y_j|x_j) dx_{i:j}. \end{aligned} \tag{3.1.2}$$

The numerical computation of (3.1.2) involves in general a $(j-i+1)$- dimensional integral. If $j-i$ is bounded by a constant that does not depend on n, the computation is likely easier compared to the full likelihood approach. In the case of pairwise likelihood with all the pairs, the integral dimension increases with n, so its evaluation might be still infeasible, depending on the structure of $f_\theta(\cdot|x)$. This is one of the motivations why people usually do not work with pairwise likelihood with all the pairs but prefer using pairwise likelihood of order L, for some $L \geq 1$.

Moreover, even if the computation of (3.1.1) were feasible, a theoretic issue comes up when we consider all the pairs. If the process has good properties and the invariant distribution is known, we expect that the normalized log pairwise likelihood

$$l_P(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=i+1}^{n} log[p_\theta(y_i, y_j)] \tag{3.1.3}$$

will be well approximated by

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=m_n+i+1}^{n} log[p_\theta(y_i) p_\theta(y_j)] \tag{3.1.4}$$

for n large enough, where m_n is chosen in such a way that, for every i, $m_n/(n-i)$

and $\frac{m_n log n}{n}$ go to zero as n goes to infinity.

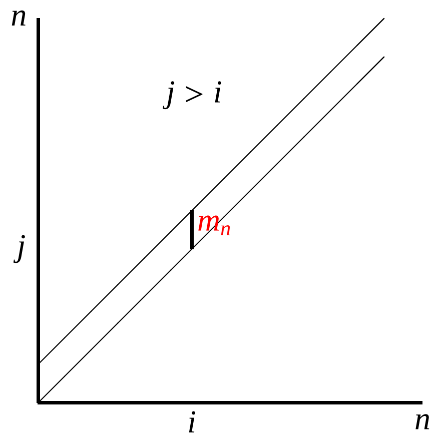

Figure 3.1.1: Heuristic proof of (3.1.4): In the likelihood (3.1.3) we are considering all the pairs (y_i, y_j) with $j > i$, i.e. we are above the bisector. The width m_n depends on the sample size n but it is constant for every i and j. Inside the stripe, observations are close. If n grows there are more pairs that are far apart than pairs that are close. Given the ergodicity of the process, the pairs that are far away act as they were independent, while the contribution to the likelihood of the the pairs that are close vanishes.

Roughly speaking, (3.1.4) tells us that if n grows there are more pairs that are far apart than pairs that are close and, if the process is ergodic, the pairs that are far away act as they were independent (see Figure 3.1.1 for an heuristic proof). In this case it is clear that all the information about the dependence structure of the model are lost, since only the marginal density is taken into account. For these reasons from here on we will concentrate on pairwise likelihood of order L.
More precisely, we prove the following theorem

Theorem 3.1.1. *Under the Assumptions* $(A1)$ *and* $(A2)$ *defined in Appendix B.2*

$$l_P(\theta; y_{1:n}) \approx \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)]$$

for n large enough, where, for every i, $m_n/(n-i)$ *and* $\frac{m_n log n}{n}$ *go to zero as n goes to infinity.*

Proof. By definition (3.1.3),

$$l_P(\theta; y_{1:n}) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\left[\sum_{j=i+1}^{m_n+i} log[p_\theta(y_i, y_j)] + \sum_{j=m_n+i+1}^{n} log[p_\theta(y_i, y_j)]\right],$$

where for every i, $m_n/(n-i)$ goes to zero as n goes to infinity to ensure that m_n does not grow 'too much' compared to n and hence the second sum makes sense. We concentrate first in the term

$$L_1(n, m_n) := \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=i+1}^{m_n+i} log[p_\theta(y_i, y_j)].$$

We have that

$$\begin{aligned}|L_1(n, m_n)| &\leq \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=i+1}^{m_n+i} |log[p_\theta(y_i, y_j)]| \\ &\leq \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{Cm_n}{n-i},\end{aligned}$$

with $C \in (0, +\infty)$. The result above follows from Assumption (A2) which ensure that $p_\theta(y_i, y_j)$ is bounded away from zero for every i, j and from the identity (B.4.1). Now

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{Cm_n}{n-i} &= \frac{Cm_n}{n-1}\sum_{i=1}^{n-1}\frac{1}{i} \\ &\approx \frac{Cm_n}{n-1}(log[n-1] + \gamma),\end{aligned}$$

where γ is the Euler constant. For our purpose, the term $L_1(n, m_n)$ has to go to zero in order to conclude that the contribution to the log pairwise likelihood of the pairs with lag distance not greater than m_n vanishes as n goes to infinity. To reach this, we need to choose m_n in such a way that $\frac{m_n logn}{n}$ goes to zero as n goes to infinity. Note that this condition holds, for example, when m_n is a constant.

We look now at

$$L_2(n, m_n) := \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i, y_j)].$$

We can rewrite $L_2(n, m_n)$ as

$$\begin{aligned}L_2(n, m_n) =&\frac{1}{n-1}\sum_{i=1}^{n-1}\Big[\frac{1}{n-i}\sum_{j=m_n+i+1}^{n}\Big(log[p_\theta(y_i, y_j)] - log[p_\theta(y_i)p_\theta(y_j)]\Big) + \\ &+\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)]\Big].\end{aligned}$$

By ergodic properties, there exist constants $\tilde{C} \in (0, +\infty)$ and $\rho \in [0, 1)$ such that, for every i, j

$$|p_\theta(y_i, y_j) - p_\theta(y_i)p_\theta(y_j)| \le \tilde{C}\rho^{j-i}.$$

Using again identity (B.4.1), the absolute value of first term in $L_2(n, m_n)$ satisfies

$$\begin{aligned}&\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n}(log[p_\theta(y_i, y_j)] - log[p_\theta(y_i)p_\theta(y_j)])\right| \\ &\le \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n}|log[p_\theta(y_i, y_j)] - log[p_\theta(y_i)p_\theta(y_j)]| \\ &\le \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} C\rho^{j-i} \le \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^{m_n} - \rho^{n-i}}{n-i} \\ &= \frac{C\rho^{m_n}}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{1}{i} - \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^i}{i},\end{aligned}$$

for a suitable constant $C \in (0, +\infty)$. For n large enough

$$\frac{C\rho^{m_n}}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{1}{i} \approx \frac{C\rho^{m_n}}{(1-\rho)(n-1)}(log[n-1] + \gamma) \overset{n\to+\infty}{\to} 0,$$

since ρ is a constant less than one. On the other hand

$$\frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^i}{i} \le \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\rho^i$$
$$= \frac{C}{(1-\rho)(n-1)}\left(\frac{1-\rho^n}{1-\rho}-1\right) \overset{n\to+\infty}{\to} 0.$$

We have that

$$L_2(n,m_n) \approx \frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)]$$

for n large enough and combining this with the result about $L_1(n,m_n)$, we are able to conclude that

$$l_P(\theta;y_{1:n}) \approx \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)].$$

□

3.2 Maximum pairwise likelihood of order L when π_θ is known

In this section, we study the properties of the estimator obtained by maximizing with respect to θ the pairwise likelihood function of order L, defined as

$$L_P^{(L)}(\theta;y_{1:n}) = \prod_{i=1}^{n-1}\prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i,y_j). \qquad (3.2.1)$$

We denote by $\hat{\theta}_P^{(L)}$ any global maximum point of $L_P^{(L)}(\theta;y_{1:n})$. Let us consider the pairwise likelihood in (3.2.1) where $p_\theta(y_i,y_j)$ is defined by (3.1.2) and $L \ge 1$ is a fixed constant (we now suppose that π_θ is known). In order to study the properties of $\hat{\theta}_P^{(L)}$ we need to point out the asymptotic behavior of the normalized log likelihood

$$l_P^{(L)}(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\frac{1}{L} \sum_{j=i+1}^{min\{i+L,n\}} log[p_\theta(y_i, y_j)] \right] \qquad (3.2.2)$$

as n goes to infinity. Since $L^{-1} \sum_{j=i+1}^{min\{i+L,n\}} log[p_\theta(y_i, y_j)]$ is a function of the observations $(y_i, \dots, y_{i+L})$ (let us denote this function as φ), under suitable ergodic assumptions

$$\begin{aligned} & \frac{1}{n-1} \sum_{i=1}^{n-1} \varphi(y_i, \dots, y_{i+L}) \overset{n\to+\infty}{\to} \mathbb{E}_{\theta^*}[\varphi(Y_1, \dots, Y_{L+1})] = \\ = & \int_{\mathcal{Y}^{L+1}} \varphi(y_1, \dots, y_{L+1}) p_{\theta^*}(y_{1:L+1}) dy_{1:L+1} \\ = & \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} log[p_\theta(y_1, y_j)] p_{\theta^*}(y_1, y_j) dy_1 dy_j, \qquad (3.2.3) \end{aligned}$$

where $\mathbb{E}_{\theta^*}[\cdot]$ is the expectation associated to the stationary process $\{Z_k; k \geq 1\}$ generated by the model defined in (2.0.1) and (2.0.2) for $\theta = \theta^* \in \Theta$.
Hence

$$\lim_{n\to+\infty} l_P^{(L)}(\theta; y_{1:n}) = l_P^{(L)}(\theta),$$

where $l_P^{(L)}(\theta)$ is defined by (3.2.3). With appropriate conditions, it can be shown that the set of parameters maximizing $l_P^{(L)}(\theta)$ includes the true parameter and hence the L-th order PL is an objective function that, when maximized, leads to a reasonable estimator of the parameter. This follows from the fact that maximizing $l_P^{(L)}(\theta)$ is equivalent to minimizing the following Kullback-Leibler divergence

$$K_P^{(L)}(\theta, \theta^*) = l_P^{(L)}(\theta^*) - l_P^{(L)}(\theta) \geq 0.$$

Varin and Vidoni [2005] called $K_P^{(L)}(\theta, \theta^*)$ *composite Kullback-Leibler divergence* since it can be seen as the linear combination of the Kullback-Leibler divergences associated with each component of the composite likelihood. In this case

$$K_P^{(L)}(\theta, \theta^*) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} \right], \qquad (3.2.4)$$

which preserves the non-negativity as soon as the ordinary Kullback-Leibler divergence does (see Appendix A).

3.2.1 Strong consistency of the pairwise likelihood estimator of order L

Following the standard technique introduced by Wald [1949] and asking that the bivariate process $\{X_k, Y_k\}$ is uniformly ergodic and that the functions f_θ and g_θ are continuous in θ, the estimator obtained by maximizing the pairwise likelihood of order L is *strongly consistent*, i. e. it converges almost surely to the true parameter value as n goes to infinity.

More precisely, we prove the following theorem (middle results can be found in Appendix A)

Theorem 3.2.1. *Assume that conditions* $(C1 - C7)$ *in Appendix A.1 hold and let* $\hat{\theta}_P^{(L)}$ *be the L-order pairwise likelihood estimator based on n observations. Then* $\hat{\theta}_P^{(L)} \to \theta^*$ P_{θ^*}*-almost surely as n* $\to \infty$.

Proof. Given an arbitrary $\epsilon > 0$, set $S_\epsilon = \{\theta \in \Theta; |\theta - \theta^*| < \epsilon\}$ and $C = \Theta \cap S_\epsilon^c$. Lemma A.2.3 allows us to choose a positive number $\overline{b}$ such that, for every $j = 2, \ldots, L+1$

$$\mathbb{E}_{\theta^*}\left[\sup_{\theta:|\theta|>\overline{b}} \log p_\theta(y_1, y_j)\right] \le \mathbb{E}_{\theta^*}\left[\log p_{\theta^*}(y_1, y_j)\right] - 1 \qquad (3.2.5)$$

and let $C_1 = C \cap \{\theta \in \Theta; |\theta| \le \overline{b}\}$. It follows from Lemma A.2.1 and Lemma A.2.2 that for each $\theta \in C_1$ there is a $\epsilon_\theta > 0$ and an open neighborhood G_θ of θ such that

$$\mathbb{E}_{\theta^*}\left[\sup_{\theta' \in G_\theta} \log p_{\theta'}(y_1, y_j)\right] \le \mathbb{E}_{\theta^*}\left[\log p_\theta(y_1, y_j)\right] \le \mathbb{E}_{\theta^*}\left[\log p_{\theta^*}(y_1, y_j)\right] - \epsilon_\theta. \qquad (3.2.6)$$

Note that C_1 is a compact set (from Assumption C2) and thus there is a finite set $\{\theta_1, \ldots, \theta_d\} \subseteq \Theta$ such that $C_1 \subseteq \cup_{i=1}^d G_i$, where $G_i = G_{\theta_i}$ and define $G_0 = \{\theta \in$

$\Theta; |\theta| > \overline{b}\}$. We have that

$$\sup_{\theta \in S_\epsilon^c} \left(\log L_P^{(L)}(\theta; y_{1:n}) - \log L_P^{(L)}(\theta^*; y_{1:n}) \right) =$$
$$= \max_{0 \le i \le d} \left(\sup_{\theta \in G_i} \log L_P^{(L)}(\theta; y_{1:n}) - \log L_P^{(L)}(\theta^*; y_{1:n}) \right).$$

From Assumption (C1), for every $i, 1 \le i \le d$

$$\sup_{\theta \in G_i} \left(l_P^{(L)}(\theta; y_{1:n}) - l_P^{(L)}(\theta^*; y_{1:n}) \right) \overset{n\to\infty}{\to}$$
$$\frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\sup_{\theta \in G_i} \log p_\theta(y_1, y_j) \right] - \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log p_{\theta^*}(y_1, y_j) \right]$$

and by Equation (3.2.6) the right term above is less or equal to $-\epsilon_{\theta_i} < 0$. Again by Assumption (C1)

$$\sup_{\theta \in G_0} \left(l_P^{(L)}(\theta; y_{1:n}) - l_P^{(L)}(\theta^*; y_{1:n}) \right) \overset{n\to\infty}{\to}$$
$$\frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\sup_{\theta \in G_0} \log p_\theta(y_1, y_j) \right] - \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log p_{\theta^*}(y_1, y_j) \right]$$

and by Equation (3.2.5) the right term above is less or equal to $-1 < 0$. This proves that

$$\max_{0 \le i \le d} \left(\sup_{\theta \in G_i} \log[L(n-1) l_P^{(L)}(\theta; y_{1:n})] - \log[L(n-1) l_P^{(L)}(\theta^*; y_{1:n})] \right)$$
$$\overset{n\to\infty}{\to} -\infty \quad \mathbb{P}_{\theta^*} - a.s.,$$

that is

$$\mathbb{P}_{\theta^*} \left\{ \lim_{n\to\infty} \sup_{\theta \in S_\epsilon^c} \left(\log L_P^{(L)}(\theta; y_{1:n}) - \log L_P^{(L)}(\theta^*; y_{1:n}) \right) = -\infty \right\} = 1. \qquad (3.2.7)$$

Now, we use the result in (3.2.7) to prove the strong consistency of $\hat{\theta}_P^{(L)}$, i.e. that $\mathbb{P}_{\theta^*} \left\{ \lim_{n\to\infty} \hat{\theta}_P^{(L)} = \theta^* \right\} = 1$. Since $\hat{\theta}_P^{(L)}$ is a global maximum point of $L_P^{(L)}(\theta; y_{1:n})$, we have that

$$L_P^{(L)}(\hat{\theta}_P^{(L)}; y_{1:n}) \geq L_P^{(L)}(\theta^*; y_{1:n})$$

for all n. It is sufficient to prove that for any $\epsilon > 0$ the probability that there exists a limit point $\hat{\theta}$ of the sequence $\{\hat{\theta}_P^{(L)}\}$ such that $|\hat{\theta} - \theta^*| > \epsilon$ is zero. If such a $\hat{\theta}$ exists than $\sup_{\theta \in S_\epsilon^c} L_P^{(L)}(\theta; y_{1:n}) \geq L_P^{(L)}(\hat{\theta}_m^{(L)}; y_{1:n})$ for infinitely many n. But then

$$\frac{\sup_{\theta \in S_\epsilon^c} L_P^{(L)}(\theta; y_{1:n})}{L_P^{(L)}(\theta^*; y_{1:n})} > 0$$

for infinitely many n. Since, according to (3.2.7), this is an event with probability zero, we have shown that the probability that all limit points $\hat{\theta}$ of $\{\hat{\theta}_P^{(L)}\}$ satisfy the inequality $|\hat{\theta} - \theta^*| \leq \epsilon$ is one. By the arbitrariness of ϵ, $\hat{\theta}_P^{(L)}$ is strongly consistent. □

3.2.2 Central limit theorem

Under suitable assumptions, a central limit theorem exists. We do not discuss here hypothesis that ensure such existence. We consider the sequence of maximum pairwise likelihood estimators $\{\hat{\theta}_P^{(n)}\}$, where

$$\hat{\theta}_P^{(n)} := \arg\max_{\theta \in \Theta} l_P^{(L)}(\theta; y_{1:n}).$$

In the definition above we explicitly underline that such estimators depend on the sample size n and we recall that $l_P^{(L)}(\theta; y_{1:n})$ is defined as in (3.2.2). In this case, if a central limit theorem exists, we establish its expression in the scalar case (for notational simplicity).

Theorem 3.2.2. *Central limit theorem: scalar case*

$$\sqrt{n-1}(\hat{\theta}_P^{(n)} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_P^{(L)}(\theta^*)),$$

where

$$\begin{aligned}
\sigma_P^{(L)}(\theta^*) &= h_L^{-2}(\theta^*)\gamma_L(\theta^*), \quad \textit{with} \\
h_L(\theta^*) &:= \mathbb{E}_{\theta^*}\left[\frac{1}{L}\sum_{j=2}^{L+1}\frac{\partial^2}{\partial\theta^2}log[p_{\theta^*}(y_1,y_j)]\right] \\
\gamma_L(\theta^*) &:= \mathbb{E}_{\theta^*}\left[\left(\frac{1}{L}\sum_{j=2}^{L+1}\frac{\partial}{\partial\theta}log[p_{\theta^*}(y_1,y_j)]\right)^2\right]+ \\
&+ 2\sum_{k=2}^{\infty}\mathbb{E}_{\theta^*}\left[\left(\frac{1}{L}\sum_{j=2}^{L+1}\frac{\partial}{\partial\theta}log[p_{\theta^*}(y_1,y_j)]\right)\left(\frac{1}{L}\sum_{j=k+1}^{L+k}\frac{\partial}{\partial\theta}log[p_{\theta^*}(y_k,y_j)]\right)\right].
\end{aligned}$$

Proof. Let $\{\hat\theta_P^{(n)}\}$ be the sequence of maximum pairwise likelihood estimators of θ^*. We have the following Taylor expansion

$$0 = \frac{\partial}{\partial\theta}l_P^{(L)}(\hat\theta_P^{(n)};y_{1:n}) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\varphi_{\hat\theta_P^{(n)}}(y_i,\dots,y_{i+L}) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\dots,y_{i+L})$$
$$+\frac{(\hat\theta_P^{(n)}-\theta^*)}{n-1}\sum_{i=1}^{n-1}\left[\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_i,\dots,y_{i+L})+\frac{1}{2}\frac{\partial^3}{\partial\theta^3}\varphi_{\bar\theta_n}(y_i,\dots,y_{i+L})(\hat\theta_P^{(n)}-\theta^*)\right], \quad (3.2.8)$$

where $\bar\theta_n$ is a point on the segment $[\hat\theta_P^{(n)},\theta^*]$ and

$$\varphi_\theta(y_i,\dots,y_{i+L}) := \frac{1}{L}\sum_{j=i+1}^{\min\{i+L,n\}}log[p_\theta(y_i,y_j)]. \quad (3.2.9)$$

Since maximum pairwise likelihood is strongly consistent, $\hat\theta_P^{(n)}-\theta^*$ converges to zero and hence the second term in squared brackets in Equation (3.2.8) vanishes as n goes to infinity. From the ergodic properties

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_i,\dots,y_{i+L}) \xrightarrow{\mathbb{P}} \mathbb{E}_{\theta^*}\left[\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_1,\dots,y_{L+1})\right]$$

and similarly

$$\frac{1}{\sqrt{n-1}}\sum_{i=1}^{n-1}\left[\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})-\mathbb{E}_{\theta^*}\left[\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})\right]\right]\xrightarrow{\mathcal{D}}\mathcal{N}(0,\gamma_L(\theta^*)).$$

But here

$$\mathbb{E}_{\theta^*}\left[\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})\right]=0$$

since pairwise likelihood is an unbiased estimating equation and hence

$$\frac{1}{\sqrt{n-1}}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})\xrightarrow{\mathcal{D}}\mathcal{N}(0,\gamma_L(\theta^*)),$$

where

$$\begin{aligned}\gamma_L(\theta^*) \;&:=\; \mathbb{E}_{\theta^*}\left[\left(\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_1,\ldots,y_{L+1})\right)^2\right]+\\ &+\; 2\sum_{k=2}^{\infty}\mathbb{E}_{\theta^*}\left[\left(\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_1,\ldots,y_{L+1})\right)\left(\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_k,\ldots,y_{L+k})]\right)\right].\end{aligned}$$

We deduce from Equation (3.2.8) that

$$-\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})=\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})(\hat{\theta}_P^{(n)}-\theta^*)$$

and hence

$$\begin{aligned}\sqrt{n-1}(\hat{\theta}_P^{(n)}-\theta^*)&=-\frac{1}{\sqrt{n-1}}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})\\ &\cdot\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_i,\ldots,y_{i+L})\right)^{-1},\end{aligned}$$

leading to

$$\sqrt{n-1}(\hat{\theta}_P^{(n)}-\theta^*)\xrightarrow{\mathcal{D}}\mathcal{N}(0,h_L^{-2}(\theta^*)\gamma_L(\theta^*)),$$

where

$$h_L(\theta^*) := \mathbb{E}_{\theta^*}\left[\frac{\partial^2}{\partial\theta^2}\varphi_{\theta^*}(y_1,\dots,y_{L+1})\right].$$

The final result follows from (3.2.9). □

Remark 3.2.3. *From expression (3.2.3) we can outline what happens if L is allowed to grow to infinity. For j large enough, by ergodicity, $p_\theta(y_1, y_j)$ is well approximated by $p_\theta(y_1)p_\theta(y_j)$, for every $\theta \in \Theta$. Hence, for j large enough*

$$\int_{\mathcal{Y}^2} log[p_\theta(y_1,y_j)]p_{\theta^*}(y_1,y_j)dy_1dy_j$$

is well approximated by

$$\int_{\mathcal{Y}^2} log[p_\theta(y_1)p_\theta(y_j)]p_{\theta^*}(y_1)p_{\theta^*}(y_j)dy_1dy_j =$$
$$= \int_{\mathcal{Y}} log[p_\theta(y_1)]p_{\theta^*}(y_1)dy_1 + \int_{\mathcal{Y}} log[p_\theta(y_j)]p_{\theta^*}(y_j)dy_j.$$

By stationarity assumption, $p_\theta(y_1) = p_\theta(y_j)$ for every j and for every $\theta \in \Theta$ and using Cesaro sum we have that

$$\lim_{L\to+\infty}\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{Y}^2} log[p_\theta(y_1,y_j)]p_{\theta^*}(y_1,y_j)dy_1dy_j = 2\int_{\mathcal{Y}} log[p_\theta(y_1)]p_{\theta^*}(y_1)dy_1.$$

If L goes to infinity, it is clear from the result above that all the information about the dependence structure of the model are lost, since only the marginal density is taken into account. Moreover, in the case where the invariant distribution is unknown, all the inference is carried out from $p_\theta(y_1|\mu) = \int_{\mathcal{X}}\mu(x_1)g_\theta(y_1|x_1)dx_1$, that might be completely wrong.

At this point, two important issues arise. First, the characterization of the bias of the estimate introduced when π_θ is unknown. In this case we need an approximation for the bivariate density (3.1.2). Second, the quantification of the loss of asymptotic efficiency introduced by the use of $l_P^{(L)}(\theta)$ in place of the full likelihood, through the evaluation of the asymptotic variance of the estimator $\hat{\theta}_P^{(L)}$.

3.3 Bias of the estimate when π_θ is replaced with μ

In many situations (exceptions are, for example, linear gaussian models for the dynamic of $\{X_k\}$ and the discrete case), invariant distribution is unknown. Denoting by $p_\theta(y_i, y_j|\mu)$ the bivariate density of the observations y_i, y_j when the process is wrongly initialized by $X_1 \sim \mu(\cdot)$ we have that

$$p_\theta(y_i, y_j|\mu) = \int_{\mathcal{X}^j} \mu(x_1)\left[\prod_{k=2}^{j} f_\theta(x_k|x_{k-1})\right] g_\theta(y_i|x_i) g_\theta(y_j|x_j) dx_{1:j}.$$

The definition above yields the following approximation of the full likelihood defined in (3.2.1)

$$L_P^{(L)}(\theta; y_{1:n}, \mu) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i, y_j|\mu). \tag{3.3.1}$$

The following result quantifies the bias of the estimate introduced when the true invariant distribution π_θ is replaced with a generic distribution μ. We denote $\hat{\theta}_P(\mu)$ (we drop the explicit dependence on L) a generic maximum of the resulting approximate pairwise likelihood (3.3.1). Assumptions under which Theorem 3.3.1 holds are summarized in the Appendix B.2. Middle results can be found in the Appendices B.4 and B.5.

Theorem 3.3.1. *There exist* $C \in (0, +\infty)$ *and* $\rho \in [0, 1)$ *such that for any* $\mu \in \mathcal{P}(\mathcal{X})$

$$|\hat{\theta}_P(\mu) - \theta^*| \leq C \left|[\nabla^2 l_P(\theta^*)]^{-1}\right| \left[\frac{||\mu - \pi_{\theta^*}||}{1-\rho} + ||\nabla\mu - \nabla\pi_{\theta^*}||\right].$$

Proof. Let us consider the following Taylor expansion around θ^* and $\mu \in \mathcal{P}(\mathcal{X})$ such that $[\theta^*, \hat{\theta}_P(\mu)] \subset \overset{\circ}{\Theta}$,

$$\begin{aligned} \nabla l_P(\hat{\theta}_P(\mu)) &= \nabla l_P(\theta^*) + (\hat{\theta}_P(\mu) - \theta^*) \int_0^1 \nabla^2 l_P(\theta^* + t(\hat{\theta}_P(\mu) - \theta^*)) dt \\ &= \nabla l_P(\theta^*) + (\hat{\theta}_P(\mu) - \theta^*)[R(\mu) + \nabla^2 l_P(\theta^*)], \end{aligned} \tag{3.3.2}$$

where

$$R(\mu) := \int_0^1 \nabla^2 l_P(\theta^* + t(\hat{\theta}_P(\mu) - \theta^*) - \nabla^2 l_P(\theta^*))dt.$$

Since the set of parameters maximizing $l_P(\theta)$ includes the true parameter, $\nabla l_P(\theta^*) = 0$. Moreover, by definition, $\nabla l_P(\hat{\theta}_P(\mu), \mu) = 0$. Hence (3.3.2) can be written as

$$\nabla l_P(\hat{\theta}_P(\mu)) = \nabla l_P(\hat{\theta}_P(\mu), \mu) + (\hat{\theta}_P(\mu) - \theta^*)[R(\mu) + \nabla^2 l_P(\theta^*)],$$

leading to

$$(\hat{\theta}_P(\mu) - \theta^*) = [R(\mu) + \nabla^2 l_P(\theta^*)]^{-1}[\nabla l_P(\hat{\theta}_P(\mu)) - \nabla l_P(\hat{\theta}_P(\mu), \mu)].$$

We have that $R(\mu)$ vanishes as $||\mu - \pi_\theta||$ goes to zero. This follows from the Theorem B.5.1 and from the continuity in θ of the function $\nabla^2 l_P(\theta)$. Using the result in Theorem B.4.2, we can easily conclude. □

In the theorem above, the constant ρ characterizes the forgetting properties of $\{X_k\}$ a priori and conditional upon $\{Y_k\}$. This result confirms the intuition that the bias introduced when using μ instead of π_{θ^*} in the pairwise likelihood depends on how close μ is to π_{θ^*} and on the ergodic properties of $\{X_k\}$.

The problem now is how to choose the distribution μ. In the cases in which the invariant distribution π_θ is unknown but transitions $f_\theta(\cdot|x)$ are simple, the idea is to approximate the invariant distribution π_θ sampling from the transition kernel $f_\theta(\cdot|x)$ and to take advantage of the geometric ergodicity of the process. More precisely, the idea is to take

$$\mu(x_{i-r} : i) = \mu(x_{i-r}) \prod_{k=i-r+1}^{i} f_\theta(x_k|x_{k-1}), \tag{3.3.3}$$

where, under geometric ergodicity, the marginal

$$\mu(x_i) \to \pi_\theta(x_i),$$

as r goes to $+\infty$.

In more complex situations, the choice of μ has to be carefully done, taking into account that this will affect the bias of the estimate.

3.4 Loss of efficiency

If L is fixed, the use of L-th order PL suggests that information about the parameter can be extracted from the dependence structure of the pairs of observations with a lag distance not greater than L. Usually, it happens that the maximum pairwise likelihood estimators tend to lose efficiency, with respect to those based on the full likelihood. Even if this behavior is obviously reliable, until now, no general results about the evaluation of this gap are available.

Instead of comparing the efficiency between the PL and the full likelihood, we would like to compare the efficiency of SDL and PL. This choice is justified by the fact that for general state- space models the full likelihood function is unavailable, as we discussed before, and hence the estimator obtained by maximizing this function is not actually a real alternative. Moreover, for non overlapping version of split data likelihood estimator, we have some theoretical results about the behavior of its variance. Anyway, maximum full likelihood estimator is the benchmark we have to refer to when we discuss efficiency of any estimator.

We would like to take into account the overlapping version of the SDL, as defined in (2.2.6) and consider the case where π_θ is known. Since in PL of order L, for every $i = 1, \ldots, n-1$, $p_\theta(y_{i:L+i})$ is approximated by $\prod_{j=i+1}^{L+i} p_\theta(y_i, y_j)$, that is the joint distribution of a block is approximated by the product of the pairs within the block, we expect that, for a general model, there will be a loss of efficiency when we use PL of order L instead of overlapping SDL. This property strongly depends on the model we consider: in the next chapter, we will empirically show that this is not necessarily true for every model. The quantification of this loss can be achieved through the evaluation of the asymptotic variance of the estimator $\hat{\theta}_P^{(L)}$. Andrieu et al. [2007] characterize the asymptotic variance in the non overlapping version of SDL, called Σ_L, and quantify the loss of efficiency by comparing Σ_L to its counterpart associated to the full likelihood based criterion. More precisely, they state that there exists a $C \in (0, +\infty)$ and $\rho \in [0, 1)$ such that for any $L \geq 2$

$$|\Sigma - \Sigma_L| \leq C\left[\frac{\log(L)^2}{L\log(\rho)^2} + \frac{\rho}{L(1-\rho)} + \frac{\rho^{L+1}}{1-\rho^L}\right], \tag{3.4.1}$$

where Σ denotes the asymptotic variance of the full likelihood estimator. Since

$$\Sigma_L = H_L^{-1}(\theta^*)G_L(\theta^*)H_L^{-T}(\theta^*),$$

where

$$\begin{aligned}
H_L(\theta^*) &= \frac{1}{L}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)],\\
G_L(\theta^*) &= \frac{1}{L}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)] + \frac{2}{L}\sum_{k=1}^{+\infty}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_k)],
\end{aligned}$$

with $\mathbf{Y}_k = (Y_{kL+1}, \ldots, Y_{(k+1)L})$, the result in (3.4.1) comes from the fact that

$$\begin{aligned}
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)] - \Sigma^{-1}| &\leq \frac{\rho}{L(1-\rho)},\\
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_1)]| &\leq C\frac{\log(L)^2}{L\log(\rho)^2},\\
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_k)]| &\leq CL\rho^{(k-1)L+1} \qquad \forall k \geq 2,
\end{aligned}$$

for a suitable $C \in (0, +\infty)$ and $\rho \in [0, 1)$. Equation (3.4.1) proves that the loss of efficiency compared to the maximum likelihood estimator vanishes as L increases and depends on the mixing properties of the model. Extending their results to the overlapping version of the maximum split data likelihood estimator is far from being easy. The difficulties arise because the dependency structure between blocks is more complex when blocks are allowed to overlap instead of being disjoint. This translates in a more complicated calculation for the counterpart of $G_L(\theta^*)$, necessary to evaluate the asymptotic variance of the estimator.

For our purpose, we shall evaluate the asymptotic variances of the estimators obtained by maximizing (2.2.6) and (3.2.1). We refer to these quantities as $\Sigma_{SD}^{(ov)}$ and $\Sigma_P^{(L)}$ respectively. As we discussed above, evaluation of $\Sigma_{SD}^{(ov)}$ and a fortiori evaluation of $\Sigma_P^{(L)}$ is not easy to obtain, even for simple models. A deep theoretical analysis of the efficiency problem in pairwise and overlapping split data likelihood inferential procedures is beyond the scope of this thesis. Anyway, while we suspect that $\Sigma_{SD}^{(ov)}$ still decreases if L grows, we do not expect that $\Sigma_{P}^{(L)}$ will do the same if L grows, unlike Σ_L does. This idea is consistent to Varin and Vidoni [2009]. In the next chapter, we give an empirical evidence of these behaviors and we suggest the existence of a 'best' lag L, in term of variance of the PL estimator. Anyway, how to determine L optimally is still a good open question.

Chapter 4

Empirical study about efficiency

In this chapter, we empirically compare the efficiency between maximum pairwise likelihood and maximum full likelihood estimators, as well as the efficiency between maximum overlapping split data likelihood and maximum pairwise likelihood estimators. Even if we do not have theoretical results that state the behavior of their variances, our intuitions, suggested in Section 3.4, are confirmed in this simple example. We consider a linear gaussian state space model, where invariant distribution is known and the likelihood function is available in a closed form. Even if it is only an empirical study in a simple context, these preliminary results may be a useful guide when we move to more complex settings where we can not compute the likelihood function in a closed form.

4.1 The model

We illustrate here by means of simulation experiments, the performance of the maximum pairwise likelihood estimator of order L and we compare it with the maximum split data likelihood estimator, where the blocks defining the likelihood function are allowed to overlap.
We consider a state space model where the latent process follows an autoregressive dynamic and the marginal distributions of the observations are explicitly known.

As defined in Example 2.0.1, we consider the following AR(1) model with additive observation noise

$$\begin{aligned} X_{n+1} &= \phi X_n + W_n, \quad W_n \sim N(0, \tau^2) \\ Y_n &= X_n + V_n, \quad V_n \sim N(0, \sigma^2). \end{aligned}$$

In this case

$$f_\theta(x'|x) = N(\phi x, \tau^2) \quad \text{and} \quad g_\theta(y|x) = N(x, \sigma^2).$$

We assume that the AR(1) evolution is stationary, so $|\phi| < 1$ and $\pi_\theta \sim N\left(0, \frac{\tau^2}{1-\phi^2}\right)$. The unknown parameter is $\theta = (\phi, \tau, \sigma)$. The full likelihood function is available in a closed form and it can be efficiently computed by the Kalman filter recursions. Thus we can compare the performance of the maximum likelihood, the maximum pairwise likelihood and the maximum split data likelihood estimators. Moreover, we can empirically compare the variance of the maximum pairwise and the split data likelihood estimators in order to study their relationship in term of efficiency. Since we set the parameter space in such a way that the process is stationary, the bivariate distribution of the pairs (Y_i, Y_j), $i = 1, \ldots, n-1$; $j = i+1, \ldots, n$ is

$$\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \frac{\tau^2}{1-\phi^2} & \phi^{j-i}\frac{\tau^2}{1-\phi^2} \\ \phi^{j-i}\frac{\tau^2}{1-\phi^2} & \sigma^2 + \frac{\tau^2}{1-\phi^2} \end{pmatrix} \right\},$$

and hence the pairwise likelihood of order L is easy to compute.

It is worthwhile to underline that the statistical model corresponding to the choice $L = 1$ is not identifiable. If $L = 1$ there exist at least two different sets of parameters values for θ which give the same value for the pairwise likelihood function. This problem can be easily overcome by adding pairs at lag distance greater than one.

On the other hand, under stationarity conditions, the marginal distribution of the blocks $(Y_i, \ldots, Y_{L+i-1})$, $i = 1, \ldots, n-L+1$ is

$$\begin{pmatrix} Y_i \\ \vdots \\ Y_{L+i-1} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \frac{\tau^2}{1-\phi^2} & \phi\frac{\tau^2}{1-\phi^2} & \cdots & \phi^{L-1}\frac{\tau^2}{1-\phi^2} \\ \phi\frac{\tau^2}{1-\phi^2} & \ddots & \ddots & \phi^{L-2}\frac{\tau^2}{1-\phi^2} \\ \vdots & \ddots & \ddots & \vdots \\ \phi^{L-1}\frac{\tau^2}{1-\phi^2} & \cdots & \cdots & \sigma^2 + \frac{\tau^2}{1-\phi^2} \end{pmatrix} \right\},$$

Table 4.2.1: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Sample mean and standard deviation (in brackets), for the maximum likelihood estimator $\hat{\theta}_{ML}$. Calculations based on 300 simulated time series of length 1000.

$\hat{\phi}_{ML}$	$\hat{\tau}_{ML}$	$\hat{\sigma}_{ML}$
0.6952	0.996	0.9932
(0.0473)	(0.0952)	(0.0798)

and hence the split data likelihood with blocks of length L turns out to be easy to compute.

4.2 Simulation study

We perform a simple simulation study with the aim of comparing the empirical properties of $\hat{\theta}_P^{(L)}$ and $\hat{\theta}_{SD}^{(L)}$, with $L = 2, \ldots, 29$. We consider 300 time series of length $n = 1000$ from the AR(1) model plus additive observation noise, with $\phi^* = 0.7, \sigma^* = 1, \tau^* = 1$ as true parameter values. Hereafter, in order to find the maximum point of the pairwise and split data likelihood functions, we adopt an optimization procedure based on the Nelder and Mead downhill simplex method, with a relative convergence tolerance of 10^{-8}. We repeat the optimization procedure starting from different values in the parameter space, finding similar results. The sample means and standard deviations for the maximum pairwise and split data likelihood estimators, for some L, are summarized in Table 4.2.2 as well as for the maximum full likelihood estimator (Table 4.2.1). The results presented here are obtained taking as starting values for the optimization procedure $\phi^0 = 0.9, \sigma^0 = 0.8, \tau^0 = 0.5$.

Table 4.2.2: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Sample means and standard deviations (in brackets) for the maximum pairwise likelihood estimator $\hat{\theta}_P^{(L)}$ and split data likelihood estimator $\hat{\theta}_{SD}^{(L)}$ as L increases. Calculations based on 300 simulated time series of length 1000.

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
2	0.6968	0.9928	0.992	0.6968	0.9929	0.9919
	(0.0561)	**(0.119)**	**(0.095)**	(0.0562)	(0.1191)	(0.0951)
3	0.6963	0.9936	0.9944	0.6956	0.9953	0.9923
	(0.0494)	**(0.1006)**	**(0.0827)**	(0.0507)	(0.1048)	(0.0866)
4	0.696	0.9939	0.9948	0.6952	0.9961	0.9924
	(0.0481)	**(0.0963)**	**(0.0803)**	(0.049)	(0.1003)	(0.0837)
5	0.6964	0.9932	0.9951	0.6951	0.9964	0.9924
	(0.0487)	(0.0983)	(0.0832)	(0.0484)	(0.0985)	(0.0825)
6	0.6954	0.9954	0.9925	0.6951	0.9965	0.9925
	(0.0505)	(0.1041)	(0.0882)	(0.048)	(0.0975)	(0.0817)
7	0.6945	0.9976	0.99	0.6951	0.9965	0.9926
	(0.052)	(0.1085)	(0.0922)	(0.0479)	(0.0971)	(0.0814)
8	0.694	0.9987	0.9885	0.6951	0.9965	0.9927
	(0.0534)	(0.113)	(0.0955)	(0.0478)	(0.0969)	(0.0811)
9	0.6937	0.9995	0.9871	0.6951	0.9965	0.9928
	(0.055)	(0.1173)	(0.0995)	(0.0477)	(0.0968)	(0.0809)
10	0.6937	0.9995	0.9869	0.6951	0.9964	0.9928
	(0.0558)	(0.12)	(0.1008)	(0.0477)	(0.0967)	(0.0807)
11	0.6935	0.9998	0.9864	0.6952	0.9964	0.9929
	(0.0566)	(0.1225)	(0.1026)	(0.0477)	(0.0966)	(0.0806)
12	0.6937	0.9993	0.9865	0.6951	0.9964	0.9929
	(0.0573)	(0.1246)	(0.1042)	(0.0476)	(0.0965)	(0.0805)

Table 4.2.2: Continued on next page

Table 4.2.2: continued from previous page

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
13	0.6943 (0.0584)	0.9977 (0.1279)	0.9875 (0.1065)	0.6952 (0.0476)	0.9963 (0.0964)	0.993 (0.0804)
14	0.6946 (0.0588)	0.9971 (0.1294)	0.9879 (0.1075)	0.6952 (0.0476)	0.9962 (0.0964)	0.9931 (0.0802)
15	0.6949 (0.0593)	0.9962 (0.1307)	0.9886 (0.1083)	0.6952 (0.0475)	0.9962 (0.0963)	0.9931 (0.0802)
16	0.695 (0.0598)	0.9957 (0.1322)	0.9889 (0.1092)	0.6952 (0.0475)	0.9962 (0.0963)	0.9931 (0.0801)
17	0.6953 (0.0601)	0.995 (0.1334)	0.9894 (0.1095)	0.6952 (0.0475)	0.9962 (0.0963)	0.9932 (0.08)
18	0.6955 (0.0605)	0.9945 (0.1345)	0.9897 (0.1105)	0.6952 (0.0475)	0.9962 (0.0963)	0.9932 (0.08)
19	0.6958 (0.0607)	0.9937 (0.1354)	0.9904 (0.111)	0.6952 (0.0475)	0.9962 (0.0963)	0.9932 (0.08)
20	0.6957 (0.061)	0.9938 (0.1362)	0.9902 (0.1113)	0.6952 (0.0475)	0.9962 (0.0963)	0.9932 (0.08)
21	0.6958 (0.0611)	0.9935 (0.1363)	0.9906 (0.1108)	0.6951 (0.0474)	0.9962 (0.0963)	0.9933 (0.0799)
22	0.696 (0.0613)	0.9932 (0.1372)	0.9907 (0.1117)	0.6951 (0.0474)	0.9962 (0.0963)	0.9933 (0.0799)
23	0.6961 (0.0612)	0.9928 (0.1371)	0.9912 (0.1117)	0.6951 (0.0474)	0.9962 (0.0964)	0.9933 (0.0799)
24	0.6959 (0.0614)	0.9933 (0.1374)	0.9906 (0.1121)	0.6951 (0.0474)	0.9962 (0.0963)	0.9933 (0.0798)
25	0.6959 (0.0612)	0.9932 (0.137)	0.9908 (0.1116)	0.6951 (0.0473)	0.9962 (0.0963)	0.9934 (0.0798)

Table 4.2.2: Continued on next page

Table 4.2.2: continued from previous page

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
26	0.696 (0.0612)	0.9931 (0.1371)	0.9908 (0.1118)	0.6951 (0.0474)	0.9961 (0.0963)	0.9934 (0.0798)
27	0.6961 (0.0611)	0.9928 (0.1366)	0.9912 (0.111)	0.6951 (0.0473)	0.9961 (0.0963)	0.9934 (0.0798)
28	0.6959 (0.0612)	0.9932 (0.1371)	0.9907 (0.1118)	0.6951 (0.0473)	0.9961 (0.0964)	0.9934 (0.0798)
29	0.6959 (0.0612)	0.9932 (0.1369)	0.9907 (0.1115)	0.6951 (0.0474)	0.9961 (0.0964)	0.9935 (0.0798)

Let us analyze the table above. We clearly see that the behavior of variance of the maximum pairwise likelihood estimator is not monotonic. More precisely, it decreases until $L = 4$ and then increases as the maximum distance between pairs of observations increases. All these results are consistent with the existence of a 'best' lag distance L^*, in terms of minimum variance. From this empirical analysis, with this actual true parameter values, we can conclude that L^* exists and it equals $L^* = 4$.

Table 4.2.2 reports also the estimates and the variances of the estimates referred to the maximum split data likelihood estimator. Our empirical study shows that the variance in this case decreases to the variance of the maximum full likelihood estimator as L grows. These results empirically prove that maximum SDL estimator goes to the maximum full likelihood estimator as the lag distance between pairs of observations goes to infinity [Andrieu et al., 2007].

Figure 4.2.1 displays the behavior of the variances of the two estimators (PL and SDL) compared to the variance of the maximum full likelihood estimator. We clearly identify L^* as $L^* = 4$ and the monotonic decreasing trend of the SDL variance.

If we display in the same plot the PL and the SDL variances as L grows a nice property arises (Figure 4.2.2). Then, if $2 < L \leq L^* = 4$, L^* being the 'best' lag distance in terms of minimum PL variance, the variance of the PL estimator is smaller then the variance of the SDL estimator. The relation turns upside-down for $L > L^* = 4$. Hitherto, we do not have a clear intuition that can justify such behavior.

We repeat the simulation changing the values of θ^*. While σ^* and τ^* do not seem to affect the optimal value L^*, increasing the value of ϕ^* results in a bigger optimal value L^* (we recover the best order equal to six given in Varin and Vidoni [2009] when $\phi^* = 0.95$). This is probably connected to the weaker or stronger dependence structure of the pairs. Anyway, the fact that the optimal choice for the lag distance between the pairs depends on the unknown true parameter values makes its investigation ambiguous in real scenarios.

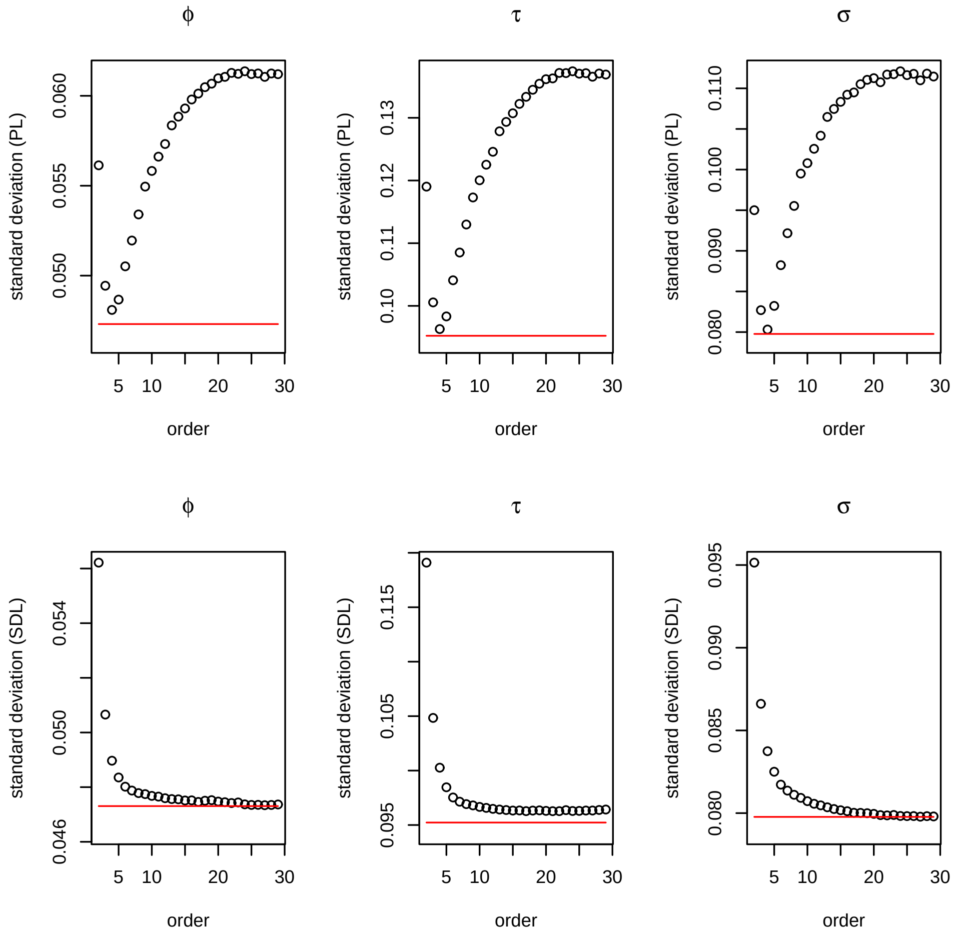

Figure 4.2.1: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Standard deviations for the maximum likelihood estimator $\hat{\theta}_{ML}$ (solid line), the maximum pairwise likelihood estimator $\hat{\theta}_P^L$ (top) and the maximum split data likelihood estimator $\hat{\theta}_{SD}^{(L)}$ (bottom), with $L = 2, \ldots, 29$ denoting the maximum distance between the observations. Calculations based on 300 simulated time series of length 1000.

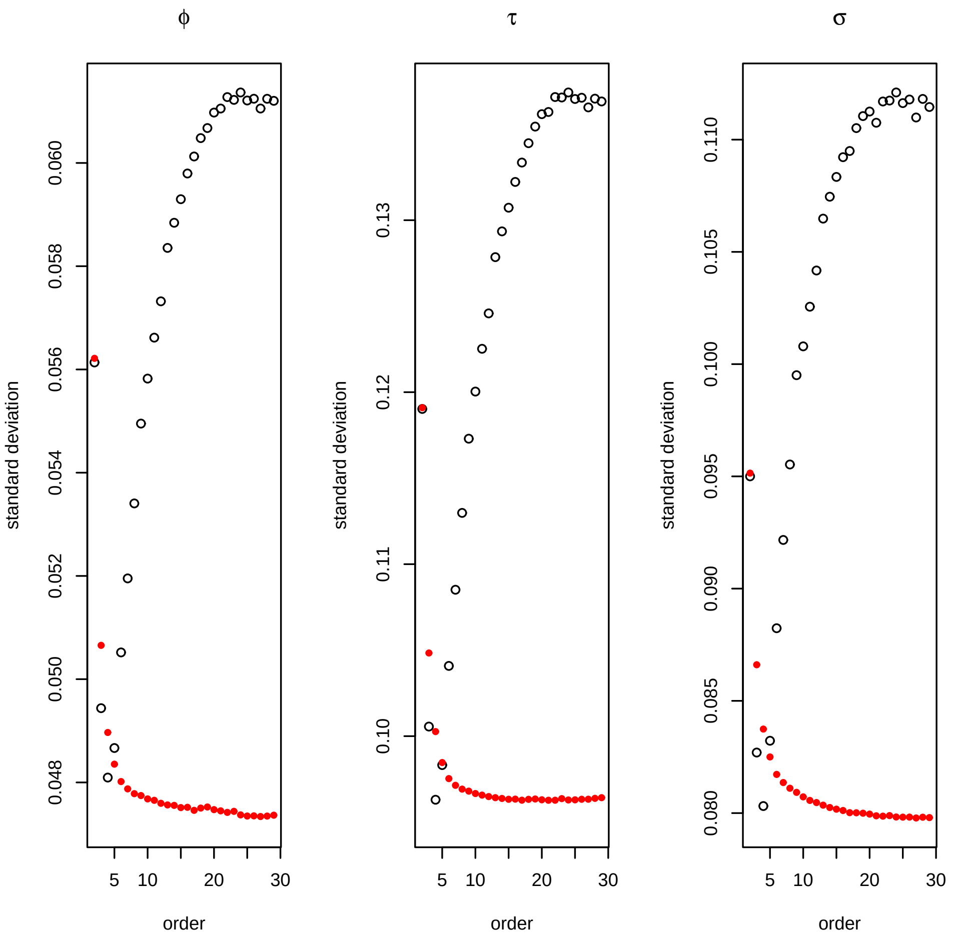

Figure 4.2.2: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Standard deviations for the maximum pairwise likelihood estimator $\hat{\theta}_P^L$ (circle) and the maximum split data likelihood estimator $\hat{\theta}_{SD}^{(L)}$ (smaller solid circle), with $L = 2, \ldots, 29$ denoting the maximum distance between the observations. Calculations based on 300 simulated time series of length 1000.

Chapter 5

Expectation- Maximization Algorithm

In this chapter we describe a possible way to obtain estimates for the parameter θ describing a general state space model. We focus on an on line Expectation- Maximization (EM) technique to minimize, with respect to θ, the Kullback- Leibler divergence $K_P^{(L)}(\theta, \theta^*)$ defined in (3.2.4), or equivalently to minimize $l_P^{(L)}(\theta)$. The key advantage of the average log pairwise likelihood function compared to the full likelihood is that it only requires the estimation of expectations with respect to distributions defined on $\mathcal{X}^{L+1}$. More precisely, this technique allows us to find

$$\min_{\theta \in \Theta} K_P^{(L)}(\theta, \theta^*),$$

where

$$\begin{aligned} K_P^{(L)}(\theta, \theta^*) &= \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} \right] \\ &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} \log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} p_{\theta^*}(y_1, y_j) dy_1 dy_j. \end{aligned}$$

This is clearly equivalent to maximize $l_P^{(L)}(\theta)$, where

$$l_P^{(L)}(\theta) = \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} \log[p_\theta(y_1, y_j)] p_{\theta^*}(y_1, y_j) dy_1 dy_j.$$

We first describe the abstract form of the EM algorithm as it is often given in the literature. We then develop the EM procedure in order to find the parameter estimates in two applications

1. Linear gaussian model [Section 5.4],

2. Jump Markov Linear System [Section 6.2].

For these models we derive the update equations in fairly explicit detail.

5.1 General EM algorithm

In this section, we discuss the EM algorithm of Dempster et al. [1977]. The presentation and the notation here are self-contained and do not have to be confused with symbols elsewhere.

The EM algorithm is a general method to find the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data are incomplete or have missing values. There are two main applications of the EM algorithm. The first occurs when the data indeed have missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but the likelihood function can be simplified by assuming the existence of values for additional but missing (or hidden) parameters. The latter application is more common in the computational pattern recognition community.

In summary, each iteration of the EM algorithm consists of two steps:

(**E-step**) In the expectation step (from now on E-step) the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology.

(**M-step**) In the maximization step (from now on M-step), the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

These steps define an efficient iterative procedure to compute the maximum likelihood estimate and convergence is assured, since the algorithm is guaranteed to increase the likelihood at each iteration.

More precisely, we assume that the data set $\mathcal{Y} = (y_1, y_2, \ldots, y_N)$ observed is generated by some distribution with density function $p(y|\theta)$, governed by the set of parameters θ. We call $\mathcal{Y}$ the *incomplete data*. We assume that a *complete data* set exists $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and also assume (or specify) a joint density function:

$$p(z|\theta) = p(x, y|\theta) = p(x|y, \theta)p(y|\theta).$$

Often the joint density comes from the marginal density function $p(y|\theta)$ and the assumption of hidden variables and parameters value guesses. In other cases (e.g., missing data values in samples of a distribution), we must assume a joint relationship between the missing and observed values.

With this density function, we can define a new likelihood function, $L(\theta|\mathcal{Z}) = L(\theta|\mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}|\theta)$, called the *complete-data likelihood*. Note that this function is in fact a random variable since the missing information $\mathcal{X}$ is unknown, random, and presumably governed by an underlying distribution. That is, we can think of $L(\theta|\mathcal{X}, \mathcal{Y}) = h_{\mathcal{Y},\theta}(\mathcal{X})$ for some function $h_{\mathcal{Y},\theta}(\cdot)$ where $\mathcal{Y}$ and θ are constant and $\mathcal{X}$ is a random variable. The original likelihood $L(\theta|\mathcal{Y})$ is referred to as the *incomplete-data likelihood* function.

The EM algorithm first finds the expected value of the complete-data log-likelihood $log[p(\mathcal{X}, \mathcal{Y}|\theta)]$ with respect to the unknown data $\mathcal{X}$ given the observed data $\mathcal{Y}$ and the current parameters estimates.

That is, we define:

$$Q(\theta, \theta^{(i-1)}) = E\left[\log[p(\mathcal{X}, \mathcal{Y}|\theta)]|\mathcal{Y}, \theta^{(i-1)}\right], \tag{5.1.1}$$

where $\theta^{(i-1)}$ are the current parameters estimates that we used to evaluate the expectation and θ are the new parameters that we optimize to increase Q. The right side of Equation (5.1.1) can be rewritten as

$$E\left[\log[p(\mathcal{X},\mathcal{Y}|\theta)]|\mathcal{Y},\theta^{(i-1)}\right] = \int_{x\in\Upsilon} \log[p(\mathcal{Y},x|\theta)]\,f(x|\mathcal{Y},\theta^{(i-1)})dx.$$

Note that $f(x|\mathcal{Y},\theta^{(i-1)})$ is the marginal distribution of the unobserved data and is dependent on both the observed data $\mathcal{Y}$ and on the current parameters, and Υ is the space of values x can take on. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameters $\theta^{(i-1)}$ and perhaps the data. In the worst of cases, this density might be very hard to obtain. Sometimes, in fact, the density actually used is $f(x,\mathcal{Y},\theta^{(i-1)}) = f(x|\mathcal{Y},\theta^{(i-1)})f(\mathcal{Y}|\theta^{(i-1)})$ but this does not affect subsequent steps, since the extra factor, $f(\mathcal{Y}|\theta^{(i-1)})$, is not dependent on θ.

The evaluation of this expectation is called the **E-step** of the algorithm. Note the meaning of the two arguments in the function $Q(\theta,\theta')$. The first argument θ corresponds to the parameters that ultimately will be optimized in an attempt to maximize the likelihood. The second argument θ' corresponds to the parameters that we use to evaluate the expectation.

The second step (the **M-step**) of the EM algorithm is to maximize the expectation we computed in the first step. That is, we find

$$\theta^{(i)} = \arg\max_{\theta} Q(\theta,\theta^{(i-1)}).$$

These two steps are repeated as necessary. Each iteration is guaranteed to increase the loglikelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function. There are many papers that analyze the rate of convergence (e.g., Dempster et al. [1977], Redner and Walker [1984], Wu [1983], Xu and Jordan [1996]), but we will not discuss them here.

Instead of maximizing $Q(\theta,\theta^{(i-1)})$, a modified form of the M-step consists of finding some $\theta^{(i)}$ such that $Q(\theta^{(i)},\theta^{(i-1)}) > Q(\theta,\theta^{(i-1)})$. This form of the algorithm is called Generalized EM (GEM) and is also guaranteed to converge.

In this section, we have described the algorithm in its most general form. The details of the steps required to compute the given quantities are very dependent on the particular model and application, so they are not discussed here where the algorithm is presented in this abstract form.

5.2 Full likelihood inference via EM algorithm

Before considering composite likelihood inference, we propose here an EM algorithm for full likelihood parameters estimation in a linear gaussian model with observation noise, which is fully automatic and gives optimal estimates. Further, we apply a Kalman smoother to obtain the estimates. We consider the following linear gaussian model, written in a state space form

$$\begin{aligned} X_{n+1} &= \phi X_n + W_n, \quad W_n \sim N(0,\tau^2) \\ Y_n &= X_n + V_n, \quad V_n \sim N(0,\sigma^2). \end{aligned}$$

The choice of the parameter vector $\theta = (\phi, \tau^2, \sigma^2) \in (-1,1) \times \mathbb{R}^+ \times \mathbb{R}^+$ ensures stationarity. So

$$\begin{aligned} f_\theta(x'|x) &= N(x'; \phi x, \tau^2) \\ g_\theta(y|x) &= N(y; x, \sigma^2). \end{aligned}$$

The Kalman filter is an optimal estimator in the mean-square sense. If the *future* measurements are available, smoothing equations can be used to further improve the estimation performance. We propose here an EM algorithm for parameters estimation in the model defined above.

Given the measurement sequence $Y_{1:T}$, we want to find estimates of the model coefficients. For this purpose we use the Kalman filter, assuming for the moment that the model parameters are available. We use the following definitions for the conditional expectations of the states and the corresponding error covariances:

$$\begin{aligned} \tilde{m}_{t|s} &= E[x_t|Y_{1:s}] \\ \tilde{P}_{t_1,t_2|s} &= E[(x_{t_1} - \tilde{m}_{t_1|s})(x_{t_2} - \tilde{m}_{t_2|s})|Y_{1:s}]. \end{aligned}$$

For convenience, when $t_1 = t_2 = t$, $\tilde{P}_{t_1,t_2|s}$ is written as $\tilde{P}_{t|s}$. The state estimate $(\tilde{m}_{t|t}, \tilde{P}_{t|t})$ can be obtained by iterating the *prediction* and *update steps* of the Kalman Filter. If the prior distribution is Gaussian, $x_0 \sim N(\mu_0, \sigma_0^2)$, then the optimal filtering equations can be evaluated in closed form:

$$\begin{aligned} p_\theta(x_t|y_{1:t-1}) &= N(x_t|\tilde{m}_{t|t-1}, \tilde{P}_{t|t-1}) \\ p_\theta(x_t|y_{1:t}) &= N(x_t|\tilde{m}_{t|t}, \tilde{P}_{t|t}) \end{aligned}$$

and the parameters of these distributions can be calculated by the following steps:

Prediction step

$$\begin{aligned} \tilde{m}_{t|t-1} &= \phi\tilde{m}_{t-1|t-1} \\ \tilde{P}_{t|t-1} &= \phi^2\tilde{P}_{t-1|t-1} + \tau^2. \end{aligned}$$

Update step

$$\begin{aligned} \tilde{e}_t &= y_t - \tilde{m}_{t|t-1} \\ \tilde{S}_t &= \tilde{P}_{t|t-1} + \sigma^2 \\ \tilde{K}_t &= \tilde{P}_{t|t-1}(\tilde{S}_t)^{-1} \\ \tilde{m}_{t|t} &= \tilde{m}_{t|t-1} + \tilde{K}_t\tilde{e}_t \\ \tilde{P}_{t|t} &= (I - \tilde{K}_t)\tilde{P}_{t|t-1}, \end{aligned}$$

with the initial condition $\tilde{m}_{1|0} = \mu_0$ and $\tilde{P}_{1|0} = \sigma_0^2$.

If the future measurements $Y_{t+1:T}$ are available, then these can be further used to improve the accuracy of the estimates. The *smoothed* estimates can be obtained as follows:

$$\begin{aligned} \tilde{J}_t &= \phi\tilde{P}_{t|t}(\tilde{P}_{t+1|t})^{-1} \\ \tilde{m}_{t|T} &= \tilde{m}_{t|t} + \tilde{J}_t(\tilde{m}_{t+1|T} - \tilde{m}_{t+1|t}) \\ \tilde{P}_{t|T} &= \tilde{P}_{t|t} + \tilde{J}_t^2(\tilde{P}_{t+1|T} - \tilde{P}_{t+1|t}), \end{aligned}$$

starting from $t = T$. We describe here the estimation of the model parameters with an EM algorithm. The objective is to compute estimates of θ given a measurement sequence. For gaussian models, maximum likelihood (ML) estimate is an obvious choice, which is given as follows:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log[p_{\theta}(Y_{1:T})],$$

where $p_{\theta}(Y_{1:T})$ is the probability density function of the observations. Note that because of the dependence on the states, which are not available, direct maximization is not possible. The problem is to maximize the likelihood with respect to two unknows: states and model parameters. The EM algorithm, as described in the previous section, takes an iterative approach by first maximizing the likelihood with respect to the states in the E-step, and then maximizing with respect the parameters in the M-step. Given the current estimate of the parameters θ_k, the E-step maximum is given by the expected value of the complete log-likelihood function as follows

$$Q(\theta, \theta_k) := E_{\theta_k}[\log[p_{\theta}(Y_{1:T}, X_{1:T})]],$$

where the expectation is taken wrt $p_{\theta_k}(x_{1:T}|y_{1:T})$. The M-step involves the direct differentiation of $Q(\theta, \theta_k)$ wrt θ to find the value of the parameters. We now describe an EM algorithm for our model.

E-step This step involves the computation of Q given the measurements $Y_{1:T}$ and the estimate of the parameters from the previous iteration, θ_k. The joint probability distribution of $X_{1:T}, Y_{1:T}$ can be written as

$$p_{\theta}(X_{1:T}, Y_{1:T}) = p(x_1) \prod_{t=2}^{T} f_{\theta}(x_t|x_{t-1}) \prod_{t=1}^{T} g_{\theta}(y_t|x_t).$$

Taking log, we have that $p_{\theta}(X_{1:T}, Y_{1:T})$ is proportional to

$$\begin{aligned}
&-\frac{1}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}\left(x_1^2 + \mu_0^2 - 2x_1\mu_0\right) + \\
&-\frac{T-1}{2}\log\tau^2 - \frac{1}{2\tau^2}\left(\sum_{t=2}^{T} x_t^2 + \phi^2 \sum_{t=2}^{T} x_{t-1}^2 - 2\phi \sum_{t=2}^{T} x_t x_{t-1}\right) +
\end{aligned}$$

$$-\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{t=1}^{T} y_t^2 + x_t^2 - 2x_t y_t\right).$$

If we take the following expectations

$$
\begin{aligned}
E_{\theta_k}[X_t] &= \tilde{m}_{t|T} && \text{(5.2.1a)}\\
E_{\theta_k}[X_t^2] &= \tilde{P}_{t|T} + \tilde{m}_{t|T}^2 := \tilde{M}_{t|T} && \text{(5.2.1b)}\\
E_{\theta_k}[X_t X_{t-1}] &= \tilde{P}_{t,t-1|T} + \tilde{m}_{t|T}\tilde{m}_{t-1|T} := \tilde{M}_{t,t-1|T}, && \text{(5.2.1c)}
\end{aligned}
$$

we get the expectation of joint log-likelihood with respect to the conditional density:

$$
\begin{aligned}
Q(\theta,\theta_k) = E_{\theta_k}[\log p_\theta(X_{1:T}, Y_{1:T})] &= -\frac{1}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}\left(M_{1|T} + \mu_0^2 - 2\tilde{m}_{1|T}\mu_0\right) + \\
&- \frac{T-1}{2}\log\tau^2 - \frac{1}{2\tau^2}\left(\sum_{t=2}^{T}\tilde{M}_{t|T} + \phi^2\sum_{t=2}^{T}\tilde{M}_{t-1|T} - 2\phi\sum_{t=2}^{T}\tilde{M}_{t,t-1|T}\right) + \\
&- \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{t=1}^{T} y_t^2 + \tilde{M}_{t|T} - 2\tilde{m}_{t|T}y_t\right).
\end{aligned}
$$

The two quantities (5.2.1a, 5.2.1b) can be obtained using the Kalman smoother described above, while (5.2.1c) can be computed as in Shumway and Stoffer [2000] with the following equation

$$\tilde{P}_{t,t-1|T} = \tilde{J}_{t-1}\tilde{P}_{t|T.}$$

M-step For the M-step, we take the derivative of Q with respect to each model parameter, and set it to zero to get the estimate:

$$
\begin{aligned}
\frac{\partial Q}{\partial\phi} &= \phi\sum_{t=2}^{T}\tilde{M}_{t-1|T} - \sum_{t=2}^{T}\tilde{M}_{t,t-1|T} = 0\\
\frac{\partial Q}{\partial\sigma^2} &= -T + \frac{\sum_{t=1}^{T} y_t^2 + \tilde{M}_{t|T} - 2\tilde{m}_{t|T}y_t}{\sigma^2} = 0\\
\frac{\partial Q}{\partial\tau^2} &= (T-1) - \frac{\sum_{t=2}^{T}\tilde{M}_{t|T} + \phi^2\sum_{t=2}^{T}\tilde{M}_{t-1|T} - 2\phi\sum_{t=2}^{T}\tilde{M}_{t,t-1|T}}{\tau^2}.
\end{aligned}
$$

The updates for the parameters can be found as

$$\begin{aligned}
\phi_{k+1} &= \frac{\sum_{t=2}^{T} \tilde{M}_{t,t-1|T}}{\sum_{t=2}^{T} \tilde{M}_{t-1|T}} \\
\sigma^2_{k+1} &= \frac{\sum_{t=1}^{T} y_t^2 + \tilde{M}_{t|T} - 2\tilde{m}_{t|T} y_t}{T} \\
\tau^2_{k+1} &= \frac{\sum_{t=2}^{T} \tilde{M}_{t|T} - \frac{\left(\sum_{t=2}^{T} \tilde{M}_{t,t-1|T}\right)^2}{\sum_{t=2}^{T} \tilde{M}_{t-1|T}}}{T-1} \\
&= \frac{\sum_{t=2}^{T} \tilde{M}_{t|T} - \phi_{k+1} \sum_{t=2}^{T} \tilde{M}_{t,t-1|T}}{T-1}.
\end{aligned}$$

Both E and M steps are iterated and convergence is monitored with the conditional likelihood function

$$\log p_{\theta_k}(Y_{1:T}) = \sum_{t=1}^{T} \log\left(N(y_t; \tilde{m}_{t|t-1}, \tilde{P}_{t|t-1} + \sigma_k^2)\right),$$

since

$$p_{\theta_k}(y_t|y_{1:t-1}) \sim N(y_t; \tilde{m}_{t|t-1}, \tilde{P}_{t|t-1} + \sigma_k^2).$$

The algorithm is said to have converged if the relative increment in the likelihood at the current time step compared to the previous time is below a certain threshold. We report here a simulation study to illustrate the approach based on the EM algorithm with Kalman smoother. Figure 5.2.1 shows the convergence of the parameter estimates. As we can see, after few iterations we reach the right value. Convergence is monitored via the relative increments in the likelihood at the current time step compared to the previous one (see Figure 5.2.2).

5.3 Pairwise likelihood inference via EM algorithm

In the previous section we have seen that the EM algorithm is an efficient iterative procedure to compute the maximum likelihood estimate in the presence of missing or hidden data. This method can be modified in order to obtain the maximum pairwise likelihood estimate in a general state space framework, provided that the algorithm increases the pairwise likelihood at each iteration.

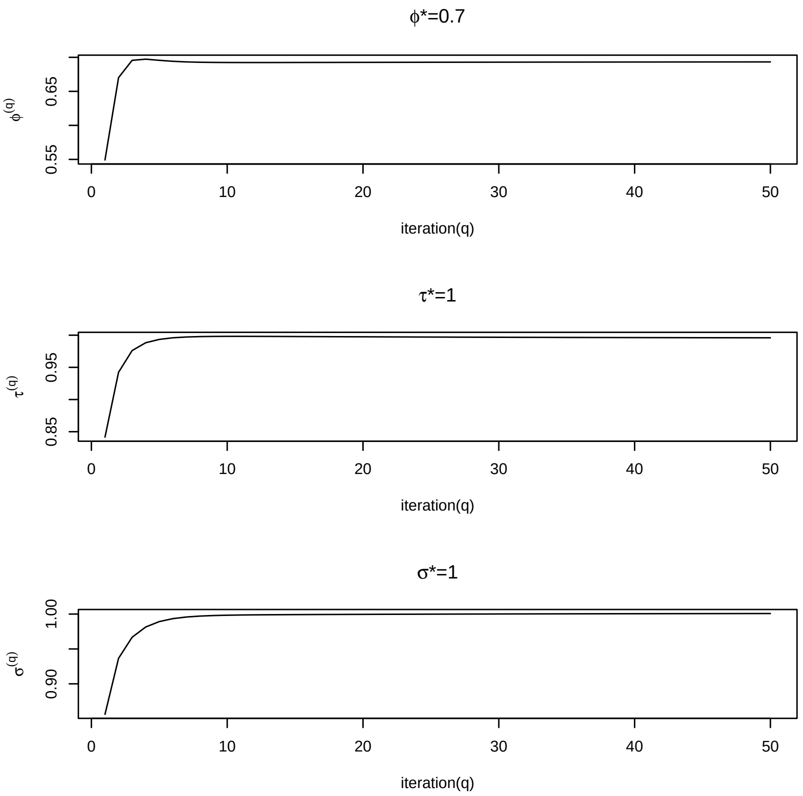

Figure 5.2.1: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Full likelihood estimation using the EM algorithm with Kalman smoother. Calculations based on a simulated series of length 10000. Initial value $\theta^{(0)} = (0.2, 0.5, 0.5)$.

Instead of the full likelihood, we want to minimize here, with respect to θ, the Kullback- Leibler divergence $K_P^{(L)}(\theta, \theta^*)$ as defined in (3.2.4), that is

$$K_P^{(L)}(\theta, \theta^*) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} \right].$$

We recall here that $p_\theta(y_1, y_j)$ is defined as (3.1.2), that is

$$\begin{aligned} p_\theta(y_1, y_j) &= \int_{\mathcal{X}^j} p_\theta(y_1, y_j, x_{1:j}) dx_{1:j} \\ &= \int_{\mathcal{X}^j} \pi_\theta(x_1) g_\theta(y_1|x_1) \left[\prod_{k=2}^{j} f_\theta(x_k|x_{k-1}) \right] g_\theta(y_j|x_j) dx_{1:j}. \end{aligned}$$

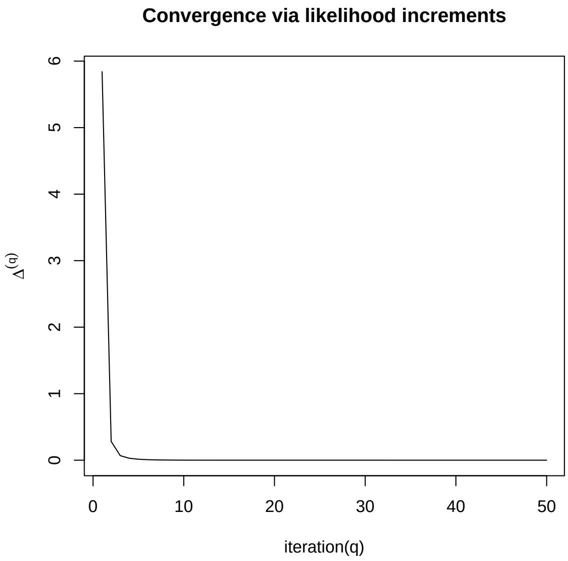

Figure 5.2.2: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Increment of the likelihood function at each iteration step.

Given an estimate θ_k of θ^*, at iteration $k + 1$ we update our estimate via

$$\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta, \theta_k),$$

where we define $Q(\theta, \theta_k)$ as

$$\begin{aligned} Q(\theta, \theta_k) &= \frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log[p_\theta(y_1, y_j, x_{1:j})]p_{\theta_k}(x_{1:j}|y_1, y_j)p_{\theta^*}(y_1, y_j)dx_{1:j}dy_1dy_j \\ &= \frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^{L+1}} \log[p_\theta(y_1, y_j, x_{1:j})]p_{\theta_k}(x_{1:j}|y_1, y_j)p_{\theta^*}(y_{1:L+1})dx_{1:j}dy_{1:L+1} \\ &= \frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^{L+1}} \log[p_\theta(y_1, y_j, x_{1:j})]p_{\theta_k}(x_{1:j}|y_1, y_j)p_{\theta^*}(\mathbf{y}_1)dx_{1:j}d\mathbf{y}_1, \end{aligned}$$

where $\mathbf{y}_s = y_{s:s+L}$ denote the s-th block of observations. For every $\theta \in \Theta$, we see that an iteration of this EM algorithm decreases the value of $K_P^{(L)}(\theta, \theta^*)$, and the stationary points correspond to the zeros of $K_P^{(L)}(\theta,\theta^*)$. More precisely,

$$
\begin{aligned}
&0 \leq Q(\theta_{k+1},\theta_k) - Q(\theta_k,\theta_k) = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log[p_{\theta_{k+1}}(y_1,y_j,x_{1:j})]p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j+ \\
&-\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log[p_{\theta_k}(y_1,y_j,x_{1:j})]p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \Big[\log[p_{\theta_{k+1}}(y_1,y_j,x_{1:j})] - \log[p_{\theta_k}(y_1,y_j,x_{1:j})]\Big]\, p_{\theta_k}(x_{1:j}|y_1,y_j)\times \\
&\qquad\qquad \times p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(y_1,y_j,x_{1:j})}{p_{\theta_k}(y_1,y_j,x_{1:j})}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)p_{\theta_{k+1}}(y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta_k}(y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)\times \\
&\qquad\qquad \times p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \left[\log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)} + \log\frac{p_{\theta_{k+1}}(y_1,y_j)}{p_{\theta_k}(y_1,y_j)}\right]p_{\theta_k}(x_{1:j}|y_1,y_j)\times \\
&\qquad\qquad \times p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j+ \\
&+\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(y_1,y_j)}{p_{\theta_k}(y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j+ \\
&+\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(y_1,y_j)}{p_{\theta_k}(y_1,y_j)}\left[\int_{\mathcal{X}^j} p_{\theta_k}(x_{1:j}|y_1,y_j)dx_{1:j}\right]p_{\theta^*}(y_1,y_j)dy_1dy_j = \\
=&\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2} \log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j+
\end{aligned}
$$

$$+\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{Y}^2}\log\frac{p_{\theta_{k+1}}(y_1,y_j)}{p_{\theta_k}(y_1,y_j)}p_{\theta^*}(y_1,y_j)dy_1dy_j.$$

Remembering the definition of $K_P^{(L)}(\theta,\theta^*)$, the quantity above equals

$$K_P^{(L)}(\theta_k,\theta^*)-K_P^{(L)}(\theta_{k+1},\theta^*)+$$
$$+\frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{X}^j\times\mathcal{Y}^2}\log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j.$$

For every $j=2,\ldots,L+1$, by Jensen inequality we have that

$$\int_{\mathcal{X}^j\times\mathcal{Y}^2}\log\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j$$
$$\leq\log\left[\int_{\mathcal{X}^j\times\mathcal{Y}^2}\frac{p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)}{p_{\theta_k}(x_{1:j}|y_1,y_j)}p_{\theta_k}(x_{1:j}|y_1,y_j)p_{\theta^*}(y_1,y_j)dx_{1:j}dy_1dy_j\right]=$$
$$=\log\left[\int_{\mathcal{Y}^2}\left[\int_{\mathcal{X}^j}p_{\theta_{k+1}}(x_{1:j}|y_1,y_j)dx_{1:j}\right]p_{\theta^*}(y_1,y_j)dy_1dy_j\right]=$$
$$=\log\left[\int_{\mathcal{Y}^2}p_{\theta^*}(y_1,y_j)dy_1dy_j\right]=\log[1]=0.$$

For these reasons

$$0\leq Q(\theta_{k+1},\theta_k)-Q(\theta_k,\theta_k)\leq K_P^{(L)}(\theta_k,\theta^*)-K_P^{(L)}(\theta_{k+1},\theta^*).\qquad(5.3.1)$$

The result above allows us to conclude that $K_P^{(L)}(\theta_{k+1},\theta^*)\leq K_P^{(L)}(\theta_k,\theta^*)$, that is at each iteration the value of the Kullback- Leibler divergence decreases.

Remark 5.3.1. *The inequality in (5.3.1) does not depend on the initial distribution. It holds even if the initial invariant distribution is replaced with any initial distribution, since all the steps can be derived exactly in the same way as before.*

In practice for the models we will consider, it is necessary to compute a set of sufficient statistics $\Phi(\theta_k,\theta^*)$ at time k in order to evaluate the function $Q(\theta,\theta_k)$. To do that we have to compute the expectation with respect to $p_{\theta_k}(x_{1:j}|y_1,y_j)\times p_{\theta^*}(\mathbf{y}_1)$. Even if it is possible to maximize $Q(\theta,\theta_k)$ analytically, in practice Q can not be computed as the expectation is with respect to a measure dependent on the unknown parameter value θ^*. However, thanks to the ergodicity and stationary assumptions, the observed process $\{Y_n\}$ provides us with a sample from $p_{\theta^*}(\mathbf{y}_1)$ which can be used for the purpose of Monte Carlo integration.

In the next section, we illustrate this method for a linear and gaussian model. It is a simple example, where the invariant distribution is known, as well as the conditional distribution of the latent states given the pairs of observations.

5.4 EM calculations for the linear gaussian model

Let us consider the linear gaussian model as defined in Example 2.0.1,

$$\begin{aligned} X_{n+1} &= \phi X_n + W_n, \quad W_n \sim N(0,\tau^2) \\ Y_n &= X_n + V_n, \quad V_n \sim N(0,\sigma^2). \end{aligned}$$

The choice of the parameter vector $\theta = (\phi, \tau^2, \sigma^2) \in (-1,1) \times \mathbb{R}^+ \times \mathbb{R}^+$ ensures stationarity. So

$$\begin{aligned} \pi_\theta(x) &= N\left(x; 0, \frac{\tau^2}{1-\phi^2}\right) \\ f_\theta(x'|x) &= N(x'; \phi x, \tau^2) \\ g_\theta(y|x) &= N(y; x, \sigma^2). \end{aligned}$$

We develop here an on line EM procedure, as suggested in the previous section. We recall that the function $Q(\theta, \theta_k)$ is defined as

$$Q(\theta,\theta_k) = \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\log[p_\theta(y_1, y_j, x_{1:j})]\right],$$

where $\mathbb{E}^{(j)}_{\theta_k,\theta^*}$ denotes the expectation with respect to $p_{\theta_k}(x_{1:j}|y_1, y_j) \times p_{\theta^*}(\mathbf{y}_1)$. In order to compute $Q(\theta, \theta_k)$, we have to derive $\log[p_\theta(y_1, y_j, x_{1:j})]$, for every $j = 2, \ldots, L+1$. We have that

$$\log[p_\theta(y_1, y_j, x_{1:j})] = \log[\pi_\theta(x_1)] + \log[g_\theta(y_1|x_1)] + \log[g_\theta(y_j|x_j)] +$$

$$+ \quad \sum_{k=2}^{j} \log[f_\theta(x_k|x_{k-1})].$$

From the definition of the model, the quantity above is proportional to

$$\begin{aligned}
&-\frac{1}{2}\log[\tau^2] + \frac{1}{2}\log[1-\phi^2] - \frac{x_1^2(1-\phi^2)}{2\tau^2} - \log[\sigma^2] - \frac{(y_1-x_1)^2}{2\sigma^2} + \\
&-\frac{(j-1)\log[\tau^2]}{2} - \frac{\sum_{k=2}^{j}(x_k-\phi x_{k-1})^2}{2\tau^2} - \frac{(y_j-x_j)^2}{2\sigma^2} = \\
&= -\frac{1}{2}\log[\tau^2] + \frac{1}{2}\log[1-\phi^2] - \frac{(j-1)\log[\tau^2]}{2} - \frac{(y_1-x_1)^2+(y_j-x_j)^2}{2\sigma^2} + \\
&-\log[\sigma^2] - \frac{1}{2\tau^2}\left(x_1^2 + \sum_{k=2}^{j} x_k^2 - \phi^2 x_1^2 + \phi^2 \sum_{k=2}^{j} x_{k-1}^2 - 2\phi \sum_{k=2}^{j} x_k x_{k-1}\right) = \\
&= -\frac{1}{2}\log[\tau^2] + \frac{1}{2}\log[1-\phi^2] - \frac{(j-1)\log[\tau^2]}{2} - \frac{y_1^2+y_j^2+x_1^2+x_j^2-2x_1y_1-2x_jy_j}{2\sigma^2} + \\
&-\log[\sigma^2] - \frac{1}{2\tau^2}\left(x_1^2 + x_j^2 + (1+\phi^2)\sum_{k=2}^{j-1} x_k^2 - 2\phi\sum_{k=2}^{j} x_k x_{k-1}\right). \qquad (5.4.1)
\end{aligned}$$

Using the linearity of Q and the expression for $\log[p_\theta(y_1, y_j, x_{1:j})]$ given by (5.4.1), we have that

$$\begin{aligned}
Q(\theta,\theta_k) &= \frac{1}{2}\log[1-\phi^2] - \frac{\log[\tau^2]}{2} - \frac{1}{2}\log[\tau^2]\left(\frac{1}{L}\frac{L(L+1)}{2}\right) - \log[\sigma^2] + \\
&-\frac{1}{2\tau^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_1^2+X_j^2] + (1+\phi^2)\sum_{k=2}^{j-1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k^2] - 2\phi\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_kX_{k-1}]\right] + \\
&-\frac{1}{2\sigma^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2+X_1^2-2X_1Y_1+Y_j^2+X_j^2-2X_jY_j]\right].
\end{aligned}$$

In practice, for this model, it is necessary to compute a set of sufficient statistics $\Phi_i(\theta_k,\theta^*), i = 1,\ldots,4$ at time k, where

$$\begin{aligned}
\Phi_1(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2+X_1^2-2X_1Y_1+Y_j^2+X_j^2-2X_jY_j] \\
\Phi_2(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_1^2+X_j^2]
\end{aligned}$$

$$\begin{aligned}\Phi_3(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=1}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=2}^{j-1}X_k^2\right]\\ \Phi_4(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=2}^{j}X_kX_{k-1}\right].\end{aligned}$$

With this definition

$$Q(\theta,\theta_k)=\frac{1}{2}\log[1-\phi^2]-\frac{\log[\tau^2]}{2}-\frac{(L+1)\log[\tau^2]}{4}-\log[\sigma^2]+ \\ -\frac{1}{2\tau^2}\Big(\Phi_2(\theta_k,\theta^*)+(1+\phi^2)\Phi_3(\theta_k,\theta^*)-2\phi\Phi_4(\theta_k,\theta^*)\Big)-\frac{1}{2\sigma^2}\Phi_1(\theta_k,\theta^*).$$

Now, dropping for simplicity the dependence on θ,θ^*,θ_k,

$$\begin{aligned}\frac{\partial Q}{\partial\phi} &= \frac{\phi}{1-\phi^2}+\frac{1}{\tau^2}(\phi\Phi_3-\Phi_4)=0\\ \frac{\partial Q}{\partial\tau^2} &= 1+\frac{L+1}{2}-\frac{1}{\tau^2}(\Phi_2+(1+\phi^2)\Phi_3-2\phi\Phi_4)=0\\ \frac{\partial Q}{\partial\sigma^2} &= 1-\frac{1}{2\sigma^2}\Phi_1,\end{aligned}$$

so

$$\begin{aligned}\sigma^2 &= \frac{\Phi_1}{2}\\ 0 &= \phi\tau^2+(1-\phi^2)(\phi\Phi_3-\Phi_4)\\ \tau^2 &= \frac{2}{L+3}(\Phi_2+(1+\phi^2)\Phi_3-2\phi\Phi_4).\end{aligned}$$

We look at the case of ϕ and τ^2:

$$\Phi_3(C-1)\phi^3-(2C-1)\Phi_4\phi^2+(C\Phi_2+(C+1)\Phi_3)\phi-\Phi_4=0 \\ \tau^2=C(\Phi_2+(1+\phi^2)\Phi_3-2\phi\Phi_4),$$

where we denote by C the quantity $\frac{2}{L+3}$. One can solve the cubic analytically, and finds three solutions

$$\phi=u-\frac{p}{3u}-\frac{(1-2C)\Phi_3}{3\Phi_2(C-1)},$$

with

$$u^3 = \frac{1}{2}\left(q \pm \sqrt{q^2 + \frac{4}{27}p^3}\right)$$

where

$$\begin{aligned} p &= \frac{3\left(\frac{(C\Phi_1+(C+1)\Phi_2)}{\Phi_2(C-1)}\right) - \left(\frac{(1-2C)\Phi_3}{\Phi_2(C-1)}\right)^2}{3} \\ q &= \frac{-27\left(\frac{\Phi_3}{\Phi_2(C-1)}\right) + 2\left(\frac{(1-2C)\Phi_3}{\Phi_2(C-1)}\right)^3 - 9\left(\frac{(1-2C)\Phi_3}{\Phi_2(C-1)}\frac{(C\Phi_1+(C+1)\Phi_2)}{\Phi_2(C-1)}\right)}{27}. \end{aligned}$$

We discard solutions that fall outside the interval $[-1, 1]$ and keep among the remaining values. The corresponding values of τ^2 that maximizes $Q(\theta, \theta_k)$ is given by

$$\tau^2(\phi) = \begin{cases} C(\Phi_2 + (1+\phi^2)\Phi_3 - 2\phi\Phi_4) & \text{if } (\Phi_2 + (1+\phi^2)\Phi_3 - 2\phi\Phi_4) > 0; \\ 0 & \text{otherwise} \end{cases}$$

since if $(\Phi_2 + (1+\phi^2)\Phi_3 - 2\phi\Phi_4) < 0$ then $\frac{\partial Q}{\partial \tau^2} > 0$ and $\tau^2 = 0$ is the solution.

Now we need to compute the sufficient statistics $\Phi_i(\theta, \theta^*)$ for $i = 1, \ldots, 4$. To do that we have to compute the expectation with respect to $p_{\theta_k}(x_{1:j}|y_1, y_j) \times p_{\theta^*}(\mathbf{y}_1)$. Even if, in this case, it is possible to maximize $Q(\theta, \theta_k)$ analytically, as we have seen above, in practice Q can not be computed. However, thanks to the ergodicity and stationary assumptions, this algorithm can be approximated using the following on line scheme. For every $i = 1, \ldots, 4$, we recursively approximate the sufficient statistics $\Phi_i(\theta_k, \theta^*)$ with the following update, given here at time k,

$$\hat{\Phi}_i^{(k)} = (1-\gamma_k)\hat{\Phi}_i^{k-1} + \gamma_k \left[\frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]\right], \tag{5.4.2}$$

where, for every function $h(\cdot)$, $\mathbb{E}_{\theta_k}^{(j)}[h(X_{1:j})|\mathbf{Y}_k]$ denotes the expectation of h with respect to $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ and for $i = 1, \ldots, 4$ we have implicitly defined

$$\Phi_i(\theta_k, \theta^*) := \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta_k,\theta^*}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})].$$

We then substitute $\hat{\Phi}_i^{(k)}$ for $\Phi_i(\theta_k, \theta^*)$ and obtain θ_k by maximizing the Q function. If θ_k was constant and $\gamma_k = k^{-1}$, then $\hat{\Phi}_i^{(k)}$ would simply compute the arithmetic average of $\{\mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]\}$ for every $j = 2, \ldots, L+1$, and converge towards $\Phi(\theta_k, \theta^*)$ by ergodicity. In fact, under mild suitable conditions, convergence is in general ensured for any non- increasing positive sequence $\{\gamma_k\}$ such that $\sum \gamma_k < \infty$ and $\sum \gamma_k^2 < \infty$. We can select $\gamma_k = Mk^{-\alpha}$ where $M > 0$ and $\frac{1}{2} < \alpha \leq 1$ thanks to the theory of stochastic approximation [Benveniste et al., 1990].

Going back to our example (linear gaussian model), $\mathbb{E}_{\theta_k}^{j}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$ is known for every $j = 2, \ldots, L+1$, since the expectation is w.r.t. $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$. In this case, we do not need to use a further Monte Carlo approximation. For a generic $\theta \in \Theta$, the density $p_\theta(x_{1:j}|y_1, y_j) \propto p_\theta(y_1, y_j|x_{1:j})p_\theta(x_{1:j})$ can be derived in the following way. Given the structure of the model

$$p_\theta(y_1, y_j|x_{1:j}) = p_\theta(y_1, y_j|x_1, x_j) = g_\theta(y_1|x_1)g_\theta(y_j|x_j),$$

with $g_\theta(y|x) = N(y; x, \sigma^2)$. For every $j = 1, \ldots, L+1$, $p_\theta(y_1, y_j|x_{1:j})$ can be written as proportional to

$$\exp\left\{-\frac{1}{2}\left[\begin{pmatrix} y_1 \\ 0 \\ \vdots \\ y_j \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \end{pmatrix}\right]^T \begin{pmatrix} \sigma^{-2} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^{-2} \end{pmatrix} \left[\begin{pmatrix} y_1 \\ 0 \\ \vdots \\ y_j \end{pmatrix} - \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \end{pmatrix}\right]\right\},$$

and we denote by Σ_y^{-1},

$$\Sigma_y^{-1} = \begin{pmatrix} \sigma^{-2} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^{-2} \end{pmatrix}.$$

On the other hand, the prior $p_\theta(x_{1:j})$ is a normal distribution $N_j(x_{1:j}; \mu_x, \Sigma_x)$. In fact, with $x = (x_1, \dots, x_j)^T$

$$\begin{aligned} x &= Ax + \tau\epsilon \\ x &= \tau M^{-1}\epsilon, \end{aligned}$$

where

$$A = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ \phi & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \phi & 0 \end{pmatrix}$$

and

$$M = I - A = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -\phi & 1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\phi & 1 \end{pmatrix}.$$

In this example, since the invariant distribution in known, $X_1 \sim N\left(x_1; 0, \frac{\tau^2}{1-\phi^2}\right)$ and hence

$$\begin{aligned} \mu_x &= \tau M^{-1}\mu_\epsilon \\ \Sigma_x &= \tau^2 M^{-1}\Sigma_\epsilon M^{-T}, \end{aligned}$$

with

$$\mu_\epsilon = 0 \quad \text{and} \quad \Sigma_\epsilon = \begin{pmatrix} \frac{1}{1-\phi^2} & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}.$$

So the posterior distribution $p_\theta(x_{1:j}|y_1, y_j)$ is also normal $N_j\left(x_{1:j}; \mu, \Sigma\right)$, with

$$\mu = (\Sigma_x^{-1} + \Sigma_y^{-1})^{-1}(\Sigma_x^{-1}\mu_x + \Sigma_y^{-1}y) \tag{5.4.3a}$$

$$\Sigma \;=\; (\Sigma_x^{-1} + \Sigma_y^{-1})^{-1} \tag{5.4.3b}$$

where $y = (y_1, 0, \ldots, y_j)^T$. We are now able to approximate the statistics $\Phi_i(\theta, \theta^*)$ for $i = 1, \ldots, 4$ using the technique described above and the posterior normal distribution with parameter given by (5.4.3a, 5.4.3b). All we need in this computation are the first and second moments of the posterior distribution $p_{\theta_k}(x_{1:j}|y_k, y_{k+j})$ for $j = 2, \ldots, L + 1$.

Remark 5.4.1. *The on line algorithm described above takes a block of observations $\mathbf{y}_k$ for each iteration of the Expectation- Maximization steps. It can be modified in order to consider more blocks in each iteration or to run more than one iteration for a single block.*

Remark 5.4.2. *Calculation of the posterior density $p_\theta(x_{1:j}|y_1, y_j)$, and in particular its first and second moments, can be achieved by a modification of the general Kalman filter and smoother outlined in the previous section. Unlike the standard Kalman filter equations, in this contest, the conditioning is on the observations y_1 and y_j and not on all the observations between 1 and j. Prediction and update steps need to be modified in order to obtain the right moments of the posterior distributions $p_\theta(x_{1:j}|y_1, y_j)$. Roughly speaking, we pretend to run a Kalman filter with all the observations $y_{1:j}$ setting an infinity variance for the missing observations from time 2 to time $j - 1$.*

More precisely, if $x_0 \sim N(m_{0|0}, P_{0|0})$, for every $j > 2$, the prediction and update steps modify as follows

- *Initialization*

$$\begin{aligned}
m_{1|0} &= \phi m_{0|0} \\
P_{1|0} &= \phi^2 P_{0|0} + \tau^2 \\
e_1 &= y_1 - m_{1|0} \\
S_1 &= P_{1|0} + \sigma^2 \\
K_1 &= P_{1|0}(S_1)^{-1} \\
m_{1|1} &= m_{1|0} + K_1 e_1 \\
P_{1|1} &= (I - \tilde{K}_1)\tilde{P}_{1|0}.
\end{aligned}$$

- *For $k = 2, \dots, j-1$*

$$\begin{aligned} m_{k|k-1} &= \phi m_{k-1|k-1} \\ P_{k|k-1} &= \phi^2 P_{k-1|k-1} + \tau^2 \\ m_{k|k} &= m_{k|k-1} \\ P_{k|k} &= P_{k|k-1}. \end{aligned}$$

- *For k=j*

$$\begin{aligned} m_{j|j-1} &= \phi m_{j-1|j-1} \\ P_{j|j-1} &= \phi^2 P_{j-1|j-1} + \tau^2 \\ e_j &= y_j - m_{j|j-1} \\ S_j &= P_{j|j-1} + \sigma^2 \\ K_j &= P_{j|j-1}(S_j)^{-1} \\ m_{j|j} &= m_{j|j-1} + K_j e_j \\ P_{j|j} &= (I - \tilde{K}_j)\tilde{P}_{j|j-1}. \end{aligned}$$

Note that for $k = 2, \dots, j-1$ we actually do not need to compute quantities that depend on the variance σ^2. Since σ^2 tends to infinity, K_k can be set equal to zero. The innovation at time k (i.e. e_k) and its covariance (i.e. S_k) do not need to be computed, and this allows us to avoid dealing with infinite quantities. Moreover, the meaning of the steps for $k = 2, \dots, j-1$ is quite sensible: if we do not take into account the observations $y_{2:j-1}$, the update step is missing and so the predicted and updated estimates coincide.

The vector μ and the matrix Σ defined in (5.4.3a, 5.4.3b) can be obtained from the smoothing recursions, i. e., starting at $k = j$

$$\begin{aligned} J_k &= \phi P_{k|k}(P_{k+1|k})^{-1} \\ m_{k|j} &= m_{k|k} + J_k(m_{k+1|j} - m_{k+1|k}) \\ P_{k|j} &= P_{k|j} + J_k^2(P_{k+1|j} - P_{k+1|k}). \end{aligned}$$

More precisely,

$$\mu := \begin{pmatrix} \mathbb{E}[x_1|Y_1, Y_j] \\ \mathbb{E}[x_2|Y_1, Y_j] \\ \vdots \\ \mathbb{E}[x_j|Y_1, Y_j] \end{pmatrix} = \begin{pmatrix} m_{1|j} \\ m_{2|j} \\ \vdots \\ m_{j|j} \end{pmatrix},$$

$$\begin{aligned} \Sigma \quad &:= \begin{pmatrix} \mathbb{V}[x_1|Y_1, Y_j] & \mathbb{C}[x_1, x_2|Y_1, Y_j] & \dots & \mathbb{C}[x_1, x_j|Y_1, Y_j] \\ \mathbb{C}[x_2, x_1|Y_1, Y_j] & \mathbb{V}[x_2|Y_1, Y_j] & \dots & \mathbb{C}[x_2, x_j|Y_1, Y_j] \\ \vdots & \ddots & \dots & \vdots \\ \mathbb{C}[x_j, x_1|Y_1, Y_j] & \dots & \dots & \mathbb{V}[x_j|Y_1, Y_j] \end{pmatrix} \\ &= \begin{pmatrix} P_{1|j} & P_{1,2|j} & \dots & P_{1,2|j} \\ P_{2,1|j} & P_{2|j} & \dots & P_{2,j|j} \\ \vdots & \ddots & \dots & \vdots \\ P_{j,1|j} & \dots & \dots & P_{j|j} \end{pmatrix}, \end{aligned}$$

where

$$P_{k,k-1|j} = J_{k-1} P_{k|j}.$$

We implement the on line EM algorithm described above in order to estimate the parameter $\theta = (\phi, \tau, \sigma)$ of the linear gaussian model. We consider a simulated time series of length $n = 10000$ from the linear gaussian model, with $\phi^* = 0.7$, $\sigma^* = 1, \tau^* = 1$ as true parameter values. Taking into account the empirical results given in Chapter 4, we decide to fix the maximum lag distance between the observations as $L = 4$. In fact, $L = 4$ seems to be the best maximum distance in terms of variance of the estimator. In order to reduce the variance of the estimate, we used the Polyak- Ruppert averaging procedure. The algorithm was ran with $\gamma_k = k^{-0.5}$ for $k \leq 2000$ and $\gamma_k = (k - 2000)^{-0.8}$ for $k > 2000$. The results of this method are displayed in Figure 5.4.1. We see that the convergence to the true value is reached in few iteration steps. The code, implemented in R, has a very low computational burden and even if we have considered a quite long time series, it takes few seconds to successfully conclude.

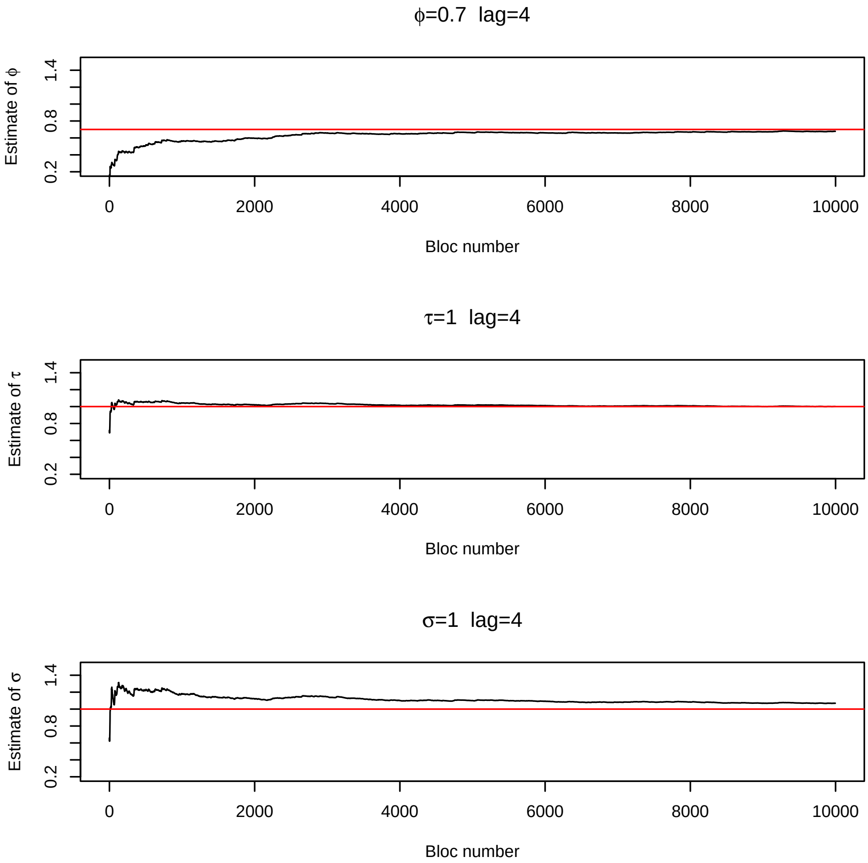

Figure 5.4.1: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Pairwise likelihood estimation using the on line EM algorithm with lag=4 denoting the maximum distance between the observations. Calculations based on a simulated series of length 10000. Initial value $\theta^{(0)} = (0.2, 0.5, 0.5)$.

5.4.1 Case with conditioning on x_1

Even if in this example the invariant distribution is known, we describe here how to modify the calculation above in order to obtain the conditional joint distribution given an initial value x_1. This is useful when invariant distribution is unknown and initial value is set equal to x_1. As we will see, calculation will be simpler, but it comes with a prize. The bias of the estimate introduced by replacing the invariant distribution with a Dirac delta mass at x_1 effects the convergence of the EM algorithm.

We report here the calculation for the linear gaussian model with observation noise, in the case where the invariant distribution is supposed to be unknown. In order to compute $Q(\theta,\theta_k)$, we have to derive $\log[p_\theta(y_1,y_j,x_{1:j})]$, for every $j = 2,\dots,L+1$. We have that

$$\begin{aligned}\log[p_\theta(y_1,y_j,x_{1:j})] &= \log[\delta_{x_1}(x_1)] + \log[g_\theta(y_1|x_1)] + \log[g_\theta(y_j|x_j)] + \\ &+ \sum_{k=2}^{j} \log[f_\theta(x_k|x_{k-1})].\end{aligned}$$

Since $\delta_{x_0}(x_1)$ does not depend on θ, the quantity above is proportional to

$$\begin{aligned}&-\log[\sigma^2] - \frac{(y_1-x_1)^2}{2\sigma^2} - \frac{(j-1)\log[\tau^2]}{2} - \frac{\sum_{k=2}^{j}(x_k-\phi x_{k-1})^2}{2\tau^2} - \frac{(y_j-x_j)^2}{2\sigma^2} = \\ &= -\frac{(j-1)\log[\tau^2]}{2} - \frac{(y_1-x_1)^2+(y_j-x_j)^2}{2\sigma^2} + \\ &-\log[\sigma^2] - \frac{1}{2\tau^2}\left(\sum_{k=2}^{j} x_k^2 + \phi^2 \sum_{k=2}^{j} x_{k-1}^2 - 2\phi\sum_{k=2}^{j} x_k x_{k-1}\right).\end{aligned} \tag{5.4.4}$$

Using the linearity of Q and the expression for $\log[p_\theta(y_1,y_j,x_{1:j})]$ given by (5.4.4), we have that

$$\begin{aligned}&Q(\theta,\theta_k) = -\frac{1}{2}\log[\tau^2]\left(\frac{L+1}{2}\right) - \log[\sigma^2] + \\ &-\frac{1}{2\tau^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k^2] + \phi^2\sum_{k=2}^{j-1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_{k-1}^2] - 2\phi\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k X_{k-1}]\right] + \\ &-\frac{1}{2\sigma^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2 + X_1^2 - 2X_1Y_1 + Y_j^2 + X_j^2 - 2X_jY_j]\right].\end{aligned}$$

In practice, for this model, it is necessary to compute a set of sufficient statistics $\tilde{\Phi}_i(\theta_k,\theta^*), i = 1,\dots,4$ at time k, where

$$\begin{aligned}\tilde{\Phi}_1(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2 + X_1^2 - 2X_1Y_1 + Y_j^2 + X_j^2 - 2X_jY_j] \\ \tilde{\Phi}_2(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k^2]\end{aligned}$$

$$\begin{aligned}
\tilde{\Phi}_3(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X^2_{k-1}] \\
\tilde{\Phi}_4(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=2}^{j} X_k X_{k-1}\right].
\end{aligned}$$

With this definition

$$Q(\theta, \theta_k) = -\frac{1}{2}\log[\tau^2]\left(\frac{(L+1)}{2}\right) - \log[\sigma^2] +$$

$$-\frac{1}{2\tau^2}\left(\tilde{\Phi}_2(\theta_k, \theta^*) + \phi^2\tilde{\Phi}_3(\theta_k, \theta^*) - 2\phi\tilde{\Phi}_4(\theta_k, \theta^*) - \frac{1}{2\sigma^2}\tilde{\Phi}_1(\theta_k, \theta^*)\right).$$

Now, dropping for simplicity the dependence on $\theta, \theta^*, \theta_k$,

$$\begin{aligned}
\frac{\partial Q}{\partial \phi} &= -\frac{1}{2\tau^2}\left(2\phi\tilde{\Phi}_3 - 2\tilde{\Phi}_4\right) = 0 \\
\frac{\partial Q}{\partial \tau^2} &= -\frac{L+1}{4\tau^2} + \frac{1}{2\tau^4}\left(\tilde{\Phi}_2 + \phi^2\tilde{\Phi}_3 - 2\phi\tilde{\Phi}_4\right) = 0 \\
\frac{\partial Q}{\partial \sigma^2} &= -\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\tilde{\Phi}_1 = 0,
\end{aligned}$$

so

$$\begin{aligned}
\phi &= \frac{\tilde{\Phi}_4}{\tilde{\Phi}_3} \\
\tau^2 &= \frac{2}{L+1}\left(\tilde{\Phi}_2 + \phi^2\tilde{\Phi}_3 - 2\phi\tilde{\Phi}_4\right) \\
\sigma^2 &= \frac{\tilde{\Phi}_1}{2}.
\end{aligned}$$

Again, we need to compute the sufficient statistics $\tilde{\Phi}_i(\theta, \theta^*)$ for $i = 1, \ldots, 4$. The technique is exactly the same as when the invariant distribution is known, the only difference concerns the derivation of the joint posterior distribution. More precisely, given a starting value x_1, the joint distribution of the latent states given the observations y_1, y_j can be obtain as before, where we substitute μ_ϵ and Σ_ϵ by

$$\mu_\epsilon = \begin{pmatrix} \frac{x_1}{\tau} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \Sigma_\epsilon = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}.$$

Remark 5.4.3. *As before, the joint distribution above can be achieved using the Kalman filter and smoothing recursion. To get the conditioning on the initial value x_1, it is sufficient to set $m_{0|0} = x_1$ and $P_{0|0} = 0$. Calculations of the first and second moments proceed exactly in the same way.*

If we implement such on line EM algorithm, we face with the convergence problem due to the bias of the estimate arising from the substitution of the unknown invariant distribution with the Dirac delta mass density function (see Figure 5.4.2).

In light of this, we develop the strategy suggested in Equation (3.3.3), since the invariant distribution is supposed to be unknown, but transitions $f_\theta(\cdot|x)$ are simple. In practice, we approximate the invariant distribution sampling from the transition kernel $f_\theta(\cdot|x)$ and we take advantage of the geometric ergodicity of the process.

5.4.2 Approximation of the invariant distribution

We take a generic initial distribution $\mu(\cdot)$ for x_{-z} and we simulate a sufficiently long Markov chain from the transition kernel. Under geometric ergodicity, the marginal distribution of x_1 converges to $\pi(x_1)$ as z goes to $+\infty$.

In order to take into account the state before time 0, we define the function $Q_z(\theta, \theta_k)$ as follows

$$Q_z(\theta, \theta_k) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta_k, \theta^*}[\log[p_\theta(y_1, y_j, x_{-z:j})]] =$$

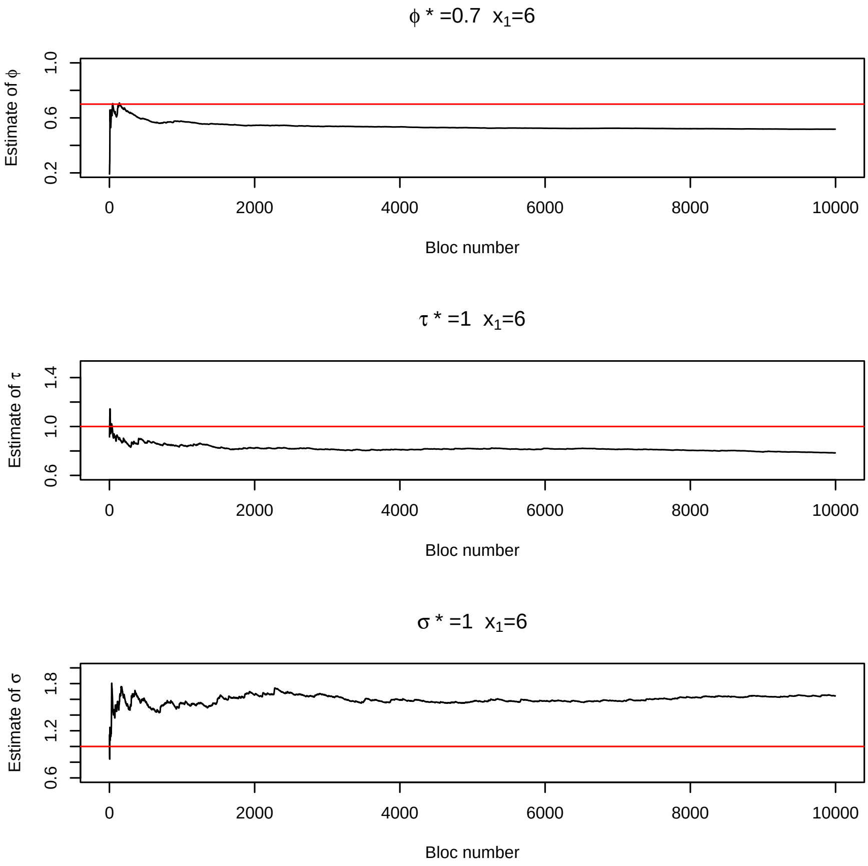

Figure 5.4.2: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Estimation of the parameters using the on line EM algorithm with lag=4 denoting the maximum distance between the observations. Calculations based on a simulated series of length 10000. Starting value $\theta^{(0)} = (0.2, 0.8, 0.5)$. We suppose here that the invariant distribution is unknown and we set as initial distribution $X_1 \sim \delta(x_1)$, where $x_1 = 6$.

$$= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\log[\mu_{x_{-z}}(x)] + \log[g_\theta(y_1|x_1)] + \log[g_\theta(y_j|x_j)] + \sum_{k=-z+1}^{j}\log[f_\theta(x_k|x_{k-1})]\right], \tag{5.4.5}$$

where $\mathbb{E}^{(j)}_{\theta_k,\theta^*}$ now denotes the expectation with respect to $p_{\theta_k}(x_{-z:j}|y_1, y_j)p_{\theta^*}(y_1, y_j)$. If we choose $\mu_{x_{-z}}(\cdot)$ independent of θ, calculation of (5.4.5) and its maximization is derived in the same way as above. Again, it is necessary to compute a set of sufficient statistics $\bar{\Phi}_i(\theta_k, \theta^*), i = 1, \dots, 4$ at time k, where

$$\begin{aligned}
\bar{\Phi}_1(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2 + X_1^2 - 2X_1Y_1 + Y_j^2 + X_j^2 - 2X_jY_j]\\
\bar{\Phi}_2(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\sum_{k=-z+1}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k^2]\\
\bar{\Phi}_3(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\sum_{k=-z+1}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_{k-1}^2]\\
\bar{\Phi}_4(\theta_k,\theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=-z+1}^{j}X_kX_{k-1}\right].
\end{aligned}$$

With this definition

$$\begin{aligned}
&Q_z(\theta,\theta_k) = -\frac{1}{2}\log[\tau^2]\left(\frac{L^2+2L+2Lz}{2}\right) - \log[\sigma^2] + \\
&-\frac{1}{2\tau^2}\left(\bar{\Phi}_2(\theta_k,\theta^*) + \phi^2\bar{\Phi}_3(\theta_k,\theta^*) - 2\phi\bar{\Phi}_4(\theta_k,\theta^*) - \frac{1}{2\sigma^2}\bar{\Phi}_1(\theta_k,\theta^*)\right).
\end{aligned}$$

Now, dropping for simplicity the dependence on $\theta, \theta^*, \theta_k$,

$$\begin{aligned}
\phi &= \frac{\bar{\Phi}_4}{\bar{\Phi}_3}\\
\tau^2 &= \frac{2}{L^2+2L+2Lz}\left(\bar{\Phi}_2 + \phi^2\bar{\Phi}_3 - 2\phi\bar{\Phi}_4\right)\\
\sigma^2 &= \frac{\bar{\Phi}_1}{2}.
\end{aligned}$$

As before, we need to compute the sufficient statistics $\bar{\Phi}_i(\theta, \theta^*)$ for $i = 1, \ldots, 4$. The technique is exactly the same as described in Remark 5.4.2, the only difference concerns the derivation of the joint posterior distribution. Calculation of the posterior density $p_\theta(x_{-z:j}|y_1, y_j)$, and in particular its first and second moments, can be achieved from the following recursions

- For $k = -z + 1, \ldots, 0$, we obtain the predicted states

$$\begin{aligned} \bar{m}_{k|k-1} &= \phi \bar{m}_{k-1|k-1} \\ \bar{P}_{k|k-1} &= \phi^2 \bar{P}_{k-1|k-1} + \tau^2 \\ \bar{m}_{k|k} &= \bar{m}_{k|k-1} \\ \bar{P}_{k|k} &= \bar{P}_{k|k-1}. \end{aligned}$$

- Initialization

$$\begin{aligned} \bar{m}_{1|0} &= \phi \bar{m}_{0|0} \\ \bar{P}_{1|0} &= \phi^2 \bar{P}_{0|0} + \tau^2 \\ \bar{e}_1 &= y_1 - \bar{m}_{1|0} \\ \bar{S}_1 &= \bar{P}_{1|0} + \sigma^2 \\ \bar{K}_1 &= \bar{P}_{1|0}(\bar{S}_1)^{-1} \\ \bar{m}_{1|1} &= \bar{m}_{1|0} + \bar{K}_1 \bar{e}_1 \\ \bar{P}_{1|1} &= (I - \bar{K}_1)\bar{P}_{1|0}. \end{aligned}$$

- For $k = 2, \ldots, j - 1$, we obtain the predicted states

$$\begin{aligned} \bar{m}_{k|k-1}(r_{-z+2:k}) &= \phi \bar{m}_{k-1|k-1} \\ \bar{P}_{k|k-1}(r_{-z+2:k}) &= \phi^2 \bar{P}_{k-1|k-1} + \tau^2 \\ \bar{m}_{k|k}(r_{-z+2:k}) &= \bar{m}_{k|k-1} \\ \bar{P}_{k|k} &= \bar{P}_{k|k-1}. \end{aligned}$$

- For k=j

$$\begin{aligned}
\bar{m}_{j|j-1} &= \phi \bar{m}_{j-1|j-1} \\
\bar{P}_{j|j-1} &= \phi^2 \bar{P}_{j-1|j-1} + \tau^2 \\
\bar{e}_j &= y_j - \bar{m}_{j|j-1} \\
\bar{S}_j &= \bar{P}_{j|j-1} + \sigma^2 \\
\bar{K}_j &= \bar{P}_{j|j-1} (\bar{S}_j)^{-1} \\
\bar{m}_{j|j} &= \bar{m}_{j|j-1} + \bar{K}_j \bar{e}_j \\
\bar{P}_{j|j} &= (I - \bar{K}_j) \bar{P}_{j|j-1}.
\end{aligned}$$

Starting from these filtering quantities, we can compute the smoothing estimates and derive the first and second moments of the latent states with respect to $p_{\theta_k}(x_{-z:j}|y_1, y_j)$. These quantities allow us to compute the sufficient statistics $\bar{\Phi}_i(\theta, \theta^*)$ for $i = 1, \ldots, 4$.

We implement this idea for the AR(1) model, where stationary distribution is supposed to be unknown. We set as initial distribution $X_{-z} \sim \delta_6(x_{-z})$ and we take $z = 100$. As we can see in Figure 5.4.3, this technique reduces the bias in the estimates for each parameter in the model.

We also report the distance between the estimate obtained taking δ_x, $x = 6$ as initial distribution and the estimate when the stationary distribution is known. In order to see how the idea suggested in (3.3.3) is useful, we compare it with the distance between the estimate obtained by approximate the invariant distribution by running a Markov chain of length $z = 100$ and the estimate when the stationary distribution is known. Reduction of the bias is displayed in Figure 5.4.4.

Figure 5.4.5 reports the distance of the estimate with respect to the true parameter values.

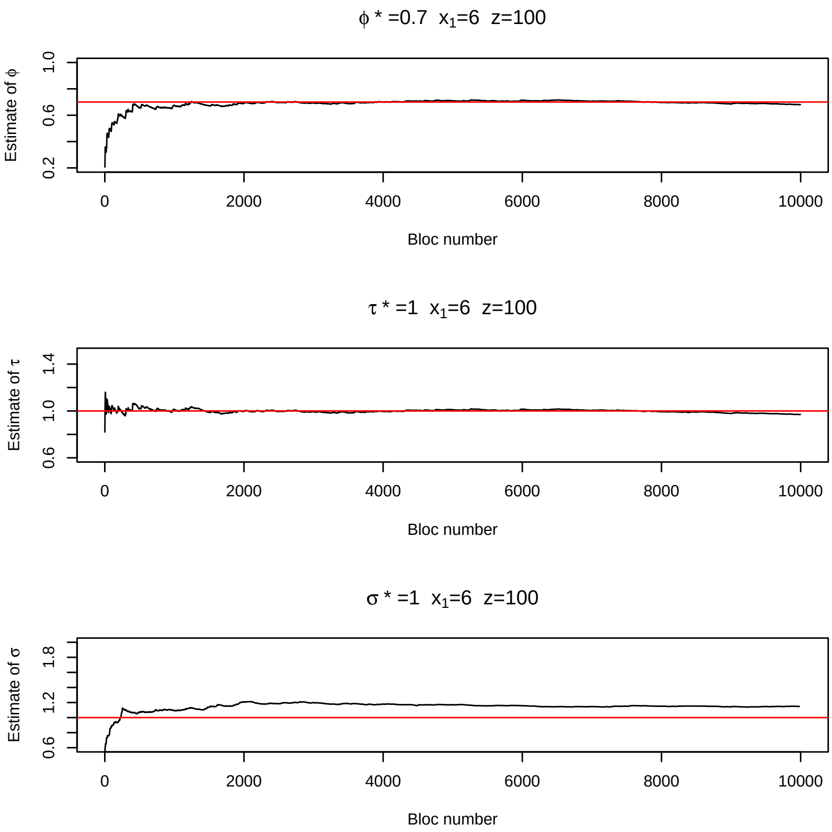

Figure 5.4.3: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Estimation of the parameter using the on line EM algorithm with lag=4 denoting the maximum distance between the observations. Calculations based on a simulated series of length 10000. We suppose here that the invariant distribution is unknown and we set as initial distribution $X_{-z} \sim \delta(x_1)$, where $x_1 = 6$ and $z = 100$.

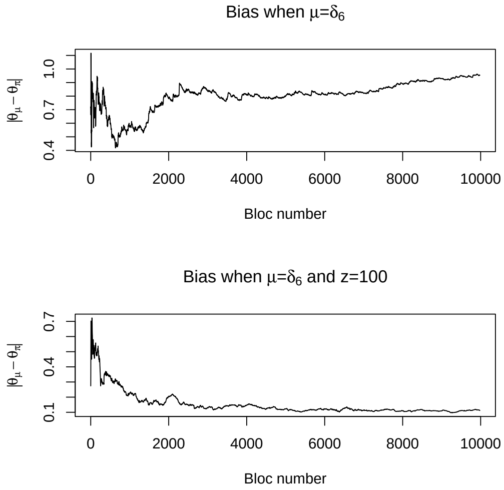

Figure 5.4.4: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Bias of the estimates when the invariant distribution is unknown and is approximated by taking as initial distribution $X_1 \sim \delta(x)$, where $x = 6$ (top) and $X_{-z} \sim \delta(x)$, where $x = 6$ and $z = 100$ (bottom).

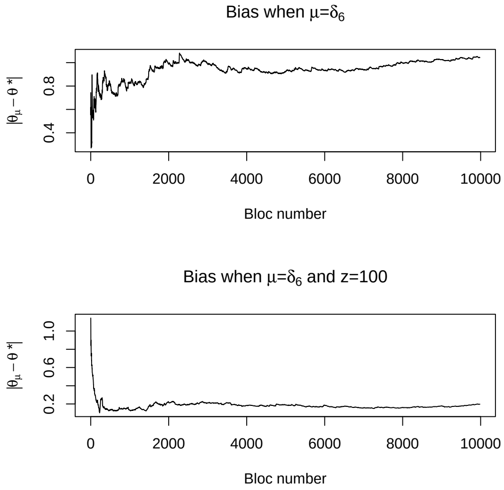

Figure 5.4.5: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Bias of the estimates with respect to the true parameter values when the invariant distribution is unknown and is approximated by taking as initial distribution $X_1 \sim \delta(x)$, where $x = 6$ (top) and $X_{-z} \sim \delta(x)$, where $x = 6$ and $z = 100$ (bottom).

Chapter 6

Jump Markov Linear System

Many variables undergo episodes in which the behavior of the series seems to change quite dramatically: this happens in almost any macroeconomic or financial time series observed for a sufficiently long period. In order to forecast this kind of series in a sensible way, the change in the process has to be seen as a random variable itself and a complete time series model would therefore include a description of the probability law governing the change in regime. The simplest model for the regime variable is a discrete-time Markov chain.

Jump Markov linear systems (JMLS) are linear systems whose parameters evolve with time according to a finite state Markov chain. These models are also used in several fields of econometrics as well as of signal processing, and include as particular cases common models in impulse deconvolution [Mendel, 1990, Doucet and Duvaut, 1997], digital communications [Krishnamurthy and Logothetis, 1999] and target tracking [Bar-Shalom and Li, 1995]. We model the latent process as an autoregressive (AR) model with switching, as introduced by Sclove [1983] and Hamilton [1989]. This process is observed with additive observation noise.

6.1 Definition of the model

Let r_t denote a discrete-time, time- homogeneous, $s-$ state first order Markov chain with transition probabilities $\lambda_{ij} := P(r_{t+1} = j | r_t = i)$ for any $i, j \in S$, where

$S := \{1, 2, \ldots s\}$. The transition probability matrix $\Lambda = [\lambda_{ij}]$ is, thus, an $s \times s$ matrix, with elements satisfying $\lambda_{ij} \geq 0$ and $\sum_{j=1}^{s} \lambda_{ij} = 1$, for each $i \in S$. We denote the initial probability distribution as $\lambda_i := P(r_1 = i)$ for $i \in S$ such that $\lambda_i \geq 0, \forall i \in S$ and $\sum_{i=1}^{s} \lambda_i = 1$. We consider the following JMLS:

$$\begin{aligned} x_{t+1} &= A(r_{t+1})x_t + B(r_{t+1})v_{t+1} + F(r_{t+1})u_{t+1}, && (6.1.1a) \\ y_t &= C(r_t)x_t + D(r_t)w_t + G(r_t)u_t && (6.1.1b) \end{aligned}$$

with $x_t \in \mathbb{R}^{n_x}$ system state, $y_t \in \mathbb{R}^{n_y}$ observation at time, $u_t \in \mathbb{R}^{n_u}$ known deterministic input, $w_t \in \mathbb{R}^{n_w}$ zero-mean white gaussian noise sequence with covariance I_{n_w} and $D(i)D^T(i) > 0$, for every $i \in S$, $v_t \in \mathbb{R}^{n_v}$ zero-mean white gaussian noise sequence with covariance I_{n_v} and $B(i)B^T(i) > 0$, for every $i \in S$. The matrices $A(\cdot), B(\cdot), C(\cdot), D(\cdot), F(\cdot)$ and $G(\cdot)$ are functions of the Markov chain r_t, i.e. $(A(\cdot),$ $B(\cdot),C(\cdot),D(\cdot),F(\cdot),G(\cdot)) \in \{(A(i), B(i),C(i),D(i),F(i),G(i)); i \in S\}$. They evolve according to the realization of the finite state Markov chain r_t.

Remark 6.1.1. *As a special case, if $A(\cdot) = 0$ and r_t is a i.i.d. discrete- valued random variable, x_t is a simple mixture of different gaussian distributions.*

Example 6.1.2. *In many situations, we want to model heterogeneity in the variance of a real time series. Let us consider a particular case of the general model in (6.1.1a, 6.1.1b), where $s = 2$ and for every t*

$$\begin{array}{rcl} n_x = 1 & \textit{and} & u_t = 0, \\ A(r_{t+1}) = \phi & \textit{and} & B(r_{t+1}) = \tau(r_{r_{t+1}}) \\ F(r_{t+1}) = 0 & \textit{and} & G(r_t) = 0, \\ C(r_t) = 1 & \textit{and} & D(r_t) = \sigma. \end{array}$$

Such a model can be written in this form

$$\begin{aligned} x_{t+1} &= \phi x_t + \tau(r_{t+1})v_{t+1}, && (6.1.2a) \\ y_t &= x_t + \sigma w_t. && (6.1.2b) \end{aligned}$$

The only dependence on the Markov chain is in the variance of the latent process noise. In this case r_t is a 2*- state first order Markov chain. If the chain r_t is in state* 1 *at time t, the process $\{X_t\}$ follows a standard AR(1) process with additive observation noise, where $\tau(1)$ is the standard deviation of the noise v_{t+1}. This specification allows us to take into account lack of homogeneity in the second moment of the series: switching to regime $r_t = 2$ affects the variance of the noise v_{t+1} that changes from $\tau^2(1)$ to $\tau^2(2)$.*

In many problems related to seismic processing and nuclear science [Mendel, 1990, Lavielle, 1993], the signal of interest can be modeled as the output of an ARMA model filter excited by a discrete time Markov chain and observed in white gaussian noise. An interesting particular case of the model (6.1.2a, 6.1.2b) can be defined as follows. We consider an AR(1) latent dynamic where $\tau(2)$ is set equal to zero

$$x_{t+1} = \phi x_t + v'_{t+1}(r_{t+1}), \tag{6.1.3a}$$

$$y_t = x_t + \sigma w_t, \tag{6.1.3b}$$

where, conditional upon $r_t = 1$, $v'_t = \tau(1)v_t$ and conditional upon $r_t = 2$, $v'_t \sim \delta_0$, where δ_0 is the Dirac delta measure in 0*. In this case, if the chain r_t is in state* 2 *at time t, the dynamic of the process $\{X_t\}$ evolves deterministically since the variance of the noise vanishes.*

The behavior of this model is described in Figure 6.1.1. For a fixed set of parameter values, the figure shows a simulated path of the finite state Markov chain r_t (top) and the continuous (in space) states processes x_t, y_t (bottom) for $t = 1, \ldots, 250$. When the Markov chain r_t is in state 2*, the signal x_t evolves without noise going deterministically towards zero (from above or below). When the chain jumps to state* 1*, we add some noise to the dynamic of the latent process $\{X_t\}$.*

6.1.1 Stationary distribution

We suppose that the bivariate latent process $\{(X_t, r_t)\}$ is stationary, and hence by hypothesis invariant distribution exists. Anyway, stationary distribution is not easy

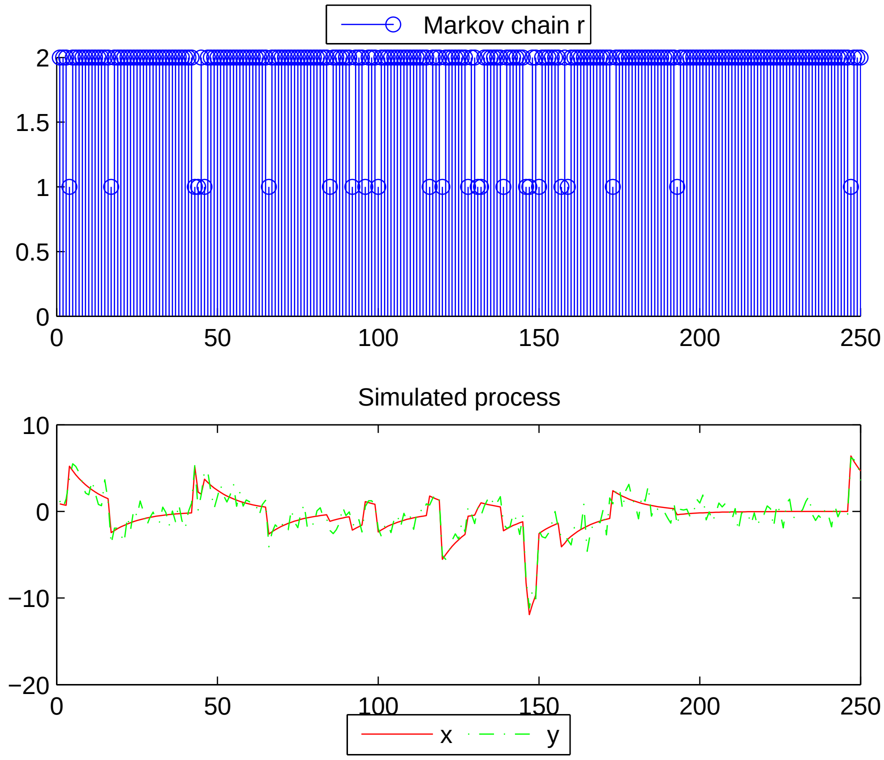

Figure 6.1.1: Signal process for the model in (6.1.3a, 6.1.3b): state of the Markov chain r_t (top) and the continuous processes x_t, y_t (bottom). Fixed value of the parameters: $\phi = 0.9, \tau(1) = 5, \sigma = 1, \lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{21} = 0.1$.

to compute and it obviously depends on the parameters of the model. Moreover, calculation will be simpler if we put as initial distribution a distribution that does not depend on the parameter values, since it becomes a constant factor when we maximize any objective function with respect to the parameter in the parameter space. Since r_t is a discrete time Markov chain with finite state space, its stationary distribution corresponds to

$$\pi^{(r)} = (\pi_1^{(r)}, \pi_2^{(r)}, \ldots, \pi_s^{(r)})$$

such that

$$\pi^{(r)} = \pi^{(r)}\Lambda$$

with the constrain that $\sum_{i=1}^{s} \pi_i^{(r)} = 1$. On the other hand, in order to compute the invariant distribution for the latent process $\{X_t\}$, we need to find a distribution $\pi(x)$ such that

$$\int \pi(x) \sum_{i=1}^{s} \left[\pi_i^{(r)} N(dy; A(i)x + F(i)u, B(i)B(i)^T) \right] = \pi(dy).$$

Solving the equation above corresponds to find a solution of an integral equation. This is not a simple task even for a specific kind of model. A simulation study (see Figures 6.1.2) shows that the shape of the empirical stationary distribution is not standard and for that reason, even if $\pi(x)$ might be derived in some way, we consider it as unknown.

Example 6.1.3. *Going back to Example 6.1.2, we have that*

$$\pi^{(r)} = (\pi_1^{(r)}, \pi_2^{(r)}) = \left(\frac{\lambda_{21}}{\lambda_{12} + \lambda_{21}}, \frac{\lambda_{12}}{\lambda_{12} + \lambda_{21}} \right),$$

while $\pi(x)$ is the solution of the following integral equation

$$\int \pi(x) \left[\pi_1^{(r)} N(y; \phi x, \tau^2(1)) + \pi_2^{(r)} \delta_{\phi x}(dy) \right] = \pi(dy).$$

where $\delta_{\phi x}(\cdot)$ is the Dirac delta measure centered in ϕx.

From the considerations above, JMLS is a framework where it should be useful exploiting the idea suggested in (3.3.3) and illustrated in Section 5.4.2 for a linear gaussian model. Here, the invariant distribution is unknown but transitions are still simple. In the next sections we will approximate the invariant distribution sampling from the transition kernel of the ergodic process.

6.1.2 Approximation of the stationary distribution

As outlined above, in this case the invariant distribution is unknown and hence has to be replaced with a suitable approximation. We take a generic initial distribution $\nu(\cdot)$ for x_{-z} and for r_{-z} and we simulate a sufficiently long chain from the transition kernel. Under geometric ergodicity, the marginal distribution of x_1 converges to

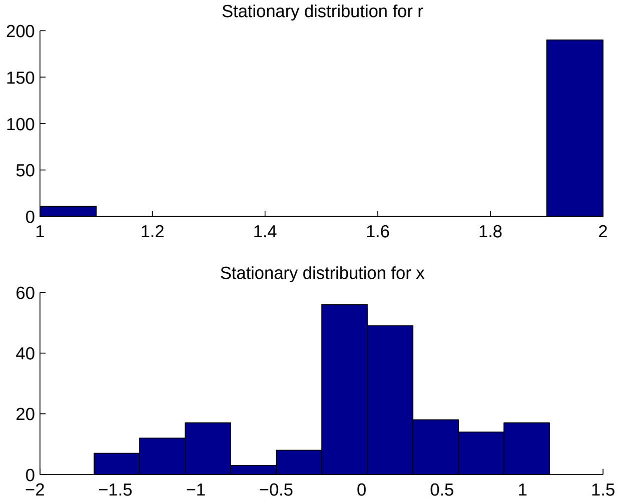

Figure 6.1.2: Simulation study- Estimation of the stationary distribution π_r (top) and of the marginal of the stationary distribution $\pi(x)$. We take as initial distribution the uniform distribution for the chain r_t and a gaussian distribution for the process x_t, with mean 15 and variance 1. Fixed value of the parameters: $n_x = 2$, $s = 2$, $\sigma = 0.25$,$\lambda_{12} = 0.9$,$\lambda_{21} = 0.1$, $x_{t+1} = (x_{t+1}, x_t)'$, $A(r_{t+1}) = \begin{pmatrix} 1.511 & -0.539 \\ 1 & 0 \end{pmatrix}$,$B(r_{t+1}) = (\tau(r_{t+1}), 0)'$, where $\tau(1) = 0.5$, $\tau(2) = 0$.

$\pi(x_1)$ as z goes to $+\infty$. More precisely, we derive the joint distribution

$$\nu(x_{-z:1}, r_{-z:1}) = \left[\nu_{r_{-z}} \prod_{i=-z+1}^{1} \lambda_{r_i r_{i+1}}\right] \nu(x_{-z:1}|z_{-z:1}),$$

where $\nu_{r_{-z}} = P(r_{-z} = r_{-z})$, $\lambda_{r_i r_{i+1}} = P(r_i|r_{i+1})$ and

$$\nu(x_{-z:1}|r_{-z:1}) = \nu(x_{-z}) \prod_{i=-z+1}^{1} f_\theta(x_i|x_{i-1}, r_i).$$

We recall that $f_\theta(x_i|x_{i-1}, r_i) = N\left(x_i; A(r_i)x_{i-1} + F(r_i)u_i, B(r_i)B(r_i)^T\right)$. Under geometric ergodicity, the marginal

$$\nu(x_1, r_1) \to \pi_\theta(x_1, r_1),$$

as z goes to $+\infty$.

6.2 Inference in a JMLS

Inference on the parameter in JMLS can be carried out via an on line EM algorithm, as described in Chapter 5. Algorithms that allow us to develop such estimates are more complex than those for linear gaussian state space model, since the latent process is bivariate and consists of a continuous and a discrete valued process. In what follows, we develop an on line EM algorithm and we describe how to sample from the required distributions. Moreover, for a special case of JMLS, we derive the update equations in fairly explicit details.

6.2.1 The $Q(\theta, \theta_k)$ function for a JMLS

Taking into account the states before time 1, the function $Q(\theta, \theta_k)$, that has to be maximized with respect to θ at each iteration of the EM algorithm, has this expression

$$Q(\theta, \theta_k) = \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta_k,\theta^*}[log[p_\theta(y_1, y_j, x_{-z:j}, r_{-z:j}|\nu)]], \tag{6.2.1}$$

where $\mathbb{E}^{(j)}_{\theta_k,\theta^*}$ denotes the expectation with respect to $p_{\theta_k}(x_{-z:j}, r_{-z:j}|y_1, y_j, \nu)\times p_{\theta^*}(\mathbf{y}_1)$, ν denoting the initial distribution. In order to compute the Q function, we need to evaluate $log[p_\theta(y_1, y_j, x_{-z:j}, r_{-z:j}|\nu)]$ for the model (6.1.1a, 6.1.1b). Taking into account the structure of the model, we have that

$$\begin{aligned}
&p_\theta(y_1, y_j, x_{-z:j}, r_{-z:j}) = \nu(x_{-z:1}, r_{-z:1})\prod_{i=2}^{j}\lambda_{r_{i-1}r_i}\times\\
&\times\prod_{i=2}^{j}\left[N(x_i; A(r_i)x_{i-1} + F(r_i)u_i, B(r_i)B(r_i)^T)\right]\times\\
&\times N\left(y_1; C(r_1)x_1 + G(r_1)u_1, D(r_1)D(r_1)^T\right)\times\\
&\times N\left(y_j; C(r_j)x_j + G(r_j)u_j, D(r_j)D(r_j)^T\right) =
\end{aligned}$$

$$
\begin{aligned}
=&\nu(r_{-z}) \prod_{i=-z+1}^{j} \lambda_{r_{i-1}r_i} \times \\
&\times \nu(x_{-z}) \prod_{i=-z+1}^{j} \left[N(x_i; A(r_i)x_{i-1} + F(r_i)u_i, B(r_i)B(r_i)^T) \right] \times \\
&\times N\left(y_1; C(r_1)x_1 + G(r_1)u_1, D(r_1)D(r_1)^T\right) \times \\
&\times N\left(y_j; C(r_j)x_j + G(r_j)u_j, D(r_j)D(r_j)^T\right).
\end{aligned}
$$

Example: The $Q(\theta, \theta_k)$ function for the model (6.1.2a, 6.1.2b)

We evaluate (6.2.1) for the model specified in Example 6.1.2, where $\tau(2) \ll \tau(1)$, but different from zero. We need this position since it is not possible to apply our algorithm in the case where $\tau(2) = 0$. In order to evaluate the Q function, we need to compute

$$
\begin{aligned}
&\log p_\theta(y_{1:j}, x_{-z,j}, r_{-z:j} = r_{-z:j}|\nu) = \log \nu(x_{-z}) + \log \nu(r_{-z}) + \sum_{i=-z+1}^{j} \log \lambda_{r_{i-1}r_i} + \\
&+ \sum_{i=-z+1}^{j} \log \left[N(x_i; \phi x_{i-1}, \tau^2(r_i)) \right] + \log N\left(y_1; x_1, \sigma^2\right) + \log N\left(y_j; x_j, \sigma^2\right).
\end{aligned}
$$

If we define $n_{12}(r_{-z:j})$ the number of times that state 1 is followed by state 2 in the sample $r_{-z}, \ldots, r_j$ (and analogously $n_{11}(r_{-z:j}), n_{21}(r_{-z:j}), n_{22}(r_{-z:j})$) and $n_1(r_{-z+1:j})$ the number of times that the chain r_t is in state 1 in the sample $r_{-z+1}, \ldots, r_j$ (and analogously $n_2(r_{-z+1:j})$), the quantity above can be written as

$$
\begin{aligned}
&\log \nu(x_{-z}) + \log \nu(r_{-z}) + \sum_{i=1}^{2} \sum_{k=1}^{2} n_{ik}(r_{-z:j}) \log \lambda_{ik} + \\
&- \frac{1}{2} \log \tau^2(1) n_1(r_{-z+1:j}) - \frac{1}{2\tau^2(1)} \sum_{k=-z+1}^{j} \left(\mathbb{I}_1(r_k)x_k^2 + \mathbb{I}_1(r_k)\phi^2 x_{k-1}^2 - \mathbb{I}_1(r_k) 2\phi x_k x_{k-1} \right) \\
&- \frac{1}{2} \log \tau^2(2) n_2(r_{-z+1:j}) - \frac{1}{2\tau^2(2)} \sum_{k=-z+1}^{j} \left(\mathbb{I}_2(r_k)x_k^2 + \mathbb{I}_2(r_k)\phi^2 x_{k-1}^2 - \mathbb{I}_2(r_k) 2\phi x_k x_{k-1} \right) +
\end{aligned}
$$

$$- \log \sigma^2 - \frac{y_1^2 + x_1^2 - 2y_1 x_1 + y_j^2 + x_j^2 - 2y_j x_j}{2\sigma^2},$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

Now we can evaluate the Q function for this particular model

$$Q(\theta, \theta') = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[\log p_\theta(y_{1:j}, x_{-z,j}, r_{-z:j} = r_{-z:j})|\nu] =$$

$$= \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[\log \nu(r_{-z})] + \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[\log \nu(x_{-z})] +$$

$$+ \log \lambda_{11} \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_{11}(r_{-z:j})] + \log \lambda_{12} \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_{12}(r_{-z:j})] +$$

$$+ \log \lambda_{21} \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_{21}(r_{-z:j})] + \log \lambda_{22} \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_{22}(r_{-z:j})] +$$

$$- \frac{1}{2} \log \tau^2(1) \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_1(r_{-z+1:j})]$$

$$- \frac{1}{2\tau^2(1)} \left[\frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_1(r_k) x_k^2 \right] + \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_1(r_k) \phi^2 x_{k-1}^2 \right] + \right.$$

$$\left. - \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_1(r_k) 2\phi x_k x_{k-1} \right] \right] - \frac{1}{2} \log \tau^2(2) \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[n_2(r_{-z+1:j})] -$$

$$\frac{1}{2\tau^2(2)} \left[\frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_2(r_k) x_k^2 \right] + \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_2(r_k) \phi^2 x_{k-1}^2 \right] + \right.$$

$$\left. - \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*} \left[\sum_{k=-z+1}^{j} \mathbb{I}_2(r_k) 2\phi x_k x_{k-1} \right] \right] +$$

$$- \log \sigma^2 - \frac{1}{2\sigma^2} \left(y_1^2 + \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[x_1^2] - 2y_1 \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[x_1] + \right.$$

$$\left. y_j^2 + \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[x_j^2] - 2y_j \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta',\theta^*}[x_j] \right).$$

The quantity above depends on the following statistics

$$\begin{aligned}
N_{11} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{11}(r_{-z:j})] \\
N_{12} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{12}(r_{-z:j})] \\
N_{21} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{21}(r_{-z:j})] \\
N_{22} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{22}(r_{-z:j})] \\
\Psi_1^{(1)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{1}(r_{-z+1:j})] \\
\Psi_1^{(2)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[n_{2}(r_{-z+1:j})] \\
\Psi_2^{(1)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_1(r_k)x_k^2\right] \\
\Psi_3^{(1)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_1(r_k)x_{k-1}^2\right] \\
\Psi_4^{(1)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_1(r_k)x_k x_{k-1}\right] \\
\Psi_2^{(2)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_2(r_k)x_k^2\right] \\
\Psi_3^{(2)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_2(r_k)x_{k-1}^2\right] \\
\Psi_4^{(2)} &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}\left[\sum_{k=-z+1}^{j}\mathbb{I}_2(r_k)x_k x_{k-1}\right] \\
\Psi_5 &= y_1^2 + \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[x_1^2 + x_j^2] - 2y_1\frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[x_1] +
\end{aligned}$$

$$+ \quad y_j^2 + \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[x_j^2] - 2y_j\frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[x_j].$$

With these definitions, the Q function can be re-written as

$$\begin{aligned}
Q = {} & \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[\log p_\theta(y_{1:j}, x_{-z,j}, r_{-z:j} = r_{-z:j})|\nu] = \\
= {} & \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[\log \nu(r_{-z})] + \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[\log \nu(x_{-z})]+ \\
& + \log\lambda_{11}N_{11} + \log\lambda_{12}N_{12} + \log\lambda_{21}N_{21} + \log\lambda_{22}N_{22} \\
& + \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta',\theta^*}[\log \nu(x_{-z})] - \frac{1}{2}\log\tau^2(1)\Psi_1^{(1)} \\
& - \frac{1}{2\tau^2(1)}[\Psi_2^{(1)} + \phi^2\Psi_3^{(1)} - 2\phi\Psi_4^{(1)}] \\
& - \frac{1}{2}\log\tau^2(2)\Psi_1^{(2)} - \\
& \frac{1}{2\tau^2(2)}[\Psi_2^{(2)} + \Psi_3^{(2)}\phi^2 - 2\phi\Psi_4^{(2)}] + \\
& - \log\sigma^2 - \frac{1}{2\sigma^2}\Psi_5.
\end{aligned}$$

In order to simplify the maximization step, we take $\nu(r_{-z})$ and $\nu(x_{-z})$ independent of θ. Since $\lambda_{12} = 1-\lambda_{11}$, $\lambda_{21} = 1-\lambda_{22}$, derivatives with respect to the parameters of the model have the following expressions

$$\begin{aligned}
\frac{\partial Q}{\partial\lambda_{11}} &= \frac{N_{11}}{\lambda_{11}} - \frac{N_{12}}{1-\lambda_{11}} = 0 \\
\frac{\partial Q}{\partial\lambda_{22}} &= \frac{N_{11}}{\lambda_{22}} - \frac{N_{12}}{1-\lambda_{22}} = 0 \\
\frac{\partial Q}{\partial\tau^2(1)} &= -\frac{\Psi_1^{(1)}}{2\tau^2(1)} + \frac{\Psi_2^{(1)} + \phi^2\Psi_3^{(1)} - 2\phi\Psi_4^{(1)}}{\tau^4(1)} = 0 \\
\frac{\partial Q}{\partial\tau^2(2)} &= -\frac{\Psi_1^{(2)}}{2\tau^2(2)} + \frac{\Psi_2^{(2)} + \phi^2\Psi_3^{(2)} - 2\phi\Psi_4^{(2)}}{\tau^4(2)} = 0 \\
\frac{\partial Q}{\partial\sigma^2} &= -\frac{1}{\sigma^2} + \frac{\Psi_5}{2\sigma^4} = 0
\end{aligned}$$

$$\frac{\partial Q}{\partial \phi} = -\frac{1}{2\tau^2(1)}[2\phi\Psi_3^{(1)} - 2\Psi_4^{(1)}] - \frac{1}{2\tau^2(2)}[2\phi\Psi_3^{(2)} - 2\Psi_4^{(2)}] = 0,$$

and hence

$$\begin{aligned}
\lambda_{11} &= \frac{N_{11}}{N_{11} + N_{12}} \\
\lambda_{22} &= \frac{N_{22}}{N_{21} + N_{22}} \\
\tau^2(1) &= -\frac{\Psi_2^{(1)} + \phi^2\Psi_3^{(1)} - 2\phi\Psi_4^{(1)}}{\Psi_1^{(1)}} \qquad (6.2.2a) \\
\tau^2(2) &= -\frac{\Psi_2^{(2)} + \phi^2\Psi_3^{(2)} - 2\phi\Psi_4^{(2)}}{\Psi_1^{(2)}} \qquad (6.2.2b) \\
\sigma^2 &= \frac{\Psi_5}{2} \\
\phi\left(\frac{\Psi_3^{(1)}}{\tau^2(1)} + \frac{\Psi_3^{(2)}}{\tau^2(2)}\right) &= \frac{\Psi_4^{(1)}}{\tau^2(1)} + \frac{\Psi_4^{(2)}}{\tau^2(2)}.
\end{aligned}$$

This last equation can be written as a third degree equation in ϕ

$$\begin{aligned}
&\phi^3\left(\frac{\Psi_3^{(1)}\Psi_3^{(2)}}{\Psi_1^{(2)}} + \frac{\Psi_3^{(1)}\Psi_3^{(2)}}{\Psi_1^{(1)}}\right) + \\
&+ \phi^2\left(-2\frac{\Psi_3^{(1)}\Psi_4^{(2)}}{\Psi_1^{(2)}} - 2\frac{\Psi_3^{(2)}\Psi_4^{(1)}}{\Psi_1^{(1)}} - \frac{\Psi_3^{(2)}\Psi_4^{(1)}}{\Psi_1^{(2)}} + \frac{\Psi_3^{(1)}\Psi_4^{(2)}}{\Psi_1^{(1)}}\right) + \\
&+ \phi\left(\frac{\Psi_2^{(2)}\Psi_3^{(1)}}{\Psi_1^{(2)}} + \frac{\Psi_2^{(1)}\Psi_3^{(2)}}{\Psi_1^{(1)}} - 2\frac{\Psi_4^{(1)}\Psi_4^{(2)}}{\Psi_1^{(2)}} - 2\frac{\Psi_4^{(1)}\Psi_4^{(2)}}{\Psi_1^{(1)}}\right) + \\
&+ \frac{\Psi_4^{(1)}\Psi_2^{(2)}}{\Psi_1^{(2)}} + \frac{\Psi_4^{(2)}\Psi_2^{(1)}}{\Psi_1^{(1)}} = 0.
\end{aligned}$$

The three solutions for ϕ can be found in the same way as in Section 5.4. Then we discard solutions that fall outside the real interval $[-1, 1]$ and keep among the remaining values, since we work with stationary process. Once we get the value for ϕ, $\tau^2(1)$ and $\tau^2(2)$ are uniquely determined by Equations (6.2.2a, 6.2.2b).

In this case, it is possible to maximize $Q(\theta, \theta_k)$ analytically, and the maximum points depend on some sufficient statistics defined as expectations with respect to a measure that depends on the true unknown parameter values. These statistics can be approximated using an on line scheme, following the same idea explained in Section 5.4.

In general, evaluation of the Q function requires the calculation of the expectation with respect to a measure that depends on the unknown parameter value θ^*, so it can not be computed. Anyway, given a block of observations $\mathbf{y}_q = (y_q, \ldots, y_{q+L})$, for every y_q, y_{q+j-1} in the block $\mathbf{y}_q$, $j = 2, \ldots, L+1$,

$$\int log[p_\theta(y_q, y_{q+j-1}, x_{-z:j}, r_{-z:j})|v]p_{\theta_k}(x_{-z:j}, r_{-z:j}|y_q, y_{q+j-1}, v)dx_{-z:j}dr_{-z:j}$$

can be approximated by

$$\frac{1}{M}\sum_{m=1}^{M} log[p_\theta(y_q, y_{q+j-1}, x_{-z:j}^{(m,q)}, r_{-z:j}^{(m,q)})|v],$$

where $\{x_{-z:j}^{(m,q)}, r_{-z:j}^{(m,q)}\}$, for $m = 1, \ldots, M$ are samples from the posterior distribution $p_{\theta_k}(x_{-z:j}, r_{-z:j}|y_q, y_{q+j-1}, v)$. When blocks of observations $\mathbf{y}_q = (y_q, \ldots, y_{q+L})$ are available one at a time, for every y_q, y_{q+j-1} in the block $\mathbf{y}_q$, $q = 1, \ldots, T-L$, $j = 2, \ldots, L+1$, we shall consider an on line scheme, as proposed in Section 5.4. We now describe how to obtain samples from the generic distribution $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j, v)$.

6.3 Sampling from $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j, v)$

Given the observations y_1, y_j, our interest relies on the joint posterior distribution $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j)$, in particular on computing integrals with respect to this joint distribution. If we were able to obtain M independent and identically distributed (i.i.d.) samples distributed according to $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j)$, then, using the law of large numbers, integrals as the minimum mean square error (MMSE) estimates could be computed by averaging. The aim of this section is to obtain, for every $j = 2, \ldots, L+1$, samples $\{r_{-z:j}^{(m)} : m = 1, \ldots, M\}$ (for large M) from the posterior distribution $p_\theta(r_{-z:j}|y_1, y_j)$. Obtaining i.i.d. samples from this distribution is not straightforward, so we use an alternative scheme based on an MCMC method. The key idea of MCMC methods is to run an ergodic Markov chain whose invariant distribution is the distribution of interest. The obtained samples are then used to compute estimates of the states. The proposed algorithm proceed as follows [Andrieu and Doucet, 2001]

1. Initialization. Set randomly $r^{(0)}_{-z:j}$

2. Iteration m, $m = 1, \dots, M$

   - For $t = -z, \dots, j$, sample $r^{(m)}_t \sim p_\theta(r_t|y_1, y_j, \mathbf{r}^{(m)}_{-t})$, where

$$\mathbf{r}^{(m)}_{-t} := (r^{(m)}_{-z}, \dots, r^{(m)}_{t-1}, r^{(m-1)}_{t+1}, \dots, r^{(m-1)}_{j}).$$

   - Compute $x^{(m)}_{-z:j} = \mathbb{E}[x_{-z:j}|y_1, y_j, r^{(m)}_{-z:j}]$ and/or other statistics of interest, given the sequence $r^{(m)}_{-z:j}$.

Once the algorithm has been iterated M times, the MMSE estimates of $r_{-z:j}$ and $x_{-z:j}$ are computed using

$$\hat{r}_{-z:j}(M) = \frac{1}{M}\sum_{m=0}^{M-1} r^{(m)}_{-z:j}, \qquad \hat{x}_{-z:j}(M) = \frac{1}{M}\sum_{m=0}^{M-1} x^{(m)}_{-z:j}.$$

The different steps of this algorithm are detailed in the next sections. In order to simplify notation, we drop the superscript (m) from all variables at iteration m when it is unnecessary.

This algorithm requires sampling from $p_\theta(r_t|y_1, y_t, \mathbf{r}_{-t})$. Direct solutions are computationally expensive, so we develop a strategy based on a key decomposition of the likelihood function (Section 6.3.2).

Furthermore, once the sequence $r_{-z:j}$ is given, a general JMLS, as defined in (6.1.1a, 6.1.1b), is linear gaussian. Therefore, estimating the sequence $x_{-z:j}$ by $\mathbb{E}[x_{-z:j}|y_1, y_j, r_{-z:j}]$ can be done using a suitable modification of the Kalman smoother, as described in the next section.

6.3.1 Modified Kalman filter : estimation of the state x_j given the sequence $r_{-z:j}$ and the observations y_1, y_j

In the framework of pairwise likelihood strategy, we need to compute expectations with respect to $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j)$ for $j = 2, \dots, L+1$. Given the sequence $r_{-z:j}$ and the observations y_1 and y_j, $p_\theta(x_{-z:j}|y_1, y_j, r_{-z:j})$ is gaussian and then we exploit here a suitable extension of the Kalman filter recursions to estimate the state $x_{-z:j}$ given the observations y_1, y_j and the state $r_{-z:j}$.

We recall here the *prediction* and *update steps* of the standard Kalman filter in the contest of jump Markov linear systems, where $z = 0$. If the prior distribution is gaussian, $x_0 \sim N(\tilde{m}_{0|0}, \tilde{P}_{0|0})$, then the optimal filtering equations can be evaluated in closed form:

$$\begin{aligned} p_\theta(x_j|y_{1:j-1}, r_{1:j}) &= N(x_j|\tilde{m}_{j|j-1}(r_{1:j}), \tilde{P}_{j|j-1}(r_{1:j})) \\ p_\theta(x_j|y_{1:j}, r_{1:j}) &= N(x_j|\tilde{m}_{j|j}(r_{1:j}), \tilde{P}_{j|j}(r_{1:j})) \end{aligned}$$

and the parameters of these distributions can be calculated by the following steps:

Prediction step

$$\begin{aligned} \tilde{m}_{j|j-1}(r_{1:j}) &= A(r_j)\tilde{m}_{j-1|j-1}(r_{1:j}) + F(r_j)u_j \\ \tilde{P}_{j|j-1}(r_{1:j}) &= A(r_j)\tilde{P}_{j-1|j-1}(r_{1:j})A(r_j)^T + B(r_j)B(r_j)^T. \end{aligned}$$

Update step

$$\begin{aligned} \tilde{e}_j(r_{1:j}) &= y_j - C(r_j)\tilde{m}_{j|j-1}(r_{1:j}) - G(r_j)u_j \\ \tilde{S}_j(r_{1:j}) &= C(r_j)\tilde{P}_{j|j-1}(r_{1:j})C(r_j)^T + D(r_j)D(r_j)^T \\ \tilde{K}_j(r_{1:j}) &= \tilde{P}_{j|j-1}C(r_j)(\tilde{S}_j(r_{1:j}))^{-1} \\ \tilde{m}_{j|j}(r_{1:j}) &= \tilde{m}_{j|j-1}(r_{1:j}) + \tilde{K}_j(r_{1:j})\tilde{e}_j(r_{1:j}) \\ \tilde{P}_{j|j}(r_{1:j}) &= (I - \tilde{K}_j(r_{1:j})C(r_j))\tilde{P}_{j|j-1}(r_{1:j}). \end{aligned}$$

If $z = 0$, our interest concerns the evaluation of $p_\theta(x_j|y_1, y_j, r_{1:j})$ where conditioning is on the observations y_1 and y_j and not on all the observations between 1 and j. The recursions above need to be modified in order to obtain the right moments of the posterior distributions $p_\theta(x_j|y_1, y_j, r_{1:j})$. We follow the same idea as in the previous chapter, setting an infinity variance for the missing observations from time 2 to time $j-1$. This means taking the matrix $D(\cdot)$ in such a way that $(D(\cdot)D(\cdot)^T)^{-1}$ is close to zero. To apply this strategy, we need to be careful about computational issues as matrix inversion.

More precisely, if $x_0 \sim N(m_{0|0}, P_{0|0})$, for every $j > 2$, the prediction and update steps modify as follows

- Initialization

$$\begin{aligned}
m_{1|0}(r_1) &= A(r_1)m_{0|0} + F(r_1)u_1 \\
P_{1|0}(r_1) &= A(r_1)P_{0|0}A(r_1)^T + B(r_1)B(r_1)^T \\
e_1(r_1) &= y_1 - C(r_1)m_{1|0}(r_1) - G(r_1)u_1 \\
S_1(r_1) &= C(r_1)P_{1|0}(r_1)C(r)^T + D(r)D(r)^T \\
K_1(r_1) &= P_{1|0}C(r_1)(S_1(r_1))^{-1} \\
m_{1|1}(r_1) &= m_{1|0}(r_1) + K_1(r_1)e_1(r_1) \\
P_{1|1}(r_1) &= (I - K_1(r_1)C(r_1))P_{1|0}(r_1).
\end{aligned}$$

- For $k = 2, \ldots, j-1$

$$\begin{aligned}
m_{k|k-1}(r_{1:k}) &= A(r_k)m_{k-1|k-1}(r_{1:k}) + F(r_k)u_k \\
P_{k|k-1}(r_{1:k}) &= A(r_k)P_{k-1|k-1}(r_{1:k})A(r_k)^T + B(r_k)B(r_k)^T \\
m_{k|k}(r_{1:k}) &= m_{k|k-1}(r_{1:k}) \\
P_{k|k}(r_{1:k}) &= P_{k|k-1}(r_{1:k}).
\end{aligned}$$

- For k=j

$$\begin{aligned}
m_{j|j-1}(r_{i:j}) &= A(r_j)m_{j-1|j-1}(r_{1:j}) + F(r_j)u_j \\
P_{j|j-1}(r_{i:j}) &= A(r_j)P_{j-1|j-1}(r_{1:j})A(r_j)^T + B(r_j)B(r_j)^T \\
e_j(r_{1:j}) &= y_j - C(r_j)m_{j|j-1}(r_{1:j}) - G(r_j)u_j \\
S_j(r_{1:j}) &= C(r_j)P_{j|j-1}(r_{1:j})C(r_j)^T + D(r_j)D(r_j)^T
\end{aligned}$$

$$\begin{aligned}
K_j(r_{1:j}) &= P_{j|j-1}(r_{1:j})C(r_j)(S_j(r_{1:j}))^{-1} \\
m_{j|j}(r_{1:j}) &= m_{j|j-1}(r_{1:j}) + K_j(r_{1:j})e_j(r_{1:j}) \\
P_{j|j}(r_{1:j}) &= (I - K_j(r_{1:j})C(r_j))P_{j|j-1}(r_{1:j}).
\end{aligned}$$

Remark 6.3.1. *Note that for $k = 2, \dots, j-1$ we actually do not need to compute quantities that depend on the variance $D(\cdot)D(\cdot)^T$. Since $(D(\cdot)D(\cdot)^T)^{-1}$ tends to zero, $K_k(r_{1:k})$ can be set equal to zero. Again, the innovation at time k (i.e. $e_k(r_{1:k})$) and its covariance (i.e. $S_k(r_{1:k})$) do not need to be computed, and this allows us to avoid matrix inversion and approximations.*

Going back to our example defined by the system (6.1.2a, 6.1.2b), we run the modified Kalman filter described above in order to estimate the latent state x_j, given the sequence $r_{1:j}$ and the observations y_1, y_j. We take $x_0 \sim N(0, 1)$. For $j = 2, \dots, 250$, we compute the minimum mean square error estimates of the continuous state of the JMLS. The results are shown in Figure 6.3.1.

We also compare the standard Kalman filter and our modified Kalman filter via the comparison between the optimal (in a mean square sense) estimates of x_j given by $\mathbb{E}[x_j|y_{1:j}, r_{1:j}]$ and $\mathbb{E}[x_j|y_1, y_j, r_{1:j}]$, respectively. The results are shown in Figure 6.3.2. Moreover, if we define $\tilde{m}_{1:T} = (\tilde{m}_{1|1}, \tilde{m}_{2|2}, \dots, \tilde{m}_{T|T})$ and $m_{1:T} = (m_{1|1}, m_{2|2}, \dots, m_{T|T})$, we can evaluate the distance between the estimates and the true values, as shown below.

$(x_{1:T} - y_{1:T})(x_{1:T} - y_{1:T})^T$	14.2611
$(\tilde{m}_{1:T} - x_{1:T})(\tilde{m}_{1:T} - x_{1:T})^T$	2.8308
$(m_{1:T} - x_{1:T})(m_{1:T} - x_{1:T})^T$	9.9984

In our case, the stationary distribution is unknown, so it becomes important to estimate $x_{-z:j}$ given the observations y_1, y_j and the state $r_{-z:j}$, for a fixed value of $z \neq 0$ and for $j = 2, \dots, L+1$. The strategy described above can be easily extended to this aim. In the next, we describe how to generalize the modified Kalman filter in order to sample from $p_\theta(r_{-z:j}|y_1, y_j, \nu)$, ν being the initial distribution. This allows us to evaluate integrals with respect to the posterior distribution $p_\theta(x_{-z:j}|y_1, y_j, r_{-z:j}, \nu)$.

More precisely, if $(x_{-z}) \sim \nu(x_{-z})$, with mean $m_{-z|-z}$ and covariance $P_{-z|-z}$, the

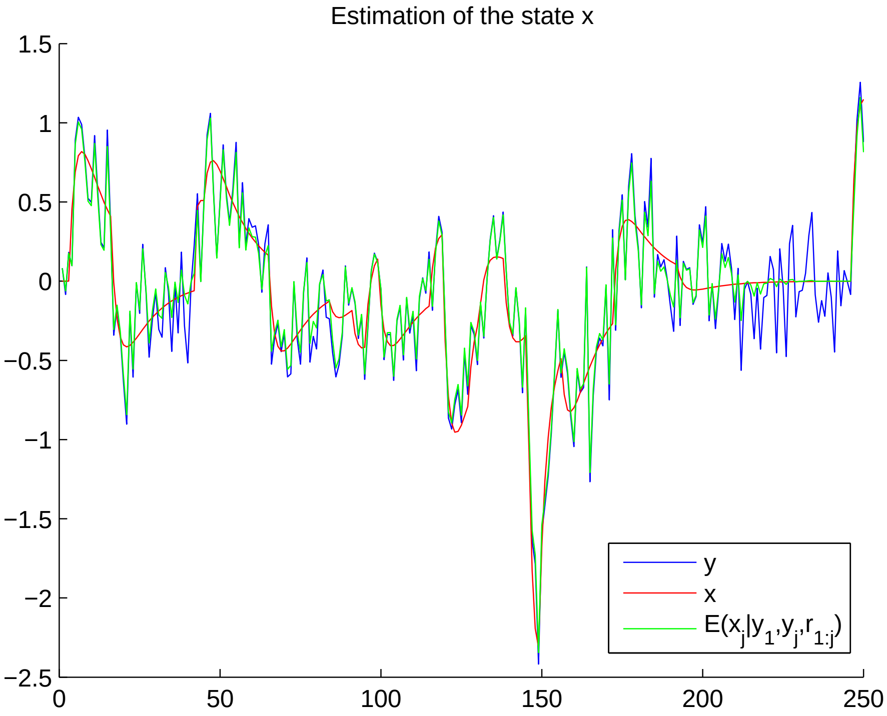

Figure 6.3.1: Modified Kalman Filter- Estimation of the state x_j, given the sequence $r_{1:j}$ and the observations y_1, y_j, for $j = 2, \ldots, 250$. Fixed value of the parameters: $\phi = 0.9, \tau(1) = 0.5, \tau(2) = 0, \sigma = 0.25, \lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{12} = 0.1, m_{0|0} = 0, P_{0|0} = 1$.

modified Kalman filter becomes

- For $k = -z, \ldots, 0$, we obtain the predicted states (*"missing observations"* $y_{-z:0}$)

$$\begin{aligned}
m_{k|k-1}(r_{-z:k}) &= A(r_k)m_{k-1|k-1}(r_k) + F(r_k)u_k \\
P_{k|k-1}(r_{-z:k}) &= A(r_k)P_{k-1|k-1}(r_k)A(r_k)^T + B(r_k)B(r_k)^T \\
m_{k|k}(r_{-z:k}) &= m_{k|k-1}(r_{-z:k}) \\
P_{k|k}(r_{-z:k}) &= P_{k|k-1}(r_{-z:k}).
\end{aligned}$$

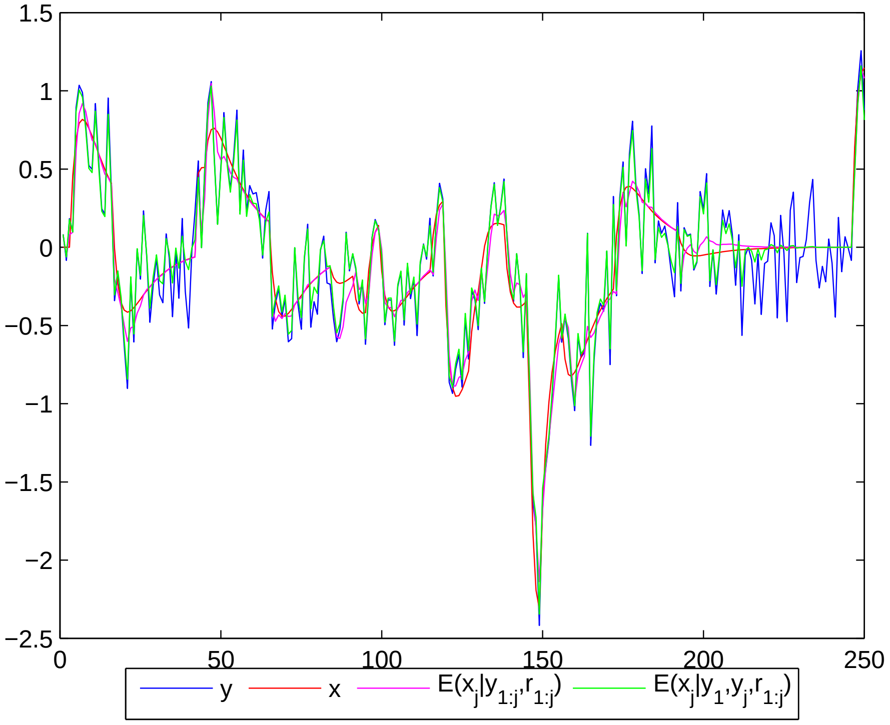

Figure 6.3.2: Comparison between the standard Kalman filter and the modified Kalman filter- MMSE of the state x_j, given the sequence $r_{1:j}$ and the observations $y_{1:j}$ (standard Kalman filter) or y_1, y_j (modified Kalman filter), for $j = 2, \ldots, 250$. Fixed value of the parameters: $\phi = 0.9, \tau(1) = 0.5, \tau(2) = 0, \sigma = 0.25, \lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{12} = 0.1, m_{0|0} = \tilde{m}_{0|0} = 0, P_{0|0} = \tilde{P}_{0|0} = 1$.

- Initialization

$$
\begin{aligned}
m_{1|0}(r_{-z:1}) &= A(r_1)m_{0|0}(r_{-z:1}) + F(r_1)u_1 \\
P_{1|0}(r_{-z:1}) &= A(r_1)P_{0|0}(r_{-z:1})A(r_1)^T + B(r_1)B(r_1)^T \\
e_1(r_{-z:1}) &= y_1 - C(r_1)m_{1|0}(r_{-z:1}) - G(r_1)u_1 \\
S_1(r_{-z:1}) &= C(r_1)P_{1|0}(r_{-z:1})C(r_1)^T + D(r_1)D(r_1)^T \\
K_1(r_{-z:1}) &= P_{1|0}(r_{-z:1})C(r_1)(S_1(r_{-z:1}))^{-1} \\
m_{1|1}(r_{-z:1}) &= m_{1|0}(r_{-z:1}) + K_1(r_{-z:1})e_1(r_{-z:1}) \\
P_{1|1}(r_{-z:1}) &= (I - K_1(r_{-z:1})C(r_1))P_{1|0}(r_{-z:1}).
\end{aligned}
$$

- For $k = 2, \ldots, j-1$, we obtain the predicted states (*"missing observations"* $y_{2:j-1}$)

$$
\begin{aligned}
m_{k|k-1}(r_{-z:k}) &= A(r_k)m_{k-1|k-1}(r_{-z:k}) + F(r_k)u_k \\
P_{k|k-1}(r_{-z:k}) &= A(r_k)P_{k-1|k-1}(r_{-z:k})A(r_k)^T + B(r_k)B(r_k)^T \\
m_{k|k}(r_{-z:k}) &= m_{k|k-1}(r_{-z:k}) \\
P_{k|k}(r_{-z:k}) &= P_{k|k-1}(r_{-z:k}).
\end{aligned}
$$

- For k=j

$$
\begin{aligned}
m_{j|j-1}(r_{-z:j}) &= A(r_j)m_{j-1|j-1}(r_{-z:j}) + F(r_j)u_j \\
P_{j|j-1}(r_{-z:j}) &= A(r_j)P_{j-1|j-1}(r_{-z:j})A(r_j)^T + B(r_j)B(r_j)^T \\
e_j(r_{-z:j}) &= y_j - C(r_j)m_{j|j-1}(r_{-z:j}) - G(r_j)u_j \\
S_j(r_{-z:j}) &= C(r_j)P_{j|j-1}(r_{-z:j})C(r_j)^T + D(r_j)D(r_j)^T \\
K_j(r_{-z:j}) &= P_{j|j-1}(r_{-z:j})C(r_j)(S_j(r_{-z:j}))^{-1} \\
m_{j|j}(r_{-z:j}) &= m_{j|j-1}(r_{-z:j}) + K_j(r_{-z:j})e_j(r_{-z:j}) \\
P_{j|j}(r_{-z:j}) &= (I - K_j(r_{-z:j})C(r_j))P_{j|j-1}(r_{-z:j}).
\end{aligned}
$$

Going back to our example defined by the system (6.1.2a, 6.1.2b), we run the modified Kalman filter described above in order to estimate the latent state x_j, given the sequence $r_{-z:j}$ and the observations y_1, y_j, for $j = 2, \ldots, 250$. We take $\nu(x_{-z}) = N(4, 1)$ as initial distribution. We compute the minimum mean square error (MMSE) estimates of the continuous state of the JMLS and the results are shown in Figure 6.3.3.

If we compute the estimates of the continuous state obtained from the above strategy, the distance between the estimates and the true values decreases from 9.9984 to 3.7902.

6.3.2 Sampling from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$

The MCMC algorithm reported at the beginning of this section requires sampling from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$ for $t = -z, \ldots, j$. Before describing how to sample from

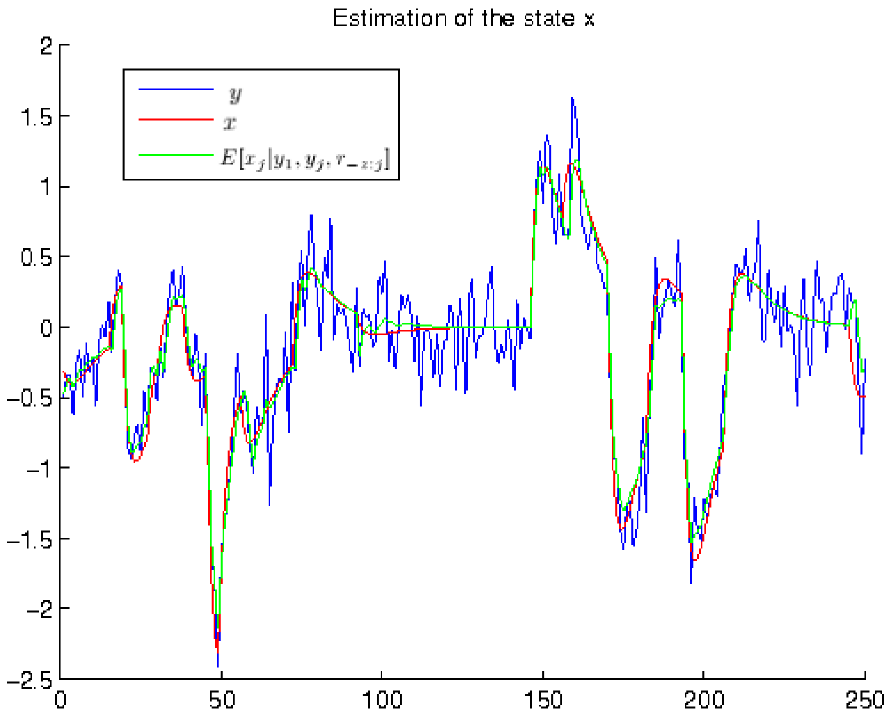

Figure 6.3.3: Modified Kalman Filter- Estimation of the state x_j, given the sequence $r_{-z:j}$ and the observations y_1, y_j, for $j = 2, \ldots, 250$ and $z = 100$. Fixed value of the parameters: $\phi = 0.9, \tau(1) = 0.5, \tau(2) = 0, \sigma = 0.25, \lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{12} = 0.1$. Initial distribution $\nu(x_{-z}) = N(4, 1)$.

this distribution, we describe a generic algorithm of computational complexity $O(T)$ that allows us to sample from $p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$, for $t = 1, \ldots, T$ (as derived in Andrieu and Doucet [2001]). A direct solution to this problem would consist of evaluating for $i \in S$ the distribution

$$p_\theta(r_t = i|y_{1:T}, \mathbf{r}_{-t}) \propto p_\theta(y_{1:T}|r_t = i, \mathbf{r}_{-t}) p_\theta(r_t = i|\mathbf{r}_{-t}),$$

using s times a Kalman filter to compute the s likelihood terms $p_\theta(y_{1:T}|r_t = i, \mathbf{r}_{-t})$ for $i = 1, \ldots, s$. As we need to sample from $p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$ for $t = 1, \ldots, T$, this would result in an algorithm of computational complexity $O(T^2)$. We describe here an algorithm of complexity $O(T)$ that relies on the following key decomposition of the likelihood $p_\theta(y_{1:T}|r_{1:T})$ that allows for the efficient computation of

$p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$ for $t = 1, \ldots, T$. Indeed, for any $t = 2, \ldots, T-1$ (the modifications needed to handle the case of boundaries are straightforward and omitted here), we have

$$\begin{aligned} p_\theta(y_{1:T}|r_{1:T}) &= p_\theta(y_{1:t-1}|r_{1:t-1})p_\theta(y_t|y_{1:t-1}, r_{1:t}) \\ &\times \int p_\theta(y_{t+1:T}|r_{t+1:T}, x_t)p_\theta(x_t|y_{1:t}, r_{1:t})dx_t \end{aligned} \tag{6.3.1}$$

where

$$p_\theta(y_{t:T}|r_{t:T}, x_{t-1}) = \int p_\theta(y_{t+1:T}|r_{t+1:T}, x_t)p_\theta(y_t, x_t|r_t, x_{t-1})dx_t. \tag{6.3.2}$$

The two first terms on the right-hand side of (6.3.1) can be computed using a forward recursion based on the Kalman filter. It appears that it is possible to evaluate the third term using a backward recursion given by (6.3.2). More precisely, $p_\theta(y_{t:T}|r_{t:T}, x_{t-1})$ turns out to be gaussian with mean $M_t(r_{t:T})x_{t-1} + \mathbb{E}[N_t(r_{t:T})]$ and covariance $cov[N_t(r_{t:T})] > 0$, where

$$\begin{aligned} M_t(r_{t:T}) &= \begin{pmatrix} C(r_t)A(r_t) \\ C(r_{t+1})A(r_{t+1})A(r_t) \\ \cdots \\ C(r_{t+i-1})\prod_{j=t+i-1}^{t} A(r_j) \\ \cdots \\ C(r_T)\prod_{j=T}^{t} A(r_j) \end{pmatrix} \\ N_t(r_{t:T}) &= \begin{pmatrix} D(r_t)w_t + C(r_t)(B(r_t)v_t + G(r_t)u_t) + G(r_t)u_t \\ \cdots \\ D(r_{t+i-1})w_{t+i-1} + \text{remaining terms} \\ \cdots \\ D(r_T)w_T + \text{remaining terms} \end{pmatrix}. \end{aligned}$$

The positiveness of $cov[N_t(r_{t:T})]$ comes from the assumption $D(i)D(i)^T > 0$ for $i \in S$. We define $L_t(r_{t:T}) = \mathbb{E}[N_t(r_{t:T})N_t^T(r_{t:T})]$ and

$$P'^{-1}_{t-1|t}(r_{t:T}) = M_t^T(r_{t:T})L_t^{-1}(r_{t:T})M_t(r_{t:T}),$$

$$P'^{-1}_{t-1|t}(r_{t:T})m'_{t-1|t}(r_{t:T}) \quad = \quad M_t^T(r_{t:T})L_t^{-1}(r_{t:T})y_{t:T}.$$

Mayne [1966] has established the algorithm to compute them recursively in time. More precisely, the quantities $P'^{-1}_{t-1|t}(r_{t:T})$ and $P'^{-1}_{t-1|t}(r_{t:T})m'_{t-1|t}(r_{t:T})$ always satisfy the following backward information filter recursions

1. Initialization

$$\begin{aligned} P'^{-1}_{T|T}(r_T) =& C^T(r_t)(D(r_T)D^T(r_T))^{-1}C(r_T) \\ P'^{-1}_{T|T}(r_T)m'_{T|T}(r_T) =& C^T(r_t)(D(r_T)D^T(r_T))^{-1}C(r_T)(y_T - G(r_T)u_T). \end{aligned}$$

2. Backward recursion. For $t = T-1, \ldots, 1$,

$$\begin{aligned} \Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:T})B(r_{t+1})\right]^{-1} \\ P'^{-1}_{t|t+1}(r_{t+1:T}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:T})\times \\ &\times \left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:T})\right)A(r_{t+1}), \\ P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:T}) &= A^T(r_{t+1})\times \\ &\times \left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:T})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times \\ &\times P'^{-1}_{t+1|t+1}(r_{t+1:T})\left(m'_{t+1|t+1}(r_{t+1|T}) - F(r_{t+1})u_{t+1}\right), \\ P'^{-1}_{t|t}(r_{t:T}) &= P'^{-1}_{t|t+1}(r_{t+1:T}) + C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t), \\ P'^{-1}_{t|t}(r_{t:T})m'_{t|t}(r_{t:T}) &= P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:T})+ \\ &+ C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t)(y_t - G(r_t)u_t). \end{aligned}$$

Now, combining (6.3.1) and the previous results, one obtains an expression for $p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$, t=2,...,T-1. In fact

$$\begin{aligned} p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t}) \propto& p_\theta(r_t|\mathbf{r}_{-t})p_\theta(y_{1:T}, \mathbf{r}_{-t}, r_t) \\ \propto& p_\theta(r_t|r_{t-1}, r_{t+1})p_\theta(y_t|y_{1:t-1}, r_{1:t})\times \\ &\times \int p_\theta(y_{t+1:T}|r_{t+1:T}, x_t)p_\theta(x_t|y_{1:t}, r_{1:t})dx_t. \end{aligned}$$

The two first terms are easy to compute as $p_\theta(r_t|r_{t-1}, r_{t+1})$ is given by the transition matrix of the Markov chain and $p_\theta(y_t|y_{1:t-1}, r_{1:t}) = N(\tilde{e}_t(r_{1:t}), \tilde{S}_t(r_{1:t}))$, where the innovation $e_t(r_{1:t})$ and its covariance $\tilde{S}_t(r_{1:t})$ are evaluated using the Kalman filter. Now, the last term is equal to

$$\int p_\theta(y_{t+1:T}|r_{t+1:T}, x_t)p_\theta(x_t|y_{1:t}, r_{1:t})dx_t = N(y_{t+1:T} - M_{t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t}),$$
$$L_{t+1}(r_{t+1:T}) + M_{t+1}(r_{t+1:T})P_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})),$$

where

$$N(m, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}m^T\Sigma^{-1}m\right).$$

If $P_{t|t}(r_{1:t}) = \mathbf{0}_{n_x\times n_x}$,

$$\begin{aligned} p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t}) \propto & \lambda_{r_{t-1}r_t}\lambda_{r_tr_{t+1}}N(\tilde{e}_t(r_{1:t}), \tilde{S}_t(r_{1:t})) \\ & \times \exp\Big(-\frac{1}{2}\Big[m^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t}) - \\ & - 2\tilde{m}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:T})\Big]\Big), \end{aligned}$$

since $y^T_{t+1:T}L^{-1}_{t+1}(r_{t+1:T})y_{t+1:T}$ does not depend on r_t. If $\tilde{P}_{t|t}(r_{1:t}) \neq \mathbf{0}_{n_x\times n_x}$ and it is symmetric, then there exist $\tilde{\Pi}_{t|t}(r_{1:t})$ and $\tilde{Q}_{t|t}(r_{1:t})$ such that

$$\tilde{P}_{t|t}(r_{1:t}) = \tilde{Q}_{t|t}(r_{1:t})\tilde{\Pi}_{t|t}(r_{1:t})\tilde{Q}^T_{t|t}(r_{1:t}).$$

The matrices $\tilde{Q}_{t|t}(r_{1:t})$ and $\tilde{\Pi}_{t|t}(r_{1:t})$ are obtained using the singular value decomposition of $\tilde{P}_{t|t}(r_{1:t})$. Matrix $\tilde{\Pi}_{t|t}(r_{1:t})$ is a diagonal matrix with the nonzero eigenvalues of $\tilde{P}_{t|t}(r_{1:t})$ as elements. Note that

$$\begin{aligned} & \left|L_{t+1}(r_{t+1:T}) + M_{t+1}(r_{t+1:T})\tilde{P}_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})\right| = \\ & = |L_{t+1}(r_{t+1:T})| \left|\tilde{\Pi}_{t|t}(r_{1:t})\tilde{Q}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{Q}_{t|t}(r_{1:t}) + I_{n_t}\right| \end{aligned}$$

and that

$$\begin{aligned} & \left[L_{t+1}(r_{t+1:T}) + M_{t+1}(r_{t+1:T})\tilde{P}_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})\right]^{-1} = \\ & \left[L_{t+1}(r_{t+1:T}) + M_{t+1}(r_{t+1:T})\tilde{Q}_{t|t}(r_{1:t})\tilde{\Pi}_{t|t}(r_{1:t})\tilde{Q}^T_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})\right]^{-1} = \\ & = L_{t+1}(r_{t+1:T})^{-1} - L_{t+1}(r_{t+1:T})^{-1}M_{t+1}(r_{t+1:T})R_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})L_{t+1}(r_{t+1:T})^{-1}, \end{aligned}$$

where

$$R_{t|t}(r_{1:t}) = \tilde{Q}_{t|t}(r_{1:t})\left[\tilde{\Pi}^{-1}_{t|t}(r_{1:t}) + \tilde{Q}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{Q}_{t|t}(r_{1:t})\right]^{-1}\tilde{Q}^T_{t|t}(r_{1:t}).$$

Therefore

$$\begin{aligned}
&(y_{t+1:T} - M_{t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t}))^T[L_{t+1}(r_{t+1:T})+\\
&+ M_{t+1}(r_{t+1:T})\tilde{P}_{t|t}(r_{1:t})M^T_{t+1}(r_{t+1:T})]^{-1}(y_{t+1:T} - M_{t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t})) =\\
=&y^T_{t+1:T}L^{-1}_{t+1}(r_{t+1:T})y_{t+1:T} + \tilde{m}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t})+\\
&- 2\tilde{m}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:T}) - (m'_{t|t+1}(r_{t+1:T}) - \tilde{m}_{t|t}(r_{1:t}))^T\times\\
&\times P'^{-1}_{t|t+1}(r_{t+1:T})R_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})(m'_{t|t+1}(r_{t+1:T}) - \tilde{m}_{t|t}(r_{1:t}))
\end{aligned}$$

and hence

$$\begin{aligned}
&p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t}) \propto \lambda_{r_{t-1}r_t}\lambda_{r_t r_{t+1}}N(\tilde{e}_t(r_{1:t}), \tilde{S}_t(r_{1:t}))\times\\
&\times\left|\tilde{\Pi}_{t|t}(r_{1:t})\tilde{Q}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{Q}_{t|t}(r_{1:t}) + I_{n_t}\right|^{-1/2}\times\\
&\times\exp\Big(- \frac{1}{2}[\tilde{m}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})\tilde{m}_{t|t}(r_{1:t})+\\
&- 2\tilde{m}^T_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:T}) - (m'_{t|t+1}(r_{t+1:T}) - \tilde{m}_{t|t}(r_{1:t}))^T\times\\
&\times P'^{-1}_{t|t+1}(r_{t+1:T})R_{t|t}(r_{1:t})P'^{-1}_{t|t+1}(r_{t+1:T})(m'_{t|t+1}(r_{t+1:T}) - \tilde{m}_{t|t}(r_{1:t}))]\Big).
\end{aligned}$$

Remark 6.3.2. *Remember that $\tilde{m}_{t|t-1}(r_{1:t}), \tilde{P}_{t|t-1}(r_{1:t}), \tilde{m}_{t|t}(r_{1:t}), \tilde{P}_{t|t}(r_{1:t}), \tilde{e}_t(r_{1:t})$ and $\tilde{S}_t(r_{1:t})$ are, respectively, the one-step ahead prediction and covariance of x_t, the filtered estimate and covariance of x_t, the innovation at time t and the covariance of this innovation. These quantities are given by the standard Kalman filter, the system (6.1.1a, 6.1.1b) being linear-gaussian until t conditional upon $r_{1:t}$.*

Figure 6.3.4 shows the performance of this algorithm. We simulate a realization of a two states Markov chain of length $T = 250$, where $\lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{21} = 0.1$. We run 100 times the algorithm to sample from the distribution of $r_{1:T}$ given all the observations between time 1 to time T, where observations come from a simulation from the model in (6.1.3a, 6.1.3b), where $\phi = 0.9, \tau(1) = 5, \sigma = 1$. We start from an uniform distribution over state 1 and 2. We clearly see that, after some iteration, we almost recover the original sequence.

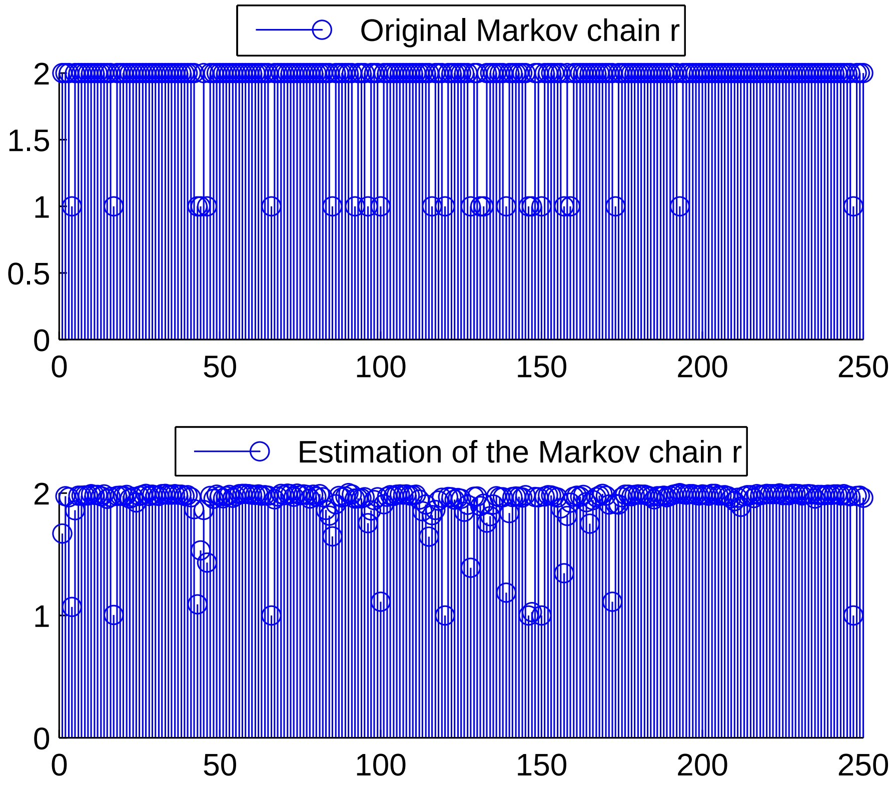

Figure 6.3.4: Sample from $p_\theta(r_{1:T}|y_{1:T})$, $T = 250$, where observations come from a simulation from the model in (6.1.3a, 6.1.3b), where $\phi = 0.9, \tau(1) = 5, \sigma = 1$ and $\lambda_1 = \lambda_2 = 0.5, \lambda_{12} = 0.9, \lambda_{21} = 0.1$. Original sequence $r_{1:T}$ (top) and estimated sequence after 100 iterations (bottom).

The algorithm above allows to sample from $p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$ for $t = 1, \ldots, T$. This strategy can be modified in order to obtain samples from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$ for $t = 1, \ldots, j$. In our contest, the conditioning is on the observations y_1 and y_j and not on all the observations between 1 and j. Using the same idea of the modified Kalman filter, we run the backward information filter recursions with all the observations $y_{1:j}$ setting an infinity variance for the missing observations from time 2 to time $j - 1$. As before, this means that $(D(\cdot)D(\cdot)^T)^{-1}$ is close to zero.

1. Initialization

$$P'^{-1}_{j|j}(r_j) = C^T(r_j)(D(r_j)D^T(r_j))^{-1}C(r_j)$$

$$P'^{-1}_{j|j}(r_j)m'_{j|j}(r_j) = C^T(r_j)(D(r_j)D^T(r_j))^{-1}C(r_j)(y_j - G(r_j)u_j).$$

2. Backward recursion without observations. For $t = j-1, \ldots, 2$,

$$\begin{aligned}
\Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\right]^{-1} \\
P'^{-1}_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\times \\
&\quad \times \left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\right)A(r_{t+1}), \\
P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})\times \\
&\quad \times \left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times \\
&\quad \times P'^{-1}_{t+1|t+1}(r_{t+1:j})\left(m'_{t+1|t+1}(r_{t+1:j}) - F(r_{t+1})u_{t+1}\right), \\
P'^{-1}_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j}), \\
P'^{-1}_{t|t}(r_{t:j})m'_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}).
\end{aligned}$$

3. Backward recursion. For $t = 1$,

$$\begin{aligned}
\Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\right]^{-1} \\
P'^{-1}_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\times \\
&\quad \times \left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\right)A(r_{t+1}), \\
P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})\times \\
&\quad \times \left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times \\
&\quad \times P'^{-1}_{t+1|t+1}(r_{t+1:j})\left(m'_{t+1|t+1}(r_{t+1:j}) - F(r_{t+1})u_{t+1}\right), \\
P'^{-1}_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j}) + C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t), \\
P'^{-1}_{t|t}(r_{t:j})m'_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:j}) + \\
&\quad + C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t)(y_t - G(r_t)u_t).
\end{aligned}$$

For $t = 1, \ldots, j$, the distribution $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$ has the same expression as $p_\theta(r_t|y_{1:T}, \mathbf{r}_{-t})$, where $\tilde{m}_{t|t-1}(r_{1:t})$, $\tilde{P}_{t|t-1}(r_{1:t})$, $\tilde{m}_{t|t}(r_{1:t})$, $\tilde{P}_{t|t}(r_{1:t})$, $\tilde{e}_t(r_{1:t})$ and $\tilde{S}_t(r_{1:t})$ are replaced by $m_{t|t-1}(r_{1:t})$, $P_{t|t-1}(r_{1:t})$, $m_{t|t}(r_{1:t})$, $P_{t|t}(r_{1:t})$, $e_t(r_{1:t})$ and $S_t(r_{1:t})$, obtained from the modified Kalman filter algorithm. Furthermore, the function $N(e_t(r_{1:t})$,$S_t(r_{1:t}))$ is computed only for $t = j$ and $t = 1$, being a constant factor in the other cases.

Remark 6.3.3. *To sum up, the algorithm to sample from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t}), t = 1, \ldots, j$ requires first the computation of the backward information filter, second, the evaluation of $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$ combining the information and the modified Kalman filter, and finally, sampling from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t})$ and storing accordingly the updated set of sufficient statistics $m_{t|t}(r^{(k)}_{1:t-1}, r^{(k)}_t)$, $P_{t|t}(r^{(k)}_{1:t-1}, r^{(k)}_t)$.*

Going back to our case, where the stationary distribution is unknown, it becomes important to sample from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t}, \nu)$, for $t = -z, \ldots, j$, for a fixed value of z and for $j = 2, \ldots, L$. The strategy described above can be easily extended to this aim. In the next, we describe how to generalize the backward information recursions in order to sample from $p_\theta(r_{-z:j}|y_1, y_j, \nu)$, ν being the initial distribution. Combining this algorithm with the modified Kalman filter recursions allows us to evaluate integrals with respect to the posterior distribution $p_\theta(x_{-z:j}, r_{-z:j}|y_1, y_j, \nu)$.

The modified Kalman filter, as described in Section 6.3.1, allows us to compute the quantities $m_{t|t-1}(r_{-z:t})$, $P_{t|t-1}(r_{-z:t})$, $m_{t|t}(r_{-z:t})$, $P_{t|t}(r_{-z:t})$, $e_t(r_{-z:t})$ and $S_t(r_{1:t})$ needed to sample from $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t}, \nu)$, for $t = -z, \ldots, j$. Backward information recursions modify as follows:

1. Initialization

$$\begin{aligned} P'^{-1}_{j|j}(r_j) =& C^T(r_j)(D(r_j)D^T(r_j))^{-1}C(r_j) \\ P'^{-1}_{j|j}(r_j)m'_{j|j}(r_j) =& C^T(r_j)(D(r_j)D^T(r_j))^{-1}C(r_j)(y_j - G(r_j)u_j). \end{aligned}$$

2. Backward recursion without observations. For $t = j-1, \ldots, 2$,

$$\begin{aligned} \Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\right]^{-1} \\ P'^{-1}_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\times \\ &\quad \times \left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\right)A(r_{t+1}), \\ P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})\times \\ &\quad \times \left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times \\ &\quad \times P'^{-1}_{t+1|t+1}(r_{t+1:j})\left(m'_{t+1|t+1}(r_{t+1:j}) - F(r_{t+1})u_{t+1}\right), \end{aligned}$$

$$\begin{aligned}
P'^{-1}_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j}),\\
P'^{-1}_{t|t}(r_{t:j})m'_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}).
\end{aligned}$$

3. Backward recursion. For $t = 1$,

$$\begin{aligned}
\Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\right]^{-1}\\
P'^{-1}_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\times\\
&\quad\times\left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\right)A(r_{t+1}),\\
P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})\times\\
&\quad\times\left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times\\
&\quad\times P'^{-1}_{t+1|t+1}(r_{t+1:j})\left(m'_{t+1|t+1}(r_{t+1:j}) - F(r_{t+1})u_{t+1}\right),\\
P'^{-1}_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j}) + C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t),\\
P'^{-1}_{t|t}(r_{t:j})m'_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:T})m'_{t|t+1}(r_{t+1:j})+\\
&\quad+ C^T(r_t)(D(r_t)D^T(r_t))^{-1}C(r_t)(y_t - G(r_t)u_t).
\end{aligned}$$

4. Backward recursion without observations. For $t = -1, \ldots, -z$,

$$\begin{aligned}
\Delta_{t+1} &= \left[I_{n_v} + B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\right]^{-1}\\
P'^{-1}_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\times\\
&\quad\times\left(I_{n_x} - B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})P'^{-1}_{t+1|t+1}(r_{t+1:j})\right)A(r_{t+1}),\\
P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}) &= A^T(r_{t+1})\times\\
&\quad\times\left(I_{n_x} - P'^{-1}_{t+1|t+1}(r_{t+1:j})B(r_{t+1})\Delta_{t+1}B^T(r_{t+1})\right)\times\\
&\quad\times P'^{-1}_{t+1|t+1}(r_{t+1:j})\left(m'_{t+1|t+1}(r_{t+1:j}) - F(r_{t+1})u_{t+1}\right),\\
P'^{-1}_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j}),\\
P'^{-1}_{t|t}(r_{t:j})m'_{t|t}(r_{t:j}) &= P'^{-1}_{t|t+1}(r_{t+1:j})m'_{t|t+1}(r_{t+1:j}).
\end{aligned}$$

Once these quantities are computed, we can evaluate $p_\theta(r_t|y_1, y_j, \mathbf{r}_{-t}, \nu)$, for $t = -z, \ldots, j$ combining the backward information recursions and the modified Kalman filter, in the case where the stationary distribution is unknown and it is replaced with a generic initial distribution ν. Again, the function $N(e_t(r_{-z:t}), S_t(r_{-z:t}))$

is computed only for $t = j$ and $t = 1$, being a constant factor in the other cases.

Remark 6.3.4. *In order to compute statistics defined on page 88, we need samples* $r_{-z:j}$ *from the distribution* $p_\theta(r_{-z:j}|y_1, y_j, v)$ *and the first two moments of the distribution* $p_\theta(x_{-z:j}|y_1, y_j, r_{-z:j}, v)$*. Samples can be obtained from the algorithm above, while moments are achieved through Kalman smoother formulas as in Chapter 5. These statistics may be recursively updated following the same idea of (5.4.2).*

Chapter 7

Discussion

The work in this thesis dealt with the problem of static parameter estimation in general state space models. Given the difficulties arising in this framework, we have focused on inferential procedures based on composite likelihood functions, in particular *pairwise* and *split data likelihood* functions. Asymptotic properties of the parameter estimators obtained by maximizing these functions in general state space scenario were investigated. We proved that standard results, as strong consistency, depend on the properties of the processes involved, in particular stationarity and ergodicity that ensure forgetting behavior of the filter.

Even if the models we considered are strictly stationary, in many situations invariant distribution is difficult (or even impossible) to compute. In this cases, it becomes important to quantify the bias in the estimate when stationary distribution is replaced with a generic approximation. When stationary distribution is unknown, objective functions need to be approximated and this leads to biased estimate of the parameters. We proved that the bias introduced when using a generic distribution instead of the stationary distribution in the pairwise likelihood function depends on how close the two distributions are, and, again, on the ergodic properties of the latent process. To prove this result, we need uniformly convergence of the pairwise likelihood function and of its gradient.

We also investigated efficiency problem in pairwise and split data likelihood framework as L, i.e. the lag distance between pairs or the length of a block, respectively, increases. We empirically proved that the loss of efficiency, with respect to maximum likelihood estimator, of the maximum split data likelihood estimator vanishes as L increases, while the variance of the maximum pairwise likelihood estimator decreases until a certain L^* and then it tends to increase. Anyway, until now, no general results about evaluation of this loss are available, even if this behavior is observed also in Varin and Vidoni [2009]. Moreover, we suggested the existence of a 'best lag' L^*, in terms of variance of the maximum pairwise likelihood estimator. However, we have not theoretically analyzed yet how to determine such value. In the future, we wish to investigate this topic through the evaluation of the asymptotic variance of the maximum pairwise likelihood estimator, in order to obtain an expression that depends on the lag distance L. We would like to follow the idea of Andrieu et al. [2007] in the non overlapping version of split data likelihood function. To do that, we need to quantify the loss of efficiency with respect to the full likelihood function. This requires a deep study of the dependence structures between pairs of observations, exploiting ergodic properties of the processes involved.

We focused on numerical methods to compute estimates of the parameter describing a general state space model. We presented an on line Expectation- Maximization algorithm in order to obtain the maximum pairwise likelihood estimate in a general state space framework. We illustrated this method for a linear gaussian model. We modified standard Kalman filter recursions in order to take into account conditioning on pairs of observations instead of all observations. In this simple example, we gave an empirical evidence of our bias theorem, i.e. starting from a generic distribution and sampling from the transition kernel reduces the bias in the estimates for each parameter in the model. We then extended this algorithm to make inference in jump Markov linear systems. In this framework, we developed new procedures to sample from the latent discrete state Markov chain given the pairs of observations.

Further research will focus on numerical methods to compute estimate of the parameter in more general contexts. In scenarios where $\mathbb{E}^{(j)}_{\theta_k}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$, i.e. the expectation of Ψ with respect to $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ as defined in (5.4.2), does not admit an analytical expression, a further Monte Carlo approximation can be used. Assume that a good approximation $q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ of $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ is available, and that it is easy to sample from $q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$. In this case, the expectation step will be altered as follows

- Sample $X_{1:j}^{(i)}$ from $q_{\theta_k}(\cdot|y_k, y_{k+j-1})$, for $i = 1, \ldots, N$
- Approximate $\Phi(\theta_k, \theta^*)$ as

$$\hat{\Phi}^{(k)} = (1 - \gamma_k)\hat{\Phi}^{k-1} + \gamma_k \left[\frac{1}{L} \sum_{j=2}^{L+1} \sum_{i=1}^{N} W_k^{(i)} \Psi(X_{1:j}^{(i)}, Y_k, Y_{k+j-1}) \right],$$

where

$$W_k^{(i)} \propto \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}, \quad \sum_{i=1}^{N} W_k^{(i)} = 1.$$

As N increases, the importance sampling approximation converges towards the true expectation. Note that if it is possible to sample from $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ exactly, then it is not necessary to have a large number N of samples and a single one might even be sufficient. Indeed it is only necessary to produce estimates of $\mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$. We underline that the algorithm above leads to asymptotically biased estimates, but that this can be corrected by considering the following recursion for the estimation of the conditional expectation

$$\begin{aligned}
\hat{F}_k &= (1 - \gamma_k)\hat{F}^{k-1} + \gamma_k \left[\frac{1}{L} \sum_{j=2}^{L+1} \frac{1}{N} \sum_{i=1}^{N} \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})} \Psi(X_{1:j}^{(i)}, Y_k, Y_{k+j-1}) \right], \\
\hat{N}_k &= (1 - \gamma_k)\hat{N}^{k-1} + \gamma_k \left[\frac{1}{L} \sum_{j=2}^{L+1} \frac{1}{N} \sum_{i=1}^{N} \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})} \right],
\end{aligned}$$

and let $\hat{\Phi}^{(k)} = \frac{\hat{F}_k}{\hat{N}_k}$. It is also possible to use rejection sampling or SMC techniques to approximate this expectation. This idea may represent a starting point for subsequent extensions to more complex models.

Appendix A

Technical results about consistency

A.1 Assumptions

In Theorem 3.2.1, we prove consistency of the pairwise likelihood estimator under the following assumptions

(C1) There exists $\underline{f_0}, \underline{g_0} > 0$ and $\overline{f}_0, \overline{g}_0 < \infty$ such that far all $x, x', y, \theta \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta$

$$\underline{f_0} \leq f_\theta(x'|x) \leq \overline{f}_0, \quad \underline{g_0} \leq g_\theta(y|x) \leq \overline{g}_0$$

(C2) Θ is a compact set

(C3) f_θ and g_θ are continuous as functions of θ

There is an integer $L \geq 1$ such that, for every $j = 2, \ldots, L+1$

(C4) $p_\theta(y_1, y_j) = p_{\theta^*}(y_1, y_j)$ if and only if $\theta = \theta^*$

(C5) for the true parameter value θ^* we have $\mathbb{E}[|log[p_{\theta^*}(y_1, y_j)]|] < \infty$

(C6) for each θ there is a $\delta > 0$ (sufficiently small) such that

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1, y_j)\right)^+\right] < \infty$$

and there is a $b > 0$ (sufficiently large) such that

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'|>b} \log p_{\theta'}(y_1, y_j)\right)^+\right] < \infty,$$

where h^+ denotes the positive part of the function h

(C7) if $\lim_{i\to\infty} |\theta_i| = \infty$ then $\lim_{i\to\infty} p_{\theta_i}(y_1, y_j) = 0$.

Condition $(C1)$ implies that the process $\{X_k, Y_k\}$ is an uniformly ergodic Markov chain.

A.2 Middle results

The composite likelihood we consider is defined as

$$L_P^{(L)}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=1+1}^{\min(i+L,n)} log[p_\theta(y_i, y_j)].$$

If we normalize and take the log, we have that the normalized log pairwise likelihood is defined as

$$l_P^{(L)}(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\frac{1}{L} \sum_{j=1+1}^{\min i+L,n} log[p_\theta(y_i, y_j)]\right].$$

We call $L_P^{(L)}(\theta; y_{1:n})$ a *pairwise likelihood*, and any global maximum point $\hat{\theta}_P^{(L)}$ of it is a *maximum pairwise likelihood estimate*. We prove here some middle results necessary to state that pairwise likelihood estimator is strongly consistent, as proved in Theorem 3.2.1. We start with a lemma concerning the L-dimensional Kullback-Leibler information

$$\begin{aligned} K_P^{(L)}(\theta, \theta^*) &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} \log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} p_{\theta^*}(y_1, y_j) dy_1 dy_j \\ &= \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*}\left[\log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)}\right]. \end{aligned}$$

If we define $l_P^{(L)}(\theta) = \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\log\frac{p_{\theta^*}(y_1,y_j)}{p_\theta(y_1,y_j)}\right]$ then $K_P^{(L)}(\theta,\theta^*) = l_P^{(L)}(\theta^*) - l_P^{(L)}(\theta)$. Under the ergodicity assumption (C1), the log pairwise likelihood $l_P^{(L)}(\theta, y_{1:n})$ satisfies

$$\lim_{n\to\infty} l_P^{(L)}(\theta, y_{1:n}) = l_P^{(L)}(\theta).$$

Lemma A.2.1. *Assume that Conditions* $(C4 - C6)$ *hold. Then* $K_P^{(L)}(\theta,\theta^*) \geq 0$ *with equality if and only if* $\theta^* = \theta$.

Proof. By Conditions $(C6)$, the expected values $l_P^{(L)}(\theta^*)$ and $l_P^{(L)}(\theta)$ exist. Because of the Assumption $(C5)$, we have that $l_P^{(L)}(\theta^*)$ is finite. If $l_P^{(L)}(\theta) = -\infty$, Lemma A.2.1 obviously holds. Thus we shall consider the case when $l_P^{(L)}(\theta)$ is finite. Then $K_P^{(L)}(\theta,\theta^*) = l_P^{(L)}(\theta^*) - l_P^{(L)}(\theta)$ exists finite. For every $j = 2,\ldots,L+1$, by Jensen inequality we have that

$$\begin{aligned}
&\int_{\mathcal{Y}^2}\log\frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)}p_{\theta^*}(y_1,y_j)dy_1dy_j\\
&\leq \log\left[\int_{\mathcal{Y}^2}\frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)}p_{\theta^*}(y_1,y_j)dy_1dy_j\right]\\
&= \log\left[\int_{\mathcal{Y}^2}p_\theta(y_1,y_j)dy_1dy_j\right] = \log[1] = 0. \qquad \text{(A.2.1)}
\end{aligned}$$

Since

$$-K_P^{(L)}(\theta,\theta^*) = \frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{Y}^2}\log\frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)}p_{\theta^*}(y_1,y_j)dy_1dy_j$$

and given the result in (A.2.1), $K_P^{(L)}(\theta,\theta^*) \geq 0$ and this proves the first part of the lemma. The equality holds if and only if, for every $j = 2,\ldots,L+1$, $p_\theta(y_1,y_j) = p_{\theta^*}(y_1,y_j)$ almost everywhere. By Condition $(C4)$, this is true if and only if $\theta = \theta^*$. □

Lemma A.2.2. *Assume that Conditions* $(C3)$ *and* $(C6)$ *hold. Then for every* $\theta \in \Theta$ *and for every* $j = 2,\ldots,L+1$

$$\lim_{\delta\to 0}\mathbb{E}_{\theta^*}\left[\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right] = \mathbb{E}_{\theta^*}\left[\log p_\theta(y_1,y_j)\right].$$

Proof. By Condition $(C3)$, $p_\theta(y_1,y_j)$ is continuous for all y_1, y_j, $j = 2,\ldots,L+1$.

Then

$$\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^+ = (\log p_\theta(y_1,y_j))^+,$$

except perhaps on a set whose probability measure is zero.

Since $\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^+$ is an increasing function of δ, it follows from Assumption $(C6)$ that

$$\begin{aligned}\lim_{\delta\to 0}\mathbb{E}_{\theta^*}&\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^+\right]\\ &= \mathbb{E}_{\theta^*}\left[\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^+\right] = \mathbb{E}_{\theta^*}\left[\left(\log p_\theta(y_1,y_j)\right)^+\right]. \qquad \text{(A.2.2)}\end{aligned}$$

Again by Condition $(C3)$

$$\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^- = (\log p_\theta(y_1,y_j))^-,$$

except perhaps on a set whose probability measure is zero, where h^- denotes the negative part of the function h. Then the relation

$$\lim_{\delta\to 0}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^-\right] = \mathbb{E}_{\theta^*}\left[(\log p_\theta(y_1,y_j))^-\right] \qquad \text{(A.2.3)}$$

is clearly satisfied in both cases, when $\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^-\right]$ is finite and when it is equal to $+\infty$. Lemma A.2.2 is a consequence of (A.2.2) and (A.2.3). □

Lemma A.2.3. *Assume that Conditions* $(C3, C6, C7)$ *hold. Then, for every* $j = 2,\dots,L+1$

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right] = -\infty.$$

Proof. From Assumptions $(C3)$ and $(C7)$

$$\lim_{b\to\infty}\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j) = \lim_{\theta\to\infty}\log p_\theta(y_1,y_j) = -\infty.$$

According to Assumption ($C6$),

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+\right]<\infty,$$

and since $\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+$ is a decreasing function of b we have that

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+\right]=0. \tag{A.2.4}$$

Since $\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-$ is an increasing function of b, in the same way we have that

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-\right]=+\infty \tag{A.2.5}$$

in both cases, when $\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-\right]$ is finite and when it is equal to $+\infty$. Lemma A.2.3 is a consequence of (A.2.4) and (A.2.5).

□

Appendix B

Technical results

B.1 Preliminary results

We recall here two results that will be used in the next. They are general and standard results in Markov chains theory (see Meyn and Tweedie [1993] for a deeper and wider treatment). For any two probability measures μ, ν we define the total variation distance $||\mu - \nu|| = \sup_A |\mu(A) - \nu(A)|$, where $A \in \mathcal{B}(\mathcal{X})$ and we also recall the identity $\sup_{0 \le f \le 1} |\mu(f) - \nu(f)| = ||\mu - \nu||$, where f is any measurable function. We use here the standard notation as in Meyn and Tweedie [1993].

Theorem B.1.1. *Let $P(x, \cdot)$ be the transition kernel of a Markov chain. Suppose there exist $\epsilon > 0$, and a measure λ on $\mathcal{X}$ such that for every $x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$*

$$P(x, A) \ge \epsilon\lambda(A). \tag{B.1.1}$$

Then for every $\mu, \nu \in \mathcal{P}(\mathcal{X})$

$$||\mu P - \nu P|| := \sup_{0 \le f \le 1} \left| \int [\mu(dx) - \nu(dx)] P(x, f) \right| \le (1 - \epsilon)||\mu - \nu||,$$

where $P(x, f) := \int P(x, dy) f(y)$.

Proof. Let us rewrite $P(x, dy)$ as

$$P(x, dy) = \epsilon\lambda(dy) + (1 - \epsilon)\frac{P(x, dy) - \epsilon\lambda(dy)}{1 - \epsilon}$$

and by condition (B.1.1)

$$R(x,dy) := \frac{P(x,dy) - \epsilon\lambda(dy)}{1-\epsilon} \geq 0.$$

Moreover

$$\begin{aligned}
|R(x,f)| &= \left|\int f(y)R(x,dy)\right| \\
&\leq \int R(x,dy) = \frac{1}{1-\epsilon}\int [P(x,dy) - \epsilon\lambda(dy)] \leq 1.
\end{aligned}$$

We have that

$$\begin{aligned}
&\left|\int [\mu(dx) - \nu(dx)]P(x,f)\right| = \left|\int\int [\mu(dx)P(x,dy) - \nu(dx)P(x,dy)]f(y)\right| \\
&= (1-\epsilon)\left|\int\int [\mu(dx)R(x,dy) - \nu(dx)R(x,dy)]f(y)\right| \\
&= (1-\epsilon)\left|\int (\mu(dx) - \nu(dx))R(x,f)\right| \\
&\leq (1-\epsilon)\|\mu - \nu\|.
\end{aligned}$$

Since the bound above does not depend on the function f, we can conclude that

$$\|\mu P - \nu P\| \leq (1-\epsilon)\|\mu - \nu\|.$$

□

When a condition such (B.1.1) holds, we say that X satisfies a *one-step minorization condition*. Under this hypothesis, X has a unique invariant measure and is uniformly ergodic (see again Meyn and Tweedie [1993] for the proof).

Corollary B.1.2. *Under the hypothesis of Theorem B.1.1, for* $k > 0$

$$\|\mu P^k - \nu P^k\| \leq (1-\epsilon)^k\|\mu - \nu\|,$$

where $P^k(x,\cdot)$ *is the k-step Markov transition kernel corresponding to P.*

Proof. By Chapman-Kolmogorov equations

$$\begin{aligned}
||\mu P^k - \nu P^k|| &= ||\mu P^{k-1}P - \nu P^{k-1}P|| \\
&\leq (1-\epsilon)||\mu P^{k-1} - \nu P^{k-1}|| \\
&\leq \dots \\
&\leq (1-\epsilon)^k||\mu - \nu||.
\end{aligned}$$

□

B.2 Assumptions

Our results hold under the following assumptions

(A1) Θ is a compact set, θ^* is a unique global maximum of $l_P(\theta)$ and belongs to the interior of Θ, denoted $\overset{\circ}{\Theta}$. Moreover $l_P(\theta)$ is twice continuously differentiable on $\overset{\circ}{\Theta}$ and $H_P(\theta^*) := \nabla^2 l_P(\theta^*)$ is positive definite.

(A2) We assume that f_θ and g_θ are twice continuously differentiable and that there exist $\underline{f_0}, \underline{g_0} > 0$ and $\overline{f_0}, \overline{g_0}, \overline{f_1}, \overline{g_1}, \overline{f_2}, \overline{g_2} < +\infty$ such that for all $x, x', y, \theta \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta$

$$\begin{aligned}
&\underline{f_0} \leq f_\theta(x'|x) \leq \overline{f_0}, && \underline{g_0} \leq g_\theta(y|x) \leq \overline{g_0} \\
&|\nabla \log f_\theta(x'|x)| < \overline{f_1}, && |\nabla \log g_\theta(y|x)| < \overline{g_1} \\
&|\nabla^2 \log f_\theta(x'|x)| < \overline{f_2} \quad \text{and} && |\nabla^2 \log g_\theta(y|x)| < \overline{g_2}.
\end{aligned}$$

In addition, we assume that $\nabla^2 \log f_\theta(x'|x)$ and $\nabla^2 \log g_\theta(y|x)$ are continuous in θ, uniformly in $x, x', y, \in \mathcal{X}^2 \times \mathcal{Y}$ and that $\sup_{\theta\in\Theta} |\nabla \log \mu| \leq \bar{\mu}$, with $\bar{\mu} \in (0, \infty)$, $\mu \in \mathcal{P}(\mathcal{X})$.

Assumptions (A2) implies that for all $x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$,

$$P(x, A) := \int_A f_\theta(x'|x)dx' \geq \underline{f_0}\lambda(A),$$

where λ denotes the Lebesgue measure. As stated in Section B.1, this means that X has a unique invariant measure π_θ and is uniformly ergodic.

B.3 Useful theorems

Theorem B.3.1. *For* $j = 2, \dots, L+1$ *and for any* $\theta \in \Theta, x_1, y_1, y_j \in \mathcal{X} \times \mathcal{Y}^2$

$$\underline{g_0}^2 \leq p_\theta(y_1, y_j | x_1) \leq \overline{g_0}^2$$

Proof. From Assumptions (A2), simple calculations yield to

$$\begin{aligned} p_\theta(y_1, y_j|x_1) &= \int g_\theta(y_1|x_1) \prod_{k=2}^{j} f_\theta(x_k|x_{k-1}) g_\theta(y_j|x_j) dx_{2:j} \\ &\leq \overline{g_0}^2 \int p_\theta(x_{2:j}|x_1) dx_{2:j} = \overline{g_0}^2. \end{aligned}$$

In the same way, we have that

$$p_\theta(y_1, y_j|x_1) \geq \underline{g_0}^2.$$

□

Theorem B.3.2. *For* $j = 2, \dots, L+1$ *and for any* $0 \leq k < j, \theta \in \Theta, x_{k+1}, y_j \in \mathcal{X} \times \mathcal{Y}$

$$\underline{g_0} \leq p_\theta(y_j | x_{k+1}) \leq \overline{g_0}$$

Proof. The result obviously holds if $k = j-1$. If $k < j-1$, from Assumptions (A2), simple calculations yield to

$$\begin{aligned} p_\theta(y_j|x_{k+1}) &= \int \prod_{l=k+2}^{j} f_\theta(x_l|x_{l-1}) g_\theta(y_j|x_j) dx_{k+2:j} \\ &\leq \overline{g_0} \int p_\theta(x_{k+2:j}|x_{k+1}) dx_{k+2:j} = \overline{g_0}. \end{aligned}$$

In the same way, we have that

$$p_\theta(y_j|x_{k+1}) \geq \underline{g_0}.$$

□

Theorem B.3.3. *For $j = 2, \dots, L+1$ and for any $\theta \in \Theta, y_1, y_j \in \times\mathcal{Y}^2, \mu \in \mathcal{P}(\mathcal{X})$*

$$\underline{g_0}^2 \le p_\theta(y_1, y_j|\mu) \le \overline{g_0^2}$$

Proof. From Assumptions (A2), simple calculations yield to

$$\begin{aligned} p_\theta(y_1, y_j|\mu) &= \int \mu(x_1) g_\theta(y_1|x_1) \prod_{k=2}^{j} f_\theta(x_k|x_{k-1}) g_\theta(y_j|x_j) dx_{1:j} \\ &\le \overline{g_0}^2 \int p_\theta(x_{1:j}|\mu) dx_{1:j} = \overline{g_0}^2. \end{aligned}$$

In the same way, we have that

$$p_\theta(y_1, y_j|\mu) \ge \underline{g_0}^2.$$

□

Theorem B.3.4. *For $j = 2, \dots, L+1$ and for any $\theta \in \Theta, y_1, y_j \in \times\mathcal{Y}^2, \mu \in \mathcal{P}(\mathcal{X})$*

$$\frac{\underline{g_0}^2}{\overline{g_0}} \le p_\theta(y_j|y_1, \mu) \le \frac{\overline{g_0}^2}{\underline{g_0}}$$

Proof. From Assumptions (A2) and Theorem B.3.3, simple calculations yield to

$$\begin{aligned} p_\theta(y_j|y_1, \mu) &= \frac{p_\theta(y_1, y_j|\mu)}{p_\theta(y_1|\mu)} \\ &\le \frac{\overline{g_0}^2}{\int \mu(x_1) g_\theta(y_1|x_1) dx_1} \le \frac{\overline{g_0}^2}{\underline{g_0}}. \end{aligned}$$

In the same way, we have that

$$p_\theta(y_j|y_1, \mu) \ge \frac{\underline{g_0}^2}{\overline{g_0}}.$$

□

Theorem B.3.5. *There exists a constant $C \in (0, +\infty)$ such that for $j = 2, \dots, L+1$*

and for any $\theta \in \Theta, y_1, y_j \in \times\mathcal{Y}^2, \mu, \nu \in \mathcal{P}(\mathcal{X})$

$$|p_\theta(y_1, y_j|\mu) - p_\theta(y_1, y_j|\nu)| \leq C\|\mu - \nu\|.$$

Proof. From Theorem B.3.1, simple calculations yield to

$$\begin{aligned} |p_\theta(y_1, y_j|\mu) - p_\theta(y_1, y_j|\nu)| &\leq \int |\mu(x_1) - \nu(x_1)| p_\theta(y_1, y_j|x_1) dx_1 \\ &\leq \overline{g_0}^2 \|\mu - \nu\|. \end{aligned}$$

□

Theorem B.3.6. *For every* $j = 2, \ldots, L+1, \mu \in \mathcal{P}(\mathcal{X})$*, the following identity holds*

$$\nabla \log p_\theta(y_1, y_j|\mu) = \mathbb{E}_{\theta^*}\left[\nabla \log p_\theta(y_1, y_j, x_{1:j}|\mu)|Y_1, Y_j, \mu\right].$$

Proof. The result follows from an application of the Fisher's identity. Under regularity assumptions

$$\begin{aligned} \nabla \log p_\theta(y_1, y_j|\mu) =& \frac{\nabla p_\theta(y_1, y_j|\mu)}{p_\theta(y_1, y_j|\mu)} = \frac{1}{p_\theta(y_1, y_j|\mu)} \nabla \int_{\mathcal{X}^j} p_\theta(y_1, y_j, x_{1:j}|\mu) dx_{1:j} \\ =& \frac{1}{p_\theta(y_1, y_j|\mu)} \int_{\mathcal{X}^j} \nabla p_\theta(y_1, y_j, x_{1:j}|\mu) dx_{1:j} \\ =& \int_{\mathcal{X}^j} \frac{\nabla p_\theta(y_1, y_j, x_{1:j}|\mu)}{p_\theta(y_1, y_j|\mu)} dx_{1:j} \\ =& \int_{\mathcal{X}^j} \frac{\nabla p_\theta(y_1, y_j, x_{1:j}|\mu)}{p_\theta(y_1, y_j, x_{1:j}|\mu)} p_\theta(x_{1:j}|y_1, y_j, \mu) dx_{1:j} \\ =& \int_{\mathcal{X}^j} \nabla \log p_\theta(y_1, y_j, x_{1:j}|\mu) p_\theta(x_{1:j}|y_1, y_j, \mu) dx_{1:j} \\ =& \mathbb{E}_{\theta^*}[\nabla \log p_\theta(y_1, y_j, x_{1:j}|\mu)|Y_1, Y_j, \mu]. \end{aligned}$$

□

Theorem B.3.7. *There exists a constant* $C \in (0, +\infty)$ *and* $\rho \in [0, 1)$ *such that for every* $\mu, \nu \in \mathcal{P}(\mathcal{X})$*,* $\theta \in \Theta$*,* $y_1, y_j \in \mathcal{Y}^2$ *and* $j = 2, \ldots, L+1, k \leq j$*,*

$$\|p_\theta(X_k \in \cdot|y_1, y_j, \mu) - p_\theta(X_k \in \cdot|y_1, y_j, \nu)\| \leq C\|\mu - \nu\|\rho^{k-1}.$$

Proof. In order to prove the theorem, some intermediate results are needed.

(i) By definition of total variation norm and under Assumptions (A2)

$$\begin{aligned}
&||p_\theta(X_1 \in \cdot|y_1,\mu) - p_\theta(X_1 \in \cdot|y_1,\nu)||\\
=&\sup_A \left| \int_A p_\theta(x_1|y_1,\mu)dx_1 - \int_A p_\theta(x_1|y_1,\nu)dx_1 \right|\\
=&\sup_A \left| \int_A [\mu(x_1) - \nu(x_1)] g_\theta(y_1|x_1)dx_1 \right|\\
\le&\overline{g_0} \quad \sup_A \left| \int_A [\mu(x_1) - \nu(x_1)]dx_1 \right| = \overline{g_0} \quad ||\mu - \nu||. \qquad \text{(B.3.1)}
\end{aligned}$$

(ii) By definition of total variation norm and under Assumptions (A2)

$$\begin{aligned}
&||p_\theta(X_k \in \cdot|y_1,\mu) - p_\theta(X_k \in \cdot|y_1,\nu)||\\
=&\sup_A \left| \int_A p_\theta(x_k|y_1,\mu)dx_k - \int_A p_\theta(x_k|y_1,\nu)dx_k \right|\\
=&\sup_A \left| \int_A \int [p_\theta(x_k|x_1,y_1,\mu)p_\theta(x_1|y_1,\mu) - p_\theta(x_k|x_1,y_1,\nu)p_\theta(x_1|y_1,\nu)]dx_1dx_k \right|\\
=&\sup_A \left| \int_A \int p_\theta(x_k|x_1)[p_\theta(x_1|y_1,\mu) - p_\theta(x_1|y_1,\nu)]dx_1dx_k \right|.
\end{aligned}$$

Using the notation and the results of Theorem B.1.1 and Corollary B.1.2

$$\begin{aligned}
&||p_\theta(X_k \in \cdot|y_1,\mu) - p_\theta(X_k \in \cdot|y_1,\nu)||\\
=&\sup_A \left| \int P^{k-1}(x,A)[p_\theta(x_1|y_1,\mu) - p_\theta(x_1|y_1,\nu)]dx_1 \right|\\
\le&(1-\underline{f_0})^{k-1}||p_\theta(X_1 \in \cdot|y_1,\mu) - p_\theta(X_1 \in \cdot|y_1,\nu)||\\
\le&\overline{g_0}(1-\underline{f_0})^{k-1}||\mu - \nu||, \qquad \text{(B.3.2)}
\end{aligned}$$

where the last inequality follows by (B.3.1).

(iii) From Assumptions (A2), Theorem B.3.2 and Equation (B.3.2)

$$|p_\theta(y_j|y_1,\mu) - p_\theta(y_j|y_1,\nu)|$$

$$\begin{aligned}
&= \left| \int p_\theta(y_j|x_{k+1}) f_\theta(x_{k+1}|x_k)[p_\theta(x_k|y_1,\mu) - p_\theta(x_k|y_1,\nu)] dx_k dx_{k+1} \right| \\
&\leq \overline{f_0}\overline{g_0} || p_\theta(X_k \in \cdot|y_1,\mu) - p_\theta(X_k \in \cdot|y_1,\nu)|| \\
&\leq C(1-\underline{f_0})^{k-1}||\mu - \nu||, \qquad \text{(B.3.3)}
\end{aligned}$$

where C is a suitable constant in $(0, +\infty)$.

Let us go back to the main theorem. If $k < j$ we can write

$$\begin{aligned}
p_\theta(x_k|y_1,y_j,\mu) &= \int p_\theta(x_k, x_{k+1}|y_1,y_j,\mu) dx_{k+1} \\
&= \int \frac{p_\theta(y_j|x_k, x_{k+1}, y_1) p_\theta(x_k, x_{k+1}|y_1,\mu)}{p_\theta(y_j|y_1,\mu)} dx_{k+1} \\
&= \int \frac{p_\theta(y_j|x_{k+1}) f_\theta(x_{k+1}|x_k) p_\theta(x_k|y_1,\mu)}{p_\theta(y_j|y_1,\mu)} dx_{k+1},
\end{aligned}$$

and hence

$$\begin{aligned}
&p_\theta(x_k|y_1,y_j,\mu) - p_\theta(x_k|y_1,y_j,\nu) \\
&= \int p_\theta(y_j|x_{k+1}) f_\theta(x_{k+1}|x_k) \left[\frac{p_\theta(x_k|y_1,\mu)}{p_\theta(y_j|y_1,\mu)} - \frac{p_\theta(x_k|y_1,\nu)}{p_\theta(y_j|y_1,\nu)} \right] dx_{k+1}.
\end{aligned}$$

The term in square brackets can be written as

$$\begin{aligned}
&\frac{p_\theta(x_k|y_1,\mu)}{p_\theta(y_j|y_1,\mu)} - \frac{p_\theta(x_k|y_1,\nu)}{p_\theta(y_j|y_1,\nu)} \\
&\qquad = \frac{(p_\theta(x_k|y_1,\mu) - p_\theta(x_k|y_1,\nu)) p_\theta(y_j|y_1,\nu)}{p_\theta(y_j|y_1,\mu) p_\theta(y_j|y_1,\nu)} \\
&\qquad\quad - \frac{p_\theta(x_k|y_1,\nu)(p_\theta(y_j|y_1,\mu) - p_\theta(y_j|y_1,\nu))}{p_\theta(y_j|y_1,\mu) p_\theta(y_j|y_1,\nu)}.
\end{aligned}$$

Under Assumptions (A2) and using the results in Theorems B.3.2, B.3.4 and Equations (B.3.2, B.3.3)

$$||p_\theta(X_k \in \cdot|y_1,y_j,\mu) - p_\theta(X_k \in \cdot|y_1,y_j,\nu)|| \leq C(1-\underline{f_0})^{k-1}||\mu - \nu||,$$

where we have used the fact that, for every k, $p_\theta(x_k|y_1,\mu)$ is bounded for any $\theta, x_k, y_1 \in \Theta \times \mathcal{X} \times \mathcal{Y}$ and for any $\mu \in \mathcal{P}(\mathcal{X})$.

If $k = j$, the reasoning is almost the same, starting from the following decomposition

$$\begin{aligned} p_\theta(x_j|y_1,y_j,\mu) &= \frac{p_\theta(y_j|x_j,y_1)p_\theta(x_j|y_1,\mu)}{p_\theta(y_j|y_1,\mu)} \\ &= \frac{g_\theta(y_j|x_j)p_\theta(x_j|y_1,\mu)}{p_\theta(y_j|y_1,\mu)}. \end{aligned}$$

□

B.4 Technical results on the convergence of $l_P^{(L)}(\theta,\mu)$ and its derivative

In this section we prove some uniform convergence results for $l_P^{(L)}(\theta,\mu)$ and its derivative. Hereafter, for simplicity, we drop the L index in $l_P^{(L)}(\cdot,\cdot) := l_P(\cdot,\cdot)$.
The first result states that $l_P(\theta,\mu)$ converges uniformly in θ to $l_P(\theta,\nu)$ as the total variation distance between μ and ν tends to zero (even if μ,ν can depend on θ, we omit the explicit dependence for notational convenience).

Theorem B.4.1. *There exists a constant* $C \in (0,+\infty)$ *such that for any* $\mu,\nu \in \mathcal{P}(\mathcal{X})$, $\theta \in \Theta$ *and* $L \geq 1$

$$|l_P(\theta,\mu) - l_P(\theta,\nu)| \leq C\|\mu-\nu\|.$$

Proof. By definition

$$l_P(\theta,\mu) - l_P(\theta,\nu) = \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\log p_\theta(y_1,y_j|\mu) - \log p_\theta(y_1,y_j|\nu)\right]$$

and using the following identity valid for any $x,y \in (0,+\infty)$,

$$|\log x - \log y| \leq \frac{|x-y|}{x \wedge y}, \tag{B.4.1}$$

we have

$$\begin{aligned}|l_P(\theta,\mu) - l_P(\theta,\nu)| &\leq \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\frac{|p_\theta(y_1,y_j|\mu) - p_\theta(y_1,y_j|\nu)|}{p_\theta(y_1,y_j|\mu)\wedge p_\theta(y_1,y_j|\nu)}\right]\\ &\leq \frac{1}{L}\sum_{j=2}^{L+1} C\|\mu-\nu\| = C\|\mu-\nu\|,\end{aligned}$$

where we have used the results in Theorems B.3.3 and B.3.5. □

Now we look at the derivative of $l_P(\theta,\mu)$. For every $\mu,\nu\in\mathcal{P}(\mathcal{X})$ the difference of the gradient of two approximated pairwise likelihood of order L is

$$\nabla l_P(\theta,\mu) - \nabla l_P(\theta,\nu) = \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\nabla\log p_\theta(y_1,y_j|\mu) - \nabla\log p_\theta(y_1,y_j|\nu)\right] \quad \text{(B.4.2)}$$

and

$$\begin{aligned}\nabla\log p_\theta(y_1,y_j,x_{1:j}|\mu) =& \nabla\log\mu(x_1) + \nabla\log g_\theta(y_1|x_1) + \\ &+\sum_{k=2}^{j}\nabla\log f_\theta(x_k|x_{k-1}) + \nabla\log g_\theta(y_j|x_j).\end{aligned}$$

We prove the following result.

Theorem B.4.2. *There exists a constant $C\in(0,+\infty)$ and $\rho\in[0,1)$ such that for every $\mu,\nu\in\mathcal{P}(\mathcal{X})$, $\theta\in\Theta$, $L\geq 1$,*

$$|\nabla l_P(\theta,\mu) - \nabla l_P(\theta,\nu)| \leq C\left[\frac{\|\mu-\nu\|}{1-\rho} + \|\nabla\mu - \nabla\nu\|\right].$$

Proof. Let us analyze the generic term of the sum (B.4.2). By Theorem B.3.6

$$\begin{aligned}&\nabla\log p_\theta(y_1,y_j|\mu) - \nabla\log p_\theta(y_1,y_j|\nu)\\ =&\mathbb{E}_{\theta^*}[\nabla\log p_\theta(y_1,y_j,x_{1:j}|\mu)|Y_1,Y_j,\mu] - \mathbb{E}_{\theta^*}[\nabla\log p_\theta(y_1,y_j,x_{1:j}|\nu)|Y_1,Y_j,\nu]\\ =&\int\nabla\log p_\theta(y_1,y_j,x_{1:j}|\mu)p_\theta(x_{1:j}|y_1,y_j,\mu)dx_{1:j}+\\ &-\int\nabla\log p_\theta(y_1,y_j,x_{1:j}|\nu)p_\theta(x_{1:j}|y_1,y_j,\nu)dx_{1:j}\end{aligned}$$

$$
\begin{aligned}
&= \int \nabla \log g_\theta(y_1|x_1)\left(p_\theta(x_{1:j}|y_1,y_j,\mu) - p_\theta(x_{1:j}|y_1,y_j,\nu)\right)dx_{1:j}\\
&\quad + \int \nabla \log g_\theta(y_j|x_j)\left(p_\theta(x_{1:j}|y_1,y_j,\mu) - p_\theta(x_{1:j}|y_1,y_j,\nu)\right)dx_{1:j}\\
&\quad + \sum_{k=2}^{j}\int \nabla \log f_\theta(x_k|x_{k-1})\left(p_\theta(x_{1:j}|y_1,y_j,\mu) - p_\theta(x_{1:j}|y_1,y_j,\nu)\right)dx_{1:j}\\
&\quad + \left[\int \nabla \log \mu(x_1)p_\theta(x_{1:j}|y_1,y_j,\mu)dx_{1:j} - \int \nabla \log \nu(x_1)p_\theta(x_{1:j}|y_1,y_j,\nu)dx_{1:j}\right]\\
&:= T_1 + T_2 + T_3 + T_4.
\end{aligned}
$$

We study the terms T_1, T_2, T_3, T_4 separately. Let us start with T_1.

$$
\begin{aligned}
T_1 &:= \int \nabla \log g_\theta(y_1|x_1)\left(p_\theta(x_{1:j}|y_1,y_j,\mu) - p_\theta(x_{1:j}|y_1,y_j,\nu)\right)dx_{1:j}\\
&= \int \nabla \log g_\theta(y_1|x_1)\left(p_\theta(x_1,x_{2:j}|y_1,y_j,\mu) - p_\theta(x_1,x_{2:j}|y_1,y_j,\nu)\right)dx_{1:j}\\
&= \int \nabla \log g_\theta(y_1|x_1)\left[\int p_\theta(x_1,x_{2:j}|y_1,y_j,\mu)dx_{2:j}\right]dx_1\\
&\quad - \int \nabla \log g_\theta(y_1|x_1)\left[\int p_\theta(x_1,x_{2:j}|y_1,y_j,\nu)dx_{2:j}\right]dx_1\\
&= \int \nabla \log g_\theta(y_1|x_1)\left[p_\theta(x_1|y_1,y_j,\mu) - p_\theta(x_1|y_1,y_j,\nu)\right]dx_1.
\end{aligned}
$$

By Theorem B.3.7 and Assumptions $(A2)$,

$$
\begin{aligned}
|T_1| &\leq \sup_{x_1}|\nabla \log g_\theta(y_1|x_1)|\cdot\|p_\theta(X_1\in\cdot|y_1,y_j,\mu) - p_\theta(X_1\in\cdot|y_1,y_j,\nu)\|\\
&\leq C\|\mu-\nu\|. \qquad \text{(B.4.3)}
\end{aligned}
$$

Analogous calculations for T_2 yield to

$$
|T_2| \leq C\rho^{j-1}\|\mu-\nu\|. \qquad \text{(B.4.4)}
$$

Now, for every $k = 2,\ldots,j$

$$
\int \nabla \log f_\theta(x_k|x_{k-1})\left(p_\theta(x_{1:j}|y_1,y_j,\mu) - p_\theta(x_{1:j}|y_1,y_j,\nu)\right)dx_{1:j}
$$

$$
\begin{aligned}
&= \int \Big(p_\theta(x_{1:k-2}, x_{k-1:k}, x_{k+1:j}|y_1, y_j, \mu) - p_\theta(x_{1:k-2}, x_{k-1:k}, x_{k+1:j}|y_1, y_j, \nu)\Big) \\
&\qquad \cdot \nabla \log f_\theta(x_k|x_{k-1}) dx_{1:j} \\
&= \int \nabla \log f_\theta(x_k|x_{k-1}) \Big[p_\theta(x_{k-1}, x_k|y_1, y_j, \mu) - p_\theta(x_{k-1}, x_k|y_1, y_j, \nu)\Big] dx_{k-1:k} \\
&= \int \Big[p_\theta(x_k|x_{k-1}, y_1, y_j, \mu) p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_k|x_{k-1}, y_1, y_j, \nu) p_\theta(x_{k-1}|y_1, y_j, \nu)\Big] \\
&\qquad \cdot \nabla \log f_\theta(x_k|x_{k-1}) dx_{k-1:k} \\
&= \int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) \Big[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\Big] dx_{k-1:k} \\
&= \int \left[\int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k\right] \\
&\qquad \cdot \Big[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\Big] dx_{k-1} \\
&= \int \Psi(x_{k-1}) \Big[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\Big] dx_{k-1},
\end{aligned}
$$

where $\Psi(x_{k-1}) := \int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k$. Moreover,

$$
\sup_{x_{k-1}} |\Psi(x_{k-1})| \leq \overline{f_1} \int p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k = \overline{f_1}. \tag{B.4.5}
$$

By Theorem B.3.7 and Equation (B.4.5), we have that

$$
|T_3| \leq \sum_{k=2}^{j} C\rho^{k-2} \|\mu - \nu\| \leq \frac{C\|\mu - \nu\|}{1 - \rho}. \tag{B.4.6}
$$

The last term in the sum can be written as

$$
\begin{aligned}
&\int \nabla \log \mu(x_1) p_\theta(x_{1:j}|y_1, y_j, \mu) dx_{1:j} - \int \nabla \log \nu(x_1) p_\theta(x_{1:j}|y_1, y_j, \nu) dx_{1:j} \\
&= \int \nabla \log \mu(x_1) p_\theta(x_1, x_{2:j}|y_1, y_j, \mu) dx_{1:j} - \int \nabla \log \nu(x_1) p_\theta(x_1 x_{2:j}|y_1, y_j, \nu) dx_{1:j} \\
&= \int \nabla \log \mu(x_1) p_\theta(x_1|y_1, y_j, \mu) dx_1 - \int \nabla \log \nu(x_1) p_\theta(x_1|y_1, y_j, \nu) dx_1 \\
&= \int \frac{\nabla \mu(x_1)}{\mu(x_1)} \frac{p_\theta(x_1, y_1, y_j|\mu)}{p_\theta(y_1, y_j|\mu)} dx_1 - \int \frac{\nabla \nu(x_1)}{\nu(x_1)} \frac{p_\theta(x_1, y_1, y_j|\nu)}{p_\theta(y_1, y_j|\nu)} dx_1 \\
&= \int \nabla \mu(x_1) \frac{p_\theta(y_1, y_j|x_1)}{p_\theta(y_1, y_j|\mu)} dx_1 - \int \nabla \nu(x_1) \frac{p_\theta(y_1, y_j|x_1)}{p_\theta(y_1, y_j|\nu)} dx_1
\end{aligned}
$$

$$= \int p_\theta(y_1, y_j|x_1)\left[\frac{\nabla\mu(x_1)}{p_\theta(y_1, y_j|\mu)} - \frac{\nabla\nu(x_1)}{p_\theta(y_1, y_j|\nu)}\right]dx_1.$$

From Theorem B.3.1, $p_\theta(y_1, y_j|x_1)$ is a bounded function of x_1 and the term in square brackets can be written as

$$\frac{\nabla\mu(x_1)}{p_\theta(y_1, y_j|\mu)} - \frac{\nabla\nu(x_1)}{p_\theta(y_1, y_j|\nu)} = \frac{\nabla\mu(x_1)p_\theta(y_1, y_j|\nu) - \nabla\nu(x_1)p_\theta(y_1, y_j|\mu)}{p_\theta(y_1, y_j|\mu)p_\theta(y_1, y_j|\nu)}$$
$$= \frac{(\nabla\mu(x_1) - \nabla\nu(x_1))p_\theta(y_1, y_j|\nu) - \nabla\nu(x_1)(p_\theta(y_1, y_j|\mu) - p_\theta(y_1, y_j|\nu))}{p_\theta(y_1, y_j|\mu)p_\theta(y_1, y_j|\nu)}.$$

Using Assumptions (A2) and Theorems B.3.3 and B.3.5, we have that

$$|T_4| \leq C(||\nabla\mu - \nabla\nu|| + ||\mu - \nu||). \qquad \text{(B.4.7)}$$

From the results (B.4.3, B.4.4, B.4.6, B.4.7), we conclude that

$$\begin{aligned} |\nabla l_P(\theta, \mu) - \nabla l_P(\theta, \nu)| &\leq \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta^*}|\nabla \log p_\theta(y_1, y_j|\mu) - \nabla \log p_\theta(y_1, y_j|\nu)| \\ &\leq C\left[\frac{||\mu - \nu||}{1 - \rho} + ||\nabla\mu - \nabla\nu||\right]. \end{aligned}$$

□

B.5 Bias when π_θ is replaced with μ

Let us define the set

$$\hat{\theta}_P(\mu) := \arg\max_{\theta\in\Theta} l_P(\theta, \mu),$$

where, as usual

$$l_P(\theta, \mu) = \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta^*}[\log p_\theta(y_1, y_j|\mu)]$$

and $l_P(\theta, \pi_\theta) := l_P(\theta)$, being π_θ the unique stationary distribution. The set $\hat{\theta}_P(\mu)$ is not empty since Θ is compact and $l_P(\theta, \mu)$ is continuous from Assumptions (A2) whenever μ is continuous. For any $\epsilon \in (0, +\infty)$ and $\theta_0 \in \Theta$, let $B(\theta_0, \epsilon) = \{\theta \in \Theta : |\theta - \theta_0| \leq \epsilon\}$ and for any set $A \in \Theta$ let $d(\theta_0, A) = \inf\{|\theta - \theta_0| : \theta \in A\}$ the distance between θ_0 and the set A. Theorem 3.3.1 quantifies the bias when the (unknown) invariant distribution π_θ is replaced with a generic μ, that is the bias of the estimate introduced by maximizing $l_P(\theta, \mu)$ instead of $l_P(\theta)$. The result says that the bias depends on how close μ_{θ^*} is to π_{θ^*} and on the ergodicity properties of $\{X_k\}$, where θ^* denotes the true parameter. We prove first the following statement.

Theorem B.5.1. *Assume (A1). Then for any sequence of measures* $\{\mu_k, k \geq 1\}$ *with uniformly continuous (in* θ*) density such that* $||\mu_k - \pi_\theta||$ *goes to zero and for any* $\epsilon > 0$ *such that* $B(\theta^*, \epsilon) \subset \overset{\circ}{\Theta}$*, there exists* $\underline{k}$ *such that* $\forall k \geq \underline{k}$*,* $l_P(\theta, \mu_k)$ *has its maxima* $\hat{\theta}(\mu_k)$ *in* $B(\theta^*, \epsilon)$ *and*

$$\lim_{||\mu_k - \pi_\theta|| \to 0} d(\theta^*, \hat{\theta}(\mu_k)) = 0. \tag{B.5.1}$$

Proof. Let ϵ be a strictly positive constant. We proceed by contradiction. Assume there exists a sequence of measures $\{\mu_k, k \geq 1\}$ with uniformly continuous (in θ) density such that $||\mu_k - \pi_\theta||$ goes to zero and $\hat{\theta}(\mu_k) \notin B(\theta^*, \epsilon)$. This means that the estimates obtained by maximizing $l_P(\theta, \mu_k)$ with respect to θ are far from the true parameter value. Hence $|\hat{\theta}(\mu_k) - \theta^*| > \epsilon \geq 0$. By definition of $\hat{\theta}(\mu_k)$, we have that

$$l_P(\theta^*, \mu_k) \leq l_P(\hat{\theta}(\mu_k), \mu_k).$$

Since $\{\hat{\theta}(\mu_k)\} \subset \Theta$, and Θ is bounded, it has at least an accumulation point $\tilde{\theta}^*$ corresponding to a subsequence of $\{\hat{\theta}(\mu_k)\}$. From Theorem B.4.1, $l_P(\theta, \mu_k)$ converges uniformly to $l_P(\theta)$ as $||\mu_k - \pi_\theta||$ goes to zero and consequently

$$l_P(\tilde{\theta}^*) \geq l_P(\theta^*)$$

with $|\tilde{\theta}^* - \theta^*| > \epsilon \geq 0$. This contradicts the fact that θ^* is the unique strong maximum of $l_P(\theta)$. Equation (B.5.1) obviously holds. □

All these results allow us to prove Theorem 3.3.1.

Bibliography

C. Andrieu and A. Doucet. Iterative algorithm for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.*, 49:1216–27, 2001.

C. Andrieu, J. F. G. De Freitas, and A. Doucet. Sequential MCMC for bayesian model selection. In *Proceedings IEEE Workshop Higher Order Statistics*, 1999.

C. Andrieu, A. Doucet, and V. B. Tadic. On-line parameter estimation in general state-space models using pseudo-likelihood. 2007. URL `http://www.maths.bris.ac.uk/ maxca/preprints/andrieu_doucet_tadic_2007`.

A. Azzalini. Maximum likelihood of order m for stationary stochastic processes. *Biometrika*, 70:367–81, 1983.

Y. Bar-Shalom and X. R. Li. *Multitarget- Multisensor Tracking: principles and techniques*. Storrs, CT: University of Connecticut Press, 1995.

A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation*. Springer-Verlag, New York, 1990.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. Ser. B*, 36:192–236, 1974.

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 64:616–18, 1977.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *Ann. Statist.*, 32:2385–411, 2004.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–37, 2004.

P. Del Moral. *Feyman- Kac formulae. Genealogical and interacting particle approximations*. Probability and Applications. Springer, New York, 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, 39(1):1–38, 1977.

R. Douc, E. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32:2254–304, 2004.

A. Doucet and P. Duvaut. Bayesian estimation of state space models applied to deconvolution of bernoulli- gaussian processes. *Signal Process.*, 57:147–61, 1997.

A. Doucet, J. F. G. de Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer- Verlag, New York, 2001.

P. Fernhead. MCMC, sufficient statistics and particle filter. *J. Comput. Graph. Statist.*, 11:848–62, 2002.

W. R. Gilks and C. Berzuini. Following a moving target- Monte Carlo inference for dynamic bayesian models. *J. R. Stat. Soc. Ser. B*, 63:127–46, 2001.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc. F*, 140:107–13, 1993.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–84, 1989.

G. Kitagawa. A self- organizing state- space model. *J. Amer. Statist. Assoc.*, 93: 1203–15, 1998.

G. Kitagawa and S. Sato. Monte Carlo smoothing and self- organizing state space model. In *Sequential Monte Carlo Methods in Practice*. 2001.

V Krishnamurthy and A. Logothetis. A bayesian expectation- maximization framework for estimating jump Markov linear system. *IEEE Trans. Signal Process.*, 47:2139–56, 1999.

M. Lavielle. Bayesian deconvolution of bernoulli- gaussian processes. *Signal Process.*, 33:67–79, 1993.

S. Le Cessie and J. C. Van Houwelingen. Logistic regression for correlated binary data. *J. R. Stat. Soc. Ser. C*, 43:95–108, 1994.

G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.*, 40:127–43, 1992.

B. G. Lindsay. Composite likelihood methods. *Contemp. Math.*, 80:221–39, 1988.

J. Liu and M. West. Combining parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*. 2001.

D. Q. Mayne. A solution of the smoothing problem for linear dynamic systems. *Automatica*, 4:73–92, 1966.

J. M. Mendel. *Maximum- Likelihood Deconvolution: a journey into model- based signal processing*. Springer- Verlag, New York, 1990.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26(2), 1984.

T. Ryden. Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.*, 22:1841–95, 1994.

S.L. Sclove. Time series segmentation: a model and a method. *Econometrica*, 29: 7–25, 1983.

R. Shumway and D. Stoffer. *Time Series Analysis and its Applications*. Springer-Verlag, New York, 2000.

G. Storvik. Particle filters in state space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.*, 50:281–89, 2002.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92:519–28, 2005.

C. Varin and P. Vidoni. Pairwise likelihood inference for general state space models. *Econometric Rev.*, 28:170–85, 2009.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 29:595–601, 1949.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95103, 1983.

L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for gaussian mixtures. *Neural Comput.*, 8:129–51, 1996.

Nadia Frigo - Curriculum Vitae

Personal Details

Date of Birth:	December 07, 1982
Nationality:	Italian

Contact Information

Address: Department of Statistics, University of Padova,
Via Cesare Battisti, 241-243 - 35121 Padova (Italy)
Phone: +39 049 827 4111
E-mail: nadia@stat.unipd.it

Current Position

Since January 2007 (admitted to final exam on December 10, 2009)
PhD in Statistical Sciences, University of Padova
Actually I am working on composite likelihood inference in state space models, under the supervision of Prof. Marco Ferrante.

Education

October 2006
Laurea Specialistica in Economic, Financial and Business Statistics, University of Padova (*"master's degree"*)
Thesis title: "Metodi Monte Carlo Sequenziali per il filtraggio stocastico - un modello per media e varianza stocastica" (in italian)
Supervisor: Prof. Marco Ferrante. Final mark: Summa cum laude

July 2004
Laurea Triennale in Mathematics, University of Padova (*"bachelor's degree"*)
Thesis title: "Statistica bayesiana e metodo Markov Chain Monte Carlo" (in italian)
Supervisor: Prof. Wolfgang J. Runggaldier. Final mark: 110/110

June 2001
Diploma Maturità Scientifica, Liceo Scientifico Statale P. Lioy, Vicenza
Final mark: 100/100

Research Periods Abroad

June 27 - July 04, 2009
Department of Mathematics, University of Bristol, Bristol, UK
Research Collaboration with Prof. Christophe Andrieu on the topic *'Composite likelihood inference in state space models'*.

March 22 - April 03, 2009
Department of Mathematics, University of Bristol, Bristol, UK

Research Collaboration with Prof. Christophe Andrieu on the topic *'Composite likelihood inference in state space models'*.

September - December 2008
Department of Mathematics, University of Bristol, Bristol, UK
Visiting Ph.D. student under the supervision of Prof. Christophe Andrieu.
Research topic : *'Asymptotic properties of the maximum pairwise likelihood estimator in general state space models'*.

February - June 2008
Department of Computer Science, University of British Columbia (UBC), Vancouver, Canada
Visiting Ph.D. student under the supervision of Prof. Arnaud Doucet.
Research topic : *'Exact Gibbs sampler for Markovian Arrival Processes (MAP)'*

Presentations and Seminars

Fleurance, France, June 25 - 29, 2007
Workshop: New directions in Monte Carlo Methods
Particle filtering approximations for a Gaussian-generalized inverse Gaussian model, with M. Ferrante

Teaching Activities

October 2009- January 2010
Facoltà di Ingegneria Chimica, Università degli Studi di Padova, Italy
Teaching assistant of the course *Modelli statistici e probabilistici per l'industria di processo (Course instructor Prof. E. Gola)*

February - March 2009
Facoltà di Economia Università Ca' Foscari, Venice, Italy
Teaching assistant of the course *Statistica I: Rilevazione, classificazione e analisi dei dati (Course instructor Prof. C. Agostinelli)*

April - May 2009
Facoltà di Economia Università Ca' Foscari, Venice, Italy
Teaching assistant of the course *Statistica II: Rilevazione, classificazione e analisi dei dati (Course instructor Prof. C. Agostinelli)*

Publications

Ferrante,M., Frigo, N., Particle filtering approximations for a Gaussian-generalized inverse Gaussian model. *Statistics and Probability Letters, vol. 79, pp. 442-449* (2009) doi:10.1016/j.spl.2008.09.017