Stony Brook University

OFFICIAL COPY

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

© All Rights Reserved by Author.

Sequential Monte Carlo Methods for Inference and Prediction of Latent Time-series

A Dissertation presented
by

Iñigo Urteaga

to
The Graduate School
in Partial Fulfillment of the
Requirements
for the Degree of

Doctor of Philosophy
in
Electrical Engineering

Stony Brook University

August 2016

Stony Brook University

The Graduate School

Iñigo Urteaga

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree,

hereby recommend acceptance of this dissertation.

Petar M. Djurić, Dissertation Advisor
Professor, Department of Electrical and Computer Engineering

Mónica F. Bugallo, Chairperson of Defense
Associate Professor, Department of Electrical and Computer Engineering

Yue Zhao
Assistant Professor, Department of Electrical and Computer Engineering

Svetlozar Rachev
Professor, Department of Applied Math and Statistics

This dissertation is accepted by the Graduate School.

Nancy Goroff

Interim Dean of the Graduate School

Abstract of the Dissertation

Sequential Monte Carlo Methods for Inference and Prediction of Latent Time-series

by

Iñigo Urteaga

Doctor of Philosophy
in
Electrical Engineering

Stony Brook University

2016

In the era of information-sensing mobile devices, the Internet-of-Things and Big Data, research on advanced methods for extracting information from data has become extremely critical. One important task in this area of work is the analysis of time-varying phenomena, observed sequentially in time. This endeavor is relevant in many applications, where the goal is to infer the dynamics of events of interest described by the data, as soon as new data-samples are acquired.

This dissertation is on novel methods for sequential inference and prediction of latent time-series. We assume that a sequence of observations is a function of a hidden process of interest and the goal is to estimate the latent process from the observed data. We consider flexible models that capture the dynamics of real phenomena and can deal with many practical burdens.

The embraced methodology is based on Bayesian theory and Monte Carlo algorithms. The former provides a consistent framework for the analysis of latent random processes. The latter allows for overcoming the inherent difficulties of nonlinear and non-Gaussian models. The goal is to propose methods that can extract the hidden dynamics from observed data in the most generic and challenging scenarios.

We start by investigating short-memory processes, that is, time-series where most of the relevant information is contained only within the most recent past. In particular, we study latent Auto-Regressive Moving-Average (ARMA) processes with independent Gaussian innovations. We first assume that the parameters are known and then, we relax the assumptions until they are all unknown.

The analysis of latent time-series is extended to processes with different memory characteristics, including those with

long-memory features. On the one hand, we investigate latent processes that show self-similarity properties and are correlated in time, such as the fractional Gaussian process (fGp). On the other, we study extensions of the ARMA(p, q) model, by considering fractional differencing, which leads to Fractional Auto-Regressive Integrated Moving-Average (FARIMA) processes.

We further generalize our work to allow for broad memory properties, by relaxing previous modeling and parameterization assumptions. We resort to wide-sense stationary (WSS) time-series in general. Within this new framework, all the previously considered models are covered. As a result, a generic Sequential Monte Carlo (SMC) method for inference of Gaussian WSS latent time-series is proposed. We broaden our work by investigating a hierarchical model where correlation amongst multiple time-series is accommodated.

Finally, we focus on model uncertainty; that is, we consider that one may not know the specific form of the underlying dynamics. For this problem, we investigate both Bayesian model selection and averaging-based solutions. The resulting outcome is a dynamically adjustable SMC method with improved accuracy.

The contribution of the work in this dissertation is both on the theoretical Bayesian analysis of the models and on its application to computational algorithms for challenging problems of interest in practice. Short- and long-memory processes are examined and the modeling assumptions relaxed as much as possible for added flexibility and applicability.

The performance of the proposed SMC methods is thoroughly evaluated via simulations of most challenging scenarios, and illustrative examples where they can be applied are provided.

Nigan sinistu duten guztiei.

The important thing is not to stop questioning.
Curiosity has its own reason for existing.

— Albert Einstein

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AIC Akaike Information Criterion

AR Auto-Regressive

ARMA Auto-Regressive Moving-Average

ARIMA Auto-Regressive Integrated Moving-Average

BIC Bayesian Information Criterion

CDF Cumulative Density Function

DA Density Assisted

ECG Electrocardiography

EKF Extended Kalman Filter

EM Expectation-Maximization

FARIMA Fractional Auto-Regressive Integrated Moving-Average

fGn fractional Gaussian noise

fGp fractional Gaussian process

FHR Fetal Heart Rate

iid independent and identically distributed

IS Importance Sampling

KF Kalman Filter

MA Moving-Average

MCMC Markov Chain Monte Carlo

MSE Mean-Squared Error

OfGp Operator fractional Gaussian process

PF Particle Filter

PDF Probability Density Function

PMF Probability Mass Function

PGM Probabilistic Graphical Model

QML quasi-maximum likelihood

SIR Sampling Importance Resampling

SMC Sequential Monte Carlo

SV Stochastic Volatility

UKF Unscented Kalman Filter

WSS wide-sense stationary

If A is a success in life, then A equals x plus y plus z.
Work is x; y is play; and z is keeping your mouth shut.

— Albert Einstein

ACKNOWLEDGMENTS

These first lines of gratitude can only be for my advisor Professor Petar M. Djurić. Five years ago, he gave me the opportunity to come to Stony Brook and join his research group. Since then, he has challenged me and made me work hard every single day. There have been many long, productive and heated discussions; and countless times where he concluded with an encouraging *"prove me wrong"* line. All this work would not be possible without his impetus. I admire his enthusiasm for teaching, his never ending curiosity, and his desire to get involved in new intellectual endeavors. For all these, I am grateful to him.

This dissertation can not be explained without the unconditional support of Professor Mónica F. Bugallo. I will be forever grateful for her advice; for always finding some time to sit, talk and help. Muchas gracias, Mónica.

I also offer my most sincere thanks to my colleagues at Stony Brook University. To Tony and Rachel for being of great help every single time; and to all my Cosine labmates for being part of the experience. Special mention goes to Çağla Taşdemir, Shishir Dash, Isaac Manuel, Yunlong Wang, Zhiyuan Weng and Kezi Yu. For the many discussions on research and others; for offering new perspectives on life; and for becoming my friends.

I am genuinely grateful for all the friendships made over the last five years. My most sincere thanks to those that have become my family here (Sara, Doug, Ainara, Bea, Adrián, Gari and Gabriel), and to the innumerable people that have been part of this amazing journey (Fede, Alberto, James, Gu, José, Pete, Maria, Liz, María Xosé and others).

Additionally, I thank those that have always been and will be there (even if not close by). Nire koadrilari, Oñativiari, Harri, Mikel eta Carlosi, eskerrik asko.

Finally, to my parents and my sister, for their unconditional love and support, muxu handi bat.

Iñigo Urteaga

PUBLICATIONS

Some of the hereby presented concepts, proposed SMC methods and results have resulted in the following publications:

[1] Iñigo Urteaga and Petar M. Djurić. "Sequential Estimation of Hidden ARMA Processes by Particle Filtering - Part I." In: *IEEE Transactions on Signal Processing* (2016).

[2] Iñigo Urteaga and Petar M. Djurić. "Sequential Estimation of Hidden ARMA Processes by Particle Filtering - Part II." In: *IEEE Transactions on Signal Processing* (2016).

[3] Iñigo Urteaga, Mónica F. Bugallo, and Petar M. Djurić. "Sequential Monte Carlo methods under model uncertainty." In: *2016 IEEE Workshop on Statistical Signal Processing*. 2016.

[4] Iñigo Urteaga, Mónica F. Bugallo, and Petar M. Djurić. "Sequential Monte Carlo sampling for correlated latent long-memory time-series." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016.

[5] Iñigo Urteaga, Mónica F. Bugallo, and Petar M. Djurić. "Filtering of nonlinear time-series coupled by fractional Gaussian processes." In: *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2015.

[6] Iñigo Urteaga, Mónica F. Bugallo, and Petar M. Djurić. "Sequential Monte Carlo sampling for systems with fractional Gaussian processes." In: *2015 23th European Signal Processing Conference (EUSIPCO)*. 2015.

[7] Iñigo Urteaga and Petar M. Djurić. "Particle filtering of ARMA processes of unknown order and parameters." In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015.

[8] Iñigo Urteaga and Petar M. Djurić. "Estimation of ARMA state processes by particle filtering." In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014.

Part I

INTRODUCTION AND PROBLEM FORMULATION

The focus of this work is on the analysis and estimation of latent time-series, motivated by a myriad of real-life applications.

Methods for inference and prediction of time-series are studied, where the state-space framework is embraced under the Bayesian data analysis methodology and sequential Monte Carlo algorithms are proposed.

A motivation to the problem of interest, its formal description, and an introduction to the foundations of the work are presented in the following chapters.

1 INTRODUCTION

In the era of information-sensing mobile devices, sensor networks, the Internet-of-Things and Big-Data, research on advanced methods for extracting information from data has gained even more relevance.

Both within academia and industry, a great amount of effort is being put into developing advanced techniques for learning and understanding of the vast amount of data being collected. Disciplines such as machine learning and data science have become trendy and they heavily influence research, business and economic activities. The goal of extracting knowledge from data concerns many disciplines of science and engineering, such as statistics, computer-science, signal-processing and pattern-recognition.

One important area of work is the analysis of time-varying phenomena, observed sequentially in time. This task is critical in many applications in science and engineering, including weather sciences, the study of time-evolution of stocks and goods prices in econometrics, biomedical signal-processing, or concentrations of pollutants in the environment.

The objective is to infer the dynamics of events of interest that the data describe, as soon as they are acquired, so that practitioners can classify events, predict the future and make informed decisions. The study of time-varying phenomena is often referred to as time-series analysis.

A time-series is nothing but a set of observations, each one being recorded at a specified time t. Research on time-series is a long and storied field of study [16, 47, 96] and examples of time-series are endless in multitude of fields of engineering, science, economics and others. We illustrate some interesting cases below.

Signal processing (both in its analog and digital variants) has been an important discipline in engineering departments, where time-series have been long studied. Not only communication signals are good examples, but the speech signal-processing community has also investigated the human voice from a time-series point of view. Speech processing has been studied for decades and, actually, is going through a revival

period because of the development of new technologies behind Apple's Siri, Google Voice, Microsoft's Cortana and others.

An example of a short human speech time-series recorded by a microphone is shown in Figure 1.

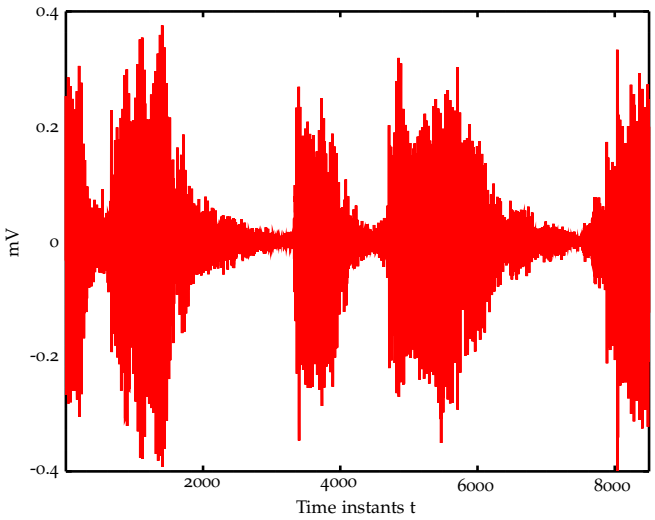
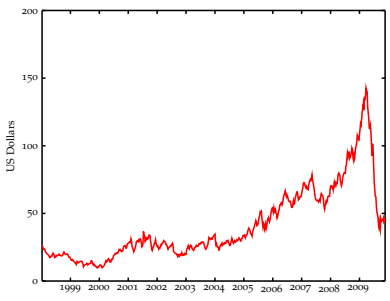
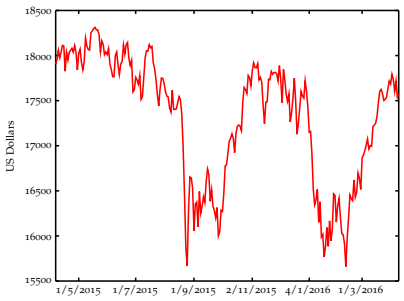


Figure 1: An example of human speech time-series.

A pronounced interest in the study of time-series is also observed in the areas of economics and finance. Two illustrative examples are the analysis of prices over time (e.g., oil price evolution as shown in Figure 2a) and the financial analysis of stock markets (the evolution of the Dow Jones index is shown in Figure 2b). In the era of algorithmic trading, complex mathematical analysis and computer algorithms for advanced trading strategies have become indispensable.



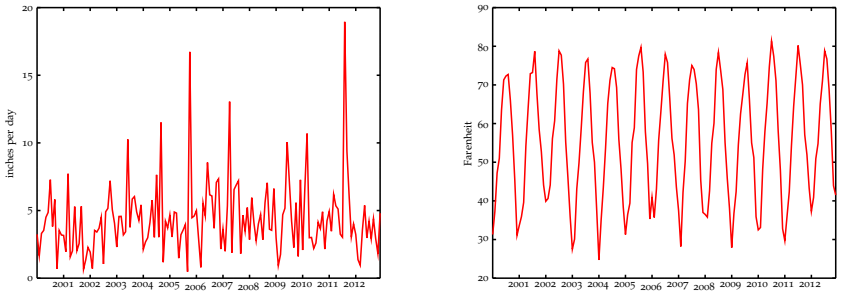
(a) Oil price over time.



(b) Dow Jones Index.

Figure 2: Examples of economics and finance time-series.

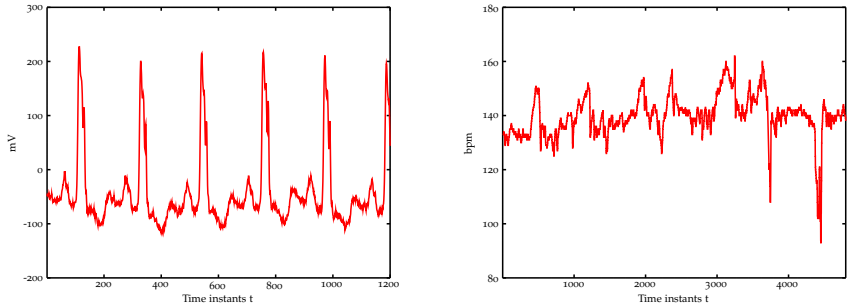
Time-varying phenomena are also critical to many other disciplines of science. For example, meteorologists have been tracking variables of their interest for decades. In order to assess the causes and evaluate the impact of Global Warming, there is an increasing need to study the underlying patterns of temperatures, precipitations and pollutants in the atmosphere. The averaged monthly precipitation and temperature recordings at Central Park are illustrated in Figure 3, for the year span between 2000 and 2012.



(a) Precipitation. (b) Temperature.

Figure 3: Examples of time-series in meteorology.

Finally, interest in time-series analysis is increasingly growing within the medical and health sciences. Time-series analysis in neuroscience is quite a mature area [79] and, lately, other fields of medicine and bio-engineering are increasingly becoming aware of this field of study. In general, doctors can benefit from the automated and machine-aided examination of bio-medical signals, e.g., electro-cardiograms or fetal heart rate monitoring (i.e., Figure 4).



(a) Sample ECG. (b) Sample Fetal Heart Rate.

Figure 4: Examples of time-series in the health sciences.

In all these applications, the interest is on learning hidden scientific truths from the acquired data. That is, practitioners need advanced techniques for obtaining insights about the observed phenomena. In a nutshell, the aim is to infer knowledge about the variables of interest and, if possible, to provide informed predictions of future values of the phenomena.

In statistical terms, the goal is to perform accurate inference and prediction (forecasting) of the variables of interest. Usually, the relevant information is latent; that is, not directly observable. For example, deciphering messages from noisy data is common for a telecommunication engineer; discovering the evolving volatility of assets is essential to financial analysts; inferring the patient's heart condition from Electrocardiography (ECG) is critical for medical doctors. Hence, the analysis of latent variables is challenging and essential in many fields of science and engineering.

In general, one discriminates between continuous- and discrete-time data series. The former are obtained when the measurements are recorded continuously over some time interval. The latter are formed by observations made at a specific (discrete) set of time instants, which can be acquired at random or periodically. The work presented here is devoted to discrete-time series.

When relevant information is latent (i.e., the variables of interest are not directly observed), flexible mathematical formulations are required. That is, the modeling framework must be able to capture the duality between the observables and the latent variables of interest.

The focus of the work presented hereafter comprises inference and prediction of latent time-series of various types, where data are available sequentially in time. The aim is to provide specific techniques, by leveraging methods from statistics and engineering, that (1) describe the data accurately, (2) model both the latent and observable information, and (3) provide mechanisms to accurately infer and predict the unknowns of interest.

2

PROBLEM FORMULATION

The analysis of latent time-series is the focal point of the work presented here. This chapter formally introduces the problem of interest and provides an overview of the foundations of the described work.

We first clarify the notation used throughout the document in Section 2.1. To accommodate for both the latent nature of the hidden dynamics of interest and the applicability of the solution to a myriad of scenarios, the state-space methodology is adopted, which is described in Section 2.2.

Inference and prediction of the latent time-series is achieved by means of the Bayesian methodology. A brief review of the Bayesian theory and its application to data analysis is provided in Section 2.3.

Since the work considers generic models, beyond linearity or Gaussianity assumptions, advanced Monte Carlo methods are in need. Specifically, and due to the sequential nature of the acquired data, an introduction to SMC methods is provided in Section 2.4.

In summary, this chapter provides a comprehensive overview of the background for understanding the solutions presented in the forthcoming chapters.

NOTATION

Throughout the document, variables and scalars of interest are represented by lower-case notation (e.g., x, y, z, θ), where the dimensionality of the variable is specifically indicated: i.e., $x \in \mathbb{R}^{d_x}$. For matrices, upper-case notation is used (e.g., mixing matrix $A \in \mathbb{R}^{d_y \times d_x}$ or covariance matrix $\Sigma \in \mathbb{R}^{d_x \times d_x}$).

Time-dependency is indicated by the subscript "$_t$", as in x_t and y_t. When referring to a sequence of variables from time t_1 to t_2, we use $y_{t_1:t_2} \equiv \{y_{t_1}, y_{t_1+1}, \cdots, y_{t_2}\}$. Unless specifically indicated, discrete-time data are considered and, therefore, $t \in \mathbb{N}_0$. Within this document, y_t represents the observed data and x_t, the latent time-series. In general, the set of parameters of the model is represented by the vector θ.

The statistical notation used for stochastic variables is summarized below:

- $a \sim f(\cdot)$ represents that a continuous random variable a is distributed according to the Probability Density Function (PDF) $f(\cdot)$. The corresponding Cumulative Density Function (CDF) is represented by the uppercase counterpart $F(\cdot)$.

- $b \sim p(\cdot)$ represents that a discrete random variable b is distributed according to the Probability Mass Function (PMF) $p(\cdot)$. The corresponding CDF is represented by $F(\cdot)$.

- $\mathcal{N}\left(\mu, \sigma^2\right)$ represents the univariate Gaussian density with mean μ and variance σ^2. Its d-dimensional multivariate counterpart is indicated by $\mathcal{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$.

- $\mathcal{T}_\nu\left(\mu, \sigma^2\right)$ represents the univariate Student's t-distribution with $\nu > 0$ degrees of freedom, location parameter μ and scale parameter σ. Its d-dimensional multivariate counterpart is represented by $\mathcal{T}_\nu(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ represents its location vector and $\Sigma \in \mathbb{R}^{d\times d}$, the scale matrix.

- $\mathrm{IW}_d(\nu, \Lambda)$ represents the d-dimensional inverse Wishart distribution with $\nu > d-1$ degrees of freedom and scale matrix $\Lambda \in \mathbb{R}^{d\times d}$.

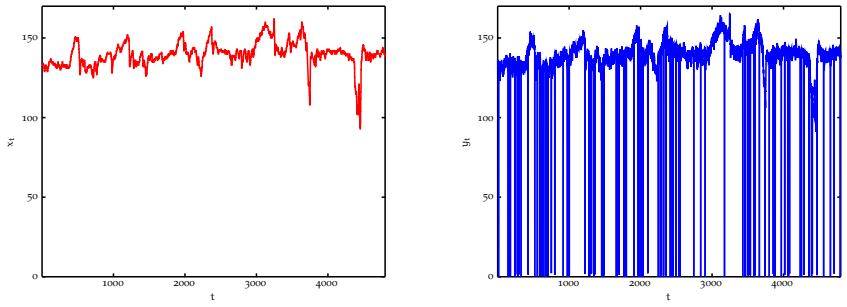
- $\mathrm{NIW}_d(\eta, \kappa, \nu, \Lambda)$ represents the d-dimensional normal-inverse-Wishart distribution with location parameter $\eta \in \mathbb{R}^d$, inverse scale matrix $\Lambda \in \mathbb{R}^{d\times d}$, and real parameters $\kappa > 0$ and $\nu > d-1$.

STATE-SPACE MODELS

State-space modeling provides a unified and robust methodology for studying a wide range of engineering problems [47]. It provides a very flexible framework for time-series analysis and they have been successfully applied to a myriad of applications [96].

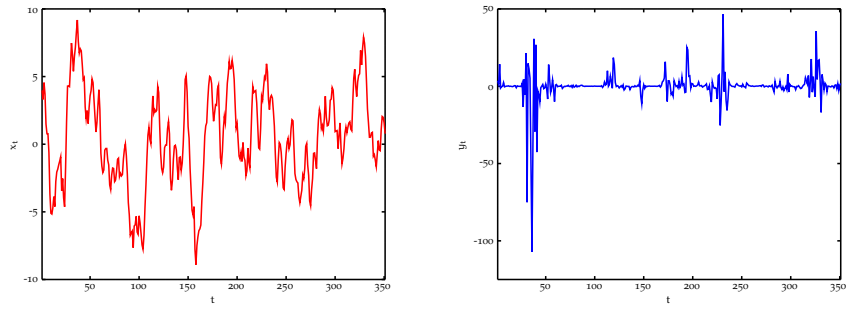
With the state-space formulation, a system is modeled over time by a series of hidden variables associated with another series of measurements. That is, it comprises a set of ordered observations y_t that depend on some latent time-evolving

unknown x_t. The hidden dynamics are not directly observed, although the observations do depend on the latent states in various forms (see Figure 5 and Figure 6 for two illustrative examples).



(a) Latent FHR. (b) Observed FHR signal.

Figure 5: Example FHR state-space model.



(a) Latent volatility process. (b) Observed stochastic signal.

Figure 6: Example stochastic log-volatility state-space model.

Within this framework, several signal processing problems can be tackled:

- **Filtering:** Updating the knowledge of the state process of the system each time a new observation is available: i.e., estimate x_t based on the newly observed y_t. Filtering is performed in a sequential manner, as new data becomes available.

- **Smoothing:** Estimating the state process of the system based on a whole set of observations: i.e., estimate x_t based on $y_{1:T}$, where $1 \leqslant t \leqslant T$. This technique requires the availability of a batch of observations and therefore, implies a delay in processing.

- **Prediction:** Forecasting the future of the state $x_{t+\tau}$ or the observations $y_{t+\tau}$ for a time horizon τ, based on a set of available observations $y_{1:t}$.

The main focus of the work presented here is on filtering and prediction, motivated by the real-life problems described in the introductory Chapter 1.

The state-space methodology mathematically models the system of interest as follows:

$$\begin{cases} x_t = g\left(x_{1:t-1}, \theta_g, u_t\right), & \text{state equation} \\ y_t = h(x_t, \theta_h, v_t), & \text{observation equation} \end{cases} \tag{1}$$

where $t = \{1, 2, \dots\}$ represents time, $x_t \in \mathbb{R}^{d_x}$ is the time-varying latent process and $y_t \in \mathbb{R}^{d_y}$ is the ordered sequence of observations available (of dimensionality d_x and d_y, respectively).

The time-dependency of the latent process is characterized by the (potentially time dependent) function $g(\cdot)$ on both the previous values of the signal, the state noise u_t (i.e., innovations) and a set of static parameters $\theta_g \in \mathbb{R}^{d_{\theta_g}}$.

The dependency of the observations on the hidden state is accounted for by the function $h(\cdot)$ (which might also vary with time and is parameterized by its static parameters $\theta_h \in \mathbb{R}^{d_{\theta_h}}$). The observation equation is stochastic in nature, with observation noise v_t.

Based on the nature of the functions and the innovations considered, there are many problems that can be presented within the state-space paradigm. Amongst them, the most studied case is the linear Gaussian one, where $g(\cdot)$ and $h(\cdot)$ are linear functions and the state and observation noises are both Gaussian:

$$\begin{cases} x_t = A x_{1:t-1} + B u_t, \\ y_t = D x_t + E v_t, \end{cases} \quad \text{with} \quad \begin{cases} u_t \sim \mathcal{N}(0, C_u) \\ v_t \sim \mathcal{N}(0, C_v). \end{cases} \tag{2}$$

Filtering based on this linear and Gaussian model has been widely studied and the closed form solution is the celebrated Kalman Filter (KF) [63]. Its application to countless signals and systems has been numerously reported. In fact, it is the optimal candidate for the small subset of linear latent time-series filtering cases with Gaussian innovations [3, 50]. However, the research described herein aims at more ambitious

settings, where the state-space models diverge from linearity and Gaussianity assumptions.

For this set of more general problems, where nonlinearity and non-Gaussianity are considered, optimal processing of the data becomes very challenging and, in practice, impossible. Therefore, approximating techniques must be considered, which result in suboptimal solutions.

There are several approaches that rely on approximating the state and space functions, minimizing the impact of the considered nonlinearities and non-Gaussianities. The Extended Kalman Filter (EKF) is considered to be the natural evolution to deal with nonlinear functions within the framework of the traditional KF [3, 50, 103]. The EKF linearizes the problem by a Taylor series approximation around the latest estimate, so that the recursive KF is applicable.

Another alternative, known as the Unscented Kalman Filter (UKF) [102] relies on the assumption of Gaussian densities in the model, instead of resorting to strict linearization of the state-space functions. A number of other variations of these, such as the Gaussian quadrature Kalman filter [57], have also been suggested. In all these, the aim is to provide a functional form that approximates the true density.

A different paradigm for nonlinear/non-Gaussian problems consists on evaluating the PDF of interest over a set of points, without relying on approximations to the nonlinear functions. This approach is orthogonal to the Kalman-based solutions, as it uses discrete random measures to represent the densities of interest. The most basic of these methods evaluate the densities over a deterministic set of points (grids) [66].

To overcome the limitations of such a rigid approach, other alternatives have been suggested, which are based on Monte Carlo sampling. These particle-based methods use random sets of points for representation of the PDFs of interest. In other words, at one time instant the PDFs are evaluated at a set of points, and at the next time instant, a completely different set of points is used. Filters based on this methodology amount to sequential Monte Carlo (SMC) sampling methods and are popularly referred to as a Particle Filter (PF) [19, 37, 41, 43].

Ever since the introduction of PFs, they have been adopted by the research community and successfully applied to a wide range of disciplines, including engineering [85], geophysical sciences [68], biology [55, 56], and economics [32].

We describe the foundations of SMC methods or PFs in Section 2.4, after first introducing the Bayesian paradigm for time-series analysis.

THE BAYESIAN METHODOLOGY

The analysis of latent time-series raises several challenges: How to fit the observed data to the mathematical model, how to learn about the hidden variable of interest, and how to make predictions on the future of the time-series.

In time-series analysis under the state-space paradigm, we acquire data (i.e., the observable variables) that depend on hidden dynamics (i.e., the latent variables). Therefore, one can formulate the problem in terms of a set of random variables that are acquired sequentially in time. The observed data are considered to be a realization of a set of random variables.

The theory of stochastic processes becomes handy to model the behavior of observed and latent data. At times, the process is modeled according to a physical mechanism that generates the data and uncertainty is included via an innovation or noise process. Other times, the data are described in a purely statistical sense, without providing any interpretation of the model parameters.

When using statistical methods for the task in hand, two related but orthogonal alternatives are available: The classical or *frequentist* statistical philosophy and the alternative *Bayesian* methodology. This study embraces Bayesian theory, as it accounts for all available information to perform statistical inference.

The Bayesian methodology [86] focuses on three main constituents: The likelihood of the data, the prior density of the model parameters and the resulting posterior, all connected by the celebrated Bayes' theorem

$$f(x|y,\theta) = \frac{f(y,x|\theta)}{f(y|\theta)} = \frac{f(y,x|\theta)}{\int f(y,x|\theta)dx} = \frac{f(y|x)f(x|\theta)}{\int f(y|x)f(x|\theta)dx}\,, \quad (3)$$

where y is the observed data, x is the unknown variable and θ the parameters of the model.

In Bayesian theory, all sources of uncertainty in the statistical model are considered to be random variables, described by their corresponding densities: $f(y|x,\theta)$ is the probability distribution modeling the observables, which depend on the unknown variable x, that has a prior distribution $f(x|\theta)$ with

parameters θ. The parameters, being random variables too, are described by their own density $f(\theta|\eta)$, that depend on the hyper-parameters η.

By applying Bayes' principles, contribution of both the experimental data (i.e., observed values) and the prior knowledge of the problem are fused. The first, via the likelihood $f(x|\theta)$ and, the second, through the prior information $f(\theta)$.

The Bayesian framework provides a unified approach to data analysis, where practitioner's knowledge is formally incorporated to the methodology. Furthermore, it allows for parameter independent evaluation of data, by using the marginalization theorem from statistics.

The parameter agnostic density of the latent variable can be derived as

$$f(x) = \int f(x|\theta) f(\theta) d\theta \tag{4}$$

and predictive densities of the observables computed by

$$f(y') = \int f(y'|x) f(x|y) dx \, . \tag{5}$$

In order to conduct a fully Bayesian analysis, specification of prior densities is critical, as model parameters are considered random variables themselves. The simplest case amounts to when the true values are known, as there is no need to express conditioning on a constant and $f(x|y)$ and $f(x)$ are employed. Thus, the Bayes formula simplifies to

$$f(x|y) = \frac{f(y,x)}{f(y)} = \frac{f(y,x)}{\int f(y,x) dx} \, . \tag{6}$$

However, there are many practical cases where the model parameters are not known. In some cases, some knowledge or belief might be available to the practitioner while, in others, no good guess can be made on the unknown θ. Nonetheless, Bayesian theory addresses all these scenarios in a unified manner, by careful determination of the prior density.

On the one hand, informative priors can be used to accommodate for field expertise (through particular shapes of the prior densities). Among all the distributional families that an informative prior $f(x|\theta)$ might belong to, a computationally convenient choice is to use conjugate priors [77]. These priors guarantee that, for a given likelihood $f(y|x,\theta)$, the posterior $f(x|y,\theta)$ falls into the same distributional family as the prior $f(x|\theta)$.

On the other hand, when no reliable prior information concerning the model parameters exists, inference based exclusively on data is preferable. The Bayesian methodology endorses such cases by means of *non-informative* priors [59], i.e., distributions $f(\theta)$ that do not favor any particular value of θ. Many efforts have focused on the study of non-informative priors and different alternatives have already been presented [13].

Furthermore, under the Bayesian methodology, the prior itself may depend on another set of unknowns, which follow some second-stage prior $f(\eta)$. This sequence of unknowns and model parameters θ and hyper-parameters η constitute a hierarchical model: $f(x|\theta)$, $f(\theta|\eta)$ and $f(\eta)$.

Since the hierarchy might grow *ad libitum*, it is commonplace to stop this at some point. When to stop and how to provide a set of known hyper-parameters is a modeling choice left to the practitioner. Estimation or marginalization of the hyper-parameters are only two of the possible alternatives.

Another possibility consists on, instead of blindly deciding on some hyper-parameter values, using observed data to provide meaningful initial values. This approach is known as the *empirical Bayes* method [21].

All in all, Bayesian statistics provide a unified framework for data analysis. In particular, and under the state-space paradigm, one can perform:

- **Filtering**, by sequentially estimating the posterior density of x_t, given the observations $y_{1:t}$; i.e., $f(x_t|y_{1:t})$.
- **Smoothing**, by estimating the distribution of x_t based on a set of observations $y_{1:T}$; i.e., $f(x_t|y_{1:T})$, where $1 \leqslant t \leqslant T$.
- **Prediction**, by estimating the density of the next state x_{t+1} based on a set of available observations $y_{1:t}$; i.e., $f(x_{t+1}|y_{1:t})$. Similarly, prediction of the space variables resorts to $f(y_{t+1}|y_{1:t})$.

The sequential processing required for time-series analysis is attained by means of factorization of the relevant densities, marginalization of unknowns and Bayes' rule:

$$\begin{cases} f(x_{1:t+1}|y_{1:t}) = f(x_{t+1}|x_{1:t}, y_{1:t}) f(x_{1:t}|y_{1:t}), \\ f(y_{t+1}|y_{1:t}) = \int f(y_{t+1}|x_{1:t+1}) f(x_{1:t+1}|y_{1:t}) dx_{1:t+1}, \\ f(x_{1:t+1}|y_{1:t+1}) = \frac{f(y_{t+1}|x_{1:t+1}) f(x_{t+1}|x_{1:t}, y_{1:t}) f(x_{1:t}|y_{1:t})}{f(y_{1:t+1})}. \end{cases} \tag{7}$$

Besides a unified methodology for data analysis, Bayesian statistics also provide model selection and averaging procedures. In brief, one can include uncertainty about the true model $\mathcal{M}$ into the analysis.

One considers some prior $p(\mathcal{M}_k)$ over the set of candidate models $\mathcal{M}_k, k = 1, \cdots, K$, and then applies Bayes' theory to update the posterior belief of each model after observing data up to time instant t, $y_{1:t}$, i.e.,

$$p(\mathcal{M}_k|y_{1:t}) = \frac{f(y_{1:t}, \mathcal{M}_k)p(\mathcal{M}_k|y_{1:t})}{f(y_{1:t})}, \tag{8}$$

where $p(\mathcal{M}_k|y_{1:0}) = p(\mathcal{M}_k)$ is the prior assumed for each model.

One can select the most likely model based on the above computation or, as an alternative, fuse information via the total probability theorem:

$$p(y_{1:t}) = \sum_{k=1}^{K} f(y_{1:t}|\mathcal{M}_k)p(\mathcal{M}_k). \tag{9}$$

The work in this dissertation is mostly on the derivation of the densities in Equation 7 (and, if applicable, Equation 8) for a myriad of considered time-series models. To do so, we follow Bayesian statistics, as it allows for consideration of known and unknown parameters (also for model uncertainty) in a unified manner.

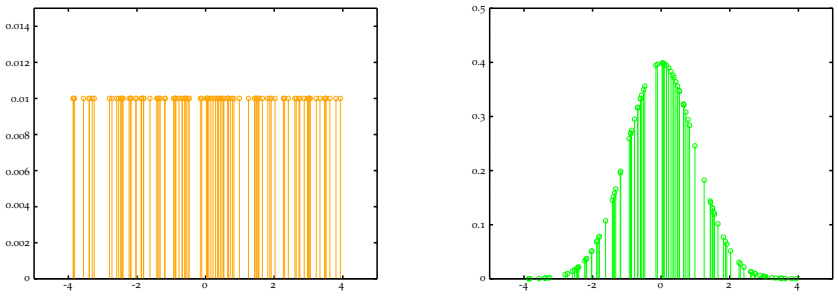
In most of the problems that are of interest here (i.e., state-space with non-Gaussian and nonlinear functions), no closed form solution is available for the densities above. Thus, as mentioned before, we resort to SMC methods, described now in detail.

SEQUENTIAL MONTE CARLO METHODS

In sequential Monte Carlo (SMC) sampling methods (also known as PFs), recursive computation of relevant probability distributions is performed by approximating them with discrete random measures that change over time. That is, the PF uses random grids (i.e., the location of the points varies randomly with time) to evaluate the PDFs of interest.

The points used to represent the relevant density are called particles. Weights are assigned to them, which can be interpreted as probability masses. Consequently, the particles (see Figure 7a) and the weights form a discrete random measure

(see Figure 7b), which approximates the density of interest as in Figure 8.



(a) Drawn samples. (b) Weighted samples.

Figure 7: Example of drawn and weighted random samples.

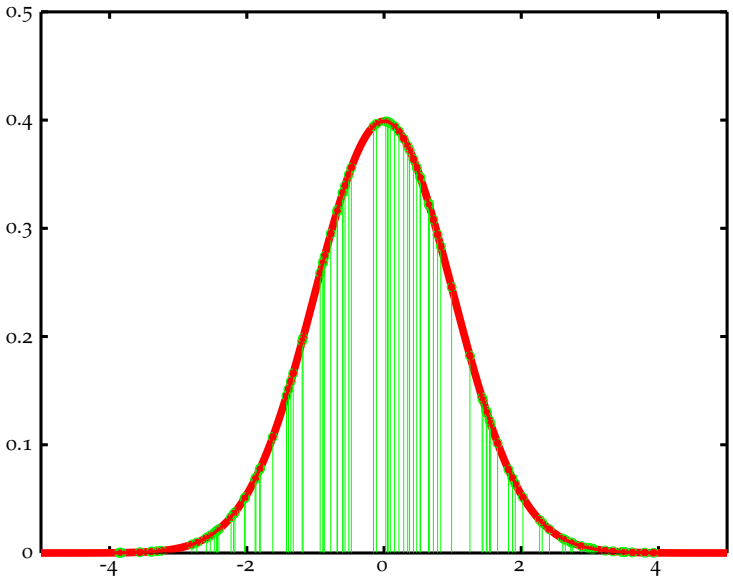


Figure 8: A random measure approximates a Gaussian density.

The random measure is thus composed of particles and their weights, i.e.,

$$f^M(x) = \sum_{m=1}^{M} w^{(m)} \delta\left(x - x^{(m)}\right) , \tag{10}$$

where $\delta(\cdot)$ is the Dirac delta function, $x^{(m)}$ is the mth particle, $w^{(m)}$ is the weight of that particle, and M is the total number of particles.

In summary, particle filtering is a Monte-Carlo method that sequentially approximates evolving densities by randomly generating candidate particles and assigning weights to them.

When applied to state-space models with sequential acquisition of data, the system aligns to the recursive procedure of the SMC methodology. The generation of particles at each time instant depends on the previous random measure; thus, each particle has a *parent*, the *parent* its own *parent*, and so on. Such sequence of particles is called a particle *stream*, and it represents one possible evolution of the latent state with time [37].

In particle filtering, the sequential processing is nothing but a recursive update of the discrete random measure upon arrival of a new observation. The update step accommodates both the generation (sampling) of new particles and the computation of their corresponding weights.

An SMC method consists of three basic steps. The first is the propagation of particles following the dynamics of the model. The second, the computation of their weights to form a proper random measure and, the last, accounts for *resampling* to avoid particle attrition over time.

A main challenge in implementing PFs is to generate and propagate particles in regions of the hidden state's space where most of the information is located. That is, the goal is to explore regions over which the densities carry significant probability masses. Ideally, the optimal distribution to sample from is the density of the state given all the available observations and past states. However, this is not usually feasible and, thus, particles are drawn from some other instrumental distribution: i.e., the *proposal* distribution. To overcome these limitations, PFs exploit a concept from statistics known as *sequential importance sampling*.

When particles are directly generated from the target distribution of interest $f(x)$, then all samples are assigned equal weights; i.e., $w^{(m)} = \frac{1}{M}, \quad m = 1, 2, \cdots, M$. However, when drawing samples directly from $f(x)$ is unfeasible, one can generate particles $x^{(m)}$ from an alternative distribution $\pi(x)$, known as the *proposal* distribution, and assign (non-normalized) weights according to

$$\widetilde{w}^{(m)} = \frac{f(x^{(m)})}{\pi(x^{(m)})} \,. \tag{11}$$

These weights have to be normalized to obtain a correct random measure approximation of the target density, i.e.,

$$w^{(m)} = \frac{\widetilde{w}^{(m)}}{\sum_{i=1}^{M} \widetilde{w}^{(i)}} \,. \tag{12}$$

Thus, by leveraging *importance sampling*, a density of interest at any given time instant is approximated by a random measure with appropriate weights. For example, for the filtering problem, a PF approximates the posterior of the state x_t given observations $y_{1:t}$, at time instant t, i.e., $f_t^M(x_t) \approx f(x_t|y_{1:t})$.

The key of a PF is the sequential computation of the random measures. The methodology proceeds recursively upon reception of a new observation y_{t+1}. Specifically, the distribution $f(x_{1:t}|y_{1:t})$ can be updated to $f(x_{1:t+1}|y_{1:t+1})$ once a new observation y_{t+1} becomes available following Bayes rule. In SMC sampling, one refurbishes the approximating random measures sequentially, i.e., it updates $f_t^M(x_t)$ to $f_{t+1}^M(x_{t+1})$.

Following the *sequential importance sampling* methodology, this is done in two steps. In the first, one propagates the particles $x_t^{(m)}$ to $x_{t+1}^{(m)}$ via

$$x_{t+1}^{(m)} \sim \pi(x_{t+1}|x_{1:t}^{(m)}, y_{1:t+1}) \,, \tag{13}$$

where $\pi(x_{t+1}|x_{1:t}^{(m)}, y_{1:t+1})$ is the proposal (instrumental, importance) density of x_{t+1}, and $x_{1:t}^{(m)}$ is the genealogical lineage of the particle (i.e., the mth particle *stream*).

The second step is the computation of the weights of $x_{t+1}^{(m)}$ according to

$$w_{t+1}^{(m)} \propto w_t^{(m)} \frac{f\left(y_{t+1}|x_{t+1}^{(m)}\right) f\left(x_{t+1}^{(m)}|x_{1:t}^{(m)}\right)}{\pi\left(x_{t+1}^{(m)}|x_{1:t}^{(m)}, y_{1:t+1}\right)} \,, \tag{14}$$

where $f\left(y_{t+1}|x_{t+1}^{(m)}\right)$ is the likelihood of $x_{t+1}^{(m)}$, and $f\left(x_{t+1}^{(m)}|x_{1:t}^{(m)}\right)$ is the transition density of the state. The computation of the weights is followed by their normalization, so that they sum to one.

One of the major drawbacks of using SMC sampling techniques is that the approximating discrete random measure degenerates quickly as time evolves. In practice, most of the particles except for a very few are assigned negligible weights and therefore, only few streams survive. Due to this particle

attrition, the performance of the particle filter considerably deteriorates as time evolves.

The impact of particle degeneracy can be reduced by using, not only good importance sampling functions, but a technique known as *resampling* too. In a nutshell, resampling eliminates particle streams with small weights and replicates particles with large weights. That is, it helps explore the region of the density's support where most of the probability mass is concentrated.

Several techniques have been introduced in the literature (e.g., regular resampling, Gaussian resampling, etc.) [42, 52, 69] and, when studying new PFs, careful consideration and evaluation of the alternatives must be explored.

A measure of the particle degeneracy is known as the effective particle size M_{eff} [37], which can be estimated as

$$M_{eff} \approx \frac{1}{\sum_{m=1}^{M} w^{(m)^2}}, \tag{15}$$

where one uses the normalized weights in the computation.

When all the particles have the same weights, the variance of the weights is zero and the particle size is equal to the total number of particles, M. The other extreme occurs when all the particles except one have negligible weights, and the particle size is equal to one. Usually, if the effective particle size is below a predefined threshold, resampling is carried out.

In the work presented here, we resort to the most common resampling technique, and do so at every time instant. We perform resampling by drawing from a categorical distribution defined by the weights of the available random measure

$$\bar{x}^{(m)} \sim f^{M}(x), \text{ where } m = 1, \cdots, M.$$

All the SMC methods in the following chapters perform resampling as above for every time instant. The extension of the proposed methods to any other resampling technique is straightforward.

It is important to note that, no matter what resampling technique is implemented, one must compute the estimates of interest based on the random measure available before resampling. That is, one computes

$$\mathbb{E}\{g(x)\} = \int_x g(x)f(x)dx \approx \int_x g(x)f^{M}(x)dx = \sum_{m=1}^{M} w^{(m)} g\left(x^{(m)}\right).$$

The estimates computed as above are guaranteed to have less variance than the alternative, where the resampled particles were to be used with equal $\frac{1}{M}$ weights.

Furthermore, one also needs to be careful with the impact of resampling on other aspects of the performance of the SMC method. More specifically, as the method evolves over time, the set of genealogical particle streams is severely depleted with resampling. That is, when using resampling to cope with importance weight degeneration, then stream depletion occurs. This phenomenon is known as path-degeneracy: i.e., in the long run, the full path particle approximation is, effectively, carried by a single particle stream only.

One needs to be careful with the potential impact of path-degeneracy in the PF. It has already been reported [27] that the Monte Carlo error of path functionals $\phi(x_{1:t}^{(m)})$ remains bounded over time if they relate to the filtering problem: i.e., $\phi(x_{1:t}^{(m)}) = x_t$. However, it explodes for the smoothing problem: i.e., $\phi(x_{1:t}^{(m)}) = x_1$.

There is, nonetheless, a lack of formal results for other functionals of interest (a preliminary analysis of a symmetric case is provided within the discussion in [28]). Therefore, any solution that mitigates the path-degeneracy issue can only be beneficial for any SMC method [64].

As noted before, the performance of the PF is critically dependent on the proposal density selected. Ideally, it must have the same support as the PDF that is being approximated. In practice, the closer the proposal function is to the target distribution, the better the approximation is.

The two most frequently used proposal functions are the transition density $f(x_{t+1}|x_{1:t})$ and the optimal importance function $f(x_{t+1}|x_{1:t}, y_{1:t+1})$, which lead to particle weight update equations $w_{t+1}^{(m)} \propto w_t^{(m)} f(y_{t+1}|x_{t+1})$ and $w_{t+1}^{(m)} \propto w_t^{(m)} f(y_{t+1}|x_t)$, respectively.

The optimal importance function minimizes the variance of the resulting weights, conditioned on the information available at each time instant: i.e., $x_{1:t}^{(m)}$ and $y_{1:t+1}$. However, the computation of the densities $f(y_{t+1}|x_t)$ and $f(x_{t+1}|x_{1:t}, y_{1:t+1})$ require integration of certain variables, which is not trivial (often intractable) for many problems of interest.

Therefore, it is of common practice to resort to the simpler transition density, as its performance is often successful enough (some illustrative results are provided in Appendix A).

Without loss of generality, the SMC methods presented in the following chapters make use of the transition density as their proposal function, as its derivation for the studied models is already a challenging task.

The proposed SMC methods are described, in general, for the filtering problem. That is, we obtain a random measure that approximates the filtering density, i.e.,

$$f(x_t|y_{1:t}) \approx f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right). \tag{16}$$

Computation of such density for different latent time-series of practical interest is the main focus of the work hereby.

However, it is important to keep in mind the power of the proposed SMC methods, as they can readily be used to compute other informative densities. One can, for example, derive parameter posteriors and predictive densities of the state and/or the observations.

At any given time instant t, one takes the provided filtering density $f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right)$ and extends it to obtain the following mixture densities

$$\begin{cases} f(\theta|y_{1:t}) \approx \sum_{m=1}^{M} w_t^{(m)} f(\theta|x_{1:t}^{(m)}), \\ f(x_{t+1}|y_{1:t}) \approx \sum_{m=1}^{M} w_t^{(m)} f(x_{t+1}|x_{1:t}^{(m)}), \\ f(y_{t+1}|y_{1:t}) \approx \sum_{m=1}^{M} w_t^{(m)} f(y_{t+1}|x_{t+1}^{(m)}), \end{cases} \tag{17}$$

where θ are model parameters and $x_{t+1}^{(m)}$ in the last equation is obtained by sampling from the transition density $f(x_{t+1}|x_{1:t}^{(m)})$ of the time-series.

Note that, with any of the above densities at hand, the proposed SMC methods allow for computation not only of point estimates, but also of more informative metrics, such as probabilities of certain events or risk metrics.

Part II

SMC METHODS FOR LATENT TIME-SERIES

The core of this dissertation is on SMC methods for inference and prediction of latent time-series. Short- and long-memory processes are studied, where an increased degree of complexity and uncertainty is considered through the following Chapters.

In Chapter 3 and Chapter 4, latent ARMA(p, q) models are examined, with known and unknown parameters, respectively. In Chapter 5 and Chapter 6, long-memory processes in the form of fGp and FARIMA models are considered.

A more generic approach to latent time-series follows, where only wide-sense stationarity is assumed in Chapter 7 and a solution for correlated time-series is provided in Chapter 8.

We conclude by studying inference under model uncertainty in Chapter 9.

3

LATENT ARMA(p, q) PROCESSES WITH KNOWN PARAMETERS

This chapter addresses the analysis of state-space models with a latent ARMA process. Specifically, the most favorable case is studied, where all the parameters of the state equation are known.

First, a description of time-series in general, ARMA models in particular, and their properties is provided in Section 3.1. Later, two SMC methods are presented, targeting latent ARMA state processes of different nature. In Section 3.2.1 the stationary ARMA(p, q) process is considered and, in Section 3.2.2, more generic ARMA models.

We conclude with Section 3.3, where the performance of the proposed methods is compared to other state-of-the-art alternatives.

TIME-SERIES AND THE ARMA(p, q) MODEL

The first step in the analysis of a time-series is the selection of a suitable mathematical model for the data. Since the nature of the observations is random, it is common to assume that the data are realizations of a stochastic process. Thus, by using the theory of statistics and stochastic processes, real-life time-series can be accurately described.

Stochastic processes have widely been used to model the behavior of time-series data [15, 16, 47, 80, 96]. At times, the process is modeled according to a physical mechanism that generates the data. Other times, the data are described in a purely statistical sense, without providing a meaningful interpretation of the model.

Among the many random processes available, this chapter focuses on a particular class of time-series, which is very flexible and widely applicable: the ARMA(p, q) model. Due to its malleable parameterization, ARMA models can be fit to any linear time-series with high accuracy.

The study of ARMA processes (and the special cases of Auto-Regressive (AR) and Moving-Average (MA) processes) has a long history. It started in the early 1950s [104], and its

popularity rose considerably in the 1970s with [15]. Their investigation in state-space form was introduced in [46].

Mathematically, the ARMA(p, q) model is described by

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \tag{18}$$

where p is the order of the AR component of the time-series with parameters a_i, $i = 1, \cdots, p$; and q is the order of the MA with parameters b_j, $j = 1, \cdots, q$. The innovations of the ARMA(p, q) model are represented by u_t. Typically, the u_ts are assumed to be independent and identically distributed (iid) samples. [1]

The ARMA(p, q) model can be rewritten as follows:

$$\begin{aligned} x_t &= \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ x_t - \sum_{i=1}^{p} a_i x_{t-i} &= \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ \sum_{i=0}^{p} a_i x_{t-i} &= \sum_{j=0}^{q} b_j u_{t-j}, \quad \text{with } a_0 = 1, b_0 = 1, \end{aligned}$$

and thus, be represented in a lag-polynomial form:

$$\begin{aligned} A(L)x_t &= B(L)u_t\,, \\ \text{where } &\begin{cases} A(L) = 1 - a_1 L - a_2 L^2 - \cdots - a_p L^p\,, \\ B(L) = 1 + b_1 L + b_2 L^2 + \cdots + b_q L^q\,. \end{cases} \end{aligned} \tag{19}$$

The autoregressive part of the model (i.e., $A(L)$) takes into account the previous values of the time-series, while the moving-average term (i.e., $B(L)$) adds correlated innovations to the time-series.

Depending on the nature of the noise u_t, different stochastic processes can be modeled: Gaussian stationary processes, mean varying processes, time-series with outliers (with heavy tailed innovations), etc.

If sequentially acquired data are understood as nothing but a realization of such generic stochastic models; then, a time-series is fully described by the mathematical formulation of the stochastic process.

1 Non iid innovation processes are considered in Chapters 5 and 6.

In fact, by leveraging statistical tools, the densities of the process (i.e., their PDFs) are sufficient to describe a time-series. Precisely, different densities can be considered, depending on the interest of the practitioner, such as the joint density $f(x_{1:t})$ of the time-series or the conditional transition density $f(x_t|x_{1:t-1})$. Specifying the full form of the joint probability distribution is sometimes quite a challenging task. However, once it is obtained, any marginal or conditional density of the series can be derived.

When necessary, it is common to make simplifying assumptions for the data observed in real-life. In fact, for many of the time-series of interest in the fields of science and engineering the stationarity of the process is usually enforced.

Time-series and stationarity

A stochastic process is said to be strictly stationary if its properties (i.e., joint probability distribution) are unaffected by a change in the time origin:

$$f(x_1, x_2, \cdots, x_t) = f(x_{1+\tau}, x_{2+\tau}, \cdots, x_{t+\tau}). \tag{20}$$

Since assuming strict sense stationarity is often too limiting, a weaker form of stationarity is commonly employed in signal processing, known as WSS.

WSS stochastic processes require that only the first and second moments exist and that they are constant with respect to time:

$$\begin{cases} \mathbb{E}\{x_1\} = \mathbb{E}\{x_2\} = \cdots = \mathbb{E}\{x_t\} = \mu\,, \\ \mathrm{Var}\{x_1\} = \mathrm{Var}\{x_2\} = \cdots = \mathrm{Var}\{x_t\} = \sigma^2\,, \\ \mathrm{Cov}\{x_t, x_{t-\tau}\} = \mathrm{Cov}\{x_{t+k}, x_{t+k-\tau}\} = \gamma_\tau\,. \end{cases} \tag{21}$$

The conditions in Equation 21 imply that the mean does not vary with time and that the autocovariance of the process is a function of only the time-difference (lag τ) and not of the actual time instants.

In order to enforce these weak-stationarity conditions for an ARMA(p, q) process, limitations on the possible values of the ARMA parameter set (a_i, b_j) must be satisfied. In fact, the stationarity of an ARMA(p, q) process depends solely on its autoregressive part (the moving average part determines the invertibility of the process).

Precisely, an ARMA(p, q) process as described in Equation 19 is stationary if and only if the roots of the AR polynomial $A(L)$ lay outside the unit circle. A quick analysis of the condition reveals that, equivalently, the stationarity of an ARMA(p, q) process is guaranteed as long as the roots of the polynomial $1 + a_1 L + a_2 L^2 + \cdots + a_p L^p$ are within the unit circle.

All in all, this implies that for a WSS stochastic process, the mean and the covariance of the process are the sufficient statistics that describe it. Thus, for an ARMA(p, q) process that fulfills the stationarity constraints, the joint density is computable in closed form.

ARMA(p, q) models: stationary densities

We now aim at deriving the joint density of a stationary ARMA(p, q) process. We start by computing the first and second order statistics of the ARMA(p, q) process, given that the roots of the AR polynomial $A(L)$ lie outside the unit circle. Without loss of generality, a zero mean iid Gaussian innovation process is assumed: i.e., $\mathbb{E}\{u_t\} = 0$ and $\mathrm{Var}\{u_t\} = \sigma_u^2$.

It can be readily concluded that the expected value of a weakly-stationary ARMA(p, q) process is zero

$$\begin{aligned}
\mathbb{E}\{x_t\} = \mu = \mathbb{E}\left\{\sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t\right\} = \\
&= \sum_{i=1}^{p} a_i \mathbb{E}\{x_{t-i}\} + \sum_{i=j}^{q} b_j \mathbb{E}\left\{u_{t-j}\right\} + \mathbb{E}\{u_t\} \\
&= \sum_{i=1}^{p} a_i \mu \,, \\
\mu\left(1 - \sum_{i=1}^{p} a_i\right) = 0 \,, \\
\boxed{\mu = 0}.
\end{aligned} \tag{22}$$

The autocovariance function (i.e., the expected value of the product of two lagged time instants of the process $\mathbb{E}\{x_t, x_{t+\tau}\} = \gamma(\tau)$), can be written in the following recursive form,

$$\begin{cases} \text{For lags } 0 \leqslant \tau < max(p, q+1): \\ \qquad \gamma(\tau) - a_1\gamma(\tau-1) - \cdots - a_p\gamma(\tau-p) = \sigma_u^2 \sum_{\tau \leqslant j \leqslant q} b_j \psi_{j-\tau}\,. \\ \text{For lags } \tau \geqslant max(p, q+1): \\ \qquad \gamma(\tau) - a_1\gamma(\tau-1) - \cdots - a_p\gamma(\tau-p) = 0. \end{cases}$$

where the $\psi_j$values are obtained from

$$\begin{cases} \psi_j - \sum\limits_{0<\tau\leqslant j} a_\tau \psi_{j-\tau} = b_j\,, & 0 \leqslant j < max(p, q+1), \\ \psi_j - \sum\limits_{0<\tau\leqslant p} a_\tau \psi_{j-\tau} = 0\,, & j > max(p, q+1). \end{cases} \tag{23}$$

The expressions in Equation 23 are recursive formulas for computation of the covariance values of an ARMA(p, q) process for any given lag τ.

These recursive equations can be rewritten in a linear system form too, where $a_i = 0$, $i > p$ and $b_j = 0$, $j > q$:

$$\begin{aligned} \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(\tau) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(\tau-1) \\ \gamma(2) & \gamma(1) & \gamma(0) & \cdots & \gamma(\tau-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(\tau) & \gamma(\tau-1) & \gamma(\tau-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_\tau \end{pmatrix} = \\ = \sigma_u^2 \begin{pmatrix} b_0 & b_1 & b_2 & \cdots & b_\tau \\ b_1 & b_2 & \cdots & b_\tau & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ b_{\tau-1} & b_\tau & 0 & \cdots & 0 \\ b_\tau & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_\tau \end{pmatrix}. \end{aligned} \tag{24}$$

By basic algebraic transformations of the above Equation 24, one obtains a matrix-based solution for computation of the autocovariance values of an ARMA(p, q) up to lag τ:

$$\begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(\tau-2) \\ \gamma(\tau-1) \\ \gamma(\tau) \end{pmatrix} = \sigma_u^2 \begin{pmatrix} 1 & -a_1 & -a_2 & \cdots & -a_{\tau-2} & -a_{\tau-1} & -a_\tau \\ -a_1 & 1-a_2 & -a_3 & \cdots & -a_{\tau-1} & -a_\tau & 0 \\ -a_2 & -a_1-a_3 & 1-a_4 & \cdots & -a_\tau & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ -a_{\tau-2} & -a_{\tau-3}-a_{\tau-1} & -a_{\tau-4}-a_\tau & \cdots & 1 & 0 & 0 \\ -a_{\tau-1} & -a_{\tau-2}-a_\tau & -a_{\tau-3} & \cdots & -a_1 & 1 & 0 \\ -a_\tau & -a_{\tau-1} & -a_{\tau-2} & \cdots & -a_2 & a_1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & b_1 & b_2 & \cdots & b_\tau \\ b_1 & b_2 & \cdots & b_\tau & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ b_{\tau-1} & b_\tau & 0 & \cdots & 0 \\ b_\tau & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & -a_1 & 1 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_\tau & -a_{\tau-1} & -a_{\tau-2} & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_\tau \end{pmatrix} \tag{25}$$

In many real world applications, observed time-series show stationarity features. That is, they can be described by sufficient statistics that do not vary with time (i.e., WSS conditions are fulfilled). Consequently, the autocovariance function $\gamma(\tau)$ of the process and the mean are the required statistics to describe such time-series.

For the ARMA(p, q) processes addressed here, the autocovariance function depends on its parameters a and b, as shown in Equation 25. Consequently, as long as the parameter set (a_i, b_j) of the ARMA(p, q) model and the sufficient statistics of the innovation process (i.e., $\mathbb{E}\{u_t\}$ and $\mathrm{Var}\{u_t\}$) are known, then an ARMA(p, q) time-series is fully described by the joint density for which the sufficient statistics are readily computed.

For a unidimensional zero-mean ARMA(p, q) process with zero-mean Gaussian innovations, the joint distribution of $x_{1:t+1} = (x_{t+1}\ x_t\ \cdots\ x_2\ x_1)^\top$ is jointly Gaussian, i.e.,

$$x_{1:t+1} \sim \mathcal{N}(x_{1:t+1}|0, \Sigma_{t+1}). \tag{26}$$

The covariance matrix $\Sigma_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ is a symmetric Toeplitz matrix of the form

$$\Sigma_{t+1} = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(t) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(t-1) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(t-1) & \gamma(t-2) & \cdots & \gamma(1) \\ \gamma(t) & \gamma(t-1) & \cdots & \gamma(0) \end{pmatrix}, \tag{27}$$

where $\gamma(\tau)$ is the autocovariance function of x_t.

We introduce the notion of a standardized autocovariance function defined as $\widetilde{\gamma}(\tau) = \frac{\gamma(\tau)}{\sigma_u^2}$. Thus, we can rewrite the covariance matrix of the vector $x_{1:t+1}$ as $\Sigma_{t+1} = \sigma_u^2 \widetilde{\Sigma}_{t+1}$, where

$$\widetilde{\Sigma}_{t+1} = \begin{pmatrix} \widetilde{\gamma}(0) & \widetilde{\gamma}_t^\top \\ \widetilde{\gamma}_t & \widetilde{\Sigma}_t \end{pmatrix}, \text{ with } \widetilde{\gamma}_t^\top = (\widetilde{\gamma}(1)\, \widetilde{\gamma}(2) \cdots \widetilde{\gamma}(t)). \tag{28}$$

Given the joint distribution of the ARMA(p, q) process and due to the properties of the Gaussian distribution (i.e., it is closed under conditioning), the transition density of the process can also be computed.

The density of the next state conditioned on the available states at any given time t, $x_{1:t}$, is a Gaussian distribution; that is,

$$f(x_{t+1}|x_{1:t}, a, b, \sigma_u^2) = \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right), \tag{29}$$

where

$$\begin{aligned} \mu_{x_{t+1}|x_{1:t}} &= \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} x_{1:t}, \\ c_{t+1} &= \widetilde{\gamma}(0) - \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t. \end{aligned} \tag{30}$$

In summary, the stationarity analysis leads to a closed-form expression for the transition density of the ARMA(p, q) process.

ARMA(p, q) models: recursive sufficient statistics

In general, the autocovariance function of an ARMA process is infinite. Therefore, as time evolves, the size of the covariance matrix in Equation 30 grows linearly in time, from dimensions $[t \times t]$ to $[t+1 \times t+1]$. This imposes a practical burden, as the computation of the Toeplitz matrix and its inverse becomes very resource challenging, specially if long time-series are considered.

However, we now show how to update recursively the required quantities. We describe this on $\widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t$, as the remaining quantities $\mu_{x_{t+1}|x_{1:t}}$, c_{t+1} and $x_{1:t}^\top \widetilde{\Sigma}_t^{-1} x_{1:t}$ can be handled analogously. In all of them, the key computation is the multiplication of a vector with the inverse of a covariance matrix.

We rewrite the key computation as $\alpha_t = -\widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t$, which relates to a system of linear equations $\widetilde{\Sigma}_t \alpha_t = -\widetilde{\gamma}_t$, that can be solved by the well-known Levinson-Durbin recursion [51]. Note that $\widetilde{\Sigma}_t$ is a symmetric Toeplitz matrix that can be generated recursively.

For the computation of the sufficient statistics of the ARMA(p, q) process at time instant $t+1$, one needs to recursively update $\widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t$ to $\widetilde{\gamma}_{t+1}^\top \widetilde{\Sigma}_{t+1}^{-1} \widetilde{\gamma}_{t+1}$.

This is achieved by

$$\widetilde{\gamma}_{t+1}^\top \alpha_{t+1} = \widetilde{\gamma}_t^\top \alpha_t + \beta_{t+1} \widetilde{\gamma}_t^\top \Pi_t \alpha_t + \widetilde{\gamma}(t+1) \beta_{t+1}, \tag{31}$$

where the values are recursively updated following these steps:

1. Initialize
   - $\beta_1 = -\widetilde{\gamma}(1)/\widetilde{\gamma}(0)$,
   - $\alpha_1 = \beta_1$, and
   - $\epsilon_1 = \widetilde{\gamma}(0)(1-\beta_1^2)$.
2. for $t = 2, \cdots$, compute:
   a) $\kappa_t = \widetilde{\gamma}(t) + \alpha_{t-1}^\top \Pi_t \widetilde{\gamma}_{t-1}$,
   b) $\beta_t = -\kappa_t/\epsilon_{t-1}$,
   c) $\epsilon_t = \epsilon_{t-1}(1-\beta_t^2)$,
   d) $\alpha_t = \begin{pmatrix} \alpha_{t-1} \\ 0 \end{pmatrix} + \beta_t \begin{pmatrix} \Pi_{t-1} \alpha_{t-1} \\ 1 \end{pmatrix}$,

where $\epsilon_t = \widetilde{\gamma}(0) + \alpha_t^\top \widetilde{\gamma}_t$ and Π_t is a $t \times t$ permutation matrix with 1's on the antidiagonal.

ARMA(p, q) models: short-memory property

In general, for a stationary ARMA(p, q) time-series at time instant t, the full past history $x_{1:t-1}$ is required to determine the sufficient statistics of the process. Even though the exposition in Section 3.1.2 follows such generic form, further insights into the nature of the ARMA(p, q) covariance matrix and the transition density are deemed relevant.

We recall here one of the prominent properties of ARMA(p, q) processes: their short-memory. Namely, that most of the relevant information in these models is contained within the most recent past only. Hence, let us elaborate on the features of the autocovariance function $\gamma(\tau)$ of any ARMA(p, q) process.

The exact form of the covariance matrix for the general ARMA(p, q) process is, in fact, intractable [93, 94]. Despite the lack of a generic analytical solution, the following statements about the autocovariance function do hold [96]:

- For AR processes, the autocovariance function decays exponentially.
- For MA processes, the autocovariance function is zero after the first q lags.
- For the general ARMA(p, q), the autocovariance function decays exponentially for lags bigger than $m = max(p,q)$.

In summary, the dependence of ARMA(p, q) models on past samples decays exponentially. Even if the sufficient statistics in Equation 30 are explicitly written in terms of the full past history $x_{1:t}$, the dependence on past samples decays exponentially. That is, it is negligible after a certain lag. This short-memory property is evident from the behavior of the recursive terms in Section 3.1.3 too: as $t \to \infty$, $\gamma_t \to 0$ and $\kappa_t \to 0$, thus $\beta_t \to 0$ and $\alpha_t \to 0$.

Furthermore, notice that even if all the information is contained in the covariance matrix, the key computation is on the transition density. As a matter of fact, the $\alpha_t = -\widetilde{\Sigma}_t^{-1}\widetilde{\gamma}_t$ term is critical in the computation of the sufficient statistics of the transition density of any ARMA(p, q) process. Let us study its properties.

On the one hand, for AR(p) models, it can be shown that the information requirement for the transition density is not the full past history, but only the last p time instants. That is, only knowledge of the previous p time-series values $x_{t-p:t-1}$ suffices.

Precisely, the sufficient statistics in Equation 30 for an AR(p) process simplify to

$$\begin{cases} \mu_{x_{t+1}|x_{1:t}} = \sum_{i=1}^{p} a_i x_{t-i} \,, \\ c_{t+1} = \sigma_u^2. \end{cases} \tag{32}$$

The transition density for the AR(p) model is linear on the past p values of the series. Particularly, for Gaussian innovations processes, we have

$$f(x_t|x_{1:t-1}) = f(x_t|x_{t-p:t-1}) = \mathcal{N}\left(x_t \left| \sum_{i=1}^{p} a_i x_{t-i}, \sigma_u^2 \right.\right). \quad (33)$$

On the other hand, the introduction of the MA(q) part complicates the problem and, for such cases, the information requirement is indeed infinite. That is, even if the MA auto-covariance function is non-zero only for lags smaller than q, the resulting conditional factors for both MA(q) and ARMA(p, q) models are not truncated. Once again, note that the analytical determination of such generic factors is not tractable.

However, not all is lost, as the key conditional factor in both sufficient statistics, that is, α_t, shows an exponential decay form for stationary (and invertible) ARMA(p, q) processes. The exact shape of the factor depends both on the model orders (p, q) and the actual ARMA parameter values.

Extensive evaluation of the key factor α_t (graphs included in Appendix B) provide meaningful insights. Precisely, these are the conclusions extracted:

1. The envelope of α_t (i.e., the curve outlining its extremes), decays to zero exponentially.
2. The rate of the decay depends on the model orders p and q but, more importantly, on the MA parameter values b_j.
3. The influence of the AR parameters a_i is primarily on the shape of α_t, but not the rate of the decay.
4. For most of the invertibility range of the ARMA model, i.e., zeros within the unit circle, α_t decays very quickly.
5. Only when the zeros are close to the unit circle, α_t decays more slowly.

These claims are further supported by results summarized in Table 1 for general ARMA(p, q) models and in Table 2, Table 3 and Table 4 for different ARMA(1, 1) parameterizations. The tables show the maximum time-lag τ_{max} after which the α_t value is smaller than a given percentage of the maximum contribution: i.e., $|\alpha_\tau| < \eta \cdot \max\{\alpha_\tau\}$, $\forall \tau > \tau_{max}$.

All in all, the short-memory property of ARMA(p, q) models has been established. In other words, most of the information

ARMA(p,q)	Impact Factor			
	$\eta = 0.1$	$\eta = 0.05$	$\eta = 0.01$	$\eta = 0.001$
AR(1)	1	1	1	1
AR(2)	2	2	2	2
MA(1)	4	5	7	10
MA(2)	4	5	8	11
ARMA(1,1) $a_1 = 0.8, b_1 = 0.5$	4	5	7	10
ARMA(1,1) $a_1 = -0.8, b_1 = 0.5$	4	5	7	10
ARMA(1,1) $a_1 = 0.8, b_1 = -0.5$	4	5	7	10
ARMA(1,1) $a_1 = -0.8, b_1 = -0.5$	4	5	7	10
ARMA(1,2)	5	6	8	11
ARMA(2,1)	3	4	7	10
ARMA(2,2)	4	5	8	11
ARMA(3,1)	4	5	7	10
ARMA(3,2)	5	6	8	11
ARMA(1,3)	2	4	6	10
ARMA(2,3)	3	5	6	10
ARMA(3,3)	2	4	7	11
ARMA(4,1)	3	3	6	9
ARMA(4,2)	4	5	7	10
ARMA(4,3)	1	1	3	7
ARMA(4,4)	3	4	7	12
ARMA(1,4)	2	4	8	12
ARMA(2,4)	2	4	7	11
ARMA(3,4)	2	4	6	10

Table 1: τ_{max} for different ARMA(p, q) models.

ARMA(1,1)	b_1																							
	-0.99	-0.95	-0.9	-0.85	-0.8	-0.75	-0.7	-0.6	-0.5	-0.4	-0.2	0	0.2	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
-0.99	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	0
-0.95	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	0	229
-0.9	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	0	45	229
-0.85	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	0	22	45	229
-0.8	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	0	15	22	45	229
-0.75	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	0	11	15	22	45	229
-0.7	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	0	9	11	15	22	45	229
-0.6	229	45	22	15	11	9	7	5	4	3	2	1	2	3	4	0	7	9	11	15	22	45	229
-0.5	229	45	22	15	11	9	7	5	4	3	2	1	2	3	0	5	7	9	11	15	22	45	229
-0.4	229	45	22	15	11	9	7	5	4	3	2	1	2	0	4	5	7	9	11	15	22	45	229
-0.2	229	45	22	15	11	9	7	5	4	3	2	1	0	3	4	5	7	9	11	15	22	45	229
0	229	45	22	15	11	9	7	5	4	3	2	0	2	3	4	5	7	9	11	15	22	45	229
0.2	229	45	22	15	11	9	7	5	4	3	0	1	2	3	4	5	7	9	11	15	22	45	229
0.4	229	45	22	15	11	9	7	5	4	0	2	1	2	3	4	5	7	9	11	15	22	45	229
0.5	229	45	22	15	11	9	7	5	0	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.6	229	45	22	15	11	9	7	0	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.7	229	45	22	15	11	9	0	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.75	229	45	22	15	11	0	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.8	229	45	22	15	0	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.85	229	45	22	0	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.9	229	45	0	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.95	229	0	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229
0.99	0	45	22	15	11	9	7	5	4	3	2	1	2	3	4	5	7	9	11	15	22	45	229

(a_1)

Table 2: τ_{max} for different ARMA(1,1) parameterizations, with impact factor $\eta = 0.1$.

ARMA(1,1)		b_1																						
		-0.99	-0.95	-0.9	-0.85	-0.8	-0.75	-0.7	-0.6	-0.5	-0.4	-0.2	0	0.2	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
a_1	-0.99	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	0
	-0.95	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	0	297
	-0.9	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	0	59	297
	-0.85	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	0	29	59	297
	-0.8	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	0	19	29	59	297
	-0.75	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	0	14	19	29	59	297
	-0.7	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	0	11	14	19	29	59	297
	-0.6	297	59	29	19	14	11	9	6	5	4	2	1	2	4	5	0	9	11	14	19	29	59	297
	-0.5	297	59	29	19	14	11	9	6	5	4	2	1	2	4	0	6	9	11	14	19	29	59	297
	-0.4	297	59	29	19	14	11	9	6	5	4	2	1	2	0	5	6	9	11	14	19	29	59	297
	-0.2	297	59	29	19	14	11	9	6	5	4	2	1	0	4	5	6	9	11	14	19	29	59	297
	0	297	59	29	19	14	11	9	6	5	4	2	0	2	4	5	6	9	11	14	19	29	59	297
	0.2	297	59	29	19	14	11	9	6	5	4	0	1	2	4	5	6	9	11	14	19	29	59	297
	0.4	297	59	29	19	14	11	9	6	5	0	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.5	297	59	29	19	14	11	9	6	0	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.6	297	59	29	19	14	11	9	0	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.7	297	59	29	19	14	11	0	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.75	297	59	29	19	14	0	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.8	297	59	29	19	0	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.85	297	59	29	0	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.9	297	59	0	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.95	297	0	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297
	0.99	0	59	29	19	14	11	9	6	5	4	2	1	2	4	5	6	9	11	14	19	29	59	297

Table 3: τ_{max} for different ARMA(1,1) parameterizations, with impact factor $\eta = 0.05$.

ARMA(1,1)		b_1																						
		-0.99	-0.95	-0.9	-0.85	-0.8	-0.75	-0.7	-0.6	-0.5	-0.4	-0.2	0	0.2	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	0.95	0.99
a_1	-0.99	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	0
	-0.95	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	0	438
	-0.9	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	0	90	434
	-0.85	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	0	44	90	433
	-0.8	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	0	29	44	90	432
	-0.75	430	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	0	21	29	44	90	432
	-0.7	431	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	0	17	21	29	44	90	432
	-0.6	431	90	44	29	21	17	13	10	7	6	3	1	3	6	7	0	13	17	21	29	44	90	431
	-0.5	431	90	44	29	21	17	13	10	7	6	3	1	3	6	0	10	13	17	21	29	44	90	431
	-0.4	431	90	44	29	21	17	13	10	7	6	3	1	3	0	7	10	13	17	21	29	44	90	431
	-0.2	431	90	44	29	21	17	13	10	7	6	3	1	0	6	7	10	13	17	21	29	44	90	431
	0	431	90	44	29	21	17	13	10	7	6	3	0	3	6	7	10	13	17	21	29	44	90	431
	0.2	431	90	44	29	21	17	13	10	7	6	0	1	3	6	7	10	13	17	21	29	44	90	431
	0.4	431	90	44	29	21	17	13	10	7	0	3	1	3	6	7	10	13	17	21	29	44	90	431
	0.5	431	90	44	29	21	17	13	10	0	6	3	1	3	6	7	10	13	17	21	29	44	90	431
	0.6	431	90	44	29	21	17	13	0	7	6	3	1	3	6	7	10	13	17	21	29	44	90	431
	0.7	432	90	44	29	21	17	0	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	431
	0.75	432	90	44	29	21	0	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430
	0.8	432	90	44	29	0	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430
	0.85	433	90	44	0	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430
	0.9	434	90	0	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430
	0.95	438	0	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430
	0.99	0	90	44	29	21	17	13	10	7	6	3	1	3	6	7	10	13	17	21	29	44	90	430

Table 4: τ_{max} for different ARMA(1,1) parameterizations, with impact factor $\eta = 0.001$.

required to predict the next state is contained in a (relatively) small number of recent samples. Cautious consideration of this assertion should be taken when MA parameters $b_j \to 1$, that is, when the zeros of the process are almost on the unit circle.

SMC METHOD FOR LATENT ARMA(p, q) PROCESSES

We now consider a state-space model where the state is represented by an ARMA process of known model order and parameters, and the observations are nonlinear functions of the state. That is, the model order (p, q); the ARMA parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$; and the state innovation's sufficient statistics, i.e., $\mathbb{E}\{u_t\}$ and $\mathrm{Var}\{u_t\}$, are all known.

Mathematically, a state-space model with a latent ARMA(p, q) time-series is represented by

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j}, \\ y_t = h(x_t, v_t), \end{cases} \tag{34}$$

where u_t and v_t are the innovations of the state and space processes, respectively.

State-space models as in Equation 34 are very generic, due to (1) the flexibility provided by the ARMA(p, q) model in capturing the dynamics of various time-series, and (2) the considered generic observation function $h(x_t, v_t)$. The only restriction made about $h(\cdot)$ is that the resulting likelihood function, $f(y_t|x_t)$, must be known up to a proportionality constant.

Given the observations up to time instant t, $y_{1:t}$, the goal is to sequentially estimate the posterior distribution of x_t, $f(x_t|y_{1:t})$ and to obtain the forecasting distribution $f(y_{t+1}|y_{1:t})$.

As previously explained in Section 2.4, SMC techniques are suitable for such inference problems in nonlinear and non-Gaussian systems.

In general, hidden linear stochastic processes with nonlinear observations have been extensively studied. Some have focused on Bayesian analysis [58], while others have used quasi-maximum likelihood (QML) type approximating techniques [17]. In engineering and statistics, hidden AR processes have been investigated with PF methods [22, 45, 73]. The importance of considering the MA part (correlated noise) is justified by the memory properties exhibited by many real-life time-series [80].

The estimation of hidden ARMA processes is a much more challenging task than the estimation of AR processes. The increased challenges are due to, amongst others, the nonlinearities induced by the MA part. Even when the ARMA process is directly observed, the derivation of the exact parameter densities is intractable and thus, approximations have to be used, as for example in [24, 75]. The problem is certainly much more difficult when the ARMA process is not observed, as in this dissertation. If in addition, observations are nonlinear functions of the ARMA process, one has to employ advanced techniques for sequential estimation.

In [47], hidden ARMA processes were studied and the objective was to estimate them. Two classes of problems were considered, one where the processes were observed via linear functions of the states and another, where the functions were nonlinear. For the first class, optimal filtering methods were presented and, for the latter, approximation methods based on model transformations and the use of the QML and the importance sampling principles were proposed.

In this dissertation, we resort to SMC methods [7], that have the capacity of overcoming these difficulties and have already been successfully applied in somewhat similar state-space models [78, 88, 97].

SMC for stationary latent ARMA(p, q) processes

We propose an SMC method for inference of latent stationary ARMA(p, q) processes with known parameters. To do so, we leverage the analysis in Section 3.1.2 and, in particular, the transition density in Equation 29 with sufficient statistics computed as in Equation 30. Such density allows for propagation of samples from time instant t to $t+1$ and thus, an SMC method can be readily derived (see Table 5 for details).

Note that both the derivation in Section 3.1.2 and the PF presented in Table 5 consider Gaussian state innovations. The generalization to other noise distributions is straightforward, as long as (1) the resulting ARMA process is stationary; (2) the innovations can be accurately described by second order statistics; and (3) the resulting densities are closed under formation of conditional distributions.

PF FOR LATENT STATIONARY ARMA, KNOWN PARAMETERS

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state (to avoid sample degeneracy) by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

(In principle, this step does not have to be performed at every time instant.)

4. Propagate the particles by sampling from the transition density, given the previous stream of (resampled) particles:

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}) = \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right),$$
$$\text{where } \begin{cases} \mu_{x_{t+1}|x_{1:t}} = \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \overline{x}_{1:t}, \\ c_{t+1} = \widetilde{\gamma}(0) - \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t. \end{cases}$$

5. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 5: PF for latent stationary ARMA with known parameters.

SMC for generic latent ARMA(p, q) processes

In this section, a more flexible case is studied, where the stationarity assumption of the previous Section 3.2.1 is dropped. Hence, generic distributions for state innovations u_t are considered. With this new approach, inference in a wider class of latent time-series that do not fall into the category of WSS processes is possible.

The foundation for this new SMC method is to drift the sampling step from the state variables to the innovation process itself. A visual justification for this perspective is provided by inspecting the problem in hand from a Probabilistic Graphical Model (PGM) point of view.

As a toy example, let us consider a hidden ARMA(1,1) state process. Mathematically, such a state-space model is formulated as

$$\begin{cases} x_t = a_1 x_{t-i} + u_t + b_1 u_{t-1} \,, \\ y_t = h(x_t, v_t) \,, \end{cases} \tag{35}$$

which is represented graphically in Figure 9.

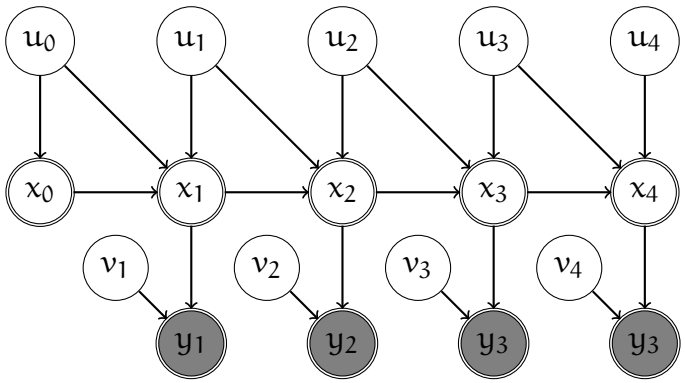


Figure 9: Graphical representation of the ARMA(1,1) model.

Recall that the single-circled variables are stochastic in nature, while the double-circled ones are deterministic.

The PGM representation of Figure 9 illustrates the deterministic nature of the state, once knowledge of the noise and the parameters is provided. That is, conditioned on the previous state x_{t-1}, the actual and past innovations u_t and u_{t-1}, and the parameters a and b, the density of the actual state is nothing but a Dirac's delta function:

$$f(x_t | x_{t-1}, u_t, u_{t-1}, a, b) = \delta\left(x_t - [a x_{t-1} + u_t + b u_{t-1}]\right). \tag{36}$$

Such a deterministic dependency readily facilitates the computation of the joint transition density of the state and innovation processes:

$$\begin{aligned} f(x_t, u_t | x_{t-1}, u_{t-1}, a, b) &= f_u(u_t) \cdot f(x_t | x_{t-1}, u_t, u_{t-1}, a, b) \\ &= f_u(u_t) \cdot \delta\left(x_t - [a x_{t-1} + u_t + b u_{t-1}]\right). \end{aligned} \tag{37}$$

Therefore, by considering the density of not only the state, but also of the driving innovation, a new SMC method is proposed for generic ARMA(p, q) models with arbitrary state noise processes.

Given a generic ARMA(p, q) latent state process

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j} \qquad \text{where } u_t \sim f_u(u), \tag{38}$$

the devised PF operates by first sampling the noise u_t, then determining the joint state-innovation distribution, conditioned on those noise samples and the given parameter values

$$f(x_t | a, b, u_{t-m:t}) = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + f_u(u), \tag{39}$$

where $m = \max(p, q)$.

The main advantages of this new proposed SMC method (fully described in Table 6) are:

- It is applicable to arbitrary state innovations, as long as samples can be drawn from its distribution.
- It does not assume stationarity.
- It provides minimal added complexity (only one additional computational step), in exchange for great flexibility.

PF FOR LATENT GENERAL ARMA, KNOWN PARAMETERS

1. At time instant t, consider the random measure

$$f_t^M(x_t, u_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\begin{pmatrix} x_t \\ u_t \end{pmatrix} - \begin{pmatrix} x_t^{(m)} \\ u_t^{(m)} \end{pmatrix}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state (to avoid sample degeneracy) by drawing from a categorical distribution defined by the random measure

$$\begin{pmatrix} \overline{x}_t^{(m)} \\ \overline{u}_t^{(m)} \end{pmatrix} \sim f_t^M(x_t, u_t), \text{ where } m = 1, \cdots, M.$$

   (In principle, this step does not have to be performed at every time instant.)

4. Propagate the innovation particles from its distribution

$$u_{t+1}^{(m)} \sim f_u(u).$$

5. Propagate the state particles, given the previous (resampled) state and noise particles:

$$x_{t+1}^{(m)} = \sum_{i=1}^{p} a_i \overline{x}_{t+1-i}^{(m)} + \sum_{j=1}^{q} b_j \overline{u}_{t+1-j}^{(m)} + u_{t+1}^{(m)}.$$

6. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}, u_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(\begin{pmatrix} x_{t+1} \\ u_{t+1} \end{pmatrix} - \begin{pmatrix} x_{t+1}^{(m)} \\ u_{t+1}^{(m)} \end{pmatrix}\right).$$

Table 6: PF for latent general ARMA with known parameters.

EVALUATION

We evaluate the performance of the proposed methods for different latent time-series. In order to provide sound evidence of the validity and convenience of the proposed SMC solution, we simulate the stochastic log-volatility (Stochastic Volatility (SV)) model, popular in the study of nonlinear state-space models [1, 40].

The observations are zero-mean with time-varying log-variance equal to the ARMA(p, q) state process. More specifically,

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ y_t = e^{\frac{x_t}{2}} v_t, \end{cases} \tag{40}$$

where v_t is a standard Gaussian variable and the state noise u_t is a zero-mean Gaussian variable. We evaluate the proposed method for both stationary ARMA and more generic cases, always under the assumption of known parameters.

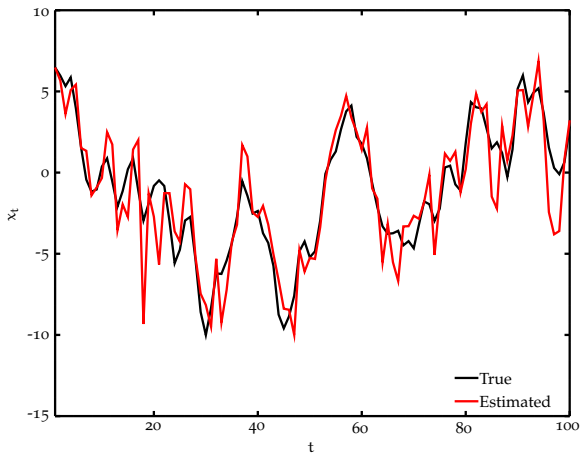
Given knowledge of all the ARMA parameters, the state equation is linear and Gaussian and thus, one only needs to deal with the nonlinearities in the observation equation. A family of very popular alternative approaches is based on the KF, such as the EKF [3], the UKF [61] and other Sigma-Point Kalman filters [74]. However, as reported in [105], these methods fail when addressing the SV model. The reason is that they never update the prior beliefs because the Kalman gain is null.

The first goal is to justify the use of the proposed SMC methods for inference of latent time-series in nonlinear systems. We evaluate the proposed methods by comparing them to an alternative based on [47] and [106]. Due to the failure of KF methods for estimation in SV models, a transformation of the model is suggested to circumvent the nonlinearity. In this transformed state-space

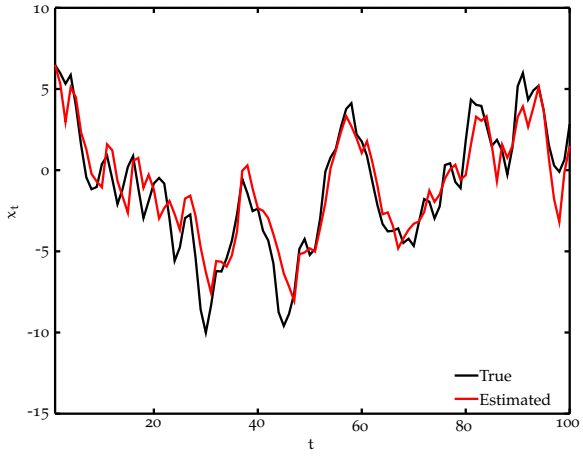
$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ \log(y_t^2) = x_t + z_t, \end{cases} \tag{41}$$

an approximated KF can be applied. Full details on the implemented approximated Kalman Filter (`Approx.KF`) are available in Appendix C.

We first compare the performance of this `Approx.KF` and the stationary PF introduced in Table 5. As shown in Figure 10, both methods track the hidden state with acceptable accuracy.



(a) `Approx.KF` state tracking.



(b) Proposed PF state tracking.

Figure 10: True (black) and estimated (red) state for `Approx.KF` and proposed SMC for a latent ARMA(1,1) model.

A more thorough evaluation of the proposed SMC is presented in Table 7. Unless otherwise indicated, 1000 particles are used for the PF method and the results are Mean-Squared Error (MSE)s of the hidden state estimates, averaged over 100 realizations of 500 samples long time-series. There, the following different ARMA(p, q) parameterizations are scrutinized:

1. AR(1) with $a_1 = 0.8$,
2. AR(2) with $a_1 = 0.8, a_2 = 0.15$,
3. MA(1) with $b_1 = 0.5$,
4. MA(2) with $b_1 = 0.8, b_2 = 0.15$,

5. ARMA(1,1) with $a_1 = 0.8, b_1 = 0.5$,
6. ARMA(1,1) with $a_1 = -0.8, b_1 = 0.5$,
7. ARMA(1,1) with $a_1 = 0.8, b_1 = -0.5$,
8. ARMA(1,1) with $a_1 = -0.8, b_1 = -0.5$,
9. ARMA(1,2) with $a_1 = 0.8, b_1 = 0.8, b_2 = 0.15$,
10. ARMA(2,1) with $a_1 = 0.8, a_2 = 0.15, b_1 = 0.5$,
11. ARMA(2,2) with $a_1 = 0.8, a_2 = 0.15, b_1 = 0.9, b_2 = 0.15$,
12. ARMA(3,1) with $a_1 = 0.5, a_2 = 0.3, a_3 = 0.15, b_1 = 0.5$,
13. ARMA(3,2) with $a_1 = 0.5, a_2 = 0.3, a_3 = 0.15, b_1 = 0.8, b_2 = 0.15$,
14. ARMA(1,3) with $a_1 = 0.8, b_1 = 0.5, b_2 = 0.3, b_3 = 0.15, b_2 = 0.15$,
15. ARMA(2,3) with $a_1 = 0.8, a_2 = 0.15, b_1 = 0.5, b_2 = 0.3, b_3 = 0.15, b_2 = 0.15$,
16. ARMA(3,3) with $a_1 = 0.5, a_2 = 0.3, a_3 = 0.15, b_1 = 0.5, b_2 = 0.3, b_3 = 0.15, b_2 = 0.15$,
17. ARMA(4,1) with $a_1 = 0.5, a_2 = 0.2, a_3 = 0.15, a_4 = 0.1, b_1 = 0.5$,
18. ARMA(4,2) with $a_1 = 0.5, a_2 = 0.2, a_3 = 0.15, a_4 = 0.1, b_1 = 0.8, b_2 = 0.15$,
19. ARMA(4,3) with $a_1 = 0.5, a_2 = 0.2, a_3 = 0.15, a_4 = 0.1, b_1 = 0.5, b_2 = 0.3, b_3 = 0.15$,
20. ARMA(4,4) with $a_1 = 0.5, a_2 = 0.2, a_3 = 0.15, a_4 = 0.1, b_1 = 0.5, b_2 = 0.2, b_3 = 0.15, b_4 = 0.1$,
21. ARMA(1,4) with $a_1 = 0.8, b_1 = 0.5, b_2 = 0.2, b_3 = 0.15, b_4 = 0.1$,
22. ARMA(2,4) with $a_1 = 0.5, a_2 = 0.15, b_1 = 0.5, b_2 = 0.2, b_3 = 0.15, b_4 = 0.1$,
23. ARMA(3,4) with $a_1 = 0.5, a_2 = 0.3, a_3 = 0.15, b_1 = 0.5, b_2 = 0.2, b_3 = 0.15, b_4 = 0.1$.

ARMA(p,q)	State estimation error (MSE)	
	`Approx.KF`	`Stationary PF`
AR(1)	1.3484	1.0891
AR(2)	1.566	1.1946
MA(1)	1.9067	1.013
MA(2)	1.3061	0.98962
ARMA(1,1) $a_1 = 0.8, b_1 = 0.8$	3.9234	1.6363
ARMA(1,1) $a_1 = 0.8, b_1 = -0.8$	0.81783	0.73751
ARMA(1,1) $a_1 = -0.8 b_1 = 0.8$	0.83402	0.74818
ARMA(1,1) $a_1 = -0.8, b_1 = -0.8$	3.9437	1.7715
ARMA(1,2)	3.4611	1.6895
ARMA(1,3)	2.3234	1.528
ARMA(1,4)	2.3356	1.5188
ARMA(2,1)	3.7798	1.7046
ARMA(2,2)	3.3571	1.7668
ARMA(2,3)	2.3192	1.6113
ARMA(2,4)	2.3422	1.61
ARMA(3,1)	3.1689	1.4607
ARMA(3,2)	2.5131	1.4801
ARMA(3,3)	1.8402	1.3985
ARMA(3,4)	1.7436	1.3415
ARMA(4,1)	3.2524	1.4442
ARMA(4,2)	2.6201	1.4749
ARMA(4,3)	1.7059	1.3142
ARMA(4,4)	1.7228	1.3086

Table 7: Filtering MSE for the `ApproxKF` and the stationary PF.

From Table 7, we see the advantage of using the proposed PF in Table 5, as it outperforms the alternative for all the studied ARMA(p, q) models. This superior accuracy, however, comes with the additional computational complexity of the Monte Carlo sampling of the method.

Note that these results have been obtained by using the full covariance matrix of the ARMA(p, q) models for the transition density in Equation 29. However, as noted in Section 3.1.4, most of the information in ARMA models is contained within only the most recent samples. Thus, one may be tempted to truncate the covariance matrix up to a maximum lag τ_{max}.

The justification comes from the short-memory feature of the studied ARMA(p, q) models, so that we save in computation and memory without sacrificing in performance. The modification consists in truncating the computation of the sufficient statistics in Equation 30, i.e., $\mu_{x_{t+1}|x_{1:t}}$ and c_{t+1}, to a maximum lag τ_{max}.

This reduces the computational cost (one computes $\gamma(\tau)$ only for a relatively short window $\tau = 0, \cdots, \tau_{max}$), while incurring a negligible information loss ($\gamma(\tau) \approx 0, \tau > \tau_{max}$).

The computation of the sufficient statistics based on a truncated sequence is approximately equal to the sufficient statistics obtained from the full sequence:

$$\begin{aligned} \mu_{x_{t+1}|x_{1:t}} &= \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} x_{1:t} \\ &\approx \widetilde{\gamma}_{\tau_{max}}^\top \widetilde{\Sigma}_{\tau_{max}}^{-1} x_{t-\tau_{max}:t}, \\ c_{t+1} &= \widetilde{\gamma}(0) - \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t \\ &\approx \widetilde{\gamma}(0) - \widetilde{\gamma}_{\tau_{max}}^\top \widetilde{\Sigma}_{\tau_{max}}^{-1} \widetilde{\gamma}_{\tau_{max}}. \end{aligned} \tag{42}$$

In Figure 11, we study the impact of truncating the covariance matrix to different lags (i.e., τ_{max}). As expected, no significant performance difference is seen for lags greater than 10 in all the studied models, while the computational burden of recursively updating the statistics in Equation 30 is avoided. Further details on the MSE for different truncation lags are collected in Table 8.

The results presented in Table 8 reinforce the statements in Section 3.1.4 about the short-memory properties of the ARMA models, i.e., most of the information is confined within the last few samples.

Note that, by looking at Table 1, informed guesses on what maximum lag to use are attainable. From the values provided there, truncation around $\tau_{max} \approx 10$ seems reasonable. As a matter of fact, there is no PF performance improvement observed for lags $\tau_{max} > 10$ in Table 8.

We reiterate that there is no loss of information with autocovariance truncation by showing the evolution of the estimation error over time in Figure 12 (averaged over 10 realizations of 10,000 samples long ARMA(1,1) processes). There, we observe that the estimation is unbiased and that it does not vary over time for any window size (the averaged error is 0.0043 for $\tau_{max} = 25$ and 0.0051 for $\tau_{max} = 100$).

ARMA(p,q)	State estimation error (MSE)									
	$\tau_{max} = m$	$\tau_{max} = 5$	$\tau_{max} = 10$	$\tau_{max} = 15$	$\tau_{max} = 20$	$\tau_{max} = 30$	$\tau_{max} = 40$	$\tau_{max} = 50$	$\tau_{max} = 75$	$\tau_{max} = t$
AR(1)	1.1029	1.1032	1.1031	1.1041	1.1031	1.1038	1.1027	1.1033	1.1033	1.1028
AR(2)	1.2305	1.2286	1.2295	1.23	1.2291	1.2301	1.2299	1.2299	1.2303	1.2294
MA(1)	1.0263	1.0099	1.01	1.0103	1.0095	1.0093	1.0095	1.0092	1.0098	1.0096
MA(2)	1.0067	1.0016	1.0024	1.0023	1.002	1.0027	1.0016	1.0021	1.0014	1.0012
ARMA(1,1) $a_1 = 0.8, b_1 = 0.8$	1.7608	1.6825	1.6834	1.6818	1.6796	1.6794	1.6815	1.6829	1.6816	1.6806
ARMA(1,1) $a_1 = -0.8, b_1 = 0.8$	0.7465	0.74623	0.74632	0.74629	0.74645	0.74647	0.7467	0.74631	0.74641	0.74599
ARMA(1,1) $a_1 = 0.8, b_1 = -0.8$	0.73987	0.74017	0.7399	0.7402	0.7408	0.74034	0.74062	0.74042	0.74042	0.74042
ARMA(1,1) $a_1 = -0.8, b_1 = -0.8$	1.8417	1.7807	1.7791	1.7791	1.7777	1.7795	1.779	1.7786	1.7779	1.7775
ARMA(1,2)	1.752	1.7365	1.7352	1.7358	1.7336	1.7353	1.736	1.7348	1.7378	1.7339
ARMA(2,1)	1.8374	1.8296	1.8283	1.8269	1.8285	1.8254	1.8292	1.8273	1.8294	1.8275
ARMA(2,2)	1.8793	1.871	1.8708	1.8681	1.8697	1.8716	1.8705	1.8691	1.8706	1.8702
ARMA(3,1)	1.58	1.5724	1.569	1.5704	1.5707	1.5707	1.5706	1.571	1.569	1.5694
ARMA(3,2)	1.5803	1.5767	1.5777	1.5777	1.5759	1.5758	1.5749	1.5758	1.5776	1.5769
ARMA(1,3)	1.5759	1.573	1.5748	1.5733	1.5723	1.5739	1.5742	1.5738	1.5732	1.5712
ARMA(2,3)	1.761	1.756	1.7581	1.7584	1.7542	1.7534	1.7538	1.7581	1.7556	1.757
ARMA(3,3)	1.4975	1.4959	1.4944	1.4943	1.4942	1.4975	1.4958	1.4954	1.4974	1.4935
ARMA(4,1)	1.5293	1.5261	1.5256	1.5253	1.5247	1.5273	1.5276	1.5267	1.5266	1.5264
ARMA(4,2)	1.6044	1.6044	1.6078	1.6027	1.6071	1.6054	1.6046	1.6041	1.6066	1.6053
ARMA(4,3)	1.4646	1.4655	1.4651	1.4664	1.4663	1.4634	1.4656	1.4634	1.4641	1.4648
ARMA(4,4)	1.5206	1.5199	1.5183	1.5193	1.518	1.5204	1.5184	1.5184	1.5221	1.5201
ARMA(1,4)	1.5503	1.5519	1.5493	1.5505	1.5513	1.5499	1.5504	1.551	1.551	1.5494
ARMA(2,4)	1.8102	1.8126	1.8087	1.8118	1.8103	1.8078	1.8095	1.8059	1.8115	1.8105
ARMA(3,4)	1.5032	1.5008	1.5012	1.5027	1.5002	1.4999	1.5033	1.503	1.499	1.5015

Table 8: Filtering MSE of the stationary PF for different truncation lags.

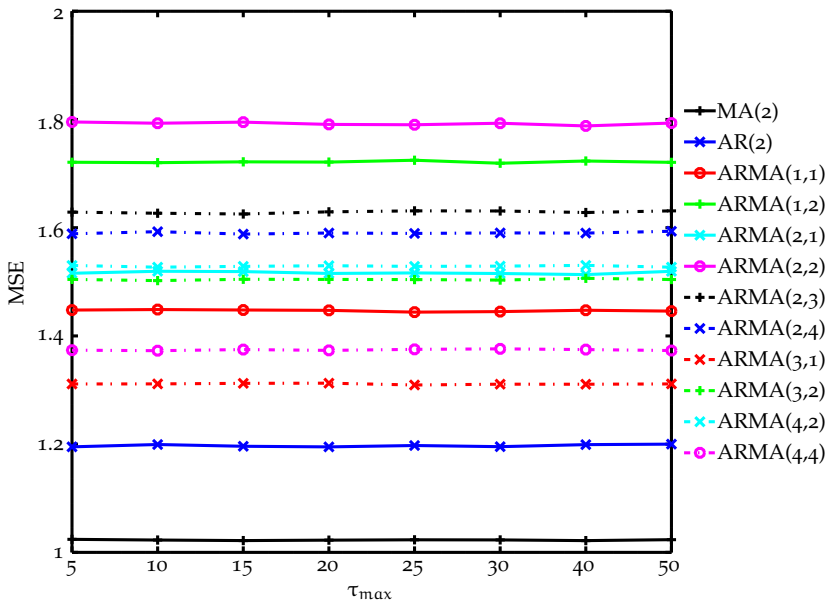
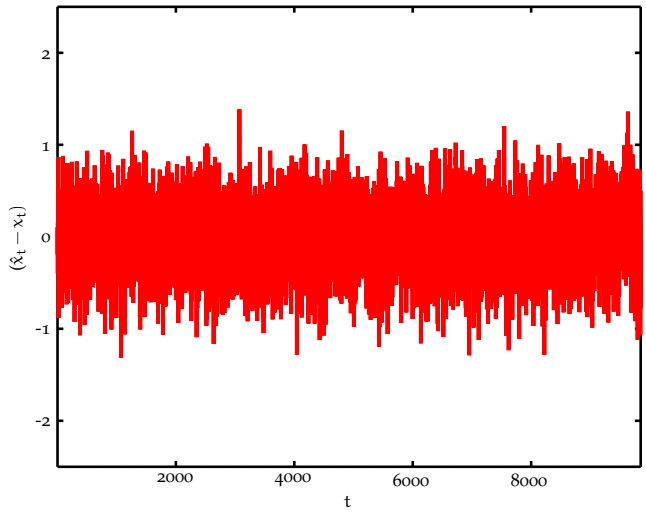
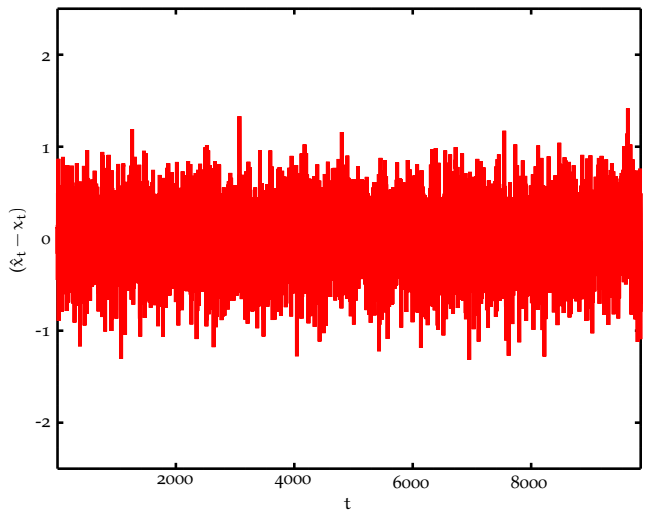


Figure 11: Impact of covariance truncation in the MSE performance of the proposed SMC.



(a) ARMA(1,1) with $\tau_{max} = 25$.



(b) ARMA(1,1) $\tau_{max} = 100$.

Figure 12: Estimation error of the proposed SMC as a function of time.

In summary, we conclude that it is reasonable to truncate the covariance matrix of ARMA(p, q) models, based on a preliminary analysis of the model at hand. The filtering accuracy is kept consistent, while both memory and computational requirements are significantly reduced.

Finally, we evaluate the alternative SMC method proposed. That is, we now study the performance of the more general PF described in Table 6.

On the one hand, we show that the performance of the SMC in Table 6 is identical to that of the SMC method in Table 5 for all the studied cases (see Table 9). Thus, we can conclude that both alternative methods are accurate solutions for inference of latent ARMA(p, q) time-series with Gaussian iid innovations.

On the other hand, it is important to emphasize the extra flexibility offered by the general PF as in Table 6. To illustrate this, three different latent ARMA(p, q) models are evaluated, each with different state innovations:

- `Sinusoidal-mean Gaussian`: The state innovation is Gaussian with a periodic mean. Specifically, a sinusoid with period 100 time-instants is used: $\mu_{u_t} = \sin(\frac{2\pi t}{100})$.
- `Time-correlated Gaussian`: The state innovation is Gaussian correlated in time with the previous samples. Specifically, $\rho_{t,t} = 1$ and $\rho_{t,t-1} = 0.5$.
- `Student-t distributed noise`: The state innovation follows a Student t-distribution with $\nu = 2$ degrees of freedom, a null location parameter $\mu = 0$ and a unit scale parameter $\sigma_u^2 = 1$.

Figure 13 provides examples of the tracking accuracy of the general SMC method presented in Table 6. The MSE evaluation for 100 realizations summarized in Table 10 illustrates how, for different Gaussian state innovations, the PF performs comparable to the iid Gaussian benchmark case.

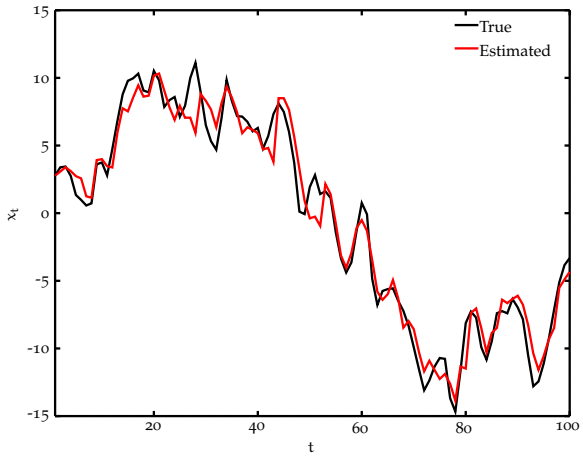
When considering non-Gaussian noises, as in Table 11 for a noise with a Student-t distribution, a performance loss in terms of MSE is observed, due to the *unexpected* events caused by the heavy tails of the distribution. Nonetheless, it is important to notice that the PF is able to recover from this rare events and track accurately the forthcoming values of the latent time-series.

ARMA(p,q)	State estimation error (MSE)	
	General PF	Stationary PF
AR(1)	1.0115	1.0126
AR(2)	1.0125	1.0135
MA(1)	1.1032	1.1031
MA(2)	1.2114	1.2133
ARMA(1,1) a=0.8, b=0.8	1.655	1.658
ARMA(1,1) a=0.8, b=-0.8	0.7377	0.73925
ARMA(1,1) a=-0.8 b=0.8	0.74593	0.74803
ARMA(1,1) a=-0.8, b=-0.8	1.7302	1.7306
ARMA(1,2)	1.7271	1.7259
ARMA(1,3)	1.5467	1.5451
ARMA(1,4)	1.5046	1.506
ARMA(2,1)	1.7336	1.735
ARMA(2,2)	1.7759	1.7795
ARMA(2,3)	1.6271	1.6241
ARMA(2,4)	1.5908	1.5906
ARMA(3,1)	1.4749	1.4746
ARMA(3,2)	1.4963	1.4991
ARMA(3,3)	1.399	1.3997
ARMA(3,4)	1.381	1.3821
ARMA(4,1)	1.4718	1.4707
ARMA(4,2)	1.4565	1.4557
ARMA(4,3)	1.3366	1.3372
ARMA(4,4)	1.306	1.3061

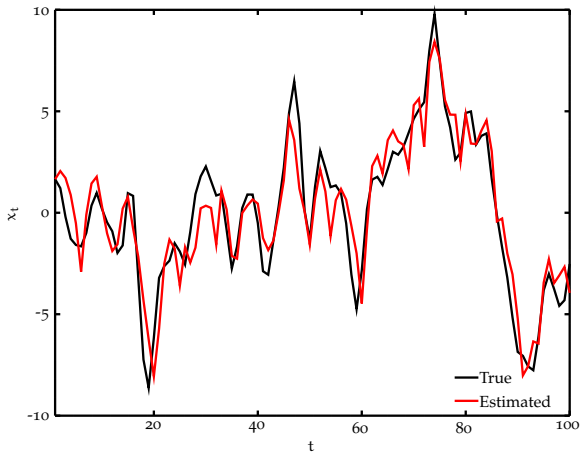
Table 9: Filtering MSE comparison for the stationary and general PFs.

ARMA(p,q)	State estimation error (MSE)		
	iid Gaussian	Sinusoidal-mean Gaussian	Time-correlated Gaussian
AR(1)	1.096	1.0961	1.325
AR(2)	1.0102	1.0114	1.221
MA(1)	1.2116	1.1902	1.4112
MA(2)	0.9955	1.0056	1.2233
ARMA(1,1) $a_1 = 0.8, b_1 = 0.8$	1.693	1.6626	2.0001
ARMA(1,1) $a_1 = 0.8, b_1 = -0.8$	0.74611	0.75225	0.70873
ARMA(1,1) $a_1 = -0.8 b_1 = 0.8$	0.74552	0.74782	0.70628
ARMA(1,1) $a_1 = -0.8, b_1 = -0.8$	1.7997	1.7413	0.60319
ARMA(1,2)	1.7068	1.7273	2.0484
ARMA(1,3)	1.5614	1.5361	1.8443
ARMA(1,4)	1.529	1.5099	1.808
ARMA(2,1)	1.7337	1.7064	2.0199
ARMA(2,2)	1.763	1.7698	2.0868
ARMA(2,3)	1.607	1.627	1.895
ARMA(2,4)	1.5876	1.6198	1.8428
ARMA(3,1)	1.4848	1.4749	1.7553
ARMA(3,2)	1.5079	1.4926	1.7807
ARMA(3,3)	1.397	1.3984	1.6251
ARMA(3,4)	1.3634	1.3744	1.6177
ARMA(4,1)	1.4428	1.4608	1.735
ARMA(4,2)	1.4848	1.4482	1.7196
ARMA(4,3)	1.3391	1.3362	1.5654
ARMA(4,4)	1.318	1.3185	1.5352

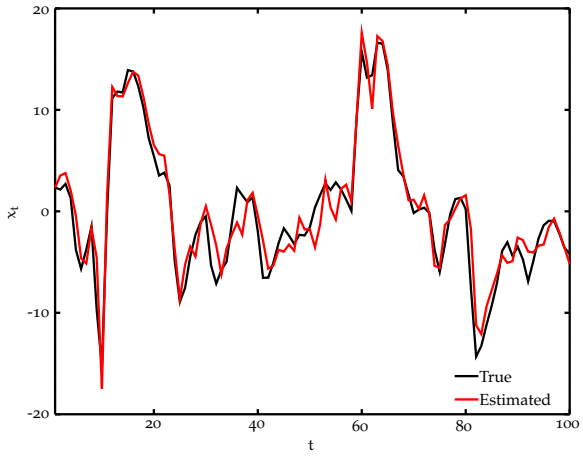
Table 10: Filtering MSE comparison for the general PF with different Gaussian innovations.



(a) Tracking with sinusoidal-mean Gaussian noise.



(b) Tracking with time-correlated Gaussian noise.



(c) Tracking with Student-t distributed noise.

Figure 13: True (black) and estimated (red) state for the proposed generic SMC.

ARMA(p,q)	State estimation error (MSE)
	Student-t noise
AR(1)	10.974
AR(2)	10.179
MA(1)	6.0782
MA(2)	4.2368
ARMA(1,1) $a_1 = 0.8, b_1 = 0.8$	21.16
ARMA(1,1) $a_1 = 0.8, b_1 = -0.8$	4.1639
ARMA(1,1) $a_1 = -0.8 b_1 = 0.8$	8.7618
ARMA(1,1) $a_1 = -0.8, b_1 = -0.8$	109.623
ARMA(1,2)	26.749
ARMA(1,3)	51.237
ARMA(1,4)	55.692
ARMA(2,1)	15.497
ARMA(2,2)	34.883
ARMA(2,3)	4.5758
ARMA(2,4)	36.234
ARMA(3,1)	38.71
ARMA(3,2)	7.2655
ARMA(3,3)	53.092
ARMA(3,4)	10.229
ARMA(4,1)	16.633
ARMA(4,2)	7.7182
ARMA(4,3)	30.58
ARMA(4,4)	21.167

Table 11: Filtering MSE for the general PF with Student-t distributed innovations.

4

LATENT ARMA(p, q) PROCESSES WITH UNKNOWN PARAMETERS

This chapter addresses the analysis of state-space models with a latent stationary ARMA process. In practice, it is unlikely (if not impossible) to have full knowledge of the underlying ARMA parameters. Thus, we relax the restrictive assumptions of Chapter 3 and consider the unknown parameter case. Nonetheless, we do assume that the model order of the process, i.e., (p, q) is known. Now, one needs to deal with the nonlinearities induced by the unknown parameters too and thus, inference of latent ARMA processes gets complicated even more.

The contribution in this chapter is in the novel SMC methods for inference of latent ARMA processes with unknown parameters, which are presented in Section 4.2.1 and Section 4.2.2.

We avoid parameter estimation and exploit Rao-Blackwellization of the unknowns, as explained in Section 4.2. With this technique, the performance of the devised methods is superior, as shown in Section 4.3.

SMC METHODS AND UNKNOWN STATIC PARAMETERS

In this chapter, the problem of estimating hidden ARMA(p, q) processes when their parameters are unknown is considered. Since the adopted SMC methodology requires special care in handling static parameters, we explore an alternative technique known as Rao-Blackwellization.

The main inconvenience for SMC methods is the presence of fixed unknown parameters. It has been extensively reported that PFs suffers when the models contain fixed parameters [72]. In some of the existing work, particles of all the unknown states and parameters are jointly generated [22, 71, 78], and then weighted according to the observations.

However, the performance of the PFs deteriorates, as the parameter does not change over time and, thus, propagating parameter particles becomes troublesome. To overcome such limitations, various methodologies have been suggested, including the use of artificial parameter evolution [49], kernel

smoothing techniques [71], density-assisted particle filtering [39], marginalization [97] or smoothed MAP estimation [90].

Here, we shift our attention from parameter estimation and exploit an alternative technique known as Rao-Blackwellization [23]. With Rao-Blackwellization, one integrates out some of these (or all) static parameters and, in principle, does not generate particles for them.

Rao-Blackwellization is a statistical procedure that guarantees reduced estimation variance of the variables of interest. Thus, it allows for improved estimation of the state process. In a nutshell, it consists of marginalizing some of the variables of the model. This technique has already been applied to PFs [44, 81, 89, 92].

In the problem of interest here, the unknown static parameters are those of the latent ARMA process. That is, the model parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$, and the driving noise variance σ_u^2 (as in Chapter 3, zero-mean innovations are considered here). The goal is to integrate out all these unknowns, so that there is no need for generating particles for them.

In the SMC sampling context, when integration of the unknown variables is carried out analytically, the accuracy of the method is improved. On the one hand, due to the reduced dimensionality of the resulting sampling problem. On the other, due to the Rao-Blackwell theorem [65]. When Rao-Blackwellization of the parameters cannot be implemented analytically, one can resort to a Monte Carlo approximation of the integral.

LATENT ARMA(p, q) PROCESSES WITH UNKNOWN PARAMETERS

In Chapter 3, inference of latent ARMA(p, q) processes has been tackled. To that end, we have resorted to the transition density of the ARMA(p, q) state as the proposal function for the proposed SMC method. This density can only be derived analytically when dealing with processes whose ARMA parameters are known.

However, assuming full knowledge of these is too restrictive. In practice, it is unlikely (if not impossible) to have full knowledge of the underlying ARMA parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$ and innovation variance σ_u^2.

We now address the problem of sequentially estimating the latent ARMA(p, q) state x_t, when the ARMA parameters are unknown. Mathematically, we write

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j}, \\ y_t = h(x_t, v_t), \end{cases} \tag{43}$$

where u_t and v_t are the innovations of the state and space processes respectively, and all the parameters of the state are unknown.

Here, we approach the problem with Rao-Blackwellization. That is, we aim at integrating out all the static unknowns: the ARMA parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$, and the innovation variance σ_u^2.

We work with the marginalized state transition density

$$f(x_{t+1}|x_{1:t}) = \int_{a,b,\sigma_u^2} f(x_{t+1}|x_t, a, b, \sigma_u^2) f(a, b, \sigma_u^2|x_{1:t}) \mathrm{d}a\mathrm{d}b\mathrm{d}\sigma_u^2. \tag{44}$$

However, one cannot obtain the analytical solution to the above integral for the general ARMA(p, q) case. Furthermore, one can separate the integration with respect to the unknown ARMA parameters a and b and the innovation variance σ_u^2.

Thus, we tackle the Rao-Blackwellization of the unknowns separately. In Section 4.2.1, we focus on the ARMA parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$ and, then, study the unknown σ_u^2 case in Section 4.2.2.

Latent ARMA(p, q) processes with unknown a and b

Here, we focus on the inner integral in Equation 44 and integrate out the unknown ARMA parameters a and b. That is, we work with

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) = \int_{a,b} f(x_{t+1}|x_t, a, b, \sigma_u^2) f(a, b|x_{1:t}, \sigma_u^2) \mathrm{d}a\mathrm{d}b. \tag{45}$$

As pointed out before, for the general ARMA(p, q) case, one cannot obtain the analytical solution to the above integral [26, 76]. However, for the AR(p) model, such solution can be derived. We first present the derivation of the AR(p) solution, and then turn our attention to the general ARMA(p, q) model.

Latent AR(p) processes with unknown a

In the latent AR(p) case, the state process x_t is of the form

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t. \tag{46}$$

We aim at marginalizing the unknown a parameter, i.e.,

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) = \int_a f(x_{t+1}|x_t, a, \sigma_u^2) f(a|x_{1:t}, \sigma_u^2) da, \tag{47}$$

where we follow the Bayesian paradigm and use the posterior of the parameters at time instant t.

If we assume an initial constant prior for a that guarantees stationarity, one can derive that the posterior of a for $t > p+1$ follows [65]

$$\begin{aligned} f(a|x_{1:t}, \sigma_u^2) &= \mathcal{N}\left(a|\mu_t, \sigma_u^2 C_t\right), \\ \text{with } &\begin{cases} \mu_t = C_t H_{t-1}^\top x_{p+1:t}, \\ C_t = \left(H_{t-1}^\top H_{t-1}\right)^{-1}, \end{cases} \end{aligned} \tag{48}$$

and

$$H_{t-1} = \begin{pmatrix} x_{t-1} & x_{t-2} & \cdots & x_{t-p} \\ x_{t-2} & x_{t-3} & \cdots & x_{t-p-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_p & x_{p-1} & \cdots & x_1 \end{pmatrix}. \tag{49}$$

Finally, the transition density of interest is

$$\begin{aligned} f(x_{t+1}|x_{1:t}, \sigma_u^2) &= \int f(x_{t+1}|x_{1:t}, a, \sigma_u^2) f(a|x_{1:t}, \sigma_u^2) da \\ &= \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right), \\ \text{where } &\begin{cases} \mu_{x_{t+1}|x_{1:t}} = h_t^\top \mu_t, \\ c_{t+1} = 1 + h_t^\top C_t h_t, \end{cases} \end{aligned} \tag{50}$$

with $h_t^\top = \begin{pmatrix} x_t & x_{t-1} & \cdots & x_{t-p+1} \end{pmatrix}$.

Consequently, an SMC method for inference of latent stationary AR(p) processes with unknown a parameters is proposed. The marginalized transition density in Equation 50 is used as the proposal density for propagating state particles. The details of the proposed SMC method are provided in Table 12.

PF FOR LATENT AR WITH UNKNOWN a

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state (to avoid sample degeneracy) by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

(In principle, this step does not have to be performed at every time instant.)

4. Propagate the particles by sampling from the transition density, given the previous stream of (resampled) particles:

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}, \sigma_u^2) = \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right),$$

$$\text{where } \begin{cases} \mu_{x_{t+1}|x_{1:t}} = h_t^\top \mu_t, \\ c_{t+1} = 1 + h_t^\top C_t h_t, \end{cases}$$

as in Equation 48 and Equation 50.

5. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 12: PF for latent AR(p) with unknown parameters a.

Latent ARMA(p, q) processes with unknown a and b

For the more general ARMA(p, q) case, the goal is to integrate out the static parameters again, in order to obtain the Rao-Blackwellized transition density $f(x_{t+1}|x_{1:t}, \sigma_u^2)$ via Equation 45.

For this case, too, we assume that the prior of parameters a and b is constant (over the region of stability of the ARMA process). However, as already pointed out, there is no analytical solution to this integral, so we proceed with a numerical approach.

Let us denote $\theta = (a^\top, b^\top)^\top$ and approximate the parameter posterior $f(\theta|x_{1:t}, \sigma_u^2)$ by a random measure

$$f(\theta|x_{1:t}, \sigma_u^2) \approx \frac{1}{J} \sum_{j=1}^{J} \delta(\theta - \theta^{(j)}), \tag{51}$$

where $\theta^{(j)}$ is a sample of θ drawn from $f(\theta|x_{1:t}, \sigma_u^2)$ and J is the total number of drawn parameter samples.

With this approximation, the integral with respect to the ARMA parameters in Equation 45 can be rewritten as

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) \approx \frac{1}{J} \sum_{j=1}^{J} f(x_{t+1}|x_{1:t}, a_t^{(j)}, b_t^{(j)}, \sigma_u^2). \tag{52}$$

Note that the subscript t of a and b indicates samples obtained at that time instant, and not that these parameters change over time. The density $f(x_{t+1}|x_{1:t}, a_t^{(j)}, b_t^{(j)}, \sigma_u^2)$ has already been derived in Chapter 3, for a given set of parameter samples.

We now address how to draw samples from the posterior of $f(\theta|x_{1:t}, \sigma_u^2)$. We explain this by assuming that at time instant $t-1$, we already had samples of the parameter θ. That is, we have samples of a and b, which we denote by $a_{t-1}^{(j)}, b_{t-1}^{(j)}$.

We approximate the posterior $f(\theta|x_{1:t}, \sigma_u^2)$ with a Gaussian density. This is justified by the asymptotic Gaussian behavior of parameter posteriors under some regularity conditions [101]. Note that the parameter posterior is indeed Gaussian for AR(p) processes.

The sufficient statistics of the distribution approximating the parameter posterior $f(\theta|x_{1:t}, \sigma_u^2)$ are computed from the particles $a_{t-1}^{(j)}, b_{t-1}^{(j)}$ and the weights $w_{t-1}^{(j)}$, conditioned on the sequence $x_{1:t}$.

Specifically, let us denote the joint state and parameter vector at time instant t as $\rho_t = (x_t\ \theta_t)^\top$. We approximate the joint posterior distribution of ρ_t with a multivariate Gaussian

$$f(\rho_t) \approx \mathcal{N}(\rho_t|\eta_t, Q_t). \tag{53}$$

The sufficient statistics of this density are computed as weighted averages of the available particles

$$\begin{aligned} \eta_t &= \sum_{i=1}^{M} w_t^{(m)} \rho_t^{(m)}, \\ Q_t &= \sum_{i=1}^{M} w_t^{(m)} (\rho_t^{(m)} - \eta_t)(\rho_t^{(m)} - \eta_t)^\top. \end{aligned} \tag{54}$$

Now, one can readily derive the conditional Gaussian distribution of the parameters to use for parameter particle propagation

$$\begin{aligned} f(\theta_t|x_t) &= \mathcal{N}(\theta_t|\eta_{\theta_t|x_t}, Q_{\theta_t|x_t}), \\ &\text{where} \begin{cases} \eta_{\theta_t|x_t} = \eta_{\theta_t} + Q_{\theta_t,x_t} Q_{x_t,x_t}^{-1} (x_t - \eta_{x_t}), \\ Q_{\theta_t|x_t} = Q_{\theta_t,\theta_t} - Q_{\theta_t,x_t} Q_{x_t,x_t}^{-1} Q_{x_t,\theta_t}, \end{cases} \\ &\text{with } \eta_t = \begin{pmatrix} \eta_{x_t} \\ \eta_{\theta_t} \end{pmatrix} \text{ and } Q_t = \begin{pmatrix} Q_{x_t,x_t} & Q_{\theta_t,x_t} \\ Q_{x_t,\theta_t} & Q_{\theta_t,\theta_t} \end{pmatrix}. \end{aligned} \tag{55}$$

One concludes that, given a set of parameter samples $a_t^{(j)}$ and $b_t^{(j)}$, the transition density for each new state sample $x_{t+1}^{(m)}$ is a mixture of J densities as provided in Equation 52. We present a SMC method for inference of latent ARMA(p, q) processes with unknown parameters a and b in Table 13.

PF FOR LATENT ARMA WITH UNKNOWN a AND b

1. At time instant t, consider the random measure for the joint state and parameter vector $\rho_t = (x_t \ \theta_t)^\top$

$$f_t^{MJ}(\rho_t) = \sum_{m=1}^{M} \sum_{j=1}^{J} w_t^{(m,j)} \delta\left(\rho_t - \rho_t^{(m,j)}\right).$$

   The superscript (m, j) indicates that for a given sample $x_t^{(m)}$, we have J sets of a and b parameters.

2. Approximate the joint posterior distribution of ρ_t

$$f(\rho_t) \approx \mathcal{N}(\rho_t | \eta_t, Q_t),$$

   with sufficient statistics as in Equation 54 with MJ samples.

3. Downsample from MJ to M and obtain a set of resampled streams $\overline{x}_t^{(m)} \sim f_t^{MJ}(\rho_t)$.

4. Draw J new parameter samples from the conditional Gaussian $\theta_{t+1}^{(m,j)} \sim \mathcal{N}(\theta_{t+1} | \eta_{\theta_t, x_t}, Q_{\theta_t | x_t})$, given each of the resampled particles $\overline{x}_t^{(m)}$. The parameters of the conditional Gaussian are given by Equation 55.

5. Propagate the state by sampling from the transition density

$$x_{t+1}^{(m,j)} \sim f(x_{t+1} | \overline{x}_{1:t}^{(m)}, a_{t+1}^{(m,j)}, b_{t+1}^{(m,j)}, \sigma_u^2),$$

   as given by Equation 29 and Equation 30.

6. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m,j)} \propto f(y_{t+1} | x_{t+1}^{(m,j)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^{MJ}(\rho_{t+1}) = \sum_{m=1}^{M} \sum_{j=1}^{J} w_{t+1}^{(m,j)} \delta\left(\rho_{t+1} - \rho_{t+1}^{(m,j)}\right).$$

Table 13: PF for stationary ARMA(p, q) with unknown parameters a and b.

Latent ARMA(p, q) processes with unknown innovation variance σ_u^2

In any practical scenario, determining the sufficient statistics of the innovation process is not a trivial task. It is of common practice to consider zero-mean ARMA processes, i.e., $\mu_u = 0$, which simplifies the problem considerably. The justification is that data can be easily adjusted to have zero-mean or a constant term can be included in the state-space model.

However, determining the variance of the innovation process is usually more challenging. Estimating the variability of a hidden process requires advanced estimation techniques or Expectation-Maximization (EM) type approaches.

We hereby deal with this uncertainty by marginalizing it. Namely, we obtain the transition density of the latent state where the unknown variance has been Rao-Blackwellized, which allows for implementation of SMC methods for inference of latent ARMA(p, q) processes with unknown variance.

We first derive the case where the ARMA parameters are known in Section 4.2.2.1, later consider the unknown AR(p) case in Section 4.2.2.2 and the general case with all unknown ARMA parameters, in Section 4.2.2.3.

ARMA(p, q) processes with unknown innovation variance: known a and b

We start by marginalizing the unknown innovation variance of an ARMA(p, q) process when the model parameters a and b are known. By means of a Bayesian analysis of the unknown variance, we obtain the transition density of x_{t+1} given $x_{1:t}$ and the parameters a and b.

That is, we are interested on computing $f(x_{t+1}|x_{1:t}, a, b)$ via

$$f(x_{t+1}|x_{1:t}, a, b) = \int_0^\infty f(x_{t+1}|x_{1:t}, a, b, \sigma_u^2) f(\sigma_u^2|x_{1:t}, a, b) d\sigma_u^2 \,, \tag{56}$$

where $f(\sigma_u^2|x_{1:t}, a, b)$ is the posterior of the unknown variance σ_u^2, given the parameters a and b, as well as the data available at time instant t, $x_{1:t}$.

First, we need to derive the posterior of the unknown state noise variance σ_u^2. We start by using Bayes rule and write

$$f(\sigma_u^2|x_{1:t}, a, b) \propto f(x_{1:t}|\sigma_u^2, a, b) f(\sigma_u^2), \tag{57}$$

where $f(\sigma_u^2)$ is the prior of the unknown variance. We use a scaled inverse Chi-squared density

$$f(\sigma_u^2) = \chi^{-2}(\sigma_u^2|\nu_0, \sigma_0^2) = \frac{\left(\sigma_0^2 \frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \frac{1}{(\sigma_u^2)^{1+\frac{\nu_0}{2}}} e^{-\frac{\nu_0 \sigma_0^2}{2\sigma_u^2}}, \quad (58)$$

where $\nu_0 > 0$ represents the degrees of freedom and $\sigma_0^2 > 0$ is the prior variance parameter.

Due to the conjugacy of the selected prior, the posterior is also a scaled inverse Chi-squared density

$$\begin{aligned} f(\sigma_u^2|x_{1:t}, a, b) &\propto f(x_{1:t}|\sigma_u^2, a, b) f(\sigma_u^2) \\ &= \chi^{-2}(\sigma_u^2|\nu_t, \sigma_t^2), \end{aligned} \quad (59)$$

with

$$\begin{aligned} \nu_t &= \nu_0 + t, \\ \sigma_t^2 &= \frac{\nu_0 \sigma_0^2 + x_{1:t}^\top \widetilde{\Sigma}_t^{-1} x_{1:t}}{\nu_t}. \end{aligned} \quad (60)$$

Now, we can proceed and integrate out the unknown variance

$$\begin{aligned} f(x_{t+1}|x_{1:t}, a, b) &= \int_0^\infty f(x_{t+1}|x_{1:t}, a, b, \sigma_u^2) f(\sigma_u^2|x_{1:t}, a, b) d\sigma_u^2 \\ &= \int_0^\infty \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right) \chi^{-2}(\sigma_u^2|\nu_t, \sigma_t^2) d\sigma_u^2, \end{aligned} \quad (61)$$

which results in the following scaled Student-t distribution:

$$\begin{aligned} f(x_{t+1}|x_{1:t}) &= \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right) \\ &= \frac{\Gamma\left(\frac{\nu_t+1}{2}\right)}{\Gamma\left(\frac{\nu_t}{2}\right)\sqrt{\pi c_{t+1}\nu_t\sigma_t^2}} \cdot \left(1 + \frac{(x_{t+1} - \mu_{x_{t+1}|x_{1:t}})^2}{c_{t+1}\nu_t\sigma_t^2}\right)^{-\left(\frac{\nu_t+1}{2}\right)}, \end{aligned} \quad (62)$$

with sufficient statistics

$$\begin{aligned} \nu_t &= \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} &= \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} x_{1:t}\,, \\ \sigma_t^2 &= \frac{\nu_0 \sigma_0^2 + x_{1:t}^\top \widetilde{\Sigma}_t^{-1} x_{1:t}}{\nu_t}\,, \\ c_{t+1} &= \widetilde{\gamma}(0) - \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t\,. \end{aligned} \quad (63)$$

Note that the covariance function and matrix above are normalized, as in Equation 28.

The SMC method for inference of latent stationary ARMA(p, q) processes with known parameters a and b, and unknown innovation variance σ_u^2 is presented in Table 14.

PF FOR LATENT ARMA WITH KNOWN a b, UNKNOWN σ_u^2

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state (to avoid sample degeneracy) by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

(In principle, this step does not have to be performed at every time instant.)

4. Propagate the particles by sampling from the transition density, given the previous stream of (resampled) particles:

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}) = \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right),$$

$$\text{where } \begin{cases} \nu_t = \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} = \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \overline{x}_{1:t}^{(m)}\,, \\ \sigma_t^2 = \frac{\nu_0\sigma_0^2 + \overline{x}_{1:t}^{(m)\top} \widetilde{\Sigma}_t^{-1} \overline{x}_{1:t}^{(m)}}{\nu_t}\,, \\ c_{t+1} = \widetilde{\gamma}(0) - \widetilde{\gamma}_t^\top \widetilde{\Sigma}_t^{-1} \widetilde{\gamma}_t\,. \end{cases}$$

5. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 14: PF for latent ARMA(p,q) with known parameters a b and unknown σ_u^2.

AR(p) processes with unknown innovation variance: unknown a

In taking care of σ_u^2 when dealing with an AR(p) process with unknown parameter a, we proceed in the same way as in the previous section. However, we now use the marginalized likelihood densities

$$f(x_{t+1}|x_{1:t}) = \int_0^\infty f(x_{t+1}|x_{1:t}, \sigma_u^2) f(\sigma_u^2|x_{1:t}) d\sigma_u^2 \,, \tag{64}$$

where $f(\sigma_u^2|x_{1:t})$ is the posterior of the unknown variance σ_u^2, where the a parameter has been marginalized.

We find such density by first writing

$$f(\sigma_u^2|x_{1:t}) \propto f(x_{p+1:t}|\sigma_u^2, x_{1:p}) f(\sigma_u^2), \tag{65}$$

where $f(\sigma_u^2)$ is the prior of the unknown variance and we again use a scaled inverse Chi-squared density $\chi^{-2}(\sigma_u^2|\nu_0, \sigma_0^2)$.

The derivation follows

$$\begin{aligned} f(\sigma_u^2|x_{1:t}) &\propto f(x_{p+1:t}|\sigma_u^2, x_{1:p}) f(\sigma_u^2) \\ &= \left(\int f(x_{p+1:t}|a, \sigma_u^2, x_{1:p}) f(a) da \right) \chi^{-2}(\sigma_u^2|\nu_0, \sigma_0^2) \\ &\propto \frac{e^{-\frac{x_{p+1:t}^\top P_t^\perp x_{p+1:t}}{2\sigma_u^2}}}{(\sigma_u^2)^{\frac{t-p}{2}}} \times \frac{e^{-\frac{\nu_0 \sigma_0^2}{2\sigma_u^2}}}{(\sigma_u^2)^{1+\frac{\nu_0}{2}}}, \end{aligned} \tag{66}$$

where $P_t^\perp = I - H_{t-1}(H_{t-1}^\top H_{t-1})^{-1} H_{t-1}^\top$ is a projection matrix and H_{t-1} has been defined in Equation 49.

Thus, we deduce that the marginalized posterior of the unknown variance $f(\sigma_u^2|x_{1:t})$ is again $\chi^{-2}(\sigma_u^2|\nu_t, \sigma_t^2)$, with

$$\begin{aligned} \nu_t &= \nu_0 + (t-p), \\ \sigma_t^2 &= \frac{\nu_0 \sigma_0^2 + x_{p+1:t}^\top P_t^\perp x_{p+1:t}}{\nu_t}. \end{aligned} \tag{67}$$

With this posterior of the unknown variance, the final expression for $f(x_{t+1}|x_{1:t})$ can be derived following the same steps as in Equation 61. That is,

$$\begin{aligned} f(x_{t+1}|x_{1:t}) &= \int_0^\infty f(x_{t+1}|x_{1:t}, \sigma_u^2) f(\sigma_u^2|x_{1:t}) d\sigma_u^2 \\ &= \int_0^\infty \mathcal{N}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma_u^2 c_{t+1}\right) \chi^{-2}(\sigma_u^2|\nu_t, \sigma_t^2) d\sigma_u^2, \end{aligned} \tag{68}$$

which results in the following scaled Student-t distribution

$$f(x_{t+1}|x_{1:t}) = \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right),$$

with

$$\begin{aligned} \nu_t &= \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} &= h_t^\top \mu_t, \\ \sigma_t^2 &= \frac{\nu_0\sigma_0^2 + x_{1:t}^\top \widetilde{\Sigma}_t^{-1} x_{1:t}}{\nu_t}\,, \\ c_{t+1} &= 1 + h_t^\top C_t h_t\,, \end{aligned} \tag{69}$$

where h_t, μ_t and C_t have been defined in Equation 48 and Equation 50.

The SMC method for inference of latent stationary AR(p) processes with unknown parameter α and innovation variance σ_u^2 is presented in Table 15.

PF FOR LATENT AR WITH UNKNOWN a AND σ_u^2

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state (to avoid sample degeneracy) by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

(In principle, this step does not have to be performed at every time instant.)

4. Propagate the particles by sampling from the transition density, given the previous stream of (resampled) particles:

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}, \sigma_u^2) = \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right),$$

$$\text{where } \begin{cases} \nu_t = \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} = h_t^\top \mu_t, \\ \sigma_t^2 = \frac{\nu_0\sigma_0^2 + x_{1:t}^\top \widetilde{\Sigma}_t^{-1} x_{1:t}}{\nu_t}\,, \\ c_{t+1} = 1 + h_t^\top C_t h_t\,, \end{cases}$$

where h_t, μ_t and C_t are defined in Equation 48 and Equation 50.

5. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 15: PF for latent AR(p) with unknown parameters a and σ_u^2.

ARMA(p, q) processes with unknown innovation variance: unknown a and b

The final case is the most challenging one, as we do not assume knowledge of any of the ARMA model's parameters. That is, a, b and σ_u^2 are all unknown. Once again, our goal is to Rao-Blackwellize all these, so that we can derive an SMC method that relies on the marginalized transition density without knowledge of model parameters.

We aim at integrating out all the static unknowns and thus, deriving the state transition density

$$f(x_{t+1}|x_{1:t}) = \int_{a,b,\sigma_u^2} f(x_{t+1}|x_t, a, b, \sigma_u^2) f(a, b, \sigma_u^2|x_{1:t}) da db d\sigma_u^2. \tag{70}$$

However, as already pointed out, one cannot analytically integrate out the ARMA parameters a and b. Thus, we approach the integration of model parameters a and b and the innovation variance σ_u^2 separately.

We rewrite Equation 70 as

$$\begin{aligned} f(x_{t+1}|x_{1:t}) = \int_{\sigma^2} \int_{a,b} & f(x_{t+1}|x_{1:t}, a, b, \sigma_u^2) \\ & \times f(a, b|x_{1:t}, \sigma_u^2) f(\sigma_u^2|x_{1:t}) da db d\sigma_u^2, \end{aligned} \tag{71}$$

where we separate the marginalization of the ARMA parameters and the noise variance. We first deal with the unknown a and b, before taking care of the unknown σ_u^2.

We first compute

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) = \int_{a,b} f(x_{t+1}|x_t, a, b, \sigma_u^2) f(a, b|x_{1:t}, \sigma_u^2) da db, \tag{72}$$

and then,

$$f(x_{t+1}|x_{1:t}) \int_0^\infty f(x_{t+1}|x_{1:t}, \sigma_u^2) f(\sigma_u^2|x_{1:t}) d\sigma_u^2. \tag{73}$$

For the Rao-Blackwellization of the unknown a and b parameters in Equation 72, we resort to the Monte Carlo solution already presented in Section 4.2.1.2. That is, we obtain a weighted sum of densities for given parameter samples, i.e.,

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) \approx \frac{1}{J} \sum_{j=1}^{J} f(x_{t+1}|x_{1:t}, a_t^{(j)}, b_t^{(j)}, \sigma_u^2). \tag{74}$$

The solution to Equation 73 when using the Monte Carlo approximation in Equation 74 results in

$$f(x_{t+1}|x_{1:t}) \approx \frac{1}{J}\sum_{j=1}^{J}\int_0^\infty f(x_t|a_{t-1}^{(j)}, b_{t-1}^{(j)}, \sigma_u^2, x_{1:t-1}) f(\sigma_u^2|x_{1:t}) d\sigma_u^2. \tag{75}$$

Thus, the solution for the transition density with unknown parameters is a weighted sum of densities

$$\begin{aligned} f(x_{t+1}|x_{1:t}) &\approx \frac{1}{J}\sum_{j=1}^{J} f^{(j)}(x_{t+1}|x_{1:t}), \\ \text{where } f^{(j)}(x_{t+1}|x_{1:t}) &= \int_0^\infty f(x_t|a_{t-1}^{(j)}, b_{t-1}^{(j)}, \sigma_u^2, x_{1:t-1}) f(\sigma_u^2|x_{1:t}) d\sigma_u^2. \end{aligned} \tag{76}$$

For given parameter samples $a_{t-1}^{(j)}$ and $b_{t-1}^{(j)}$, this is of the exactly the same form as the one in Equation 61. Thus, we readily conclude that

$$\begin{aligned} f(x_{t+1}|x_{1:t}) &\approx \frac{1}{J}\sum_{j=1}^{J} f^{(j)}(x_{t+1}|x_{1:t}), \\ \text{where } f^{(j)}(x_{t+1}|x_{1:t}) &= \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}^{(j)}, c_{t+1}^{(j)}\sigma_t^{(j)^2}\right) \\ \text{with } &\begin{cases} \nu_t = \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}}^{(j)} = \widetilde{\gamma}_t^{(j)^\top}\widetilde{\Sigma}_t^{(j)^{-1}} x_{1:t}\,, \\ \sigma_t^{(j)^2} = \frac{\nu_0\sigma_0^2 + x_{1:t}^\top \widetilde{\Sigma}_t^{(j)^{-1}} x_{1:t}}{\nu_t}\,, \\ c_{t+1}^{(j)} = \widetilde{\gamma}^{(j)}(0) - \widetilde{\gamma}_t^{(j)^\top}\widetilde{\Sigma}_t^{(j)^{-1}}\widetilde{\gamma}_t^{(j)}\,. \end{cases} \end{aligned} \tag{77}$$

Note the superscript $^{(j)}$ in all the sufficient statistics of the solution, as they are computed for each parameter sample $\left\{a_t^{(j)}, b_t^{(j)}\right\}$. The Rao-Blackwellization of the parameters is thus performed by averaging over all the sample parameters J.

The SMC method for inference of latent stationary ARMA(p, q) processes with unknown model parameters a, b and σ_u^2 is presented in Table 16.

PF FOR LATENT ARMA WITH UNKNOWN a, b AND σ_u^2

1. At time instant t, consider the random measure for the joint state and parameter vector $\rho_t = (x_t \ \theta_t)^\top$

$$f_t^{MJ}(\rho_t) = \sum_{m=1}^{M} \sum_{j=1}^{J} w_t^{(m,j)} \delta\left(\rho_t - \rho_t^{(m,j)}\right).$$

   The superscript (m, j) indicates that for a given sample $x_t^{(m)}$, we have J sets of a and b parameters.

2. Approximate the joint posterior distribution of ρ_t

$$f(\rho_t) \approx \mathcal{N}(\rho_t | \eta_t, Q_t),$$

   with sufficient statistics as in Equation 54 with MJ samples.

3. Downsample from MJ to M and obtain a set of resampled streams $\overline{x}_t^{(m)} \sim f_t^{MJ}(\rho_t)$.

4. Draw J new parameter samples from the conditional Gaussian $\theta_{t+1}^{(m,j)} \sim \mathcal{N}(\theta_{t+1} | \eta_{\theta_t, x_t}, Q_{\theta_t | x_t})$, given each of the resampled particles $\overline{x}_t^{(m)}$. The parameters of the conditional Gaussian are given in Equation 55.

5. Propagate the state by sampling from the transition density

$$x_{t+1}^{(m,j)} \sim f(x_{t+1} | \overline{x}_{1:t}^{(m)}, a_{t+1}^{(m,j)}, b_{t+1}^{(m,j)}),$$

   as given by Equation 77, computed per particle.

6. Compute the non-normalized weights for the drawn particles according to

$$\widetilde{w}_{t+1}^{(m,j)} \propto f(y_{t+1} | x_{t+1}^{(m,j)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^{MJ}(\rho_{t+1}) = \sum_{m=1}^{M} \sum_{j=1}^{J} w_{t+1}^{(m,j)} \delta\left(\rho_{t+1} - \rho_{t+1}^{(m,j)}\right).$$

Table 16: PF for latent ARMA(p, q) with unknown parameters a, b and σ_u^2.

EVALUATION

In this chapter we, too, evaluate the performance of the proposed methods by simulating the SV model. As explained in Section 3.3, this model is popular in the study of nonlinear state-space models [1, 105], as it makes inference very challenging for Kalman based filters, even with known state parameters.

The problem gets further complicated when the parameters are unknown (as is the case in this chapter for the latent ARMA process), due to the nonlinearities introduced by the MA parameters and the Student-t distributions resulting from the Bayesian analysis of the unknown σ_u^2.

For this reason, we hereby use the performance of the PF proposed in Chapter 3 for the known parameter case (i.e., Table 5) as a benchmark for comparison with the more challenging scenarios considered.

Mathematically, we write

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ y_t = e^{\frac{x_t}{2}} v_t, \end{cases} \tag{78}$$

where v_t is a standard Gaussian variable and the state noise u_t is a zero-mean Gaussian variable. Here, the ARMA parameters $a = (a_1 \ \cdots \ a_p)^\top$ and $b = (b_1 \ \cdots \ b_q)^\top$, and the innovation variance σ_u^2 are unknown.

First, we show the performance of the SMC method for AR(p) processes, as their unknown parameters can be analytically integrated out (see Section 4.2.1.1 and Section 4.2.2.2).

The results are presented in Table 17, where the entries are the averaged MSEs of the estimates of the latent state x_t. Unless otherwise indicated, 1000 particles are used for the PFs and the shown MSEs are averages over 100 realizations of 500 samples long time-series. The results demonstrate a mild deterioration in estimation performance as our knowledge about the AR processes decreases.

We observe how the unknown a parameters have a more pronounced impact on the performance of the methods, when compared to the unknown innovation variance. The explanation is two-fold.

On the one hand, the a parameters are directly determining the evolution of the state. On the other, the integration of the unknown variance results in a Student-t distributed transition density, which is different from a Gaussian density only on its

PF Type	State estimation error (MSE)		
	AR(1)	AR(2)	AR(3)
`Known AR, known` σ_u^2	1.10969	1.18849	0.92265
`Known AR, unknown` σ_u^2	1.16164	1.20956	0.96545
`Unknown AR, known` σ_u^2	1.32100	1.41613	0.98582
`Unknown AR, unknown` σ_u^2	1.35901	1.43513	0.99948

Table 17: MSE filtering performance of the proposed SMC methods for AR processes.

tails. As a matter of fact, the former gets more and more similar to the latter as the degrees of freedom increase; this is indeed the case as we observe more data over time (see Equation 67 for example).

We also evaluate the impact of the unknown innovation variance for the general ARMA(p, q) case. Results are illustrated in Figure 14, where we also study the impact of truncating the covariance matrix in the computed sufficient statistics.

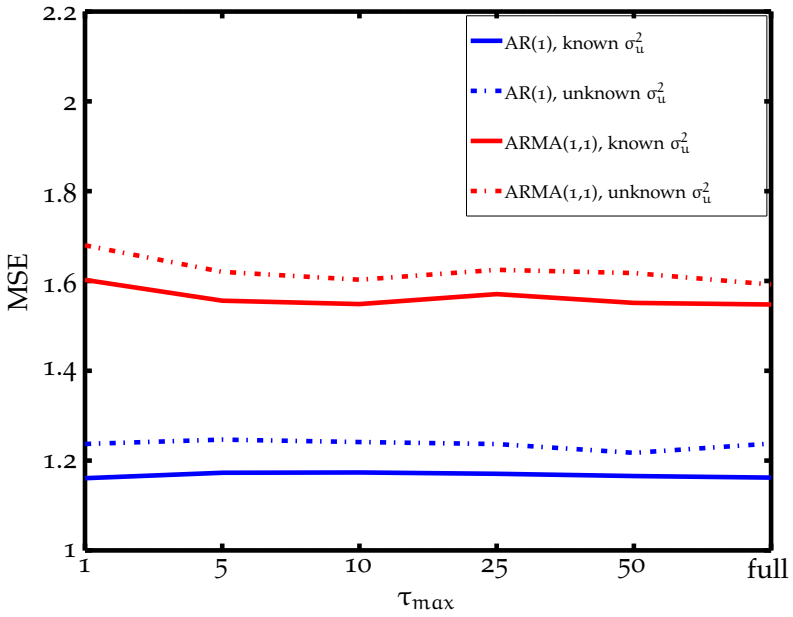


Figure 14: MSE for SMC methods with known and unknown σ_u^2 as a function of τ_{max}.

On the one hand, we observe the loss in performance induced by not knowing the driving noise variance, which is minimal as explained before for the AR(p) case. On the

other, the negligible effect of truncating the covariance matrix is noticed, which is justified by the short-memory claims in Section 3.1.4 and coincides with the results provided in Section 3.3 for the truncation with known ARMA parameters.

Now, we evaluate the more general case of ARMA models with unknown a and b parameters. First, we illustrate the tracking accuracy of the methods described in Section 4.2.1.2.

In general, the method is able to successfully estimate the log-volatility x_t for different ARMA(p, q) processes, as shown in Figure 15 for specific realizations. Note the difficulty in accurately estimating MA(q) processes, as plain correlated noise is not very informative. We further illustrate the method's potential by plotting the evolution of the posterior densities of the unknown ARMA(1,1) parameters in Figure 16.

Before further evaluating the proposed SMC for the general ARMA(p, q) model with all unknown parameters, we first use the unknown AR case to validate the numerical approximation to the integrals in Equation 52 and Equation 76. That is, we study the unknown AR case to provide insights into the benefits of Rao-Blackwellization and validate the numerical approximation to the integration of the unknown parameters.

We implement the method in Section 4.2.2.3 for AR processes with unknown parameters and compare it to both the analytically Rao-Blackwellized AR solution (i.e., PF in Table 15), and the alternative joint state and parameter estimation approach. In all cases, the variance of the driving noise of the process is assumed unknown.

The MSE results for estimation of the hidden process are shown in Table 18. The performance of the Rao-Blackwellized methods is in between the case when the parameters are known (`Known AR`) and the case when the PF method estimates the parameters and the process jointly (`Unknown AR, param.est`). Furthermore, a better performance of the analytically integrated out solution (`Unknown AR, RB analytical`) compared to the numerical Rao-Blackwellization (`Unknown AR, RB numerical J=10`) is observed.

These results demonstrate that Rao-Blackwellization of the unknown parameters provides a superior performance. Namely, (1) it always outperforms the PF that jointly estimates all the parameters, (2) it provides estimation accuracy comparable to that of the benchmark (`Known AR`), and (3) the numerical approximation to the analytical solution is accurate.

PF Type	State estimation error (MSE)		
	AR(1)	AR(2)	AR(3)
`Known AR`	1.1616	1.20956	0.96545
`Unknown AR, RB analytical`	1.35901	1.43513	0.99948
`Unknown AR, RB numerical` $J = 10$	1.49248	1.67136	1.02314
`Unknown AR, param. est.`	1.50258	1.70387	1.05567

Table 18: MSE filtering performance of variants of the PF for AR(p) processes.

Once we have investigated the benefits and drawbacks of the method proposed Section 4.2.2.3, we now evaluate the accuracy of the PF for different ARMA(p, q) processes. Results are presented in Table 19 for both known and unknown σ_u^2, with PFs implemented as in Table 13 and Table 16, respectively.

Note that the accuracy of the proposed methods is reasonably close to the benchmark, which is now the known parameter case (`Known ARMA` in Table 19), as there is no closed-form Rao-Blackwellized solution available.

We evaluate the quality of the Monte Carlo integral for different J values. Even though, in general, the more particles (larger J) used to approximate the integral, the more accurate the estimate becomes, the improvement becomes negligible compared to the computational burden for $J > 20$.

To conclude with the evaluation section, we emphasize the importance of considering ARMA(p, q) processes instead of using higher order AR(p) models to approximate them. For the AR case, the analytical Rao-Blackwellization of its unknowns has been derived in closed form and, thus, this alternative might seem appropriate.

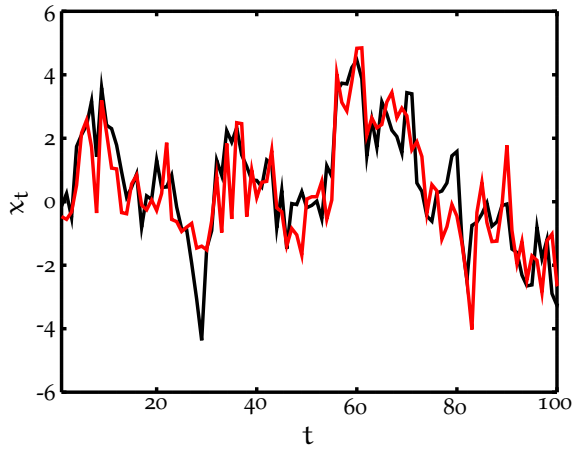
However, the results presented in Table 20 show how considering a true ARMA(p, q) state outperforms the alternative approximation. The ARMA model provides a more parsimonious solution whereas the AR approximation is less accurate. This is mainly due to the high dimension of the approximating model and its impact on SMC sampling.

PF Type	State estimation error (MSE)				
	MA(1)	ARMA(1,1)	ARMA(1,1)	ARMA(1,2)	ARMA(2,1)
Known ARMA, known σ_u^2	0.84729	1.51002	0.83048	1.74338	1.71039
Known ARMA, unknown σ_u^2	0.89428	1.56405	0.88028	1.78301	1.77861
Unknown ARMA $J = 1$, known σ_u^2	0.86743	2.05675	0.91206	2.28310	2.34009
Unknown ARMA $J = 1$, unknown σ_u^2	0.93760	2.08892	0.99929	2.38283	2.35899
Unknown ARMA $J = 10$, known σ_u^2	0.86317	2.01788	0.91414	2.23473	2.22838
Unknown ARMA $J = 10$, unknown σ_u^2	0.93959	2.09366	0.98532	2.32099	2.33336
Unknown ARMA $J = 20$, known σ_u^2	0.86274	2.00809	0.91642	2.23436	2.23242
Unknown ARMA $J = 20$, unknown σ_u^2	0.94077	2.06118	0.99258	2.31816	2.34098
Unknown ARMA $J = 30$, known σ_u^2	0.86292	1.99660	0.91219	2.23251	2.21125
Unknown ARMA $J = 30$, unknown σ_u^2	0.93067	2.08272	0.99819	2.31880	2.35163

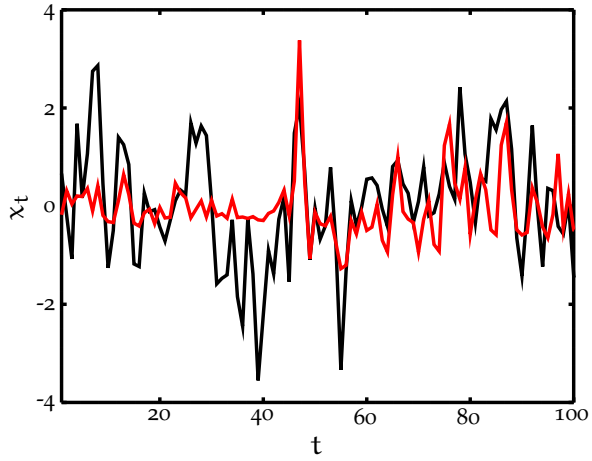
Table 19: MSE performance of proposed PFs for different ARMA(p, q) processes.

PF Type	State estimation error (MSE)	
	$a_1 = 0.85, b_1 = 0.5$	$a_1 = 0.85, b_1 = 0.75$
Known ARMA	1.68605	1.83369
Unknown ARMA $J = 10$	2.22736	2.33239
Unknown AR(2)	2.23292	2.34956
Unknown AR(3)	2.35514	2.40280
Unknown AR(4)	2.49513	2.52133
Unknown AR(5)	2.82545	2.68774

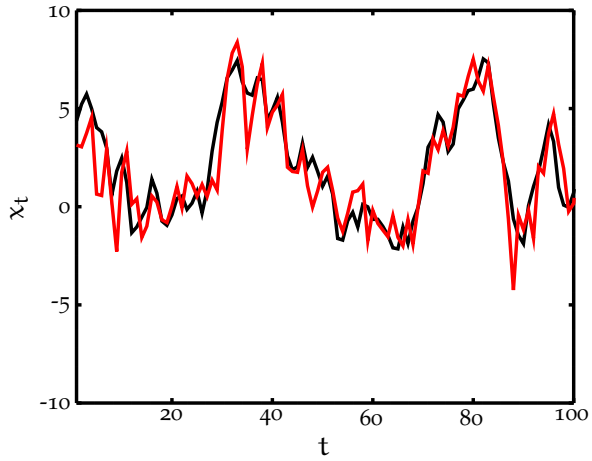
Table 20: MSE of the PFs assuming true ARMA(1,1) and different high-order approximations to the latent process.



(a) AR(1) tracking.

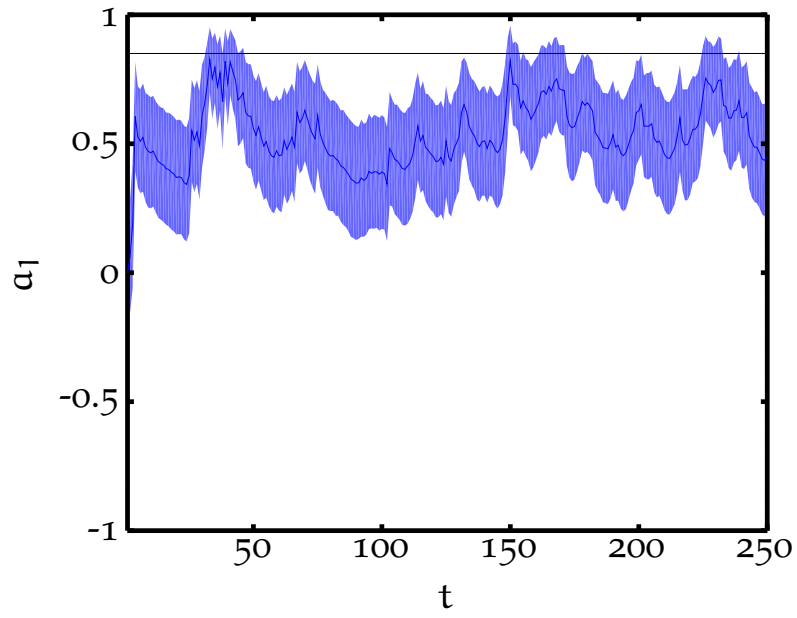


(b) MA(1) tracking.

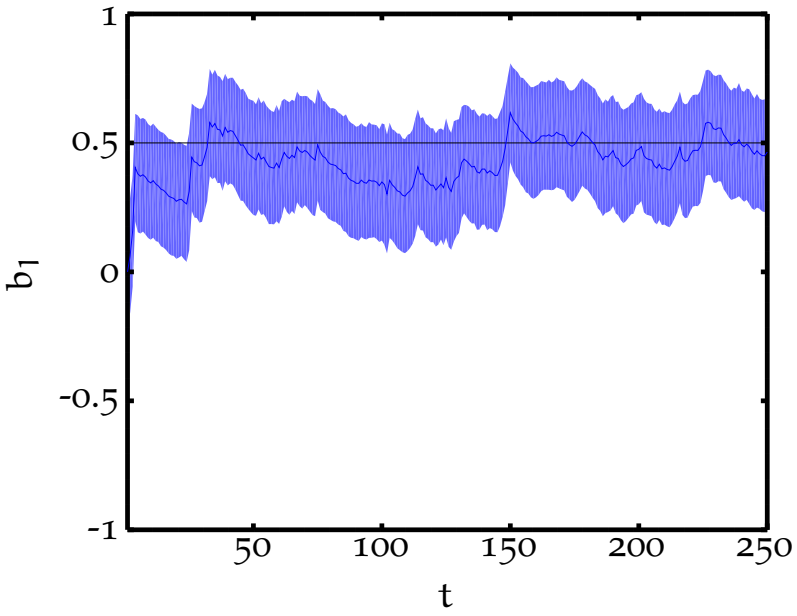


(c) ARMA(1,1) tracking.

Figure 15: True (black) and estimated (red) state for the proposed SMC method.



(a) $f(a_1|y_{1:t})$.



(b) $f(b_1|y_{1:t})$.

Figure 16: Posteriors of the unknown parameters a_1 and b_1 with 95% confidence interval in blue (the true value in black).

5

LATENT CORRELATED INNOVATION PROCESSES

Among the relevant features of the studied models in previous Chapters 3 and 4, the "memory" of the time-series is one of the most important characteristics. As shown in Section 3.1.4, for an ARMA(p, q) process, the decay of the autocovariance is asymptotically exponential. Thus, these time-series are referred to as short-memory processes.

In this chapter, we address time-series where the dependencies amongst samples goes further into the past. That is, we study long-memory processes and, in particular, inference of latent fGps. We provide a detailed overview of long-memory processes and describe two types of time-series that fall into this category, i.e., the fGp and the FARIMA model, in Section 5.1. We focus on the former and present an SMC method for inference of latent fGps in Section 5.2.

TIME-SERIES AND MEMORY PROPERTIES

Experts in many scientific areas have studied short-memory processes in the form of ARMA (or their AR and MA variants) and other Markov processes. In simple words, in short-memory time-series only few past samples affect the present data values.

On the contrary, there are other time-series where the present value is dependent on samples far into the past. These are often refered to as long-memory processes. The intuitive interpretation of such a process is that the dependence between events that are far apart diminishes slowly with increasing lag.

The groundwork on long-memory processes was laid by Hurst [54], when he found that the Nile river data manifested long-range dependence. Later, many other hydrological, geophysical, and climatological records were reported to describe similar characteristics. Motivated by the verification that a plethora of real-life time-series manifest such properties [8, 12, 80], we are interested on the study of processes with long-memory properties.

First, let us understand the features of data with long-memory properties [12]. The *qualitative* features of long-memory time-series are as follows:

1. There are relatively long periods where data tend to stay at similar values. That is, once at a high level, the next values are usually high too and, similarly, there are long periods with low level values.

2. When one looks at short time intervals, there seems to be cycles or local trends. However, when looking at the whole series, there is no apparent persisting trend or cycle. It rather seems that cycles of (almost) all frequencies occur, superimposed and in random sequence.

3. Overall, the series looks stationary.

The memory of data can be mathematically studied by means of their sample variance and correlation function. For data that can be modeled with short-memory properties, the decay of the sample autocorrelation is asymptotically exponential. On the contrary, a slower decay is observed in long-memory data.

Specifically, the *quantitative* features of a long-memory process are described by [12]:

1. The variance of the sample mean decays to zero at a slower rate than n^{-1}. In a good approximation, the rate is proportional to $n^{-\alpha}$ for some $0 < \alpha < 1$.

2. The sample correlations $\hat{\rho}(\tau) = \hat{\gamma}(\tau)/\hat{\gamma}(0)$ decay to zero at a rate that is in good approximation proportional to $\tau^{-\alpha}$ for some $0 < \alpha < 1$.

3. Near the origin, the logarithm of the periodogram $I(f)$ plotted against the logarithm of the frequency appears to be scattered around a straight line with negative slope.

As for short-memory data, stochastic processes are useful in modeling long-memory data. For these processes, too, stationarity is a common assumption. Thus, one considers processes for which the first two moments exist. In statistical terms, the memory of any process is accurately described by the autocovariance function $\gamma(\tau)$ of the process. Recall that, because we consider stationarity, the autocovariance function only depends on the time-lag τ.

As previously reported in Chapter 3, for ARMA processes driven by iid noise, the decay of the autocovariance is asymptotically exponential. That is, an upper bound of the form $|\gamma(\tau)| \propto b a^{\tau}$ for some real scalars $0 < b < \infty$ and $0 < a < 1$ exists for short-memory processes.

On the contrary, a slower decay of the autocovariance function is observed in long-memory processes. Specifically, one can establish the following statistical properties for such processes:

1. The variance of the sample mean $\text{Var}\{\bar{x}\}$ is asymptotically equal to $c_{var} \cdot n^{-\alpha}$ for some $0 < \alpha < 1$.
2. The correlations $\rho(\tau)$ are asymptotically equal to $c_{\rho} \cdot \tau^{-\alpha}$ for some $0 < \alpha < 1$.
3. The spectral density $S(f)$ has a pole at zero that is equal to $c_f \cdot \tau^{-\beta}$ for some $0 < \beta < 1$.

All in all, we use the above to define a stationary process with long-memory or long-range dependence.

Definition 1 *x_t is called a stationary process with long-memory if there exists a real number $\alpha \in (0, 1)$ and a constant $c_{\rho} > 0$ such that the following holds:*

$$\lim_{\tau \to \infty} \frac{\rho(\tau)}{c_{\rho} \tau^{-\alpha}} = 1. \tag{79}$$

Due to the one to one mapping between the autocovariance function and the spectral density function, an equivalent definition in terms of the spectral density $S(f)$ follows:

Definition 2 *x_t is called a stationary process with long-memory if there exists a real number $\beta \in (0, 1)$ and a constant $c_f > 0$ such that the following holds:*

$$\lim_{f \to 0} \frac{S(f)}{c_f |f|^{-\beta}} = 1. \tag{80}$$

These two definitions above are equivalent, as shown in the theorems below.

Theorem 1 *If the autocorrelation function in Equation 79 holds with* $0 < \alpha = 2 - 2H < 1$*, then the spectral density* $S(f)$ *exists and*

$$\lim_{f \to 0} \frac{S(f)}{c_f(H)|f|^{1-2H}} = 1, \tag{81}$$

where for $\sigma^2 = Var\{x_t\}$,

$$c_f(H) = \frac{\sigma^2}{\pi} c_\rho \Gamma(2H-1) \sin(\pi - \pi H). \tag{82}$$

Theorem 2 *If the spectral density in Equation 80 holds with* $0 < \beta = 2H - 1 < 1$*, then the autocorrelation function* $\rho(\tau)$ *exists and*

$$\lim_{\tau \to \infty} \frac{\rho(\tau)}{c_\rho(H)\tau^{2-2H}} = 1, \tag{83}$$

where for $\sigma^2 = Var\{x_t\}$,

$$c_\rho(H) = \frac{2}{\sigma^2} c_f \Gamma(2-2H) \sin\left(\pi H - \frac{1}{2}\pi\right). \tag{84}$$

Note that the above definitions for long-memory processes are formulated in asymptotic terms. Thus, Equation 79 only describes the behavior of the correlations as the lag tends to infinity, and does not specify the correlations for any fixed finite lag. Moreover, it only determines the rate of convergence, and not the magnitude of the function, i.e., each individual correlation can be arbitrarily small or large, only the rate of decay is slow. Equivalent asymptotic insights are described by Equation 80 for the frequency domain.

Among the potential models for which Equation 79 might hold, there are two classes of models that are of special interest, because they arise in a natural way from limit theorems and classic models. On the one hand, those described as stationary increments of self-similar processes and, in particular fractional Gaussian noise (fGn). On the other, fractional Auto-Regressive Integrated Moving-Average (ARIMA) processes or FARIMA, which are built upon ARMA models by differentiation with fractional values.

Long-memory processes: the fractional Gaussian noise

Processes with long-memory properties naturally arise when studying self-similar processes and stationary increments. We formally define self-similarity of a stochastic process below.

Definition 3 *Let* y_t *be a stochastic process with continuous-time parameter* t. y_t *is called self-similar with self-similarity parameter* H, *if for any positive stretching factor* c, *the rescaled process with time scale* ct, $c^{-H}y(ct)$, *is equal in distribution to the original process* y_t*:*

$$y_t \overset{d}{=} c^{-H}y(ct). \tag{85}$$

Typical sample paths of a self-similar process look qualitatively the same, irrespective of the distance from which we look at them. The motivation for stochastic self-similarity, apart from aesthetic appeal and mathematical elegance, arises from limit theorems for sums of random variables. As a matter of fact [12], whenever a process is the limit of normalized partial sums of random variables, it is necessarily self-similar.

By replacing $c = t^{-1}, t > 0$ in Equation 85 for a self-similar process y_t with parameter H, one can study $y_t \overset{d}{=} t^H y_1$ and conclude the asymptotic behavior of such processes.

- When $t \to \infty$,
  1. if $H < 0$, then $y_t \overset{d}{\to} 0$;
  2. if $H = 0$, then $y_t \overset{d}{=} y_1$;
  3. if $H > 0$ and $y_t \neq 0$, then $|y_t| \overset{d}{\to} \infty$.
- When $t \to 0$,
  1. if $H < 0$ and $y_t \neq 0$, then $|y_t| \overset{d}{\to} \infty$,
  2. if $H = 0$, then $y_t \overset{d}{=} y_1$,
  3. if $H > 0$, then $y_t \overset{d}{\to} 0$.

For modeling self-similar data that are stationary, one needs to consider processes with stationary increments, which are formally defined as follows.

Definition 4 y_t *has stationary increments if for* $k \geqslant 1$ *and any* k *time points* $t_1, \cdots, t_k$, *the distribution of*

$$\left\{x_{t_1+c} - x_{t_1+c-1}, \cdots, x_{t_k+c} - x_{t_k+c-1}\right\}$$

does not depend on $c \in \mathbb{R}$.

The covariance function $\gamma_y(t,s) = \mathrm{Cov}\{y_t, y_s\}$ of a self-similar process y_t with stationary increments can be analytically derived.

Let us assume a zero-mean process, i.e., $\mathbb{E}\{y_t\} = 0$, and let $\sigma^2 = \mathbb{E}\left\{(y_t - y_{t-1})^2\right\} = \mathbb{E}\left\{y_1^2\right\}$ be the variance of the increment process $x_t = y_t - y_{t-1}$.

For a self-similar process y_t we can write $y_t \overset{d}{=} t^H y_1$ and if $s < t$, derive

$$\mathbb{E}\left\{(y_t - y_s)^2\right\} = \mathbb{E}\left\{(y_{t-s} - y_0)^2\right\} = \sigma^2 (t-s)^{2H}. \tag{86}$$

One can similarly write

$$\begin{aligned}\mathbb{E}\left\{(y_t - y_s)^2\right\} &= \mathbb{E}\left\{y_t^2\right\} + \mathbb{E}\left\{y_s^2\right\} - 2\mathbb{E}\{y_t y_s\} \\ &= \sigma^2 t^{2H} + \sigma^2 s^{2H} - 2\gamma_y(t,s),\end{aligned} \tag{87}$$

and by combination of Equation 86 and Equation 87, conclude

$$\gamma_y(t,s) = \frac{1}{2}\sigma^2\left[t^{2H} - (t-s)^{2H} + s^{2H}\right]. \tag{88}$$

We now focus on the stationary increments $x_t = y_t - y_{t-1}$, in order to obtain the covariance function of such process too.

Let us first write some identities for later use:

$$\begin{cases} x_t = y_t - y_{t-1}, \\ y_t = y_{t-1} + x_t = y_0 + \sum_{i=1}^{t} x_t, \\ \sum_{i=1}^{t} x_t = y_t - y_0. \end{cases} \tag{89}$$

$$\begin{aligned}&\left(\sum_{i=1}^{k+1} x(i)\right)^2 + \left(\sum_{i=2}^{k} x(i)\right)^2 - \left(\sum_{i=1}^{k} x(i)\right)^2 - \left(\sum_{i=2}^{k+1} x(i)\right)^2 = \\ &\quad= \left(X(1) + \sum_{i=2}^{k} x(i) + x(k+1)\right)^2 + \left(\sum_{i=2}^{k} x(i)\right)^2 \\ &\qquad - \left(x(1) + \sum_{i=2}^{k} x(i)\right)^2 - \left(\sum_{i=2}^{k} x(i) + x(k+1)\right)^2 \\ &\quad= \left(x(1) + \sum_{i=2}^{k} x(i)\right)^2 + 2\left(x(1) + \sum_{i=2}^{k} x(i)\right) x(k+1) \\ &\qquad + x(k+1)^2 + \left(\sum_{i=2}^{k} x(i)\right)^2 - \left(x(1) + \sum_{i=2}^{k} x(i)\right)^2 \\ &\qquad - \left(\sum_{i=2}^{k} x(i)\right)^2 - 2\left(\sum_{i=2}^{k} x(i)\right) x(k+1) - x(k+1)^2 \\ &\quad= 2x(1)x(k+1).\end{aligned}$$

The derivation of the covariance function $\gamma_x(k)$ for the stationary increments x_t follows

$$\begin{aligned}
\gamma_x(k) &= \mathbb{E}\{x(i)x(k+i)\} = \mathbb{E}\{x(1)x(k+1)\} = \frac{1}{2}\mathbb{E}\{2x(1)x(k+1)\} \\
&= \frac{1}{2}\mathbb{E}\left\{\left(\sum_{i=1}^{k+1} x(i)\right)^2 + \left(\sum_{i=2}^{k} x(i)\right)^2\right\} \\
&\quad - \frac{1}{2}\mathbb{E}\left\{\left(\sum_{i=1}^{k} x(i)\right)^2 - \left(\sum_{i=2}^{k+1} x(i)\right)^2\right\} \\
&= \frac{1}{2}\mathbb{E}\left\{(y(k+1)-y_0)^2 + (y(k-1)-y_0)^2\right\} \\
&\quad - \frac{1}{2}\mathbb{E}\left\{(y(k)-y_0)^2 - (y(k)-y_0)^2\right\} \\
&= \frac{1}{2}\sigma^2\left[(k+1)^{2H} + (k-1)^{2H} - 2k^{2H}\right].
\end{aligned} \tag{90}$$

Thus, we conclude that the covariance function for the stationary increments of a self-similar process at any lag τ follows

$$\gamma_x(\tau) = \frac{1}{2}\sigma^2\left[(\tau+1)^{2H} - 2\tau^{2H} + (\tau-1)^{2H}\right]. \tag{91}$$

We further elaborate on the analysis of such function, in order to explicitly observe its long-memory behavior. First, we rewrite the above function as

$$\gamma_x(\tau) = \frac{1}{2}\sigma^2\tau^{2H}\left[\left(1+\frac{1}{\tau}\right)^{2H} - 2 + \left(1-\frac{1}{\tau}\right)^{2H}\right], \tag{92}$$

and compute its Taylor series expansion, i.e., $g(a) \approx g(a) + g'(a)(x-a) + \frac{g''(a)}{2}(x-a)^2$, at the origin

$$\begin{aligned}
\gamma_x(0) &\approx g(0) + g'(0)(x) + \frac{g''(0)}{2}(x)^2 \\
&\approx (1-2+1) + (2H-2H)\,x + \frac{2H(2H-1)+2H(2H-1)}{2}(x)^2 \\
&\approx 2H(2H-1)x^2.
\end{aligned} \tag{93}$$

All in all, when τ tends to infinity, $\frac{1}{\tau} \to 0$, and thus

$$\gamma_x(\tau) \approx \frac{1}{2}\sigma^2\tau^{2H}(2H)(2H-1)\tau^{-2} = \sigma^2 H(2H-1)\tau^{2H-2}. \tag{94}$$

We conclude that if a process with such an autocovariance function exists, and $\lim_{\tau\to\infty} \gamma_x(\tau) = 0$ is met, then

1. for $0 < H < \frac{1}{2}$, the process has short-memory;
2. for $H = \frac{1}{2}$, the process is uncorrelated; and
3. for $\frac{1}{2} < H < 1$, the process has long-memory.

We further study the properties of these stationary increment processes. For $0 < H < 1$, one can show [12] that its spectral density is, for $f \in [-\pi, \pi]$,

$$\begin{aligned} S_x(f) &= 2c_f(H, \sigma^2)\,(1 - \cos(f)) \sum_{i=\infty}^{\infty} |2\pi i + f|^{-2H-1}, \\ \text{with } c_f(H, \sigma^2) &= \frac{\sigma^2}{2\pi} \sin(\pi H)\Gamma(2H + 1). \end{aligned} \tag{95}$$

We evaluate the spectrum again at null frequencies by Taylor expansion of the above and obtain

$$S_x(f) \approx c_f(H, \sigma^2)|f|^{1-2H} + O\left(|f|^{\min(3-2H,2)}\right). \tag{96}$$

The numerical study of such function shows that it can be approximated with a straight line (a detailed analysis is provided in [12]).

So far, we have only considered self-similarity and stationarity in the most general sense. We now consider Gaussian processes and thus, we assume that the stationary increments x_t are Gaussian. Thus, the mean and the covariance are the sufficient statistics of the process x_t.

For each value of $H \in (0, 1)$, also known as the Hurst parameter, there is exactly one Gaussian process x_t that is the stationary increment of a self-similar process y_t. This process x_t is called fractional Gaussian noise (fGn). Its covariance and spectrum are given by Equation 91 and Equation 95, respectively. The corresponding self-similar process y_t is commonly referred to as the fractional Brownian motion B_t^H.

That is, for B_t^H a self-similar process with self-similarity parameter H with stationary Gaussian increments u_t, these increments $u_t = B_t^H - B_{t-1}^H$ are the fGn. For $H \in (\frac{1}{2}, 1)$, the fGp shows long-memory properties.

Long-memory processes: the FARIMA model

In applied time-series analysis, ARMA models and their extensions have been widely studied. FARIMA models are a natural extension of the classic ARMA models. We briefly describe in the following how one elaborates on ARMA models into first, ARIMA and then, FARIMA models.

As explained in detail in Section 3.1, the ARMA(p, q) model is mathematically described by

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \tag{97}$$

or, in lag-polynomial form

$$A(L)x_t = B(L)u_t \,, \qquad \text{where } \begin{cases} A(L) = 1 - a_1 L - a_2 L^2 - \cdots - a_p L^p \,, \\ B(L) = 1 + b_1 L + b_2 L^2 + \cdots + b_q L^q \,. \end{cases} \tag{98}$$

One extends these models by considering differentiation of a process. In discrete-time statistics, differencing refers to a transformation applied to a time-series, where the difference between consecutive data points is computed

$$x_t^{'} = x_t - x_{t-1}. \tag{99}$$

This operation removes the changes in the level of a time-series, thus eliminating trend and seasonality. Consequently, it stabilizes the mean of the time-series and usually results in stationarity.

Thus, if one computes differences of a time-series d times before adding the autoregressive and moving average parts, then the resulting process x_t is referred to as ARIMA(p, d, q) process, i.e.,

$$A(L)(1 - L)^d x_t = B(L)u_t. \tag{100}$$

This equation can also be understood as if the autoregressive polynomial has d unit roots. For this reason, every ARIMA model with $d > 0$ is not wide-sense stationary. For ARIMA models, the differencing has to be done d times and thus, only integer values of d are considered.

However, if d is allowed to take any real value in $(-\frac{1}{2}, \frac{1}{2})$, then the resulting stationary process is called a fractional

ARIMA(p, d, q) or FARIMA(p, d, q) process. For $0 \leqslant d < \frac{1}{2}$, the process has long-memory and when $d \geqslant \frac{1}{2}$, the process is not stationary.

The differentiation with non-integer values is mathematically explained by using, for $d \geqslant 0$,

$$(1-L)^d = \sum_{k=0}^{d} \binom{d}{k} (-1)^k L^k, \tag{101}$$

where the binomial coefficients follow

$$\binom{d}{k} = \frac{d!}{k!(d-k)!} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}. \tag{102}$$

The gamma function $\Gamma(\cdot)$ is defined for all real numbers and thus, Equation 100 can be extended to all real numbers d for the definition of FARIMA(p, d, q) models.

There are two different but complementary interpretations of the FARIMA(p, d, q) model:

1. An ARMA(p, q) process is obtained by d differencing a FARIMA(p, d, q) process

$$(1-L)^d x_t = \tilde{x}_t, \tag{103}$$

   where $\tilde{x}_t$ is an ARMA(p, q) process $\tilde{x}_t = A^{-1}(L)B(L)u_t$.

2. A FARIMA(p, d, q) process is obtained by filtering a FARIMA$(0, d, 0)$ process through an ARMA(p, q) filter

$$x_t = A^{-1}(L)B(L)x_t^*, \tag{104}$$

   where x_t^* is FARIMA$(0, d, 0)$ process $x_t^* = (1-L)^{-d}u_t$.

The interpretation in Equation 103 allows for derivation of the spectrum density of FARIMA(p, d, q) models. One simply needs to multiply the spectrum of the ARMA(p, q) model with that of a linear differencing filter. Thus, given $h(t) = \sum c_s x_{t-s}$ and its spectrum density $H(f) = \sum c_s e^{isf}$, one derives

$$S_X(f) = |H(f)|^2 \cdot S_{\tilde{X}}(f) = |1 - e^{if}|^{-2d} \cdot \frac{\sigma^2}{2\pi} \frac{|B(f)|^2}{|A(f)|^2}. \tag{105}$$

By replacing $|1 - e^{if}| = 2\sin(\frac{f}{2})$, and taking the limit

$$\lim_{f \to 0} \frac{2\sin(\frac{f}{2})}{f} = 1, \tag{106}$$

the behavior of the spectral density at the origin follows

$$S_X(0) \approx \frac{\sigma^2}{2\pi}\frac{|B(1)|^2}{|A(1)|^2}|f|^{-2d}. \tag{107}$$

The derivation of explicit formulas for the covariance function of the general FARIMA(p, d, q) model is very challenging. Nonetheless, we can, for $0 < d < \frac{1}{2}$, obtain asymptotic formulas when $\tau \to \infty$,

$$\begin{aligned}\gamma_x(\tau) &\approx c_\gamma(d, A, B)|\tau|^{2d-1},\\ &\text{where } c_\gamma(d, A, B) = \frac{\sigma^2}{\pi}\frac{|B(1)|^2}{|A(1)|^2}\Gamma(1-2d)\sin(\pi d).\end{aligned} \tag{108}$$

More specifically, for the interesting FARIMA$(0, d, 0)$ case we have

$$\begin{cases}\gamma_x(\tau) = \sigma^2\frac{(-1)^\tau\Gamma(1-2d)}{\Gamma(\tau-d+1)\Gamma(1-\tau-d)},\\ \rho_x(\tau) = \frac{\Gamma(1-d)\Gamma(\tau+d)}{\Gamma(d)\Gamma(\tau+1-d)}.\end{cases} \tag{109}$$

Note that with this equation and by following the intuition in Equation 104, one can derive the covariances for any particular FARIMA(p, d, q) model. That is, one needs to combine the above with the ARMA(p, q) covariance function.

Finally, due to Equation 102, one can express the FARIMA$(0, d, 0)$ model as an infinite AR or MA process

1. $\sum_{k=0}^{\infty} a_k x_{t-k} = u_t$, with $\quad a_k = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}$.
2. $x_t = \sum_{k=0}^{\infty} b_k u_{t-k}$, with $\quad b_k = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)}$.

Long-memory processes: model comparison

From the descriptions in Section 5.1.1 and Section 5.1.2, it is evident that both processes are not identical. Nonetheless, there are many shared features.

Let us summarize the properties, differences and similarities of the fGp and the FARIMA models.

- Both the fGp and the FARIMA model have long-memory if $\frac{1}{2} < H < 1$ and $0 \leqslant d < \frac{1}{2}$, respectively:
  1. For $\tau \to \infty$, the covariance function follows long-memory as in Equation 79.

2. For $f \to 0$, the spectrum follows long-memory as in Equation 80.

- The autocorrelation function for a fGp is defined for all lags

$$\rho_x(\tau) = \frac{1}{2}\left[(\tau+1)^{2H} - 2\tau^{2H} + (\tau-1)^{2H}\right], \tag{110}$$

and is shown for different H values in Figure 17.

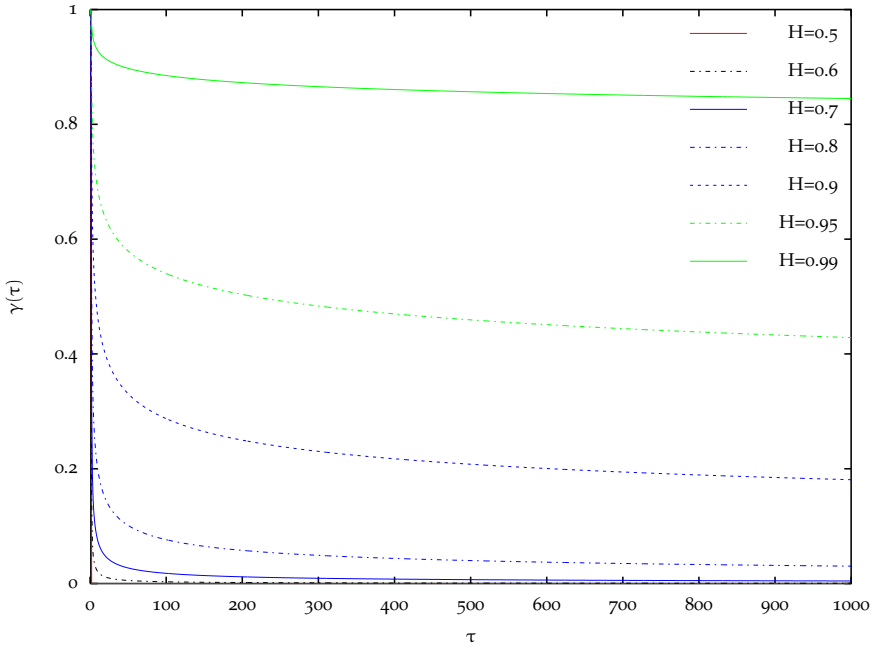


Figure 17: $\gamma(\tau)$ for fGps, as a function of the Hurst parameter.

The autocorrelation for the FARIMA$(0, d, 0)$ model is

$$\rho_x(\tau) = \frac{\Gamma(1-d)\Gamma(\tau+d)}{\Gamma(d)\Gamma(\tau+1-d)}, \tag{111}$$

where for big τ values, we can approximate it with

$$\rho_x(\tau) \approx \frac{\Gamma(1-d)}{\Gamma(d)}\tau^{2d-1}. \tag{112}$$

We illustrate the differences in the autocorrelation functions for the fGp and FARIMA$(0, d, 0)$ in Figure 18. Further comparison details are included in Section D.2.

- The FARIMA$(0, d, 0)$ model spectrum has a simple expression for all frequencies

$$S_x(f) = |1 - e^{if}|^{-2d} = |2\sin(\frac{f}{2})|^{-2d}. \tag{113}$$

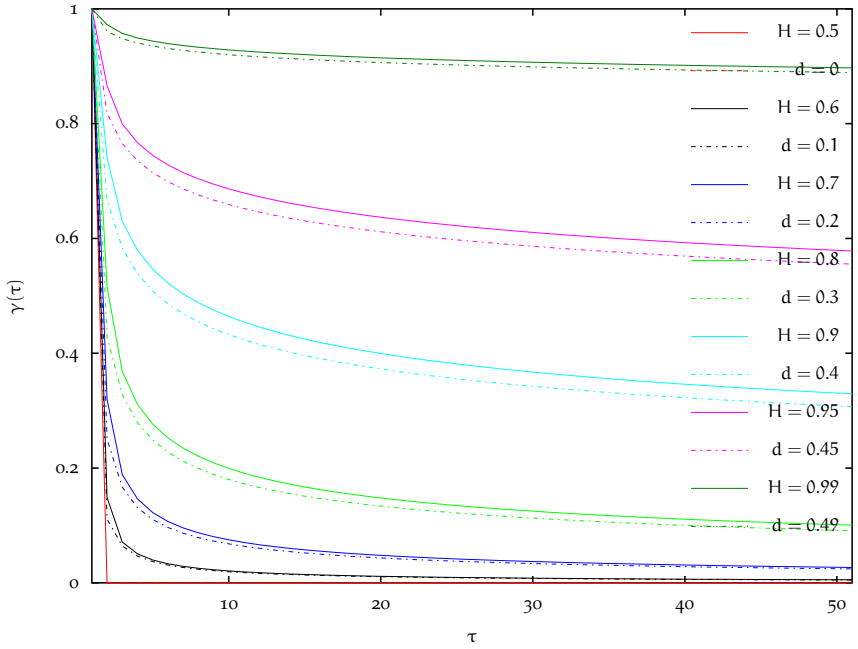


Figure 18: $\gamma(\tau)$ for fGps and FARIMA$(0, d, 0)$, as a function of the Hurst parameter.

The spectrum of the fGp is

$$S_x(f) = 2\frac{\sigma^2}{2\pi}\sin(\pi H)\Gamma(2H+1)\,(1-\cos(f))\sum_{i=\infty}^{\infty}|2\pi i + f|^{-2H-1}. \tag{114}$$

- The low-frequency spectrum is identical for both processes. If $d = H - \frac{1}{2}$ and $f \to 0$, then
  1. $S_x(f) \propto |\sin \pi f|^{-2d} \approx f^{-2d}$ for the FARIMA$(0, d, 0)$ model.
  2. $S_x(f) \approx c_f(H, \sigma^2)|f|^{1-2H}$ for the fGp.

SMC METHOD FOR LATENT FRACTIONAL GAUSSIAN PROCESSES

In this chapter, we are interested in the study of latent long-memory processes observed through nonlinear functions. In particular, we study the stationary increments of the fractional Brownian motion, that is, the fGp described in Section 5.1.1. Inference of the alternative FARIMA(p, d, q) model (as in Section 5.1.2) is left aside for now, as it is addressed in the forthcoming Chapter 6.

Here, we focus on the inference of latent fGps with nonlinear observations. Mathematically, we model the problem at hand by the following state-space representation

$$\begin{cases} x_t = u_t, \\ y_t = h(x_t, v_t), \end{cases} \tag{115}$$

where u_t represents a zero-mean fGp with autocovariance function

$$\gamma_u(\tau) = \frac{\sigma_u^2}{2}\left[|\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H}\right], \tag{116}$$

where σ^2 is the variance of the process.

For $\frac{1}{2} < H < 1$, the process has long-range dependence; and, for $H = 0.5$, the observations are uncorrelated, as the expression in Equation 116 simplifies to the Kronecker delta function: i.e., $\gamma(\tau) = \sigma^2\delta(\tau)$. We illustrated the memory properties of the fGp by plotting its autocovariance function for different H values in Figure 17.

If short-memory processes, such as ARMA(p, q) models, were used for modeling non-quickly decaying autocovariance functions, one would need to increase the model orders p and q. Actually, the number of parameters tends to infinity. In practice, an excessive number of parameters is undesirable for many reasons and, especially, because it increases the uncertainty of the statistical inference. Therefore, the parsimonious description provided by the fractional Gaussian process is a more satisfactory model to describe such autocovariance decays (see Figure 17 and the long-term dependence as $H \to 1$).

Our goal hear is to sequentially infer the hidden state as new observations become available. To do so, we are interested on the filtering density $f(x_{t+1}|y_{1:t+1})$, which, due to the nonlinearities considered, is approximated by a random measure updated sequentially by SMC sampling. The challenge is on deriving the densities for fractional Gaussian processes that result in an efficient SMC solution.

The latent state is a fGp and thus, due to the Gaussianity and stationarity of the process, the sufficient statistics are its mean and the autocovariance function. For simplicity, we assume here a zero-mean stationary fGp and thus, its autocovariance function follows Equation 116.

Therefore, at time instant $t+1$, the joint distribution of the whole process is a zero-mean multivariate Gaussian, i.e., $x_{1:t+1} \sim \mathcal{N}(x_{1:t+1}|0, C_{t+1})$, where

$$C_{t+1} = \begin{pmatrix} \gamma_u(0) & c_{t+1} \\ c_{t+1}^\top & C_t \end{pmatrix}, \tag{117}$$

and

$$\begin{cases} c_{t+1} = \begin{pmatrix} \gamma_u(1) & \gamma_u(2) & \cdots & \gamma_u(t-1) & \gamma_u(t) \end{pmatrix}, \\ C_t = \begin{pmatrix} \gamma_u(0) & \gamma_u(1) & \gamma_u(2) & \cdots & \gamma_u(t-2) & \gamma_u(t-1) \\ \gamma_u(1) & \gamma_u(0) & \gamma_u(1) & \cdots & \gamma_u(t-3) & \gamma_u(t-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_u(t-3) & \gamma_u(t-4) & \gamma_u(t-5) & \cdots & \gamma_u(1) & \gamma_u(2) \\ \gamma_u(t-2) & \gamma_u(t-3) & \gamma_u(t-4) & \cdots & \gamma_u(0) & \gamma_u(1) \\ \gamma_u(t-1) & \gamma_u(t-2) & \gamma_u(t-3) & \cdots & \gamma_u(1) & \gamma_u(0) \end{pmatrix}. \end{cases}$$

Note that the covariance matrix of the joint distribution of the fractional Gaussian process at time $t+1$ is determined by computing the symmetric Toeplitz matrix of the autocovariance function up to lag $\tau = t$.

From this joint distribution and, given the last t samples of the process, we can compute the conditional distribution of the next sample x_{t+1}, given $x_{1:t}$.

It readily follows that the resulting transition density is the univariate Gaussian distribution

$$\begin{aligned} &f(x_{t+1}|x_{1:t}, \sigma_u^2) = \mathcal{N}(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma^2_{x_{t+1}|x_{1:t}}), \\ &\text{where} \begin{cases} \mu_{x_{t+1}|x_{1:t}} = c_{t+1} C_t^{-1} x_{1:t}, \\ \sigma^2_{x_{t+1}|x_{1:t}} = \gamma_u(0) - c_{t+1} C_t^{-1} c_{t+1}^\top. \end{cases} \end{aligned} \tag{118}$$

By relating the autocovariance and autocorrelation functions as follows

$$\gamma_u(\tau) = \sigma_u^2 \rho(\tau) = \sigma_u^2 \left(\frac{1}{2} \left[|\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H} \right] \right), \tag{119}$$

one can readily derive, as in previous Chapter 4, a normalized covariance matrix $C_{t+1} = \sigma_u^2 \widetilde{C}_{t+1}$, where

$$\widetilde{C}_{t+1} = \begin{pmatrix} \widetilde{\gamma}_u(0) & \widetilde{c}_{t+1} \\ \widetilde{c}_{t+1}^\top & \widetilde{C}_t \end{pmatrix}, \tag{120}$$

with

$$\begin{cases} \widetilde{\gamma}_u(0) = \rho_u(0), \\ \widetilde{c}_{t+1} = \begin{pmatrix} \rho_u(1) & \rho_u(2) & \cdots & \rho_u(t-1) & \rho_u(t) \end{pmatrix}, \\ \widetilde{C}_t = \begin{pmatrix} \rho_u(0) & \rho_u(1) & \rho_u(2) & \cdots & \rho_u(t-2) & \rho_u(t-1) \\ \rho_u(1) & \rho_u(0) & \rho_u(1) & \cdots & \rho_u(t-3) & \rho_u(t-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_u(t-3) & \rho_u(t-4) & \rho_u(t-5) & \cdots & \rho_u(1) & \rho_u(2) \\ \rho_u(t-2) & \rho_u(t-3) & \rho_u(t-4) & \cdots & \rho_u(0) & \rho_u(1) \\ \rho_u(t-1) & \rho_u(t-2) & \rho_u(t-3) & \cdots & \rho_u(1) & \rho_u(0) \end{pmatrix}. \end{cases} \tag{121}$$

This new notation allows for consideration of the unknown noise variance σ_u^2 case. Following the same approach as in previous chapters, we apply Rao-Blackwellization and thus, one can readily derive the marginalized transition density (details are equivalent to those in Section 4.2.2)

$$f(x_{t+1}|x_{1:t}) = \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right)$$

$$\text{with} \begin{cases} \nu_t = \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} = \widetilde{c}_{t+1}^\top \widetilde{C}_t^{-1} x_{1:t}\,, \\ \sigma_t^2 = \frac{\nu_0\sigma_0^2 + x_{1:t}^\top \widetilde{C}_t^{-1} x_{1:t}}{\nu_t}\,, \\ c_{t+1} = \rho_u(0) - \widetilde{c}_{t+1}\widetilde{C}_t^{-1}\widetilde{c}_{t+1}^\top\,. \end{cases} \tag{122}$$

We leverage the transition densities in Equation 118 and Equation 122 to present an SMC sampling method for latent fractional Gaussian processes in Table 21.

PF FOR LATENT FGP WITH KNOWN H PARAMETER

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

4. Propagate the particles by sampling from the transition density, given the (resampled) stream of particles:
   - If σ_u^2 is known,

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}\sigma_u^2) = \mathcal{N}(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, \sigma^2_{x_{t+1}|x_{1:t}}),$$

$$\text{with} \begin{cases} \mu_{x_{t+1}|x_{1:t}} = c_{t+1}C_t^{-1}\overline{x}_{1:t}, \\ \sigma^2_{x_{t+1}|x_{1:t}} = \gamma_u(0) - c_{t+1}C_t^{-1}c_{t+1}^\top. \end{cases}$$

   - If σ_u^2 is unknown,

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}) = \mathcal{T}_{\nu_t}\left(x_{t+1}|\mu_{x_{t+1}|x_{1:t}}, c_{t+1}\sigma_t^2\right),$$

$$\text{with} \begin{cases} \nu_t = \nu_0 + t\,, \\ \mu_{x_{t+1}|x_{1:t}} = \widetilde{c}_{t+1}^\top \widetilde{C}_t^{-1}\overline{x}_{1:t}^{(m)}\,, \\ \sigma_t^2 = \frac{\nu_0\sigma_0^2 + \overline{x}_{1:t}^{(m)\top}\widetilde{C}_t^{-1}\overline{x}_{1:t}^{(m)}}{\nu_t}\,, \\ c_{t+1} = \rho_u(0) - \widetilde{c}_{t+1}\widetilde{C}_t^{-1}\widetilde{c}_{t+1}^\top\,. \end{cases}$$

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 21: PF for latent fGp with known H parameter.

EVALUATION

We evaluate the proposed SMC sampling method under the SV model, as in previous chapters (see Section 3.3 and Section 4.3). However, we now consider a latent fGp, that is, the log-volatility of the observed time-series is a fractional Gaussian process. Volatility models with long-memory characteristics have been widely studied in finance [8]. The goal there is to estimate the evolving volatility of an observed series of stock prices.

Mathematically, the state-space model is written as

$$\begin{cases} x_t = u_t, \\ y_t = e^{x_t/2} v_t, \end{cases} \tag{123}$$

where u_t is a zero-mean fractional Gaussian process with autocovariance function as in Equation 116 and $v_t \sim \mathcal{N}(v_t|0, \sigma_v^2)$. For the following results, we assume $\sigma^2 = 1$ and $\sigma_v^2 = 1$ and evaluate the performance of the proposed SMC method in Table 21 for different Hurst parameter values. The presented results are MSEs averaged over 100 realizations of 200 instants long timeseries, and $M = 1000$ particles are used.

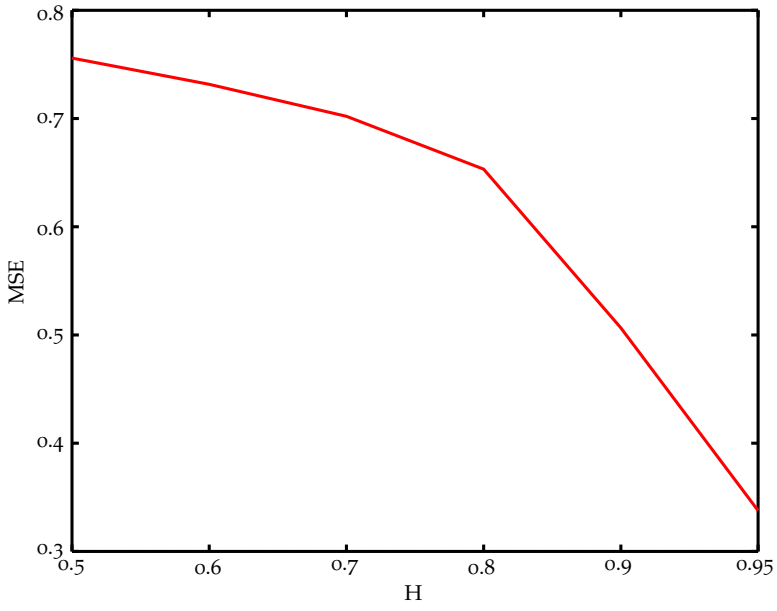


Figure 19: MSE of the proposed SMC method for latent fGp with different Hurst parameters.

The state filtering accuracy of the proposed SMC for the known Hurst parameter case is presented in Figure 19. The

results illustrate the implicit benefits of tracking long-memory processes. As $H \to 1$, more information is provided by past samples in estimating the present value of the series and thus, the performance of the SMC sampling technique improves. That is, it is possible to more accurately estimate the time-evolving log-variance of the observed time-series, because of the slow decay of the autocovariance function in such long-memory processes.

Note that the proposed SMC and the results in Figure 19 require computation of the full autocovariance function of the fGp process up to time instant t, as shown in Equation 121. Thus, in order to reduce computational cost, one may be tempted to truncate the autocovariance function up to a maximum lag τ_{max} (along the lines of what was proposed for ARMA models in Chapters 3 and 4).

We implemented a truncated version of the SMC method in Table 21, where the truncated sufficient statistics follow

$$\begin{cases} \mu_{x_{t+1}|x_{t-\tau_{max}:t}} = c_{\tau_{max}+1} C^{-1}_{\tau_{max}} x_{t-\tau_{max}:t}, \\ \sigma^2_{x_{t+1}|x_{t-\tau_{max}:t}} = \gamma_u(0) - c_{\tau_{max}+1} C^{-1}_{\tau_{max}} c^{\top}_{\tau_{max}+1}. \end{cases} \tag{124}$$

This reduces the computational cost of the method, as one computes $\gamma(\tau)$ only for a short window $\tau = 0, \cdots, \tau_{max}$. The question however, is how much information is lost when truncating the sufficient statistics. We provide MSE results for latent fGps with different Hurst parameters as a function of the truncation lag τ_{max} in Figure 20, Figure 21, Figure 22, Figure 23, Figure 24 and Figure 25.

The results indicate that truncation is, indeed, only feasible for short-memory processes (i.e., $0.5 \leqslant H < 0.75$), while non-negligible information loss is observed as the memory of the process increases (i.e., $H \to 1$). These findings in terms of MSE are backed up by the analysis provided in Section D.1. There, we can observe how quickly the fGp forgets the past, both in terms of the autocovariance function $\gamma(\tau)$ or the key $-\alpha_t = \widetilde{c}_t^{-1} \widetilde{C}^{\top}_{t+1}$ term. Only for $0.5 \leqslant H < 0.75$ values one can consider the process to forget quickly.

Thus, one can only truncate the sufficient statistics for processes with short-memory features, as we save in computation and memory resources without sacrificing in performance. On the contrary, truncation should be avoided when dealing with long-memory processes.

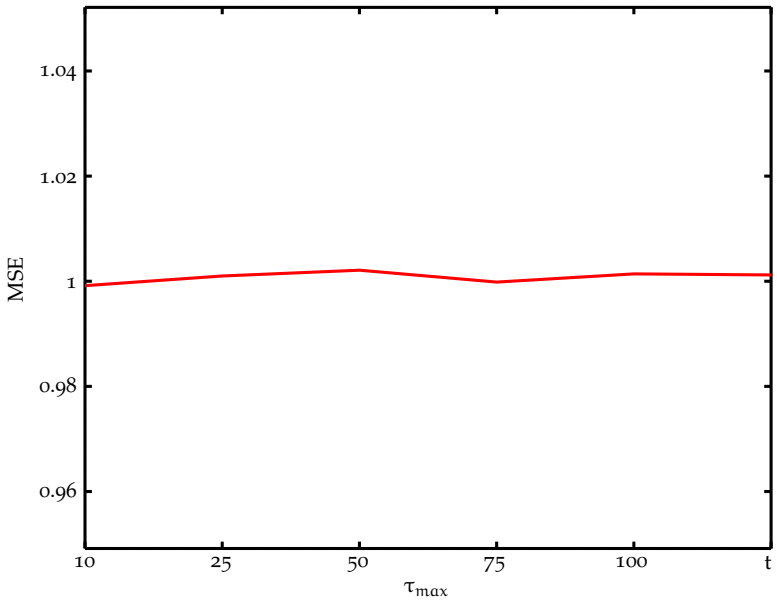


Figure 20: MSE of SMC method with truncation for fGp with $H = 0.5$.

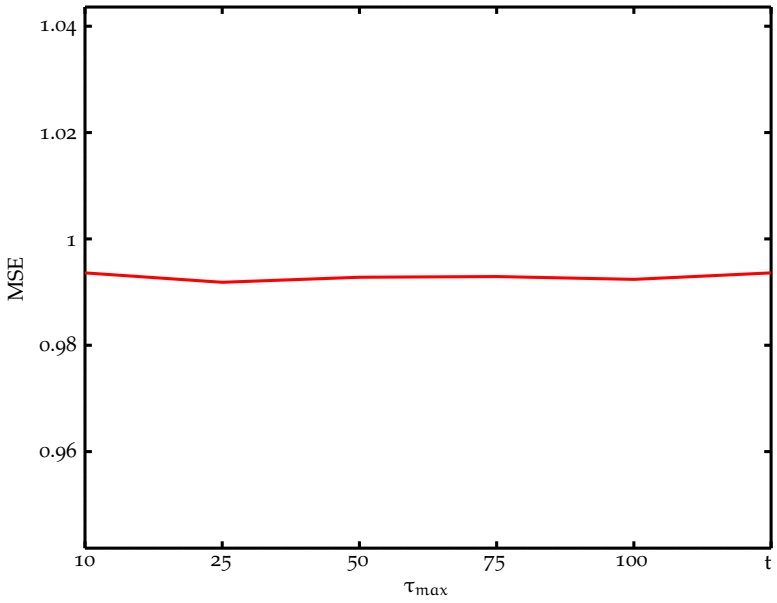


Figure 21: MSE of SMC method with truncation for fGp with $H = 0.6$.

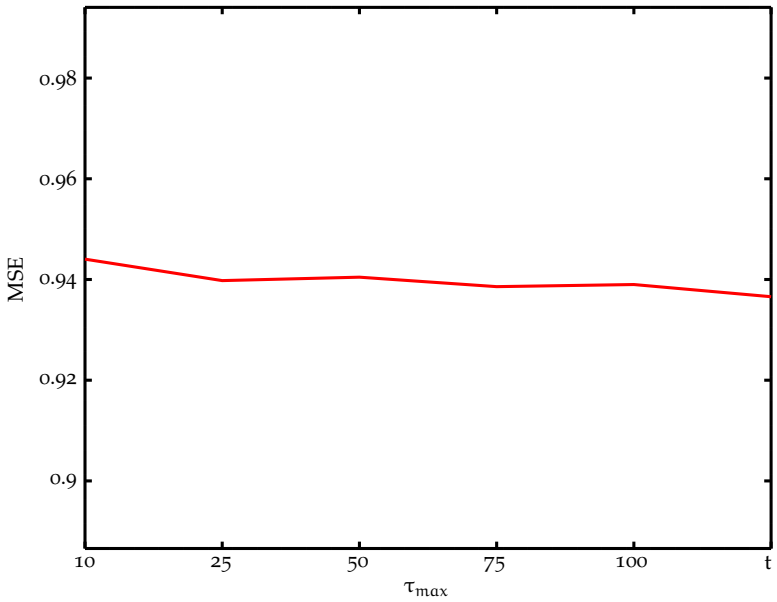


Figure 22: MSE of SMC method with truncation for fGp with $H = 0.7$.

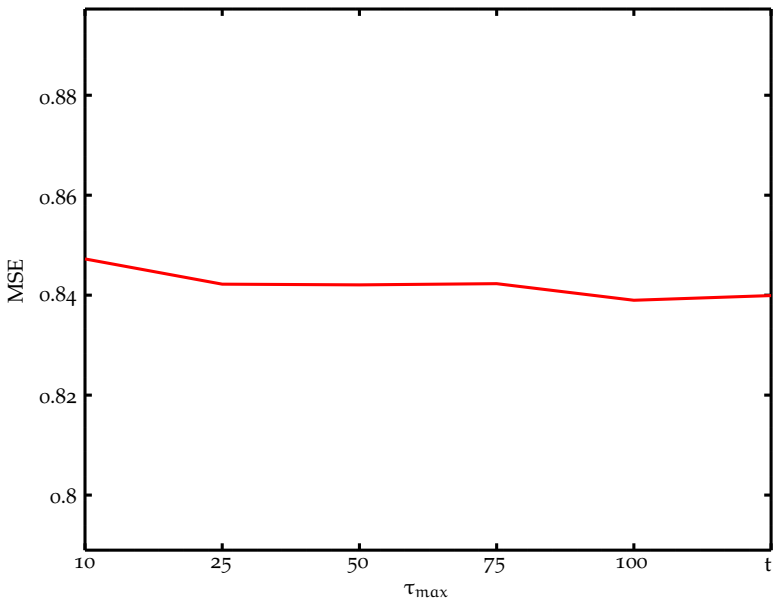


Figure 23: MSE of SMC method with truncation for fGp with $H = 0.8$.

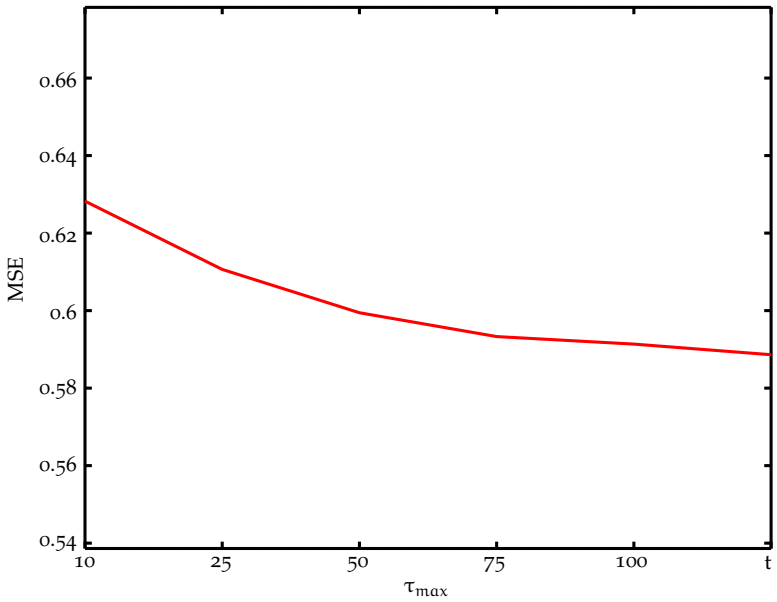


Figure 24: MSE of SMC method with truncation for fGp with $H = 0.9$.

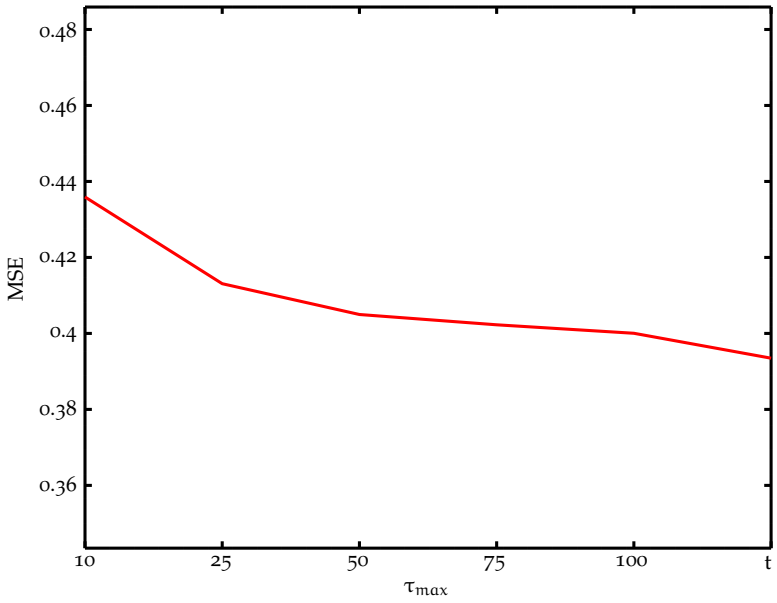


Figure 25: MSE of SMC method with truncation for fGp with $H = 0.95$.

As of now, it is clear that accurately computing the sufficient statistics of the fGp is critical for the performance of the proposed method. That is, one needs to compute the autocovariance function as in Equation 116, which is parameterized by the Hurst parameter H. In practice though, knowing the exact memory properties of a latent time-series is not possible and thus, one needs to consider alternatives when the autocovariance function is unknown.

However, estimation of all the unknowns in $\gamma(\tau)$ is too ambitious of a task. The alternative, that is, to estimate the unknown Hurst parameter and, from there, derive the autocovariance function up to time-instant t, is also a very challenging task, due to the nonlinear dependency of the parameter with the data.

On the one hand, the most popular methods on estimating the Hurst parameter H rely on direct observation of x_t [98, 100]. However, this is not the case when dealing with latent fGps, as is the case here.

On the other, since H is an unknown static parameter, the proposed SMC will have problems in dealing with it [72]. To overcome such limitations on estimating static parameters with PFs, various methodologies have been suggested [39, 72]. In these, an artificial model is assumed for the fixed parameter, either by assuming a slowly varying evolution of the parameter or by approximating the joint distribution of the parameter and the state.

However, these approaches raise several concerns in our problem of interest, i.e., the estimation of the Hurst parameter of a latent fGp. Assuming a varying H would not only break important properties of the process (i.e., stationarity and self-similarity), but would also affect the mixing properties of the states, thus endangering the convergence of the SMC method. Besides, determining a suitable density for H is very challenging, due to its complicated dependency on the state.

Thus, we only consider here the SMC that assumes knowledge of the autocovariance function of the latent fGp. Nonetheless, alternatives to deal with the unknown Hurst parameter are later presented in Chapter 9.

6

LATENT ARMA(p, q) PROCESSES WITH CORRELATED INNOVATIONS

In this chapter, we extend the analysis of time-series models that are not limited to short-memory properties. We study a very flexible set of time-series, which are ARMA(p, q) models that are driven by non-iid noise. That is, we allow for the innovation process to be correlated over time.

This new modeling approach generalizes previously studied models (i.e., ARMA(p, q) and fGp) and, furthermore, allows for modeling of a wide range of memory properties. Note that, amongst the processes that this new model can describe are the FARIMA(p, d, q) models. As explained in Section 5.1.2, these can be understood as fGps or FARIMA(0, d, 0) processes filtered by an ARMA(p, q) model. We hereby generalize these notion and consider any correlated innovation driving an ARMA(p, q) filter.

In the following, we analyze the properties of these time-series and derive the joint and transition densities of such processes in Section 6.1. In Section 6.2, we leverage from the previous analysis to propose, for different parameter knowledge assumptions, an SMC method for inference of latent ARMA(p, q) processes with correlated innovations. We conclude by evaluating in Section 6.3 the performance of the proposed methods in the illustrative SV model with latent ARMA with fGp innovations.

BAYESIAN ANALYSIS OF THE CORRELATED TIME-SERIES

We hereby investigate latent time-series with correlated innovations, observed via nonlinear functions. That is, we consider hidden dynamics that follow an ARMA(p, q) model with non-iid innovations.

Mathematically, we represent the model of interest as

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j} \,, \\ y_t = h(x_t, v_t) \,, \end{cases} \tag{125}$$

where u_t is a zero-mean Gaussian innovation process that is correlated in time. It is described by an arbitrary autocovariance function $\gamma_u(\tau)$. For added modeling flexibility, we do not enforce any restriction on the form of such autocovariance function, as long as stationarity conditions are met.

We proceed with a Bayesian analysis of ARMA processes driven by a time-correlated innovation process, as in Equation 125. We aim at deriving the joint distribution of the time-series up to time instant t, i.e., $x_{1:t} \sim f(x_{1:t})$. This density contains all the relevant information of the time-series, from which any marginal and conditional of interest can be directly derived.

We start by reformulating the state in Equation 125, where we fix $b_0 = 1$,

$$\begin{aligned} x_t &= \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=0}^{q} b_j u_{t-j} \,, \\ x_t - \sum_{i=1}^{p} a_i x_{t-i} &= \sum_{j=0}^{q} b_j u_{t-j} \,. \end{aligned} \tag{126}$$

Now, we rewrite the above recursion for a time-series up to time instant t in matrix form

$$\begin{pmatrix} 1 & -a_1 & -a_2 & \cdots & -a_p & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -a_1 & \cdots & -a_{p-1} & -a_p & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & -a_{p-2} & -a_{p-1} & -a_p & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -a_1 & -a_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -a_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \\ x_3 \\ x_2 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 & b_1 & b_2 & \cdots & b_q & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & b_1 & \cdots & b_{q-1} & b_q & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & b_{q-2} & b_{q-1} & b_q & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & b_1 & b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & b_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_t \\ u_{t-1} \\ u_{t-2} \\ \vdots \\ u_3 \\ u_2 \\ u_1 \end{pmatrix} . \tag{127}$$

If we define matrices

$$\begin{aligned} A_t &= \begin{pmatrix} 1 & -a_1 & -a_2 & \cdots & -a_p & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -a_1 & \cdots & -a_{p-1} & -a_p & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & -a_{p-2} & -a_{p-1} & -a_p & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -a_1 & -a_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -a_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{t \times t}, \\ B_t &= \begin{pmatrix} 1 & b_1 & b_2 & \cdots & b_q & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & b_1 & \cdots & b_{q-1} & b_q & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & b_{q-2} & b_{q-1} & b_q & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & b_1 & b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & b_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{t \times t}, \end{aligned} \tag{128}$$

where the subscript t indicates matrix dimensionality, then one can immediately write $A_t x_{1:t} = B_t u_{1:t}$ and, thus

$$x_{1:t} = A_t^{-1} B_t u_{1:t}. \tag{129}$$

Note that in this chapter, the innovation process u_t can be correlated in time. We now consider a Gaussian innovation process where its sufficient statistics are the mean (usually assumed to be zero) and an arbitrary autocovariance function $\gamma_u(\tau) = \sigma_u^2 \rho_u(\tau)$.

Consequently, the distribution of the vector of innovations up to time instant t (i.e., $u_{1:t}$) is a zero-mean Gaussian density with covariance matrix $C_{u_t} = \sigma_u^2 R_{u_t}$, where

$$R_{u_t} = \begin{pmatrix} \rho_u(0) & \rho_u(1) & \cdots & \rho_u(t-2) & \rho_u(t-1) \\ \rho_u(1) & \rho_u(0) & \cdots & \rho_u(t-3) & \rho_u(t-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_u(t-2) & \rho_u(t-3) & \cdots & \rho_u(0) & \rho_u(1) \\ \rho_u(t-1) & \rho_u(t-2) & \cdots & \rho_u(1) & \rho_u(0) \end{pmatrix}. \tag{130}$$

Note the symmetric Toeplitz structure of the matrix, where the fundamental row (column) contains the values of the autocorrelation $\rho_u(\tau)$, for lags $\tau = 0, 1, \cdots, t-1$.

We leverage these sufficient statistics of the correlated innovations and the formulation of the state equation as in Equation 129 to derive the joint density of the time-series of interest.

ARMA(p, q) with correlated innovations: joint density

Let us consider a time-series at time instant t with correlated innovations, i.e., we write

$$x_{1:t} = A_t^{-1} B_t u_{1:t}\,, \qquad u_{1:t} \sim \mathcal{N}(0, \sigma_u^2 R_{u_t}). \tag{131}$$

Due to the properties of the Gaussian distribution, the joint density of the time-series $x_{1:t}$ is a zero-mean Gaussian with a covariance matrix dependent on the matrices A_t, B_t and R_{u_t}. Specifically,

$$\begin{aligned} x_{1:t} \sim f(x_{1:t}|\sigma_u^2) = \mathcal{N}\left(x_{1:t}|0, \sigma_u^2 \Sigma_t\right)\,, \\ \Sigma_t = A_t^{-1} B_t R_{u_t} B_t^\top (A_t^{-1})^\top \in \mathbb{R}^{t\times t}\,. \end{aligned} \tag{132}$$

Because in any practical scenario knowledge of the underlying innovation variance is troublesome, it is of interest to

consider the unknown noise variance σ_u^2 case. Once again, we resort to marginalization of the unknown (i.e., we Rao-Blackwellize σ_u^2). We do so by considering conjugate priors, due to their convenient analytical properties.

We assume a scaled inverse Chi-squared prior for the unknown variance,

$$f(\sigma_u^2) = \chi^{-2}(\sigma_u^2|\nu_0, \sigma_0^2) = \frac{\left(\sigma_0^2\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\frac{1}{(\sigma_u^2)^{1+\frac{\nu_0}{2}}}e^{-\frac{\nu_0\sigma_0^2}{2\sigma_u^2}} \ . \quad (133)$$

The marginalization of σ_u^2 in the joint density follows

$$\begin{aligned}
f(x_{1:t}) &= \int_0^\infty f(x_{1:t}|\sigma_u^2)f(\sigma_u^2)d\sigma_u^2 \\
&= \int_0^\infty (2\pi)^{-\frac{t}{2}}(\sigma_u^2)^{-\frac{t}{2}}\,|\Sigma_t|^{-\frac{1}{2}}\,e^{-\frac{1}{2\sigma_u^2}x_{1:t}^\top\Sigma_t^{-1}x_{1:t}} \\
&\qquad \times \frac{\left(\sigma_0^2\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\frac{1}{(\sigma_u^2)^{1+\frac{\nu_0}{2}}}e^{-\frac{\nu_0\sigma_0^2}{2\sigma_u^2}}\,d\sigma_u^2 \\
&\propto \int_0^\infty (\sigma_u^2)^{-\left(1+\frac{\nu_0+t}{2}\right)}e^{-\frac{1}{2\sigma_u^2}\left(\nu_0\sigma_0^2+x_{1:t}^\top\Sigma_t^{-1}x_{1:t}\right)}d\sigma_u^2 \quad (134) \\
&\propto \left[\left(\nu_0\sigma_0^2 + x_{1:t}^\top\Sigma_t^{-1}x_{1:t}\right)\right]^{-\frac{\nu_0+t}{2}} \\
&\propto \left[\left(1+\frac{1}{\nu_0}x_{1:t}^\top\left(\sigma_0^2\Sigma_t\right)^{-1}x_{1:t}\right)\right]^{-\frac{\nu_0+t}{2}} \\
&= \mathcal{T}_{\nu_0}(x_{1:t}|0, \sigma_0^2\Sigma_t) \ .
\end{aligned}$$

Consequently, the joint density of the time-series at time instant t after marginalization of the unknown variance is a multivariate Student-t density

$$\begin{aligned}
&f(x_{1:t}) = \mathcal{T}_\nu\left(x_{1:t}|\mu_{1:t}, \Phi_{1:t}\right) \\
&= \frac{\Gamma\left(\frac{\nu+t}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{\frac{t}{2}}\nu^{\frac{t}{2}}\,|\Phi|^{\frac{1}{2}}}\cdot\left(1+\frac{\left(x_{1:t}-\mu_{1:t}\right)^\top\Phi^{-1}\left(x_{1:t}-\mu_{1:t}\right)}{\nu}\right)^{-\left(\frac{\nu+t}{2}\right)},
\end{aligned} \quad (135)$$

with $\nu = \nu_0$ degrees of freedom, location parameter $\mu_{1:t} = 0$ and scale matrix $\Phi_{1:t} = \sigma_0^2\Sigma_t$.

ARMA(p, q) with correlated innovations: transition density

For inference and prediction of time-series, it is of practical interest to derive the transition density of the time-series considered. To do so, we leverage the joint densities from Section 6.1.1.

Let us write the joint density of the time-series at time instant $t+1$, i.e.,

$$\begin{cases} f(x_{1:t+1}|\sigma_u^2) = \mathcal{N}\left(x_{1:t+1}|0, \sigma_u^2\Sigma_{t+1}\right), & \text{if } \sigma_u^2 \text{ is known,} \\ f(x_{1:t+1}) = \mathcal{T}_{\nu_0}(x_{1:t+1}|0, \sigma_0^2\Sigma_{t+1}), & \text{if } \sigma_u^2 \text{ is unknown,} \end{cases} \tag{136}$$

and rewrite the covariance matrix Σ_{t+1} in block form

$$\Sigma_{t+1} = A_{t+1}^{-1}B_{t+1}R_{u_{t+1}}B_{t+1}^\top(A_{t_1}^{-1})^\top = \begin{pmatrix} h_{t+1} & \lambda_t \\ \lambda_t^\top & \Sigma_t \end{pmatrix}, \tag{137}$$

where $h_{t+1} \in \mathbb{R}^{1\times 1}$, $\lambda_t \in \mathbb{R}^{1\times t}$ and $\Sigma_t \in \mathbb{R}^{t\times t}$.

By means of the expressions for the conditional densities of the Gaussian and Student-t distributions [13], one can readily derive the transition densities of the time-series:

- If σ_u^2 is known,

$$f(x_{t+1}|x_{1:t}, \sigma_u^2) = \mathcal{N}\left(x_{t+1}|\mu_{t+1}, \sigma_{t+1}^2\right), \quad \text{with} \begin{cases} \mu_{t+1} = \lambda_t\Sigma_t^{-1}x_{1:t}, \\ \sigma_{t+1}^2 = \sigma_u^2\left(h_{t+1} - \lambda_t\Sigma_t^{-1}\lambda_t^\top\right). \end{cases} \tag{138}$$

- If σ_u^2 is unknown,

$$f(x_{t+1}|x_{1:t}) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1}|\mu_{t+1}, \phi_{t+1}^2\right), \quad \text{with} \begin{cases} \nu_{t+1} = \nu_0 + t, \\ \mu_{t+1} = \lambda_t\Sigma_t^{-1}x_{1:t}, \\ \phi_{t+1}^2 = \frac{\nu_0\sigma_0^2 + x_{1:t}^\top\Sigma_t^{-1}x_{1:t}}{\nu_0+t}\left(h_{t+1} - \lambda_t\Sigma_t^{-1}\lambda_t^\top\right). \end{cases} \tag{139}$$

SMC METHOD FOR ARMA(p, q) WITH CORRELATED INNOVATIONS

We now propose SMC sampling methods for ARMA processes with correlated innovations, based on the transition densities derived in Equation 138 and Equation 139. We first consider the known ARMA parameter case and later relax these assumptions.

In order to use the transition density for propagating state samples, one needs to have knowledge of the matrices A_{t+1},B_{t+1} and $R_{u_{t+1}}$ in Equation 137, so that the covariance matrix Σ_{t+1} can be computed. We present an SMC method that assumes knowledge of those parameters in Table 22.

Nonetheless, assuming knowledge of such parameters is impractical in many scenarios. Since analytical marginalization of all the parameters is not possible (same arguments as in Chapter 4), we resort to a parameter sampling scheme.

On the one hand, sampling for the unknowns in matrix $R_{u_{t+1}}$ is a too ambitious task and thus, we hereby assume knowledge of the autocovariance function of the correlated innovation process (alternatives to deal with uncertainty on $R_{u_{t+1}}$ are presented later in Chapter 9).

On the other, one can easily compute the matrices A_{t+1} and B_{t+1} once the ARMA parameters a and b are known. Thus, a sampling approach for the unknown ARMA model parameters can be devised, as explained in detail in the following section.

PF FOR ARMA WITH CORRELATED u_t

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Perform resampling of the state, if necessary.

4. Propagate the particles by sampling from the transition density, given the (resampled) stream of particles:

   - If σ_u^2 is known,

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}, \sigma_u^2) = \mathcal{N}\left(x_{t+1}|\mu_{t+1}, \sigma_{t+1}^2\right),$$

$$\text{with} \begin{cases} \mu_{t+1} = \lambda_t \Sigma_t^{-1} \overline{x}_{1:t}^{(m)}, \\ \sigma_{t+1}^2 = \sigma_u^2 \left(h_{t+1} - \lambda_t \Sigma_t^{-1} \lambda_t^\top\right). \end{cases}$$

   - If σ_u^2 is unknown,

$$x_{t+1}^{(m)} \sim f(x_{t+1}|x_{1:t}) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1}|\mu_{t+1}, \phi_{t+1}^2\right),$$

$$\text{with} \begin{cases} \nu_{t+1} = \nu_0 + t, \\ \mu_{t+1} = \lambda_t \Sigma_t^{-1} x_{1:t}, \\ \phi_{t+1}^2 = \frac{\nu_0 \sigma_0^2 + x_{1:t}^\top \Sigma_t^{-1} x_{1:t}}{\nu_0 + t} \left(h_{t+1} - \lambda_t \Sigma_t^{-1} \lambda_t^\top\right). \end{cases}$$

   with sufficient statistics computed as in Equation 137.

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 22: PF for latent ARMA with correlated innovations, known a and b.

SMC method for ARMA(p, q) with correlated innovations: unknown a and b parameters

We extend the SMC method presented in Table 22 to the case where the parameters of the ARMA model $\theta = (a \; b)^\top$ are unknown. We propose a sampling approach for the unknown ARMA model parameters. That is, we consider a joint state and parameter vector $\rho_t = (x_t \; \theta_t)^\top$. Here, too, the subscript t in θ_t does not imply that the parameter evolves over time.

The state samples are propagated by drawing from the state transition density. For parameter propagation, two alternatives are explored, one where the principles of Density Assisted (DA) PFs [39] are followed, and another where Importance Sampling (IS) is used.

In the former, one approximates the posterior of the unknown parameter θ, given the current time-series $x_{1:t}$ with a Gaussian distribution, i.e., $f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)}) = \mathcal{N}(\theta_{t+1}|\mu_{\theta_t}, \Sigma_{\theta_t})$, where

$$\begin{aligned} \mu_{\theta_t} &= \sum_{i=1}^{M} w_t^{(m)} \theta_t^{(m)}, \\ \Sigma_{\theta_t} &= \sum_{i=1}^{M} w_t^{(m)} (\theta_t^{(m)} - \mu_{\theta_t})(\theta_t^{(m)} - \mu_{\theta_t})^\top. \end{aligned} \tag{140}$$

One can use such approximation to the posterior of the parameters to propagate parameter samples from one time instant to the next. In this case, the resulting weight computation results in

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}) \cdot \frac{f(x_{t+1}^{(m)}|x_{1:t}^{(m)})}{\pi(x_{t+1})} \cdot \frac{f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)})}{\pi(\theta_{t+1})} = f(y_{t+1}|x_{t+1}^{(m)}) \; . \tag{141}$$

Details of an SMC sampling method following this approach are presented in Table 23 and Table 24, for the known and unknown σ_u^2 cases, respectively.

DA-PF: ARMA WITH CORRELATED u_t, UNKNOWN a AND b

1. At time instant t, consider the joint random measure

$$f_t^M(\rho_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\rho_t - \rho_t^{(m)}\right),$$

and compute the parameter sufficient statistics

$$\begin{cases} \mu_{\theta_t} = \sum_{i=1}^{M} \theta_t^{(m)} w_t^{(m)}, \\ \Sigma_{\theta_t} = \sum_{i=1}^{M} (\theta_t - \mu_{\theta_t})(\theta_t - \mu_{\theta_t})^\top w_t^{(m)} . \end{cases}$$

2. Upon reception of a new observation at time instant $t+1$, perform resampling of the state (if necessary).

3. Propagate parameter samples from the approximated Gaussian posterior density

$$\theta_{t+1}^{(m)} \sim f(\theta_t | x_{1:t}) \approx \mathcal{N}(\theta_{t+1} | \mu_{\theta_t}, \Sigma_{\theta_t}) ,$$

and compute a per $\theta_{t+1}^{(m)}$ sample covariance matrix

$$\Sigma_{t+1}^{(m)} = A_{t+1}^{-1^{(m)}} B_{t+1}^{(m)} R_{u_t} B_{t+1}^{\top^{(m)}} A_{t_1}^{-1^{\top(m)}} = \begin{pmatrix} h_{t+1}^{(m)} & \lambda_t^{(m)} \\ \lambda_t^{\top(m)} & \Sigma_t^{(m)} \end{pmatrix}.$$

4. Propagate state particles by sampling from

$$x_{t+1}^{(m)} \sim f(x_{t+1} | \overline{x}_{1:t}^{(m)}, \sigma_u^2) = \mathcal{N}\left(x_{t+1} | \mu_{t+1}, \sigma_{t+1}^2\right) ,$$

$$\text{with} \begin{cases} \mu_{t+1} = \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)} , \\ \sigma_{t+1}^2 = \sigma_u^2 \left(h_{t+1}^{(m)} - \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \lambda_t^{(m)^\top}\right) . \end{cases}$$

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1} | x_{t+1}^{(m)}), \quad \text{and normalize them}$$

to obtain a new random measure $f_{t+1}^M(\rho_{t+1})$.

Table 23: Density assisted PF for latent ARMA process with correlated innovations, unknown a and b and known σ_u^2.

DA-PF: ARMA WITH CORRELATED u_t, UNKNOWN a, b AND σ_u^2

1. At time instant t, consider the joint random measure

$$f_t^M(\rho_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\rho_t - \rho_t^{(m)}\right),$$

and compute the parameter sufficient statistics

$$\begin{cases} \mu_{\theta_t} = \sum_{i=1}^{M} \theta_t^{(m)} w_t^{(m)}, \\ \Sigma_{\theta_t} = \sum_{i=1}^{M} (\theta_t - \mu_{\theta_t})(\theta_t - \mu_{\theta_t})^\top w_t^{(m)} . \end{cases}$$

2. Upon reception of a new observation at time instant $t+1$, perform resampling of the state (if necessary).

3. Propagate parameter samples from the approximated Gaussian posterior density

$$\theta_{t+1}^{(m)} \sim f(\theta_t | x_{1:t}) \approx \mathcal{N}(\theta_{t+1} | \mu_{\theta_t}, \Sigma_{\theta_t}) ,$$

and compute a per $\theta_{t+1}^{(m)}$ sample covariance matrix

$$\Sigma_{t+1}^{(m)} = A_{t+1}^{-1^{(m)}} B_{t+1}^{(m)} R_{u_t} B_{t+1}^{\top^{(m)}} A_{t_1}^{-1^{\top^{(m)}}} = \begin{pmatrix} h_{t+1}^{(m)} & \lambda_t^{(m)} \\ \lambda_t^{\top^{(m)}} & \Sigma_t^{(m)} \end{pmatrix}.$$

4. Propagate state particles by sampling from

$$x_{t+1}^{(m)} \sim f(x_{t+1} | \overline{x}_{1:t}^{(m)}) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1} | \mu_{t+1}, \phi_{t+1}^2\right) ,$$

$$\text{with} \begin{cases} \nu_{t+1} = \nu_0 + t , \\ \mu_{t+1} = \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)} , \\ \phi_{t+1}^2 = \frac{\nu_0 \sigma_0^2 + \overline{x}_{1:t}^{(m)^\top} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)}}{\nu_0 + t} \left(h_{t+1}^{(m)} - \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \lambda_t^{(m)^\top}\right) . \end{cases}$$

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1} | x_{t+1}^{(m)}), \quad \text{and normalize them}$$

to obtain a new random measure $f_{t+1}^M(\rho_{t+1})$.

Table 24: Density assisted PF for latent ARMA process with correlated innovations, unknown a, b and σ_u^2.

Instead of approximating the parameter posterior, an alternative is to use a Gaussian proposal density for sampling the unknown parameters. Then, one applies IS to properly weight the joint state and parameter samples.

That is, at every time instant, one propagates parameter particles by sampling from the proposal $\pi(\theta_{t+1}) = \mathcal{N}(\theta_{t+1}|\mu_{\theta_t}, \Sigma_{\theta_t})$ (with sufficient statistics as in Equation 140). The corresponding weight computation results in

$$\begin{aligned}\widetilde{w}_{t+1}^{(m)} &\propto f(y_{t+1}|x_{t+1}^{(m)}) \cdot \frac{f(x_{t+1}^{(m)}|x_{1:t}^{(m)})}{\pi(x_{t+1})} \cdot \frac{f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)})}{\pi(\theta_{t+1})} \\ &= f(y_{t+1}|x_{t+1}^{(m)}) \cdot \frac{f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)})}{\mathcal{N}(\theta_{t+1}|\mu_{\theta_t}, \Sigma_{\theta_t})} \,.\end{aligned} \tag{142}$$

Since the posterior of the parameters is analytically intractable, we use

$$f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)}) = \frac{f(x_{1:t}^{(m)}|\theta_{t+1}^{(m)}) f(\theta_{t+1}^{(m)})}{f(x_{1:t}^{(m)})} = \frac{f(x_{1:t}^{(m)}|\theta_{t+1}^{(m)})}{f(x_{1:t}^{(m)})} \,, \tag{143}$$

which results in

$$\begin{cases} f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)}) = \dfrac{\mathcal{N}\left(x_{1:t}^{(m)} \middle| 0, \sigma_u^2 \Sigma_t^{(m)}\right)}{\mathcal{N}\left(x_{1:t}^{(m)} \middle| 0, \sigma_u^2 \Sigma_t^{\left(\mu_{\theta_t}\right)}\right)}, & \text{if } \sigma_u^2 \text{ is known,} \\ f(\theta_{t+1}^{(m)}|x_{1:t}^{(m)}) = \dfrac{\mathcal{T}_{\nu_0}\left(x_{1:t}^{(m)} \middle| 0, \sigma_0^2 \Sigma_t^{(m)}\right)}{\mathcal{T}_{\nu_0}\left(x_{1:t}^{(m)} \middle| \nu_0, 0, \sigma_0^2 \Sigma_t^{\left(\mu_{\theta_t}\right)}\right)}, & \text{if } \sigma_u^2 \text{ is unknown.} \end{cases} \tag{144}$$

Note that, with $\Sigma_t^{\left(\mu_{\theta_t}\right)}$ we refer to the covariance matrix computed using the parameter estimates μ_{θ_t} as in Equation 140; while with $\Sigma_t^{(m)}$ we refer to the covariance matrix evaluated per drawn parameter sample $\theta_{t+1}^{(m)}$.

Full details of the SMC sampling method that follows this approach are presented in Table 25 and Table 26, for the known and unknown σ_u^2 cases, respectively.

The proposed SMC methods rely on different a and b parameter sampling approaches. They are both presented in terms of *one* parameter sample per state sample. However, the proposed solutions can be readily extended to $J \geqslant 1$ parameter samples per state sample, and resort to a numerical Rao-Blackwellization, in a similar fashion to the solutions proposed in Chapter 4.

IS-PF: ARMA WITH CORRELATED u_t, UNKNOWN a AND b

1. At time instant t, consider the joint random measure

$$f_t^M(\rho_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\rho_t - \rho_t^{(m)}\right),$$

and compute the parameter sufficient statistics

$$\begin{cases} \mu_{\theta_t} = \sum_{i=1}^{M} \theta_t^{(m)} w_t^{(m)}, \\ \Sigma_{\theta_t} = \sum_{i=1}^{M} (\theta_t - \mu_{\theta_t})(\theta_t - \mu_{\theta_t})^\top w_t^{(m)} . \end{cases}$$

2. Upon reception of a new observation at time instant $t+1$, perform resampling of the state (if necessary).

3. Propagate parameter samples from the Gaussian proposal

$$\theta_{t+1}^{(m)} \sim \pi(\theta_{t+1}) = \mathcal{N}\left(\theta_{t+1} | \mu_{\theta_t}, \Sigma_{\theta_t}\right),$$

and compute a per $\theta_{t+1}^{(m)}$ sample covariance matrix

$$\Sigma_{t+1}^{(m)} = A_{t+1}^{-1^{(m)}} B_{t+1}^{(m)} R_{u_t} B_{t+1}^{\top(m)} A_{t_1}^{-1^{\top(m)}} = \begin{pmatrix} h_{t+1}^{(m)} & \lambda_t^{(m)} \\ \lambda_t^{\top(m)} & \Sigma_t^{(m)} \end{pmatrix}.$$

4. Propagate state particles by sampling from

$$x_{t+1}^{(m)} \sim f(x_{t+1} | \overline{x}_{1:t}^{(m)}, \sigma_u^2) = \mathcal{N}\left(x_{t+1} | \mu_{t+1}, \sigma_{t+1}^2\right),$$

$$\text{with} \begin{cases} \mu_{t+1} = \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)}, \\ \sigma_{t+1}^2 = \sigma_u^2 \left(h_{t+1}^{(m)} - \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \lambda_t^{(m)^\top}\right). \end{cases}$$

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1} | x_{t+1}^{(m)}) \frac{\mathcal{N}\left(x_{1:t}^{(m)} \middle| 0, \sigma_u^2 \Sigma_t^{(m)}\right)}{\mathcal{N}\left(\theta_{t+1}^{(m)} | \mu_{\theta_t}, \Sigma_{\theta_t}\right) \mathcal{N}\left(x_{1:t}^{(m)} \middle| 0, \sigma_u^2 \Sigma_t^{(\mu_{\theta_t})}\right)},$$

and normalize them to obtain a new random measure $f_{t+1}^M(\rho_{t+1})$.

Table 25: Importance sampling PF for latent ARMA process with correlated innovations, unknown a and b and known σ_u^2.

IS-PF: ARMA WITH CORRELATED u_t, UNKNOWN a, b AND σ_u^2

1. At time instant t, consider the joint random measure

$$f_t^M(\rho_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\rho_t - \rho_t^{(m)}\right),$$

and compute the parameter sufficient statistics

$$\begin{cases} \mu_{\theta_t} = \sum_{i=1}^{M} \theta_t^{(m)} w_t^{(m)}, \\ \Sigma_{\theta_t} = \sum_{i=1}^{M} (\theta_t - \mu_{\theta_t})(\theta_t - \mu_{\theta_t})^\top w_t^{(m)} . \end{cases}$$

2. Upon reception of a new observation at time instant $t+1$, perform resampling of the state (if necessary).

3. Propagate parameter samples from the Gaussian proposal

$$\theta_{t+1}^{(m)} \sim \pi(\theta_{t+1}) = \mathcal{N}(\theta_{t+1} | \mu_{\theta_t}, \Sigma_{\theta_t}) ,$$

and compute a per $\theta_{t+1}^{(m)}$ sample covariance matrix

$$\Sigma_{t+1}^{(m)} = A_{t+1}^{-1^{(m)}} B_{t+1}^{(m)} R_{u_t} B_{t+1}^{\top^{(m)}} A_{t_1}^{-1^{\top^{(m)}}} = \begin{pmatrix} h_{t+1}^{(m)} & \lambda_t^{(m)} \\ \lambda_t^{\top^{(m)}} & \Sigma_t^{(m)} \end{pmatrix}.$$

4. Propagate state particles by sampling from

$$x_{t+1}^{(m)} f(x_{t+1} | \overline{x}_{1:t}^{(m)}) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1} | \mu_{t+1}, \phi_{t+1}^2\right) ,$$

$$\text{with} \begin{cases} \nu_{t+1} = \nu_0 + t , \\ \mu_{t+1} = \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)} , \\ \phi_{t+1}^2 = \frac{\nu_0 \sigma_0^2 + \overline{x}_{1:t}^{(m)^\top} \Sigma_t^{(m)^{-1}} \overline{x}_{1:t}^{(m)}}{\nu_0 + t} \left(h_{t+1}^{(m)} - \lambda_t^{(m)} \Sigma_t^{(m)^{-1}} \lambda_t^{(m)^\top}\right) . \end{cases}$$

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1} | x_{t+1}^{(m)}) \frac{\mathcal{T}_{\nu_0}\left(x_{1:t}^{(m)} \middle| 0, \sigma_0^2 \Sigma_t^{(m)}\right)}{\mathcal{N}\left(\theta_{t+1}^{(m)} | \mu_{\theta_t}, \Sigma_{\theta_t}\right) \mathcal{T}_{\nu_0}\left(x_{1:t}^{(m)} \middle| \nu_0, 0, \sigma_0^2 \Sigma_t^{(\mu_{\theta_t})}\right)} ,$$

and normalize them to obtain a new random measure $f_{t+1}^M(\rho_{t+1})$.

Table 26: Importance sampling PF for latent ARMA process with correlated innovations, unknown a, b and σ_u^2.

The only modification consists of first approximating the joint posterior distribution of the state and parameters $\rho_t = (x_t\ \theta_t)^\top$ with a multivariate Gaussian

$$f(\rho_t) \approx \mathcal{N}(\rho_t|\eta_t, Q_t), \tag{145}$$

with sufficient statistics computed as weighted averages of the available particles

$$\begin{aligned} \eta_t &= \sum_{i=1}^{M} w_t^{(m)} \rho_t^{(m)}, \\ Q_t &= \sum_{i=1}^{M} w_t^{(m)} (\rho_t^{(m)} - \eta_t)(\rho_t^{(m)} - \eta_t)^\top. \end{aligned} \tag{146}$$

Then, one samples J samples per state sample $x_t^{(m)}$ from the conditional Gaussian distribution of the parameters

$$\begin{aligned} f(\theta_t|x_t) &= \mathcal{N}(\theta_t|\eta_{\theta_t|x_t}, Q_{\theta_t|x_t}), \\ &\text{where} \begin{cases} \eta_{\theta_t|x_t} = \eta_{\theta_t} + Q_{\theta_t,x_t} Q_{x_t,x_t}^{-1} (x_t - \eta_{x_t}), \\ Q_{\theta_t|x_t} = Q_{\theta_t,\theta_t} - Q_{\theta_t,x_t} Q_{x_t,x_t}^{-1} Q_{x_t,\theta_t}, \end{cases} \\ &\text{with } \eta_t = \begin{pmatrix} \eta_{x_t} \\ \eta_{\theta_t} \end{pmatrix} \text{ and } Q_t = \begin{pmatrix} Q_{x_t,x_t} & Q_{\theta_t,x_t} \\ Q_{x_t,\theta_t} & Q_{\theta_t,\theta_t} \end{pmatrix}. \end{aligned} \tag{147}$$

This alternative method implements numerical Rao-Blackwellization of the unknown parameters, where downsampling from MJ to M particles is necessary at every time instant.

EVALUATION

We now evaluate the proposed SMC methods for latent ARMA models with correlated innovations. To do so, we focus on a family of processes that show various memory properties: the ARMA model with fGn (which is equivalent to FARIMA(p, d, q) models). This is a natural extension of the classical ARMA model, where instead of iid innovations, the ARMA(p, q) filters a fGp or a FARIMA(0, d, 0) (see Section 5.1.2 for more details).

In this chapter we, too, consider the SV model, where the latent time-series is now an ARMA model with fGn. Mathematically, we have

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ y_t = e^{\frac{x_t}{2}} v_t, \end{cases} \tag{148}$$

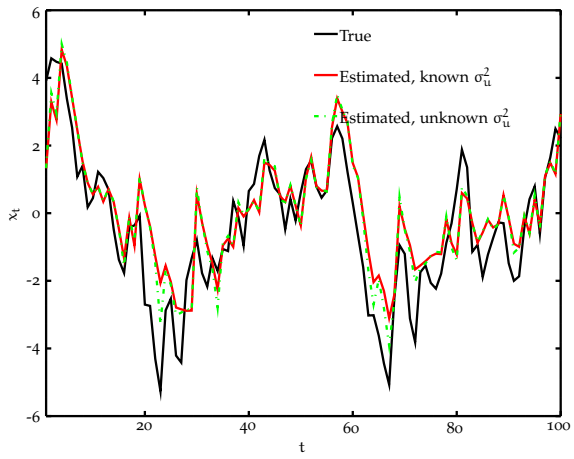
where v_t is a standard Gaussian variable and the state innovation u_t is a zero-mean Gaussian process with autocovariance function

$$\gamma(\tau) = \frac{\sigma_u^2}{2}\left[|\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H}\right], \tag{149}$$

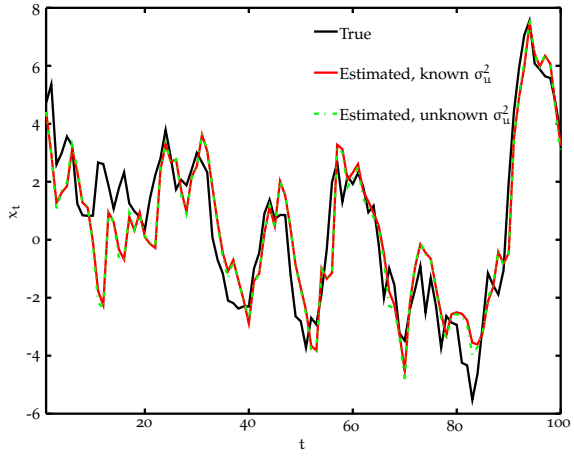
parameterized by a Hurst parameter H and variance σ_u^2. As explained in Section 5.1.1, when $H = 0.5$, the process is uncorrelated, while the memory of the innovations increases as $H \to 1$.

We evaluate the proposed method in this nonlinear model first under the known ARMA parameter case, and then, drop the assumption.

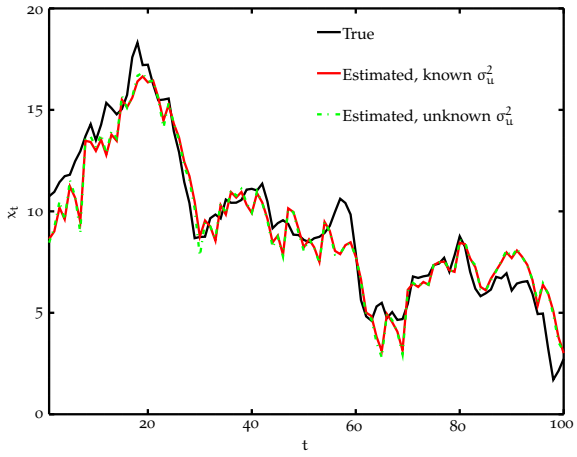
We show in Figure 26, Figure 27 and Figure 28 how the proposed method is able to accurately track different FARIMA(p, d, q) models, for both the known and unknown σ_u^2 cases. The SMC methods were run with $M = 1000$ particles and different memory properties of the fGp were evaluated for each of the studied ARMA parameterizations: AR(1) in Figure 26, MA(1) in Figure 27 and ARMA(1,1) in Figure 28.



(a) AR(1), H = 0.5.

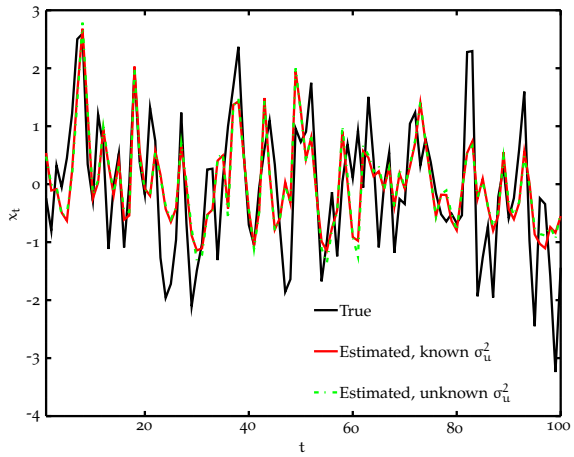


(b) AR(1), H = 0.7.

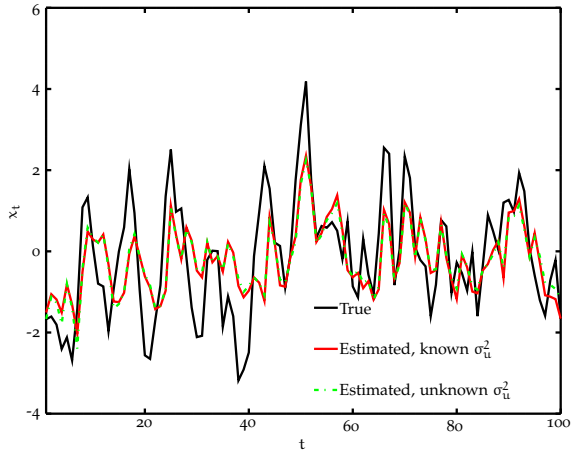


(c) AR(1), H = 0.9.

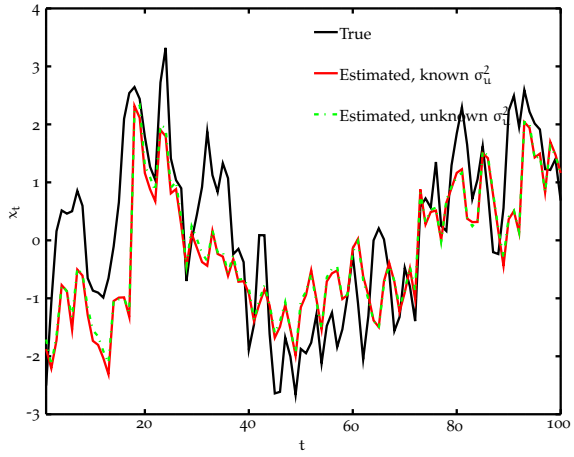
Figure 26: True (black) and estimated state for the proposed SMC methods with known AR(1) parameter $a_1 = 0.85$.



(a) MA(1), H = 0.5.

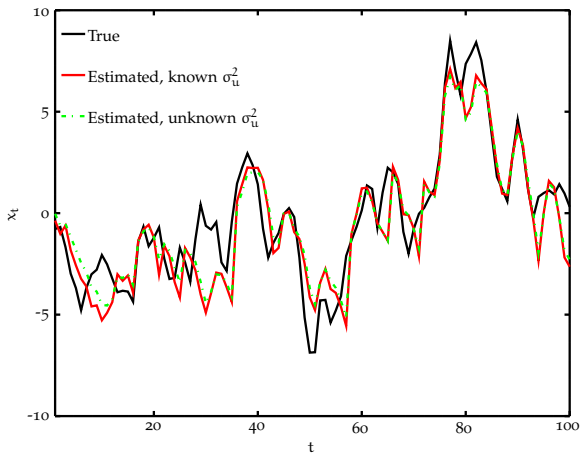


(b) MA(1), H = 0.7.

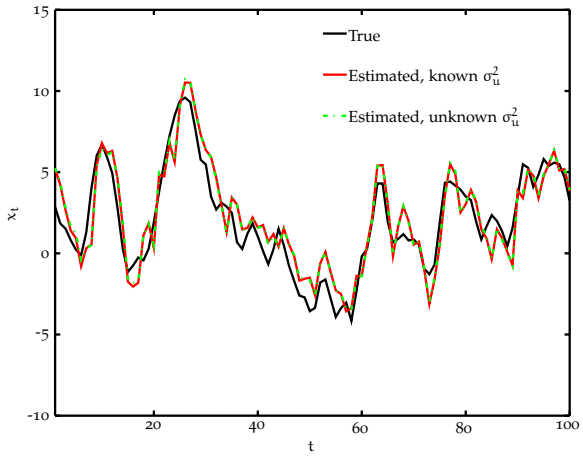


(c) MA(1), H = 0.9.

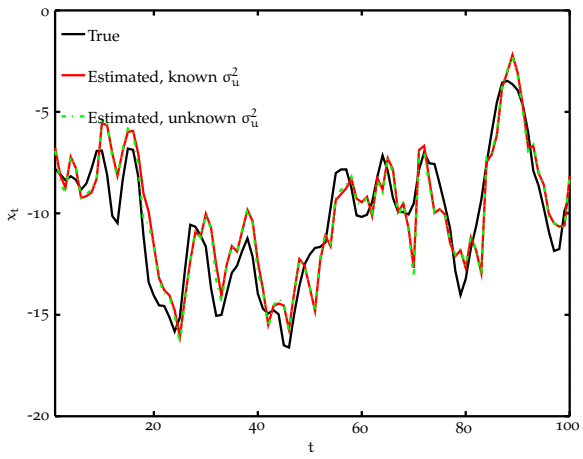
Figure 27: True (black) and estimated state for the proposed SMC methods with known MA(1) parameter $b_1 = 0.8$.



(a) ARMA(1,1), H = 0.5.



(b) ARMA(1,1), H = 0.7.



(c) ARMA(1,1), H = 0.9.

Figure 28: True (black) and estimated state for the proposed SMC methods with known ARMA(1,1) parameters $a_1 = 0.85$ and $b_1 = 0.8$.

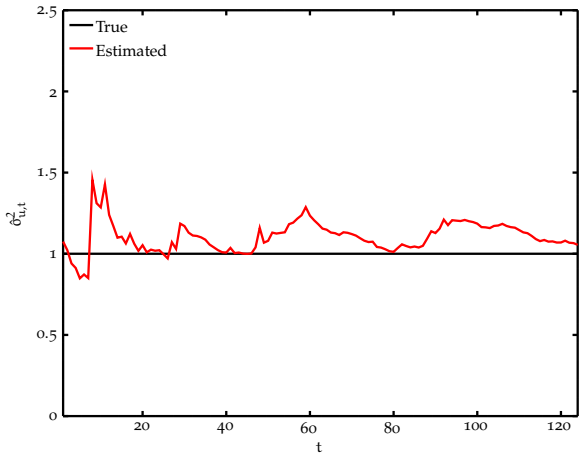
PF Type	State estimation error (MSE)	
	Known σ_u^2	unknown σ_u^2
AR(1), H = 0.5	1.1081	1.1945
AR(1), H = 0.7	1.3946	1.4397
AR(1), H = 0.9	1.1195	1.1970
MA(1), H = 0.5	1.0223	1.0686
MA(1), H = 0.7	1.0585	1.1136
MA(1), H = 0.9	0.87374	0.94053
ARMA(1,1), H = 0.5	1.5947	1.6197
ARMA(1,1), H = 0.7	1.7852	1.8516
ARMA(1,1), H = 0.9	1.7214	1.7362

Table 27: MSE performance of the proposed SMC methods for ARMA models (known a and b) with fGn, known and unknown σ_u^2.

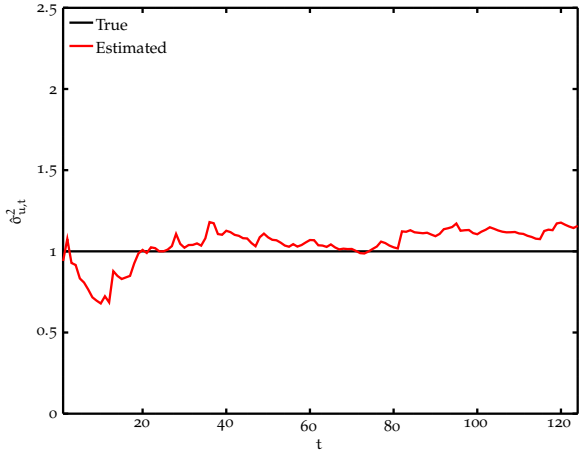
We further study the filtering performance of the methods described in Table 22 and conclude, based on results summarized in Table 27, that the proposed SMC method successfully estimates the latent ARMA with fGn, both for the known and unknown innovation variance cases.

Note that there is a slight accuracy loss of the unknown σ_u^2 cases. The justification for such a good performance relies on the form of the derived marginalized density. As more data are observed, the density in Equation 139 becomes very similar to the one in Equation 138, i.e., a Student-t distribution with high degrees of freedom is very similar to a Gaussian distribution. Thus, the proposal densities in both SMC methods become almost identical with time.

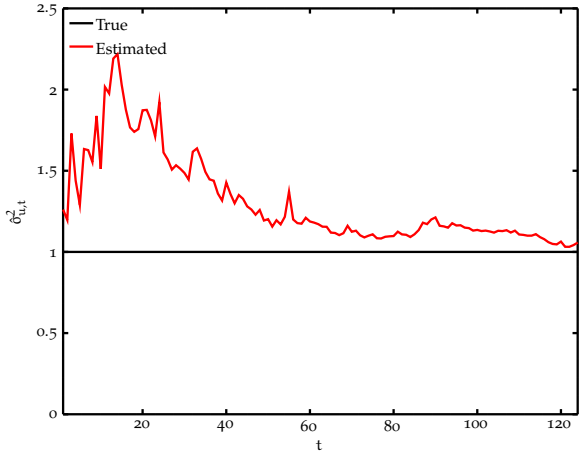
Furthermore, we plot the evolution of the scale factor $\frac{\nu_0\sigma_0^2+x_{1:t}^\top\Sigma_t^{-1}x_{1:t}}{\nu_0+t}$ in Equation 139 over time and observe that the estimate approaches the true value σ_u^2 (see Figure 29, Figure 30 and Figure 31). The estimation accuracy improves with time for all the evaluated ARMA parameterizations and memory properties of the fGp: AR(1) in Figure 29, MA(1) in Figure 30 and ARMA(1,1) in Figure 31.



(a) AR(1), $H = 0.5$.

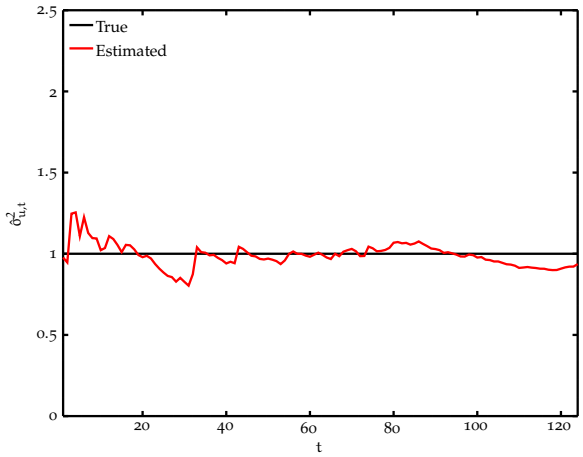


(b) AR(1), $H = 0.7$.

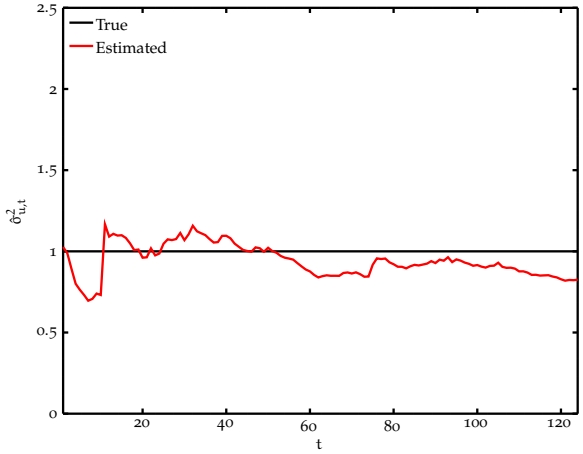


(c) AR(1), $H = 0.9$.

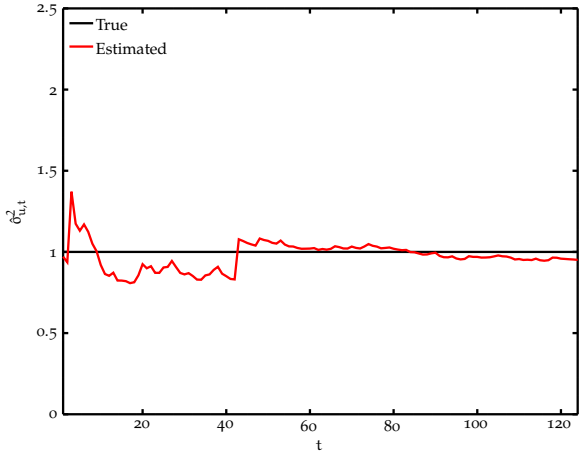
Figure 29: True (black) and estimated (red) scale factor for the proposed SMC method with known AR(1) parameter $a_1 = 0.85$.



(a) MA(1), H = 0.5.

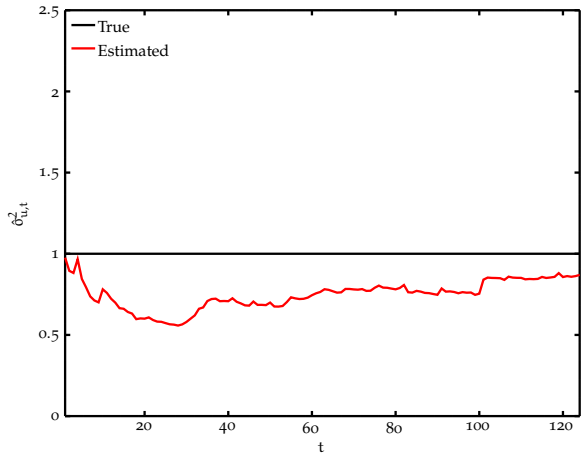


(b) MA(1), H = 0.7.

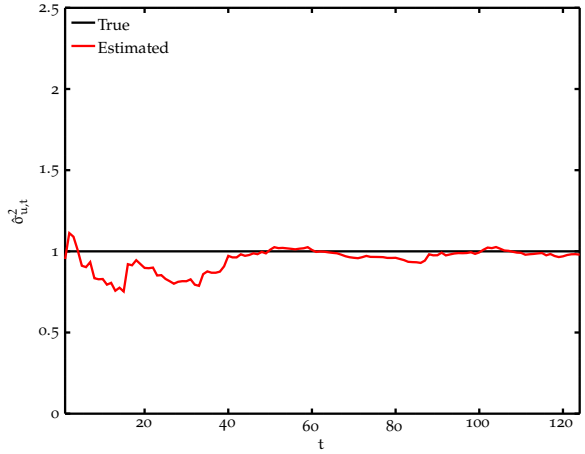


(c) MA(1), H = 0.9.

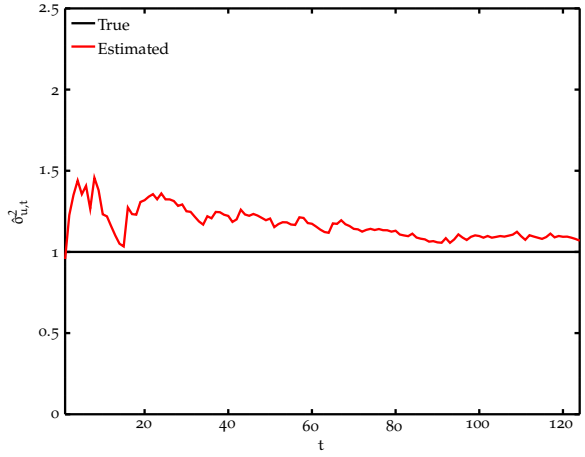
Figure 30: True (black) and estimated (red) scale factor for the proposed SMC method with known MA(1) parameter $b_1 = 0.8$.



(a) ARMA(1,1), H = 0.5.



(b) ARMA(1,1), H = 0.7.



(c) ARMA(1,1), H = 0.9.

Figure 31: True (black) and estimated (red) scale factor for the proposed SMC method with known ARMA(1,1) parameters $a_1 = 0.85$ and $b_1 = 0.8$.

We now turn our attention to the more challenging scenario where the ARMA parameters are unknown. We evaluate both proposed approaches, i.e., the DA- and IS-based SMC methods in Section 6.2.1, and provide averaged state MSE results in Table 28 for different latent processes.

PF Type	State estimation error (MSE)					
	Known a,b	Known a,b	Unknown a,b, DA	Unknown a,b, , IS	Unknown a,b, DA	Unknown a,b IS
	Known σ_u^2	Unknown σ_u^2	Known σ_u^2	Known σ_u^2	Unknown σ_u^2	Unknown σ_u^2
AR(1), H = 0.5	1.0991	1.127	1.6689	1.7337	1.4549	1.5903
AR(1), H = 0.7	1.4077	1.4375	2.5759	5.9889	1.9272	3.191
AR(1), H = 0.9	1.1336	1.1774	2.4334	6.5974	1.7795	6.4853
MA(1), H = 0.5	1.0348	1.0758	1.1033	1.185	1.1384	1.3831
MA(1), H = 0.7	1.0878	1.1138	1.1857	1.2688	1.1884	1.3748
MA(1), H = 0.9	0.88045	0.90841	0.96348	1.1124	0.97747	1.2517
ARMA(1,1), H = 0.5	1.638	1.6512	2.8563	3.6266	2.3157	2.3619
ARMA(1,1), H = 0.7	1.7452	1.7926	3.0939	4.1174	2.7807	2.4627
ARMA(1,1), H = 0.9	1.7374	1.7533	4.3466	20.617	2.5818	2.569

Table 28: MSE performance of the proposed SMC methods for ARMA models (unknown a and b) with fGn, known and unknown σ_u^2.

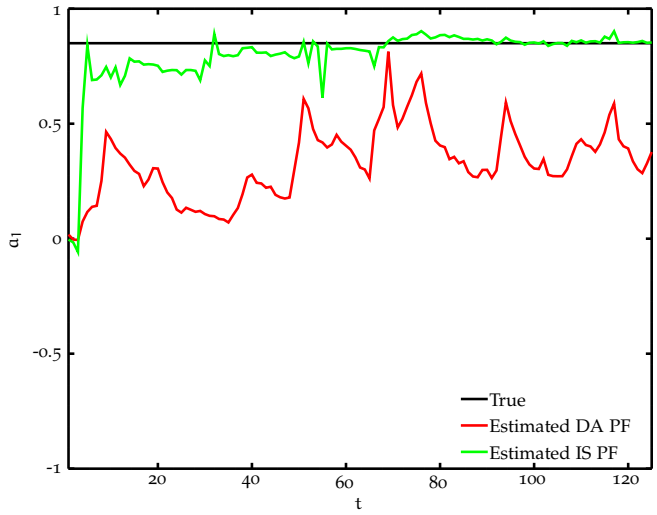
By analyzing the state filtering performance, we conclude that both proposed approaches are suitable solutions for the unknown ARMA case.

It is again remarkable that the impact of not knowing parameters a and b is more pronounced in filtering, than not knowing σ_u^2.

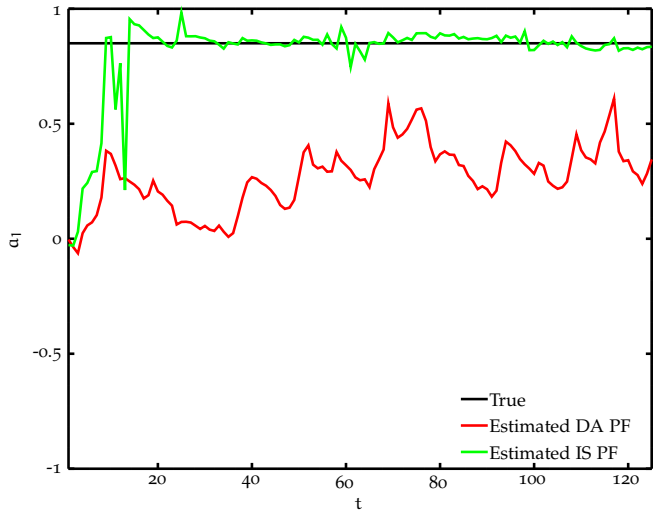
When comparing the DA PFs (i.e., Table 23 and Table 24) with the IS-based PFs (i.e., Table 25 and Table 26), we observe a slightly better filtering performance for the former when compared to the latter.

However, this improved state filtering accuracy comes with a cost, as the estimation of the unknown parameters is worse for the DA-based SMC. We provide in the following pages some illustrative examples of various parameter estimates for the proposed PFs.

In Figure 32, Figure 33 and Figure 34 we show the estimation accuracy for the unknown a_1 parameter in an ARMA(1,1) process with different driving fGns. Similarly, the estimation of b_1 is shown in Figure 35, Figure 36 and Figure 37. Finally, the estimation of the unknown variance σ_u^2 is plotted in Figure 38.

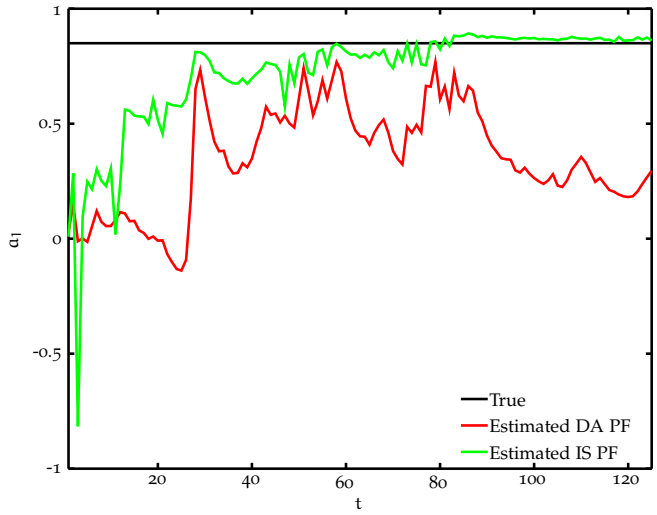


(a) Known σ_u^2.

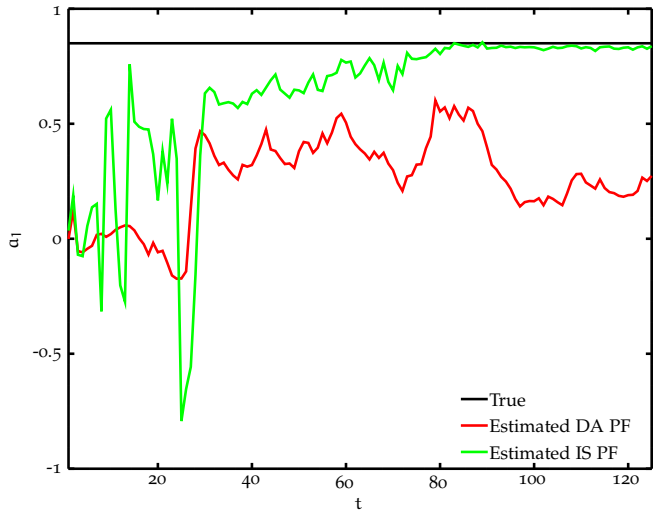


(b) Unknown σ_u^2.

Figure 32: True (black) and estimated (DA PF in red, IS PF in green) a_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.5$).

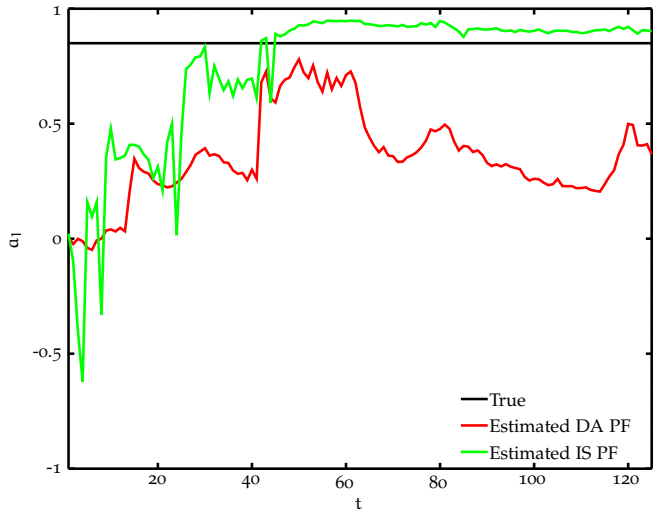


(a) Known σ_u^2.

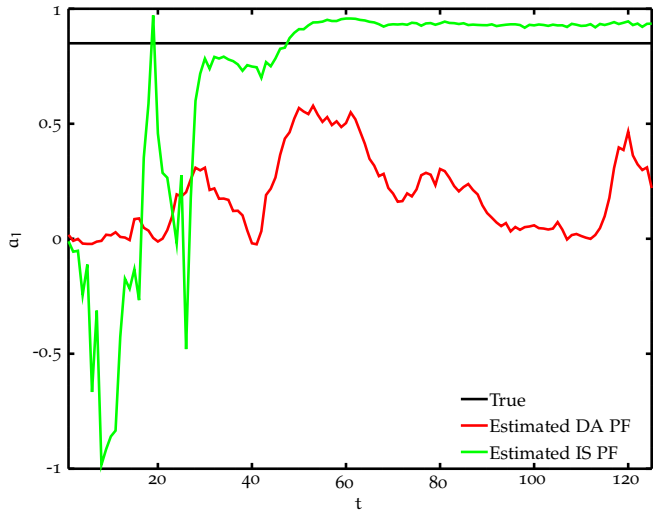


(b) Unknown σ_u^2.

Figure 33: True (black) and estimated (DA PF in red, IS PF in green) a_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.7$).

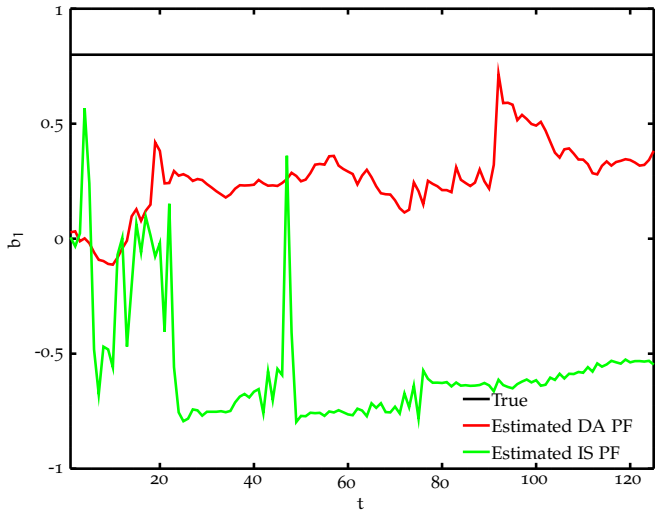


(a) Known σ_u^2.

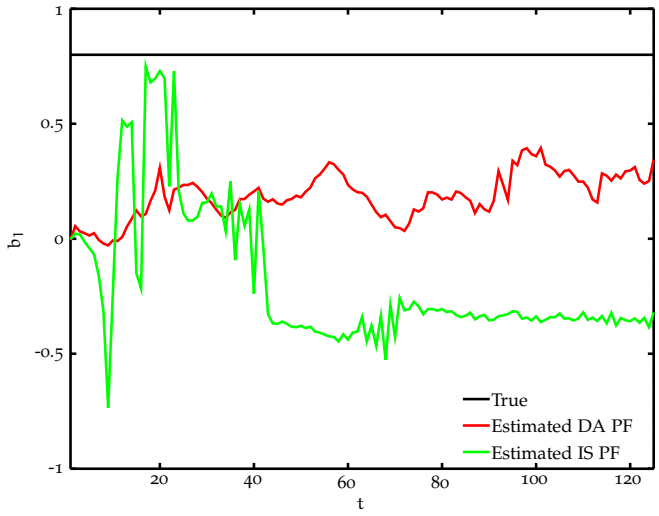


(b) Unknown σ_u^2.

Figure 34: True (black) and estimated (DA PF in red, IS PF in green) a_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.9$).

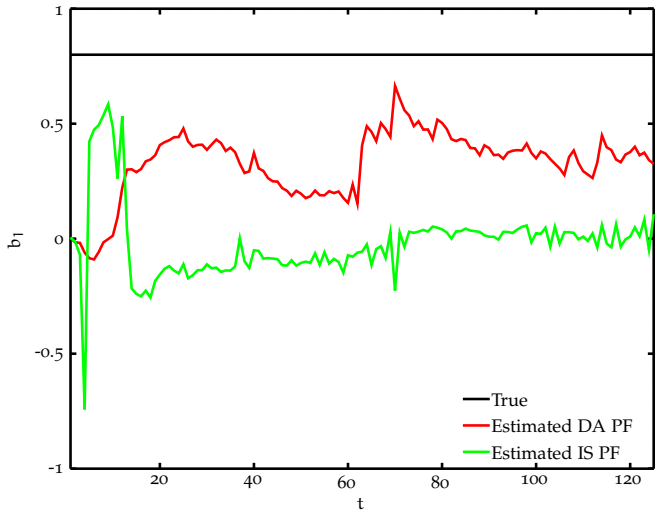


(a) Known σ_u^2.

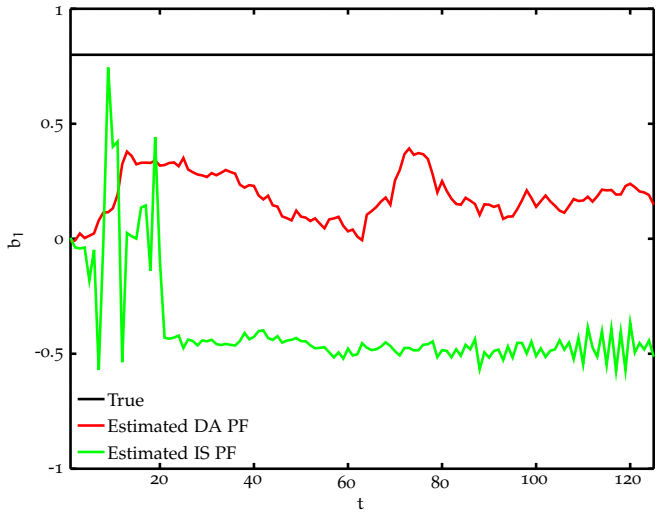


(b) Unknown σ_u^2.

Figure 35: True (black) and estimated (DA PF in red, IS PF in green) b_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.5$).

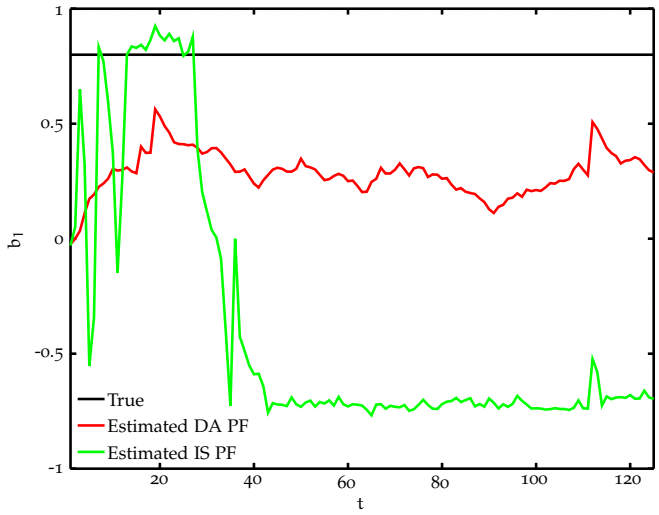


(a) Known σ_u^2.

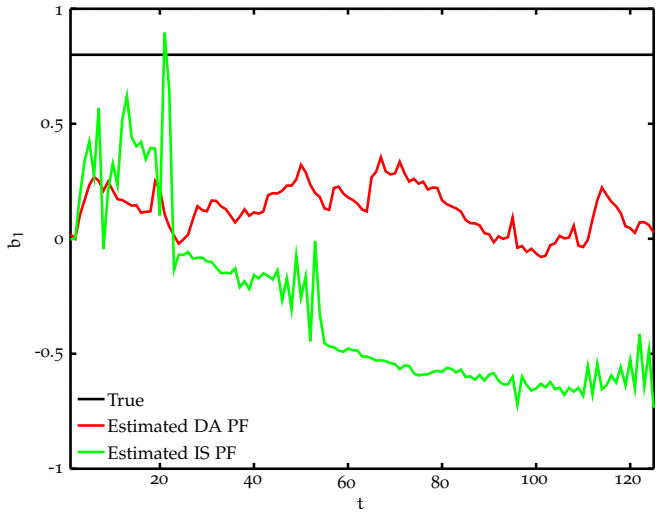


(b) Unknown σ_u^2.

Figure 36: True (black) and estimated (DA PF in red, IS PF in green) b_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.7$).

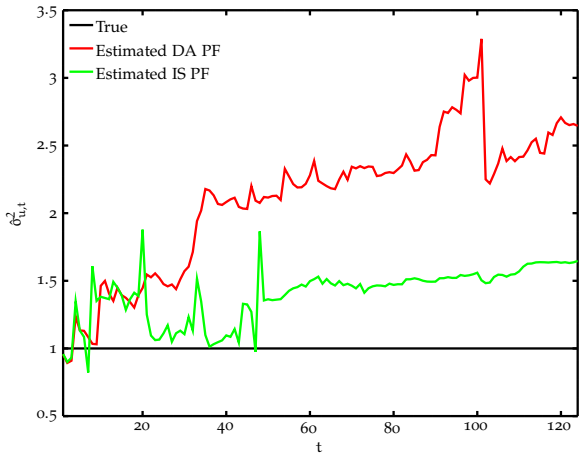


(a) Known σ_u^2.

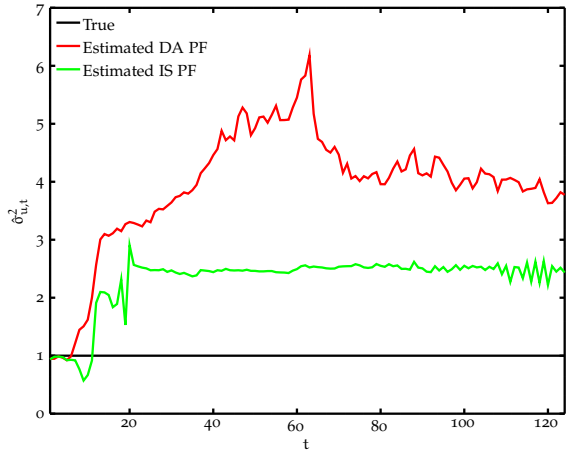


(b) Unknown σ_u^2.

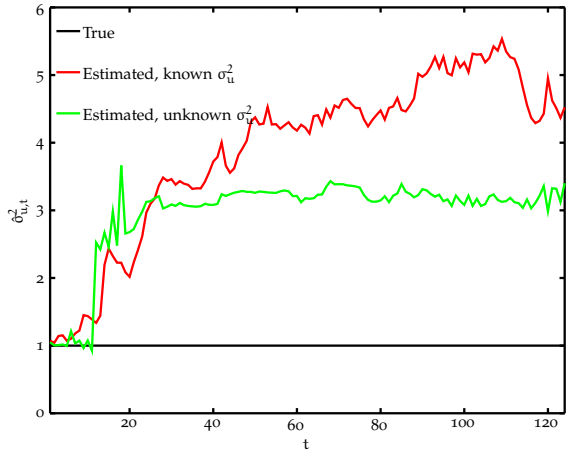
Figure 37: True (black) and estimated (DA PF in red, IS PF in green) b_1 for the proposed SMC methods with unknown ARMA(1,1) parameters (fGn with $H = 0.9$).



(a) ARMA(1,1), H = 0.5.



(b) ARMA(1,1), H = 0.7.



(c) ARMA(1,1), H = 0.9.

Figure 38: True (black) and estimated (red for DA PF and green for IS PF) scale factor for the proposed SMC methods with unknown ARMA(1,1) parameters.

The analysis of the parameter estimation performance allows us to conclude that, for both proposed SMC methods, estimation of the AR parameters is more accurate than the MA parameters. We also note that, due to how these parameters are used (inversion and multiplication of matrices is required for computation of the sufficient statistics), numerical and identifiability issues arise.

Furthermore, the DA PF overestimates the unknown variance of the process (see Figure 38). Although this might seem irrelevant for the filtering problem, variance overestimation is critical when predicting future instances of the time-series, as the density becomes too wide to be informative.

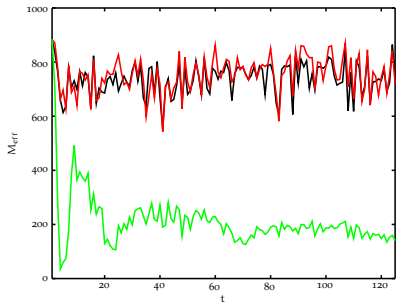
The poor parameter estimation accuracy for the DA PFs can be explained by inspection of the weight computation for each of the proposed alternatives. For the DA approach (i.e., Equation 141) only state samples are involved, while for the IS approach as in Equation 142, both state and parameter samples are taken into account.

That is, when applying IS, one explicitly computes weights based on both the state and parameter samples; while with the DA approach, one hopes that the best state particles are linked with good parameters too (although this is not explicitly accounted for).

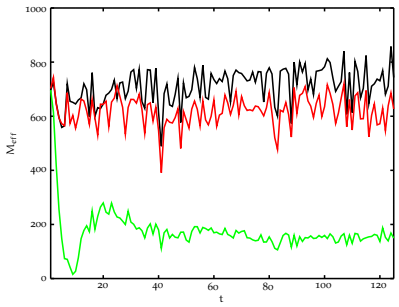
The parameter-explicit weight computation in IS PFs implies that, as a result, the number of particles with non-negligible weights is much reduced at every time instant (as one looks for both good state and parameter samples).

Consequently, the effective particle size of the PFs in Table 25 and Table 26 is quite low and thus, the obtained results much more volatile. Averaged effective particle sizes for the proposed SMC methods are depicted in Figure 39, Figure 40 and Figure 41.

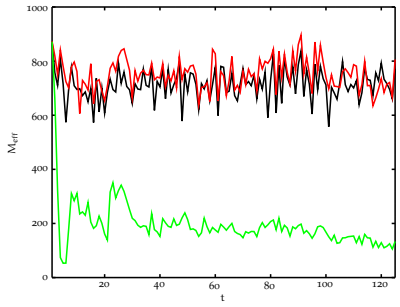
For reference, the effective particle size of the PF with known parameters as in Table 22 is shown in black, with the effective particle size for the DA PF in red. The IS-based PF's effective particle size is always the lowest (shown in green in Figure 39, Figure 40 and Figure 41).



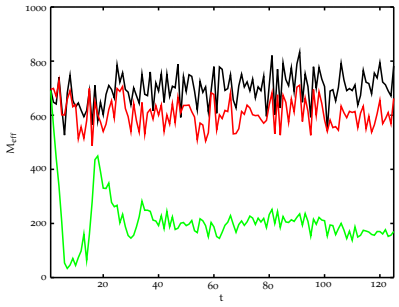
(a) $H = 0.5$, known σ_u^2.



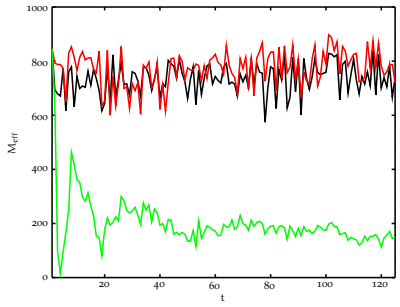
(b) $H = 0.5$, unknown σ_u^2.



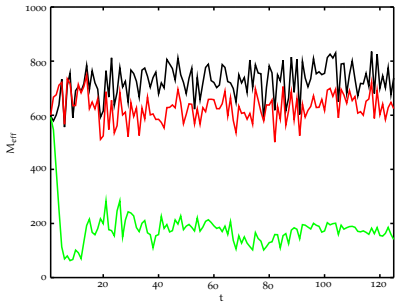
(c) $H = 0.7$, known σ_u^2.



(d) $H = 0.7$, unknown σ_u^2.

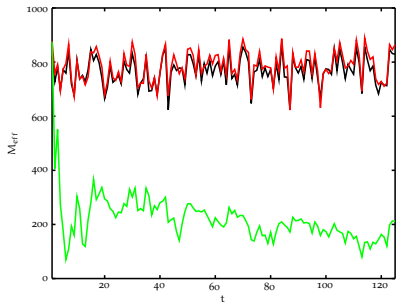


(e) $H = 0.9$, known σ_u^2.

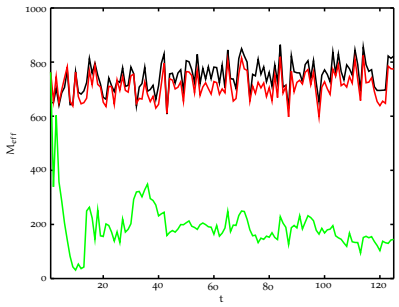


(f) $H = 0.9$, unknown σ_u^2.

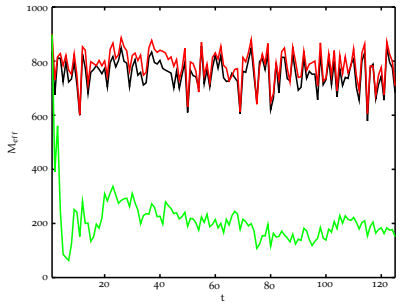
Figure 39: Average effective particle size for proposed SMCs with latent AR(1) with fGn: PF with known a and b (black), DA PF (red), IS PF (green).



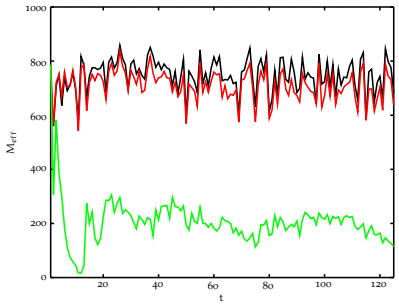
(a) $H = 0.5$, known σ_u^2.



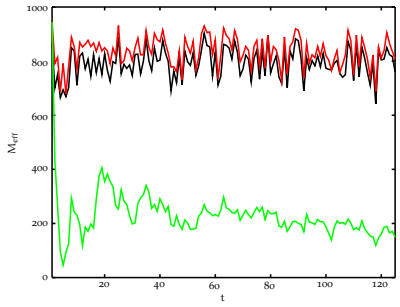
(b) $H = 0.5$, unknown σ_u^2.



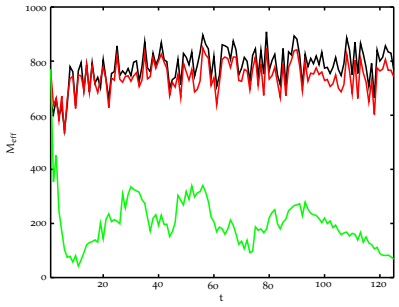
(c) $H = 0.7$, known σ_u^2.



(d) $H = 0.7$, unknown σ_u^2.

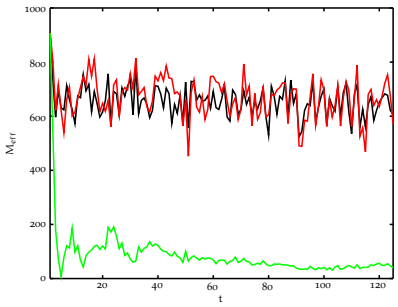


(e) $H = 0.9$, known σ_u^2.

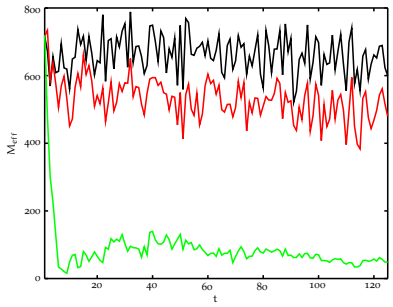


(f) $H = 0.9$, unknown σ_u^2.

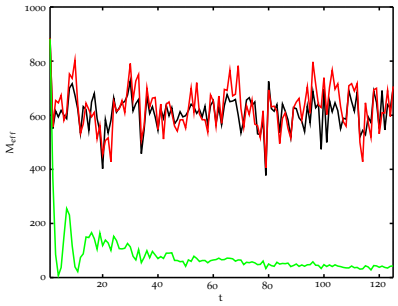
Figure 40: Average effective particle size for proposed SMCs with latent MA(1) with fGn: PF with known a and b (black), DA PF (red), IS PF (green).



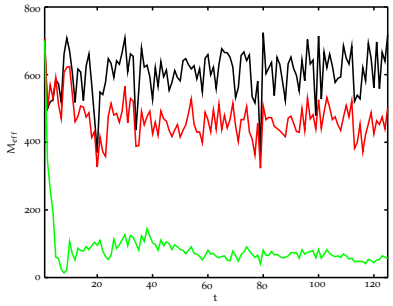
(a) $H = 0.5$, known σ_u^2.



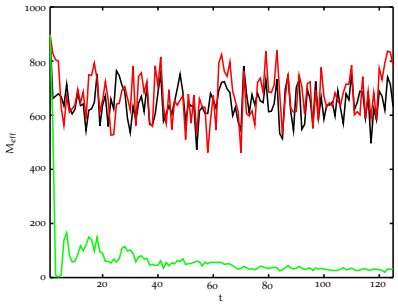
(b) $H = 0.5$, unknown σ_u^2.



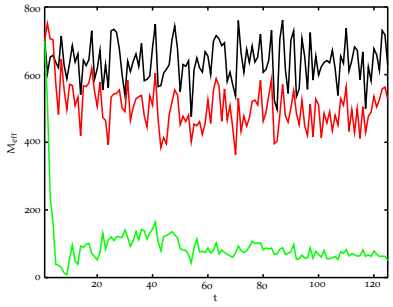
(c) $H = 0.7$, known σ_u^2.



(d) $H = 0.7$, unknown σ_u^2.



(e) $H = 0.9$, known σ_u^2.



(f) $H = 0.9$, unknown σ_u^2.

Figure 41: Average effective particle size for proposed SMCs with latent ARMA(1,1) with fGn: PF with known a and b (black), DA PF (red), IS PF (green).

7

LATENT WIDE-SENSE STATIONARY TIME-SERIES

In this chapter, we study a more general problem where the restrictions on the latent time-series model are relaxed. We aim at a generic SMC method for hidden processes, where the modeling assumptions are minimal.

By inspection of the properties of the models considered in previous chapters, i.e., ARMA(p, q) processes with iid innovations, fGps and FARIMA models, one readily concludes that a common feature is their stationarity. That is, their statistical properties do not change with time-shifts. In other words, one can fully describe them by means of their first and second order statistics, which are not functions of time.

We now focus on these generic latent time-series, where the only assumption is that they are wide-sense stationary. Consequently, these time-series are fully described by a constant mean vector and a covariance matrix that is a function of the time lag τ.

A Bayesian analysis of these WSS time-series is provided in Section 7.1. This allows us to propose a generic SMC method in Section 7.2, which is evaluated in Section 7.3 for latent time-series that are wide-sense stationary.

WIDE-SENSE STATIONARY TIME-SERIES

In the following, we keep our assumptions on the model of the latent process to bare minimums, as we aim at a generic SMC method for hidden time-series. As a matter of fact, we only assume wide-sense stationary of the latent process.

Consequently, we require that the first and second moments of the process exist and that they are not functions of time. That is, the mean of the process is constant and the autocovariance of the process is only a function of the time-difference τ, and not of the actual time instants t.

If the innovation process driving the state is Gaussian and the wide-sense stationarity requirements are met, one can

show that the joint distribution of the time-series $x_{1:t+1}$ is a multivariate Gaussian. That is,

$$\begin{pmatrix} x_{t+1} \\ x_t \\ \vdots \\ x_2 \\ x_1 \end{pmatrix} \sim \mathcal{N}(x_{1:t+1}|\mu_{t+1}, \Sigma_{t+1}). \tag{150}$$

The sufficient statistics of the process are a constant mean $\mu_{t+1} \in \mathbb{R}^{t+1}$ and a covariance matrix $\Sigma_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ given by

$$\Sigma_{t+1} = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(t) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(t-1) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(t-1) & \gamma(t-2) & \cdots & \gamma(1) \\ \gamma(t) & \gamma(t-1) & \cdots & \gamma(0) \end{pmatrix}, \tag{151}$$

where $\gamma(\tau)$ is the autocovariance of the latent process for lag τ.

In general, the mean and the autocovariance function depend on the particularities of the process. That is, based on the specifics of the underlying model, their functional form and values vary. In this chapter, we do not restrict ourselves to any parameterization and thus, we consider sufficient statistics that fulfill the WSS requirements.

The key paradigm shift here is to acknowledge that many different time-series parameterizations can be, in general, described by a joint density as in Equation 150. That is, ARMA(p, q) processes with iid or correlated innovations, fGps and FARIMA models, all result in a joint density of the above form. However, one usually needs to know the specific type of model and its parameters to be able to derive the particular values of the sufficient statistics in Equation 150.

Consequently, either because one does not know the details of the model or the true values of its parameters, a method that can deal with such generic time-series is of interest.

Hereby, we focus on time-series that are described by the joint distribution in Equation 150. The primary goal is, once again, to infer the evolution of the states over time, without knowledge of the specific form of the mean and autocovariance function of the latent process. In other words, all the parameters of the model are considered of secondary importance, as we are exclusively interested on the state's evolution.

Because the interest is on the state only, we proceed with a hierarchical Bayesian approach and marginalize out the

sufficient statistics of the process. That is, we obtain the marginal joint distribution of $x_{1:t+1}$ in Equation 150, so that we can condition on previous samples to derive the marginalized transition density $f(x_{t+1}|x_{1:t})$.

The proposed marginalization implies that one does not require knowledge of the mean and covariance matrix in Equation 150. Therefore, there is no need to know the specific underlying model or its parameterization. Thus, the marginalized densities that are derived in Section 7.1.1 are applicable for any type of model that fulfills the WSS conditions, without requiring knowledge of their parameters, model orders or the specifics of the autocovariance function.

In the following, we describe how one can marginalize the unknown mean and covariance matrix of such WSS processes via the Bayesian methodology.

Marginalized densities for wide-sense stationary time-series

Given a vector $x \in \mathbb{R}^d$ generated according to

$$x \sim f(x|\mu, \Sigma) = \mathcal{N}(x|\mu, \Sigma), \tag{152}$$

with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, we are interested on integrating the parameters out and obtaining the marginal density of the vector. That is, we compute

$$f(x) = \int\int f(x|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma, \tag{153}$$

where $f(\mu, \Sigma)$ is the prior of the unknown sufficient statistics μ and Σ.

In the following, we make use of conjugate priors due to their convenient analytical properties [77, 86]. The derivation is outlined here, while further details are provided in Appendix E.

The conjugate prior of the covariance matrix is the inverse Wishart distribution $\mathrm{IW}_d(\Sigma|\nu_w, \Lambda)$, where $\nu_w > d-1$ represents degrees of freedom, and $\Lambda \in \mathbb{R}^{d\times d}$ is a scale matrix.

The conjugate prior of μ given Σ is a Gaussian with hyperparameters η and κ, $\mathcal{N}(\mu|\eta, \frac{\Sigma}{\kappa})$. From $f(\mu, \Sigma) = f(\mu|\Sigma)f(\Sigma)$, we deduce that the joint conjugate prior is a normal-inverse-Wishart distribution $\mathrm{NIW}(\mu, \Sigma|\eta, \kappa, \nu_w, \Lambda)$.

With this prior, we integrate out the parameters of the density of x and find its marginal

$$\begin{aligned} f(x) &= \int\int f(x|\mu,\Sigma) f(\mu,\Sigma|\eta,\kappa,\Lambda,\nu_w) d\mu d\Sigma \\ &\propto \left| 1 + (x-\eta)^\top \left(\frac{1+\kappa}{\kappa}\Lambda\right)^{-1} (x-\eta) \right|^{-\frac{\nu_w+1}{2}}, \end{aligned} \tag{154}$$

which has the functional form of a multivariate Student-t distribution (see Section E.1).

Consequently, one concludes that the joint marginal density of the vector follows

$$f(x) = \mathcal{T}_{\nu_w - d+1}\left(x \left| \eta, \frac{(1+\kappa)\Lambda}{\kappa(\nu_w - d + 1)}\right.\right), \tag{155}$$

where $(\nu_w - d + 1)$ indicates the degrees of freedom, η is the location parameter and $\frac{(1+\kappa)\Lambda}{\kappa(\nu_w - d+1)}$ is the scale matrix.

The marginal density for the case when the mean of the process is known, but the covariance matrix Σ is unknown, is derived similarly and results in a density as in Equation 155 with $\kappa \to \infty$ (see Appendix E for full details).

The result in Equation 155 is directly applicable to the time-series of interest in this chapter, i.e., for a process described by the joint distribution in Equation 150. Based on the above derivation, the joint density of the process when its sufficient statistics are unknown follow Equation 155.

Specifically, if the mean and the covariance of a time-series as in Equation 150 are unknown, then the joint marginal density of the process is

$$f(x_{1:t+1}) = \mathcal{T}_{\nu_w - t}\left(x_{1:t+1} \left| \eta, \frac{(1+\kappa)\Lambda}{\kappa(\nu_w - t)}\right.\right), \tag{156}$$

where $d = t + 1$ in Equation 155, $\nu_w > t$, and in block form,

$$\eta = \begin{pmatrix} \zeta_{t+1} \\ \zeta_{1:t} \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \lambda_{t+1} & l_{1:t}^\top \\ l_{1:t} & L_{1:t} \end{pmatrix}.$$

Now, we derive the conditional density of the next state x_{t+1} given the past samples $x_{1:t}$ [87]. We conclude that the transition density is a univariate Student-t distribution

$$f\left(x_{t+1}|x_{1:t}\right) = \mathcal{T}_{\nu_w}\left(x_{t+1}|\zeta_{t+1|1:t}, \sigma^2_{t+1|1:t}\right), \tag{157}$$

where

$$\begin{cases} \zeta_{t+1|1:t} = \zeta_{t+1} + l_{1:t}^{\top} L_{1:t}^{-1} \left(x_{1:t} - \zeta_{1:t} \right), \\ \sigma_{t+1|1:t}^2 = h_{t+1|1:t} \left(\lambda_{t+1} - l_{1:t}^{\top} L_{1:t}^{-1} l_{1:t} \right), \\ h_{t+1|1:t} = \frac{\frac{(1+\kappa)}{\kappa} + (x_{1:t} - \zeta_{1:t})^{\top} L_{1:t}^{-1} (x_{1:t} - \zeta_{1:t})}{\nu_w}. \end{cases} \tag{158}$$

The above derivation considers the general case where both the mean and the covariance are unknown. However, it is of common practice to assume stationarity of the process. This entails that, even if unknown, the mean vector μ is composed of identical elements (which are often assumed to be equal to zero). This deviates from the assumptions used in deriving Equation 157.

Consequently, when zero-mean stationary processes are considered, the joint mean is known to be zero, i.e., $\mu_{t+1} = 0$. This assumption usually holds because either the data are adjusted or zero-mean Gaussian innovations are assumed. For these cases, the transition density is similarly derived by applying $\kappa \to \infty$ (details are provided in Section E.2) which results in a simplified transition density

$$\begin{aligned} f\left(x_{t+1} | x_{1:t}\right) &= \mathcal{T}_{\nu_w} \left(x_{t+1} | \zeta_{t+1|1:t}, \sigma_{t+1|1:t}^2 \right), \\ \text{with} & \begin{cases} \zeta_{t+1|1:t} = \zeta_{t+1} + l_{1:t}^{\top} L_{1:t}^{-1} \left(x_{1:t} - \zeta_{1:t} \right), \\ \sigma_{t+1|1:t}^2 = h_{t+1|1:t} \left(\lambda_{t+1} - l_{1:t}^{\top} L_{1:t}^{-1} l_{1:t} \right), \\ h_{t+1|1:t} = \frac{1 + (x_{1:t} - \zeta_{1:t})^{\top} L_{1:t}^{-1} (x_{1:t} - \zeta_{1:t})}{\nu_w}. \end{cases} \end{aligned} \tag{159}$$

Marginal distributions for generic time-series: hyper-parameters

The marginalized densities of the time-series derived in Equation 156 and Equation 157 depend on the set of hyper-parameters $\eta \in \mathbb{R}^{t+1}$, $\Lambda \in \mathbb{R}^{(t+1)\times(t+1)}$, $\kappa \in \mathbb{R}$ and $\nu_w \in \mathbb{R}$. We now elaborate on the determination of these hyper-parameters of the Student-t distributions.

Under the used Bayesian hierarchical model, the normal-inverse-Wishart prior distribution can be understood as our prior belief about the mean and covariance of the process. These are parameterized by η and Λ and, by definition, $\kappa > 0$ and $\nu_w > t$. It is important to note that these hyper-parameters depend on the time index t.

Determination of the hyper-parameters is critical for a successful description of the time-series. Here, we leverage the

properties of the process and adopt the empirical Bayesian paradigm. Unlike the standard Bayes approach, which uses a prior that is independent from the observed data, the empirical Bayes method [20] relies on estimating priors from the data.

That is, we combine a-priori knowledge about the process (i.e., stationarity) with the a-posteriori information learned from the observed data. All this information is put into the hyper-parameters of the assumed hierarchical model via the empirical Bayes paradigm.

Suppose we have data up to time instant t. Then, we want to provide meaningful hyper-parameters for the priors for the next time instant $t+1$, by using data available up to time instant t. We do so by embracing a two-fold approach: (1) impose the WSS properties and (2) learn the shape of the autocovariance function from the data. We now explain the specifics.

On the one hand, we take advantage of the wide-sense stationarity of the process and include prior knowledge into the hyper-parameters. Specifically, that the mean does not change over time, i.e., $\mu_t = \mu \mathbf{1}_t$, where $\mathbf{1}_t$ is a $t \times 1$ vector with elements equal to one; and that the covariance matrix of the process is a symmetric Toeplitz matrix dependent on the autocovariance function $\gamma(\tau)$ as in Equation 151.

On the other, the empirical Bayes principles suggest that we compute the empirical mean and autocovariance function of the sequence $x_{1:t}$ and use it for the next set of hyper-parameters η and Λ.

We compute the empirical stationary mean at time instant t by

$$\mu \approx \hat{x}_t = \frac{1}{t} \sum_{i=1}^{t} x_i, \tag{160}$$

and the empirical autocovariance function by

$$\gamma(\tau) \approx \hat{\gamma}(\tau) = \frac{1}{t-\tau} \sum_{i=1}^{t-\tau} (x_i - \hat{x}_t)\,(x_{i+\tau} - \hat{x}_t)\,, \tag{161}$$

for $\tau = \{0, 1, \cdots, t-1\}$.

Following the empirical Bayes approach, we plug these estimated values in the hyper-parameters of the mean and the covariance matrix for the next time instant $t+1$ as follows:

$$\begin{cases} \eta_{t+1} = \hat{x}_t 1_{t+1}, \\ \lambda_{t+1} = \hat{\gamma}(0), \\ l_{1:t} = \begin{pmatrix} \hat{\gamma}(1) & \hat{\gamma}(2) & \cdots & \hat{\gamma}(t-1) & 0 \end{pmatrix}^\top, \\ L_{1:t} = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(t-2) & \hat{\gamma}(t-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(t-3) & \hat{\gamma}(t-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\gamma}(t-2) & \hat{\gamma}(t-3) & \cdots & \hat{\gamma}(0) & \hat{\gamma}(1) \\ \hat{\gamma}(t-1) & \hat{\gamma}(t-2) & \cdots & \hat{\gamma}(1) & \hat{\gamma}(0) \end{pmatrix}. \end{cases} \tag{162}$$

Obviously, when zero-mean stationary processes are considered, there is no need for computation of the empirical mean $\hat{x}_t$ and thus, one can simply assume $\eta = 0, \forall t$ and use the simplified Equation 159.

SMC METHOD FOR LATENT WIDE-SENSE STATIONARY TIME-SERIES

We leverage the transition density in Equation 157 with sufficient statistics in Equation 158 computed as explained in Section 7.1.1.1 to present an SMC sampling method for generic latent WSS time-series in Table 29.

We emphasize that the transition density used for particle propagation does not depend on any of the parameters of the underlying model. Thus, one can apply the SMC in Table 29 to ARMA(p, q) and FARIMA(p, d, q) models with unknown parameters and model orders, to fGps with unknown Hurst parameter, and to many other processes, as long as the WSS conditions are met. That is, one needs to be certain that the joint density of the underlying process is of the form as in Equation 150, but no knowledge on the specific values of the sufficient statistics is required.

Furthermore, note that the derivation of the transition density of a given stream only relies on previous states. That is, it does not involve any observations at all.

On the construction of the marginalized density, the hyper-parameters of the priors for η and Λ are obtained in a per-stream basis, i.e., with information from each stream only.

PF FOR LATENT GAUSSIAN WSS TIME-SERIES

1. At time instant t, consider the random measure

$$f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

2. Estimate the hyper-parameter values for each particle stream $x_{1:t}^{(m)}$

$$\begin{cases} \hat{x}_t^{(m)} = \frac{1}{t}\sum_{i=1}^{t} x_i^{(m)}, \\ \hat{\gamma}(\tau)^{(m)} = \frac{1}{(t-\tau)}\sum_{i=1}^{t-\tau}\left(x_i^{(m)} - \hat{x}_t^{(m)}\right)\left(x_{i+\tau}^{(m)} - \hat{x}_t^{(m)}\right), \\ \qquad \text{where } \tau = \{0, 1, \cdots, t-1\}. \end{cases}$$

3. Upon reception of a new observation at time instant $t+1$.

4. Perform resampling of the state by drawing from a categorical distribution defined by the random measure

$$\overline{x}_t^{(m)} \sim f_t^M(x_t), \text{ where } m = 1, \cdots, M.$$

5. Propagate the particles by sampling from the transition density, given the (resampled) stream of particles:

$$x_{t+1}^{(m)} \sim f(x_{t+1}|\overline{x}_{1:t}^{(m)}) = \mathcal{T}_{\nu_w}\left(x_{t+1}|\zeta_{t+1|1:t}^{(m)}, \sigma_{t+1|1:t}^{2(m)}\right),$$

$$\text{with } \begin{cases} \zeta_{t+1|1:t}^{(m)} = \zeta_{t+1}^{(m)} + l_{1:t}^{(m)^\top} L_{1:t}^{(m)^{-1}}\left(\overline{x}_{1:t}^{(m)} - \zeta_{1:t}^{(m)}\right), \\ \sigma_{t+1|1:t}^{2(m)} = h_{t+1|1:t}^{(m)}\left(\lambda_{t+1}^{(m)} - l_{1:t}^{(m)^\top} L_{1:t}^{(m)^{-1}} l_{1:t}^{(m)}\right), \\ h_{t+1|1:t}^{(m)} = \frac{\frac{(1+\kappa)}{\kappa} + \left(\overline{x}_{1:t}^{(m)} - \zeta_{1:t}^{(m)}\right)^\top L_{1:t}^{(m)^{-1}}\left(\overline{x}_{1:t}^{(m)} - \zeta_{1:t}^{(m)}\right)}{\nu_w}. \end{cases}$$

   for each stream $m = 1, \cdots, M$ as in Equation 162.

6. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m)} \propto f(y_{t+1}|x_{t+1}^{(m)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^M(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 29: PF for latent Gaussian WSS time-series.

The computed $\hat{\gamma}^{(m)}(\tau)$ values are hyper-parameters of the inverse-Wishart prior and not estimates of the covariance matrix of the underlying process. Thus, the observations have no role in the construction of the prior nor the transition density. The only place where the observations play a role is in the weight computation and in the selection of which streams to keep and which to discard from further processing.

EVALUATION

We evaluate the proposed method by considering latent WSS time-series and the stochastic log-volatility (SV) model.

In particular, and without loss of generality, we hereby examine the latent ARMA(p, q) process. Note that many other models could be considered, as long as the WSS conditions are met.

We focus on the ARMA(p, q) model because we can readily compare the performance of the proposed method to those in Chapter 3 and Chapter 4. The added flexibility of the SMC proposed in this chapter can also be evaluated by considering ARMA(p, q) models of unknown order (that is, any short-memory process).

We mathematically write the considered state-space as

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j u_{t-j} + u_t, \\ y_t = e^{\frac{x_t}{2}} v_t, \end{cases} \tag{163}$$

where the only assumptions made are that v_t is the standard Gaussian and the state noise u_t is a zero-mean Gaussian. That is, we consider a wide-sense stationary ARMA state process in its most generic form, where the orders p and q are not known. Thus, the parameters a, b, and σ_u^2 of the latent process are all unknown.

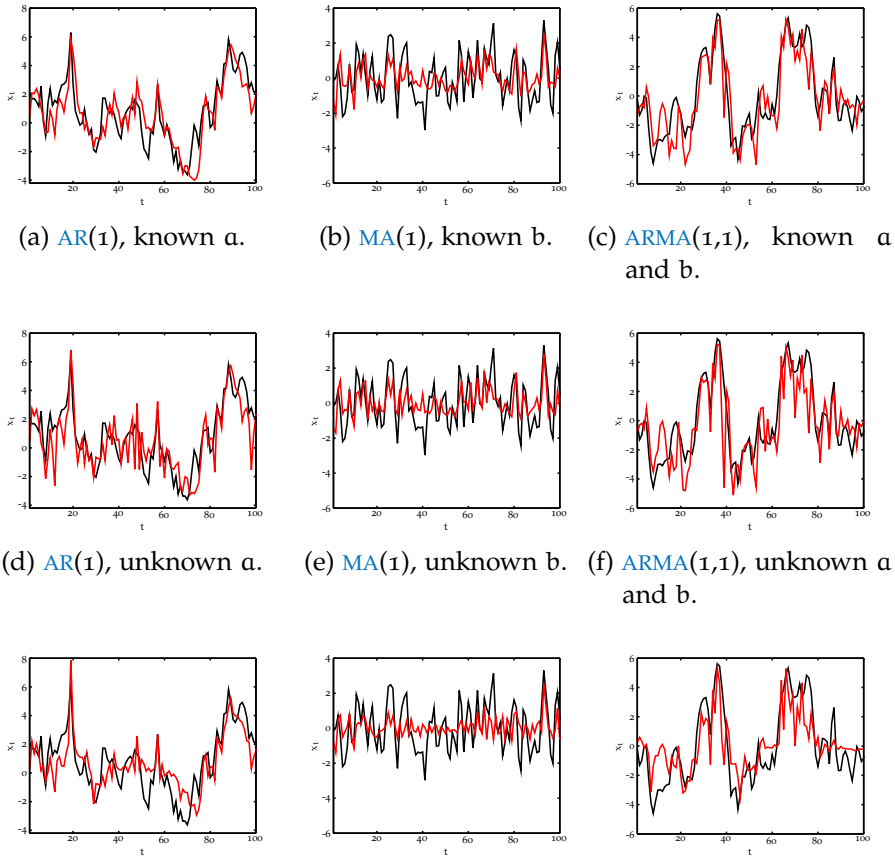
We follow the derivation in Section 7.1.1.1 and, since zero-mean innovations are considered, use the transition density in Equation 159 with $\nu_w = t + 1$ for all time instants. This implies that, as time evolves, the transition Student-t distribution has more degrees of freedom, as we are conditioning on more available data samples.

First, we illustrate the accuracy of the sequential estimates provided by the proposed method (i.e., the evolution of the state estimates). In Figure 42, we compare state estimates for particular realizations of an AR(1) with $a_1 = 0.8$, a MA(1) with

$b_1 = 0.8$ and an ARMA(1,1) with $a_1 = 0.8$ and $b_1 = 0.5$ processes, respectively.

We evaluate the proposed method by comparing it with the previous alternatives: (1) when the parameters of the ARMA process are known as in Chapter 3, (2) when the parameters of the latent process are unknown but the model order is known as in Chapter 4, and (3) when both, the model order and parameters are unknown, i.e., the setting considered in Chapter 7.

On the top row of Figure 42, we show the results of the estimation when everything is known (i.e., PF in Table 5); on the second, when only the model order is known (i.e., PF in Table 16); and on the third, when nothing is known (i.e., PF in Table 29).



(a) AR(1), known a. (b) MA(1), known b. (c) ARMA(1,1), known a and b.

(d) AR(1), unknown a. (e) MA(1), unknown b. (f) ARMA(1,1), unknown a and b.

(g) AR(1), unknown p and q. (h) MA(1), unknown q. (i) ARMA(1,1), unknown p and q.

Figure 42: True (black) and estimated (red) state for three different proposed SMC methods.

The figure shows that all the PFs are able to track the hidden process. As observed in Chapter 4, estimating pure correlated noise as in MA(q) processes is the most challenging case, while the estimation accuracy improves with the presence of strong autoregressive components.

We provide the MSE of the latent ARMA process estimates for the same three SMC methods in Table 30. The results clearly indicate how the increased lack of knowledge about the latent process reflects on the estimation performance of the PFs.

The results are averaged over 50 realizations, which are 100 samples long. We use $M = 500$ particles and the parameters of the processes were set to:

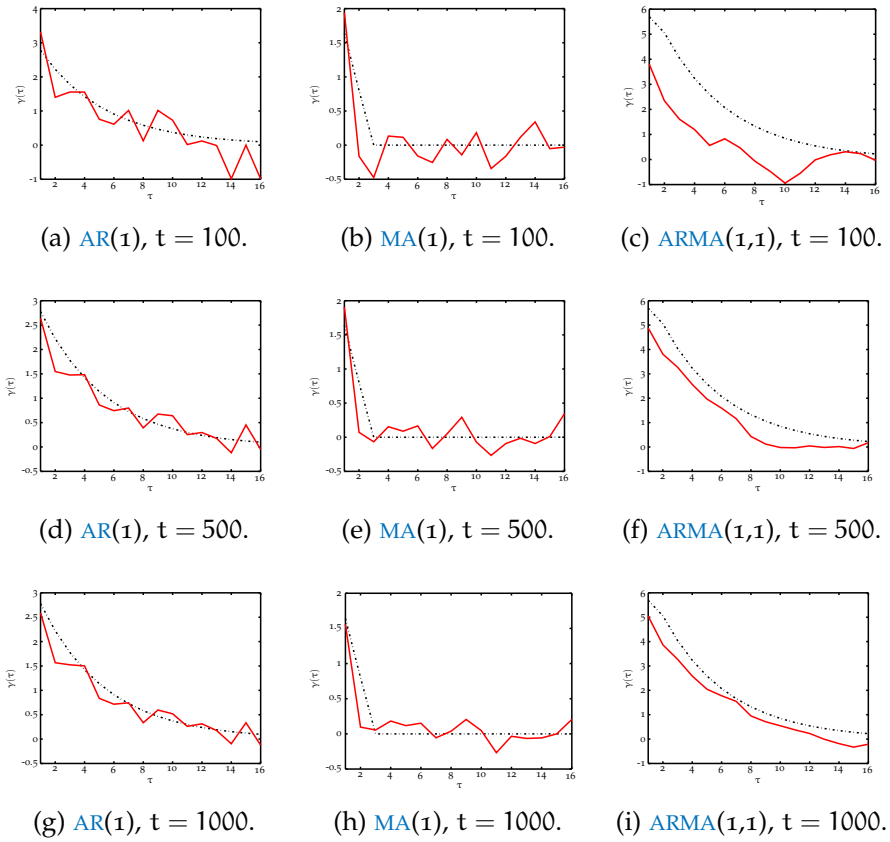
- AR(1) with $a_1 = 0.85$;
- AR(2) with $a_1 = 0.8$, $a_2 = 0.1$;
- AR(3) with $a_1 = 0.65$, $a_2 = -0.2$, $a_3 = 0.1$;
- MA(1) with $b_1 = 0.5$;
- ARMA(1,1) with $a_1 = 0.85$, $b_1 = 0.5$;
- ARMA(1,1) with $a_1 = 0.85$, $b_1 = -0.5$;
- ARMA(1,2) with $a_1 = 0.85$, $b_1 = 0.5$, $b_2 = 0.1$; and
- ARMA(2,1) with $a_1 = 0.8$, $a_2 = 0.1$, $b_1 = 0.5$.

ARMA(p,q)	State estimation error (MSE)		
	`Known Params`	`Unknown Params`	`All Unknown`
AR(1)	1.16164	1.49248	1.99644
AR(2)	1.20956	1.67136	2.26017
AR(3)	0.96545	1.02314	1.23342
MA(1)	0.89428	0.93959	1.07060
ARMA(1,1)	1.56405	2.09366	2.91670
ARMA(1,1)	0.88028	0.98532	1.16023
ARMA(1,2)	1.78301	2.32099	3.36742
ARMA(2,1)	1.77861	2.33336	3.74361

Table 30: MSE comparison for proposed SMC methods with relaxed assumptions.

The estimation of the hyper-parameters is critical for the performance of our method. Clearly, the quality of the estimated

values in Equation 162 heavily relies on the amount of available data. We illustrate in Figure 43 how, as more data are used for estimation, the autocovariance estimates become more stable and accurate.



(a) AR(1), t = 100. (b) MA(1), t = 100. (c) ARMA(1,1), t = 100.

(d) AR(1), t = 500. (e) MA(1), t = 500. (f) ARMA(1,1), t = 500.

(g) AR(1), t = 1000. (h) MA(1), t = 1000. (i) ARMA(1,1), t = 1000.

Figure 43: True (black) and estimated (red) autocovariance function of the latent process.

For the method proposed in Table 29, one can again consider to approximate the sufficient statistics by means of truncation. The benefits are two-pronged: (1) by truncating the computations of the sufficient statistics to a maximum lag τ_{max}, the computational (and memory) cost of the method is reduced; and (2), the accuracy of the estimates in Equation 160 and Equation 161 improves as more data are available.

With truncation, one needs to compute the autocovariance function $\hat{\gamma}(\tau)$ only for a relatively short window $\tau = 0, \cdots, \tau_{max}$. Besides, the sufficient statistics become dependent on only the most recent past, i.e., $\phi\left(x_{1:t}^{(m)}\right) \approx \phi\left(x_{t-\tau_{max}:t}^{(m)}\right)$.

We note that one must be careful with truncation, as the memory features of the process will determine how much

information is lost. In general, for short-memory processes (e.g., ARMA(p, q)), one incurs in negligible information loss, because $\gamma(\tau) \approx 0, \tau > \tau_{max}$. However, the truncation might imply significant information loss when long-memory processes are in play (e.g., fGp with $H \to 1$).

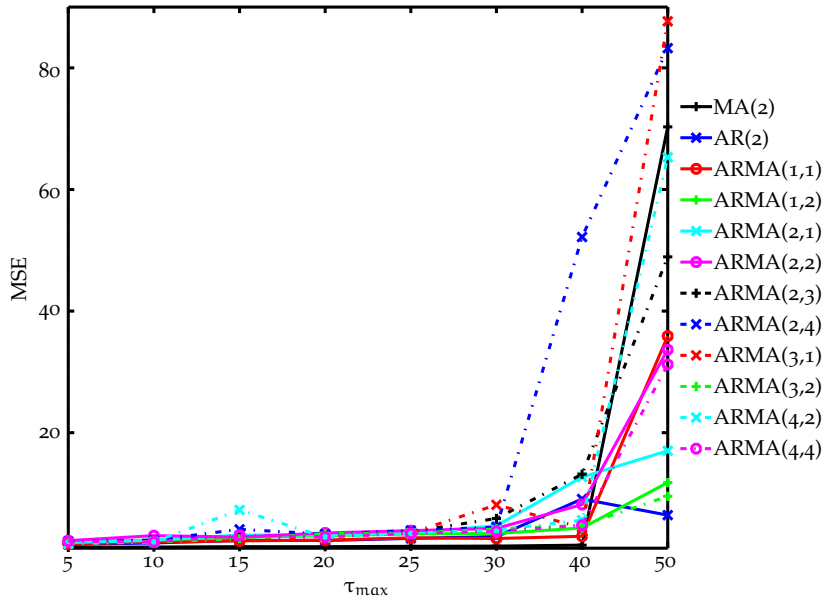
In the following, we evaluate the MSE for different values of truncation lags for several ARMA(p, q) processes, with parameters as follows:

- MA(2) with $b_1 = 0.8$, $b_2 = 0.15$;
- AR(2) with $a_1 = 0.8$, $a_2 = 0.15$;
- ARMA(1,1) with $a_1 = 0.9$, $b_1 = 0.5$;
- ARMA(1,2) with $a_1 = 0.8$, $b_1 = 0.8$, $b_2 = 0.15$;
- ARMA(2,1) with $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$;
- ARMA(2,2) with $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.8$, $b_2 = 0.15$;
- ARMA(2,3) with $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$, $b_2 = 0.3$, $b_3 = 0.15$;
- ARMA(2,4) with $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$;
- ARMA(3,1) with $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.5$;
- ARMA(3,2) with $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.8$, $b_2 = 0.15$;
- ARMA(4,2) with $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.8$, $b_2 = 0.15$; and
- ARMA(4,4) with $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$.

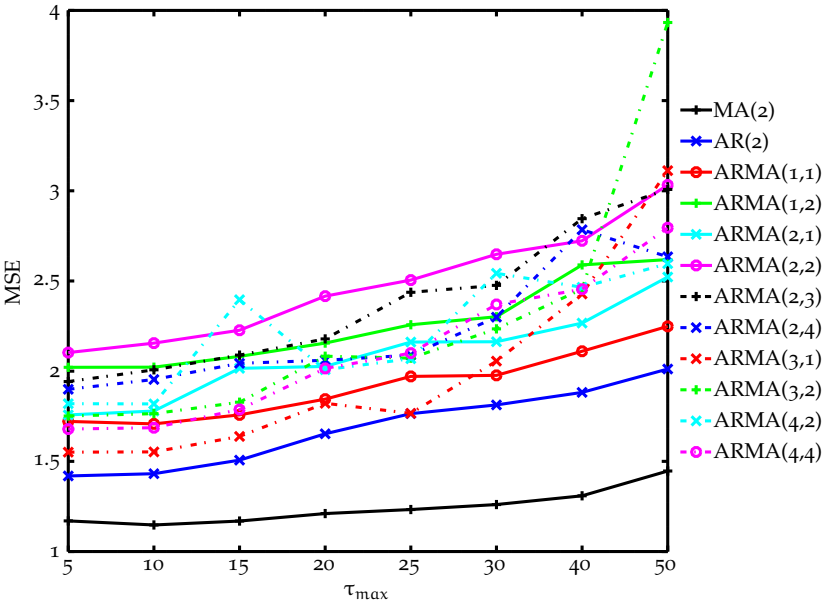
We illustrate in Figure 44, for time-series of length 100 time instants (top) and 1,000 time instants (bottom), the influence of truncating the sufficient statistics for the SMC method in Table 29 (tested with $M = 1000$ particles).

We observe that when short time-series are used, the hyperparameters for windows with $\tau_{max} > 30$ are not accurate enough and, thus, the resulting filtering performance deteriorates considerably (see Figure 44a).

On the contrary, as more data become available, the hyperparameter estimates improve (as shown in Figure 43) and so the inferred states do too (see Figure 44b).



(a) MSE for time-series of length $t = 100$.



(b) MSE for time-series of length $t = 1000$.

Figure 44: Impact of the covariance truncation in the MSE performance of the proposed SMC.

This improvement in accuracy is explained not only by the better quality of the $\gamma(\tau)$ estimation, but also by the more degrees of freedom ($\nu_w = t + 1$) used for the transition density

(i.e., the density becomes more similar to a Gaussian as more data are observed).

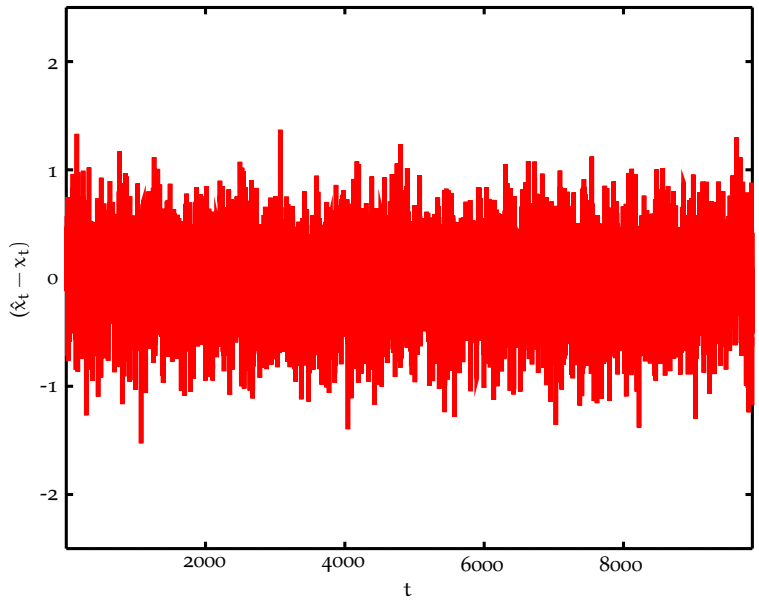
As a matter of fact, these results also justify using a burn-out period for the SMC method. That is, one should allow for the hyper-parameter estimates to stabilize with some initial data, so that the subsequent state estimates are computed with improved accuracy.

The results in Figure 44b emphasize the importance of selecting an appropriate window size. Windows that are too long hinder the estimation performance and, at the same time, increase the computational complexity of the method. Results in Figure 44b indicate that for the simulated processes, small truncation lags ($\tau_{max} \leqslant 25$) provide good estimates of the autocovariance function and, thus, lead to good filtering performance.

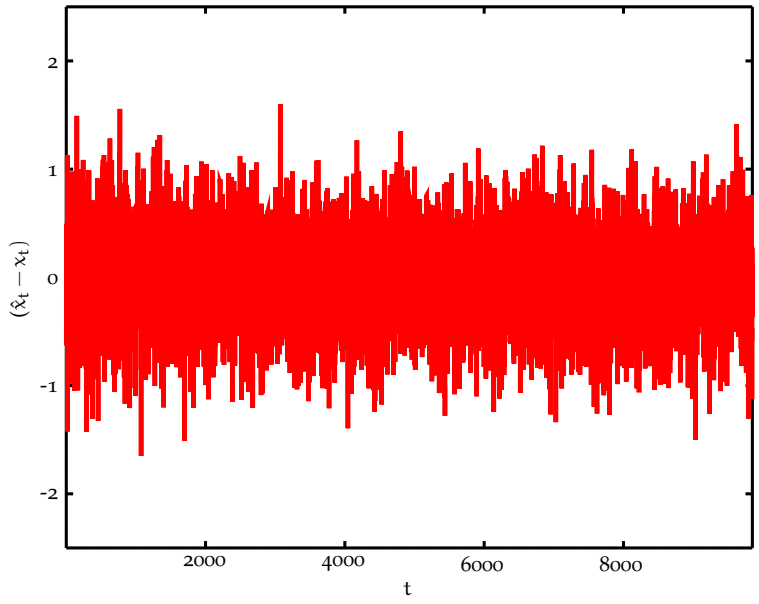
Nonetheless, we reiterate that the consideration of such small windows makes sense only due to the short-memory of the latent ARMA processes studied here. Careful consideration of which window to use must be taken when dealing with long-memory process, if truncation is used at all.

In other words, there is an important trade-off between computational cost and estimation accuracy of the proposed SMC method. The ratio between available data and the number of variables to estimate (i.e., $\hat{\gamma}(\tau)$ for $\tau = 0, \cdots, \tau_{max}$) plays a very important role. Also, one must carefully study the memory properties of the latent state before applying any truncation.

We illustrate this trade-off in Figure 45, where we observe how the performance worsens when considering long windows (MSE of 1.51 for $\tau_{max} = 25$ and 1.80 for $\tau_{max} = 100$). These results, however, reinforce our claims about the unbiasedness and consistency of the proposed method, as the estimation error does not degrade with time.



(a) ARMA(1,1) with $\tau_{max} = 25$.



(b) ARMA(1,1) $\tau_{max} = 100$.

Figure 45: State estimation error over time for long time-series.

8

LATENT CORRELATED TIME-SERIES

In this chapter, we extend on the studied models by considering multiple correlated time-series. This can be understood as a multivariate time-series, which we model as a set of (potentially) correlated mixtures of independent processes embedded in noise. We embrace a hierarchical model, where each independent time-series is correlated via a mixing matrix.

We provide a Bayesian analysis of the hierarchical model in Section 8.1, where Rao-Blackwellization of the unknowns is implemented, in order to propose an SMC method in Section 8.2. We conclude with an evaluation Section 8.3.

CORRELATED TIME-SERIES

We are interested in making inference of correlated latent processes observed through nonlinear functions. Specifically, we consider state-space models that are described by a set of latent Gaussian independent processes correlated via a mixing matrix and embedded in noise.

Let $s_t \in \mathbb{R}^{d_s}$ be a vector of independent time-series, where each of them is described by their own density. That is, we consider a set of independent processes $s_{i,t}, i = 1, \cdots, d_s$; where the only assumption is that the transition density for each of them is known, i.e.,

$$s_{i,t+1} \sim f(s_{i,t+1}|s_{i,1:t}), \qquad i = 1, \cdots, d_s. \tag{164}$$

These latent independent time-series denoted as s_t here can, for example, be any of the processes studied in the previous chapters: i.e., ARMA processes, fGps or ARMA models with correlated innovations. Again, the assumptions on the model specifics and knowledge of its parameters is not relevant here, as long as the transition density can be computed. As such, the methodology below works for any given distribution.

We model a set of latent correlated processes x_t as a mixture of those independent time-series s_t. That is, let $x_t \in \mathbb{R}^{d_x}$ be a set of latent correlated processes and $y_t \in \mathbb{R}^{d_y}$ the observed vector at time t.

Mathematically, we represent the hierarchical model as

$$\begin{cases} s_{i,t} \sim f(s_{i,t+1}|s_{i,1:t}), & i = 1, \cdots, d_s, \\ x_t = As_t + \epsilon_t, \\ y_t = h(x_t, v_t), \end{cases} \tag{165}$$

where $t = 1, 2, \cdots$; $\epsilon_t \in \mathbb{R}^{d_x}$ is a zero mean Gaussian vector with covariance matrix C_ϵ; $A \in \mathbb{R}^{d_x \times d_s}$ is a mixing matrix; $v_t \in \mathbb{R}^{d_v}$ denotes an independent white Gaussian noise; and $h(x_t, v_t) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_v} \to \mathbb{R}^{d_y}$, is some nonlinear function.

Given a set of observations $y_{1:t} \equiv \{y_1, y_2, \cdots, y_t\}$, we want to sequentially estimate the posterior distribution of x_t, $f(x_t|y_{1:t})$. To do so, we need to proceed sequentially and the challenge is to estimate $f(x_{t+1}|x_{1:t})$ for the state-space model as in Equation 165, given its hierarchical nature and the properties of the latent process.

By following the Bayesian paradigm, we now derive the state transition density under different assumptions for the parameters in Equation 165.

Correlated time-series: joint and transition densities

We are interested in the joint filtering density of both latent states in Equation 165, i.e., $f(s_t, x_t|y_{1:t})$ and thus, we need to derive the joint transition density $f(s_{t+1}, x_{t+1}|s_{1:t}, x_{1:t})$.

Due to the hierarchical structure of the model, we can factorize it as

$$f(s_{t+1}, x_{t+1}|s_{1:t}, x_{1:t}) = f(s_{t+1}|s_{1:t}) f(x_{t+1}|s_{t+1}). \tag{166}$$

That is, we can derive the joint transition density of the states by separately studying the latent time-series s_t and the correlated counterpart x_t.

On the one hand, the transition density of the independent time-series, given all their sufficient statistics, $f(s_{t+1}|s_{1:t})$, follows a multivariate distribution, as long as the sufficient statistics of each of them are computable.

For example, if Gaussian transition densities are in place, i.e.,

$$s_{i,t+1} \sim f(s_{i,t+1}|s_{i,1:t}) = \mathcal{N}\left(s_{i,t+1} \,\middle|\, \mu_{i,t+1}, \sigma^2_{i,t+1}\right), i = 1, \cdots, d_s, \tag{167}$$

then the joint distribution of the latent states is a multivariate Gaussian $f(s_{t+1}|s_{1:t}) = \mathcal{N}(s_{t+1}|\mu_{t+1}, C_{t+1})$, with parameters

$$\begin{cases} \mu_{t+1} = \begin{pmatrix} \mu_{1,t+1} & \cdots & \mu_{d_s,t+1} \end{pmatrix}^\top, \\ C_{t+1} = \begin{pmatrix} \sigma^2_{1,t+1} & 0 & \cdots \\ 0 & \sigma^2_{2,t+1} & \cdots \\ \vdots & \vdots & \ddots \\ 0 & 0 & \sigma^2_{d_s,t+1} \end{pmatrix}. \end{cases} \tag{168}$$

On the other, we need to compute the conditional density of the correlated states, given the independent ones, i.e., $f(x_{t+1}|s_{t+1})$. By inspection of the model in Equation 165, we deduce that the transition density of x_{t+1} given s_{t+1} is a multivariate Gaussian

$$f(x_{t+1}|s_{t+1}) = \mathcal{N}(x_{t+1}|As_{t+1}, C_\epsilon), \tag{169}$$

for known A and C_ϵ.

Nevertheless, it is unrealistic to know their true values in practice and thus, we further study this challenging case. As already pointed out, we resort to marginalizing out or Rao-Blackwellizing the unknowns.

Consider the estimate of the mixing matrix at time instant t, $\hat{A}_t = X_t U_t^\top (U_t U_t^\top)^{-1}$, where the following historical data matrices have been defined [65]:

$$\begin{cases} X_t = [x_1 x_2 \cdots x_t] \in \mathbb{R}^{d_x \times t}, \\ U_t = [s_1 s_2 \cdots s_t] \in \mathbb{R}^{d_s \times t}. \end{cases} \tag{170}$$

In the following, instead of using such point estimates, we leverage them to determine full prior distributions. We integrate out the unknowns A and C_ϵ to derive the transition density of x_{t+1}, given s_{t+1}, X_t and U_t, by following the derivation in [48].

The resulting density is a multivariate Student-t distribution

$$f(x_{t+1}|s_{t+1}, X_t, U_t) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1}|\mu_{x_{t+1}}, R_{x_{t+1}}\right), \tag{171}$$

with ν_{t+1} degrees of freedom, location parameter $\mu_{x_{t+1}} \in \mathbb{R}^{d_x}$ and scale matrix $R_{x_{t+1}} \in \mathbb{R}^{d_x \times d_x}$ [13], computed by

$$\begin{cases} \nu_{t+1} = t - d_x - d_s + 1, \\ \mu_{x_{t+1}} = \hat{A}_t s_{t+1}, \\ R_{x_{t+1}} = \frac{(X_t - \hat{A}_t U_t)(X_t - \hat{A}_t U_t)^\top}{\nu_{t+1}\left(1 - s_{t+1}^\top (U_{t+1} U_{t+1}^\top)^{-1} s_{t+1}\right)}. \end{cases} \tag{172}$$

Note that this density does not depend on any of the parameters A and C_ϵ.

SMC METHOD FOR CORRELATED TIME-SERIES

We leverage the densities derived in Section 8.1.1 to propose a SMC sampling method for inference of correlated latent time-series. As explained before, we proceed sequentially, i.e., for a new observation y_{t+1}, we update $f(x_t|y_{1:t})$ to a new distribution $f(x_{t+1}|y_{1:t+1})$. This density is derived using the Bayes' rule

$$f(x_{t+1}|y_{1:t+1}) \propto f(y_{t+1}|x_{t+1}) \int f(x_{t+1}|x_{1:t}) f(x_{1:t}|y_{1:t}) dx_{1:t} \quad (173)$$

and thus, one needs the state transition density $f(x_{t+1}|x_{1:t})$.

For the model in Equation 165, we have shown that we can factorize the transition density

$$f(s_{t+1}, x_{t+1}|s_{1:t}, x_{1:t}) = f(s_{t+1}|s_{1:t}) f(x_{t+1}|s_{t+1}). \quad (174)$$

Thus, the SMC method we propose relies on hierarchical sampling. That is, one first propagates from s_t to s_{t+1}, then conditions on those samples to propagate x_{t+1}.

In summary, we propose to:

1. Propagate M_s samples of the latent process s_t from its transition density

$$s_{t+1}^{(m_s)} \sim f(s_{t+1}|s_{1:t}^{(m_s)}), \qquad m_s = 1, \cdots, M_s. \quad (175)$$

2. Propagate M_x state samples per each independent process sample $s_{t+1}^{(m_s)}$ (to improve diversity) from the conditional transition density

$$x_{t+1}^{(m_s, m_x)} \sim f(x_{t+1}|s_{t+1}^{(m_s)}), \qquad m_x = 1, \cdots, M_x. \quad (176)$$

3. Compute weights of the particles based on the likelihood of the data

$$\widetilde{w}_{t+1}^{(m_s, m_x)} \propto f(y_{t+1}|x_{t+1}^{(m_s, m_x)}). \quad (177)$$

4. Downsample from $M_s \cdot M_x$ to M_s to prevent the growth of the number of samples with time. That is, draw a tuple $\left\{\overline{s}_{1:t+1}^{(m_s)}, \overline{x}_{1:t+1}^{(m_s)}\right\}$ from a categorical distribution defined by the random measure

$$\left\{\overline{s}_{1:t+1}^{(m_s)}, \overline{x}_{1:t+1}^{(m_s)}\right\} \sim f_{t+1}^{M_s M_x}(s_{t+1}, x_{t+1}). \quad (178)$$

The details of the proposed SMC method for the known and unknown parameter case are described in Table 31 and Table 32 respectively.

EVALUATION

We illustrate the applicability of the proposed SMC method by studying the inference of latent correlated time-series that exhibit long-memory properties. These processes have attracted the attention of many practitioners [12] and, for instance in finance, the interest in multivariate long-memory volatility processes is of increasing interest [8–10, 25, 31, 60, 83].

Thus, we investigate state-space models where the latent states represent correlated mixtures of independent fGps embedded in white Gaussian noise and the observed data are nonlinear functions of the states.

The paradigm of self-similarity and scale-invariance has recently attracted a lot of attention within the finance community due to the multi-scale nature of econometric data. For capturing these features, both fractal-system analysis [18] and self-similar processes [91] have been popularized.

However, econometric data are often multivariate and, thus, the concept of scale invariance needs to be generalized. To that end, a recently suggested approach is based on the use of Operator fractional Gaussian process (OfGp)s [2, 35], which are multivariate Gaussian self-similar processes. The approach to describing these processes coincides with the model addressed in this chapter, i.e., that of correlated mixtures of independent processes embedded in noise.

Specifically, OfGps model multivariate self-similar and scale-invariant data based on linearly mixing a set of independent fGps, where each of them may have a different Hurst parameter H. The Hurst parameter relates, amongst others, to the memory properties of the time-series (see Chapter 5 for details). The resulting stationarity, self-similarity and other properties of such OfGps have been studied in [34] and references therein.

Here, we evaluate the SMC methods presented in Table 31 and Table 32 for studying OfGps embedded in noise (and observed through nonlinear functions) for known and unknown mixing parameters.

Let us write the specifics of the state-space in Equation 165 for the OfGp considered. We start by defining $u_t \in \mathbb{R}^{d_u}$ as a vector of independent fGps, each with a Hurst parameter H_i,

PF FOR LATENT CORRELATED TIME-SERIES, KNOWN A AND C_ϵ

1. At time instant t, consider the random measure

$$f_t^{M_s}(s_t, x_t) = \sum_{m=1}^{M_s} w_t^{(m_s)} \delta\left(\begin{pmatrix} s_t \\ x_t \end{pmatrix} - \begin{pmatrix} s_t^{(m_s)} \\ x_t^{(m_s)} \end{pmatrix}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Propagate M_s samples of the latent process s_t from its transition density, given (resampled) streams

$$s_{t+1}^{(m_s)} \sim f(s_{t+1}|\overline{s}_{1:t}^{(m_s)}), \qquad m_s = 1, \cdots, M_s.$$

4. Propagate M_x samples per each process sample $s_{t+1}^{(m_s)}$ (to improve diversity) from the conditional transition density

$$x_{t+1}^{(m_s,m_x)} \sim f(x_{t+1}|s_{t+1}^{(m_s)}) = \mathcal{N}(x_{t+1}|As_{t+1}^{(m_s)}, C_\epsilon),$$

for $m_x = 1, \cdots, M_x$.

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m_s,m_x)} \propto f(y_{t+1}|x_{t+1}^{(m_s,m_x)}),$$

and normalize them to obtain a new random measure

$$f_{t+1}^{M_s M_x}(s_{t+1}, x_{t+1}) = \sum_{m_s=1}^{M_s} \sum_{m_x=1}^{M_x} w_{t+1}^{(m_s,m_x)} \delta\left(\begin{pmatrix} s_{t+1} \\ x_{t+1} \end{pmatrix} - \begin{pmatrix} s_{t+1}^{(m_s)} \\ x_{t+1}^{(m_s,m_x)} \end{pmatrix}\right).$$

6. Downsample from $M_s \cdot M_x$ to M_s to prevent the growth of the number of samples with time. That is, draw a tuple $\left(\overline{s}_{1:t+1}^{(m_s)}, \overline{x}_{1:t+1}^{(m_s)}\right)^\top$ from a categorical distribution defined by the random measure

$$\left\{\overline{s}_{1:t+1}^{(m_s)}, \overline{x}_{1:t+1}^{(m_s)}\right\} \sim f_{t+1}^{M_s M_x}(s_{t+1}, x_{t+1}).$$

Table 31: PF for latent correlated time-series with known A and C_ϵ.

PF FOR LATENT CORRELATED TIME-SERIES, UNKNOWN A AND C_ϵ

1. At time instant t, consider the random measure

$$f_t^{M_s}(s_t, x_t) = \sum_{m=1}^{M_s} w_t^{(m_s)} \delta\left(\begin{pmatrix} s_t \\ x_t \end{pmatrix} - \begin{pmatrix} s_t^{(m_s)} \\ x_t^{(m_s)} \end{pmatrix}\right).$$

2. Upon reception of a new observation at time instant $t+1$.

3. Propagate M_s samples of the latent process s_t from its transition density, given (resampled) streams

$$s_{t+1}^{(m_s)} \sim f(s_{t+1}|\overline{s}_{1:t}^{(m_s)}), \qquad m_s = 1, \cdots, M_s.$$

4. Propagate M_x samples per each process sample $s_{t+1}^{(m_s)}$ (to improve diversity) from the conditional transition density

$$x_{t+1}^{(m_s,m_x)} \sim f(x_{t+1}|s_{t+1}^{(m_s)}, X_t^{(m_s)}, U_t^{(m_s)}) = \mathcal{T}_{\nu_{t+1}}\left(x_{t+1}|\mu_{x_{t+1}}^{(m_s)}, R_{x_{t+1}}^{(m_s)}\right),$$

$$\text{with} \begin{cases} \nu_{t+1} = t - d_x - d_s + 1, \\ \mu_{x_{t+1}}^{(m_s)} = \hat{A}_t^{(m_s)} s_{t+1}^{(m_s)}, \\ R_{x_{t+1}}^{(m_s)} = \dfrac{\left(X_t^{(m_s)} - \hat{A}_t^{(m_s)} U_t^{(m_s)}\right)\left(X_t^{(m_s)} - \hat{A}_t^{(m_s)} U_t^{(m_s)}\right)^\top}{\nu_{t+1}\left(1 - s_{t+1}^{(m_s)\top}(U_{t+1}^{(m_s)} U_{t+1}^{(m_s)\top})^{-1} s_{t+1}^{(m_s)}\right)}. \end{cases}$$

   for $m_x = 1, \cdots, M_x$ and using Equation 170 per sample.

5. Compute the non-normalized weights for the particles

$$\widetilde{w}_{t+1}^{(m_s,m_x)} \propto f(y_{t+1}|x_{t+1}^{(m_s,m_x)}),$$

   and normalize them to obtain a new random measure

$$f_{t+1}^{M_s M_x}(s_{t+1}, x_{t+1}) = \sum_{m_s=1}^{M_s} \sum_{m_x=1}^{M_x} w_{t+1}^{(m_s,m_x)} \delta\left(\begin{pmatrix} s_{t+1} \\ x_{t+1} \end{pmatrix} - \begin{pmatrix} s_{t+1}^{(m_s)} \\ x_{t+1}^{(m_s,m_x)} \end{pmatrix}\right).$$

6. Downsample from $M_s \cdot M_x$ to M_s from a categorical distribution defined by the random measure

$$\left\{\overline{s}_{1:t+1}^{(m_s)}, \overline{x}_{1:t+1}^{(m_s)}\right\} \sim f_{t+1}^{M_s M_x}(s_{t+1}, x_{t+1}).$$

Table 32: PF for latent correlated time-series with unknown A and C_ϵ.

variance σ_i^2 and autocovariance function as in Equation 116. Note that, for $\frac{1}{2} < H_i < 1, i = 1, \cdots, d_u$, the process has long-range dependence; and, for $H_i = 0.5$, it is uncorrelated. Now, let $x_t \in \mathbb{R}^{d_x}$ be a set of latent correlated processes (i.e., the OfGp) and $y_t \in \mathbb{R}^{d_y}$ the observed vector at time t.

The resulting hierarchical model for inference of a latent OfGp is mathematically written as

$$\begin{cases} u_{i,t} \sim \mathrm{fGp}(H_i, \sigma_i^2), & i = 1, \cdots, d_u, \\ x_t = A u_t + \epsilon_t, \\ y_t = h(x_t, \nu_t), \end{cases} \tag{179}$$

where same notation and conditions as in Equation 165 apply.

Now, for evaluation purposes, we specifically address the SV model with correlated trend and volatility observations. That is, we simplify Equation 179 to (1) a latent two-dimensional OfGp state and (2) the SV model with correlated trend and volatility.

The model is given by

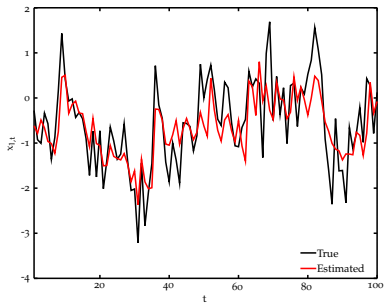
$$\begin{cases} u_{1,t} \sim \mathrm{fGp}(H_1, \sigma_1^2), \\ u_{2,t} \sim \mathrm{fGp}(H_2, \sigma_2^2), \\ \\ \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2 I), \\ \\ y_t = x_{1,t} + e^{x_{2,t}/2} \nu_t, \end{cases} \tag{180}$$

where ν_t represents a standard Gaussian variable.

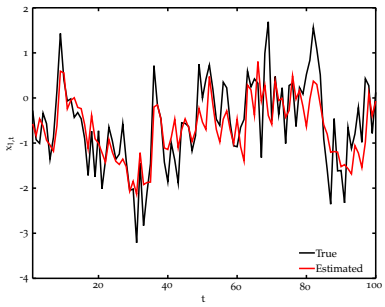
Note that ρ is the idiosyncratic correlation between the trend $x_{1,t}$ and log-volatility $x_{2,t}$ of the return y_t observed over time, while σ_ϵ^2 is the variance of the additive noise ϵ_t.

The latent OfGp allows for modeling of clustering in the return and volatility [83], i.e., the tendency of asset returns to show large magnitudes in periods of high volatility and calmness in periods of low volatility. Furthermore, we relate the expected asset return to its risk or volatility. We consider an idiosyncratic correlation between the return and its volatility through the mixing matrix A and, at the same time, allow for some random perturbations. The model is used to illustrate the performance of the suggested method, but we do not claim that it fits to any particular instance of real data.

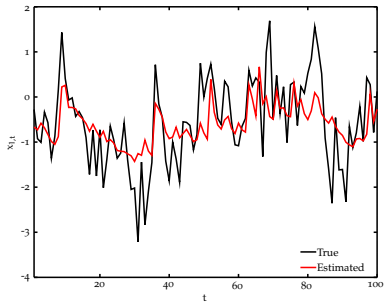
We show on a particular realization the capability of the proposed SMC methods in Section 8.2 to track the latent state under different settings in Figure 46. We present the estimates of only $x_{1,t}$ here, but similar results are obtained for all the latent variables. The plots show a good tracking accuracy and suggest that the impact of not knowing σ_i^2 is less severe than not knowing the mixing parameters.



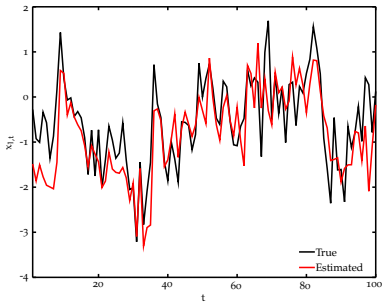
(a) All parameters are known.



(b) The variances σ_i^2 are unknown.



(c) The mixing parameters are unknown.



(d) All parameters are unknown.

Figure 46: True (black) and estimated (red) states $x_{1,t}$.

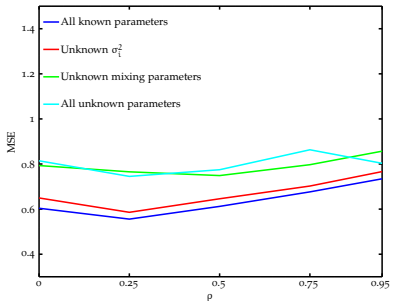
We present averaged filtering MSE results in Figure 47, Figure 48, Figure 49 and Figure 50, where the proposed method is evaluated for different combinations of Hurst parameters and ρ values. All the results have been averaged over 25 realizations and obtained with known H_i values and $M_u = 500$, $M_x = 20$, $\sigma_\epsilon^2 = 0.01$. We note that one expects increased accuracy (specially for x_1 and x_2) for bigger values of M_x, at a higher computational cost (i.e., cost of oversampling from M_u to $M_u \cdot M_x$).

The study of the results show that, with the increase of memory of the latent process (i.e., when $H_i \to 1$), the MSE of the estimated latent states decreases. This effect is more

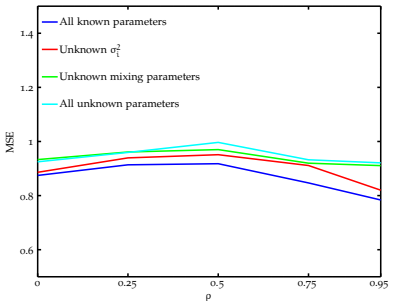
evident for the variables x_1 and x_2, but it is also observed for the independent fGps u_1 and u_2. This finding is aligned with those in Section 5.3.

Regarding the estimation of the trend and the log-volatility, one immediately concludes that estimation accuracy for the log-volatility is worse, when compared to the trend of the observations. That is, the method is able to track the trend (i.e., u_1 and x_1) much better than the log-volatility (i.e., u_2 and x_2). The explanation is that the dependency of the trend x_1 with the observations y_t is linear, while it is nonlinear for the log-volatility x_2.

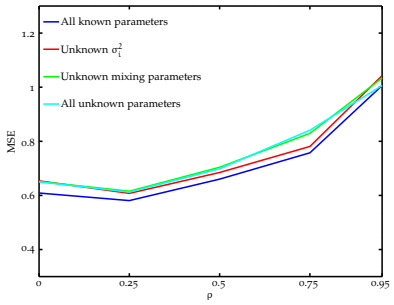
Finally, the results reveal that the performance of the SMC methods is consistent for different values of ρ. The justification comes from the way the proposed approach deals with the mixing parameters (by integrating them out). We reiterate that the marginalization is generic, as we do not assume any particular value or structure for A and C_ϵ.



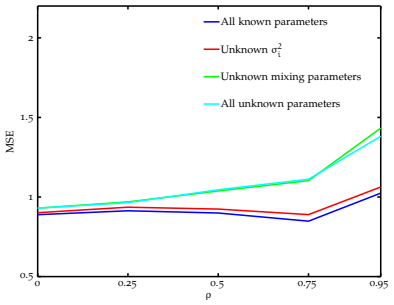
(a) u_1 for $H_1 = 0.5$, $H_2 = 0.5$.



(b) u_2 for $H_1 = 0.5$, $H_2 = 0.5$.

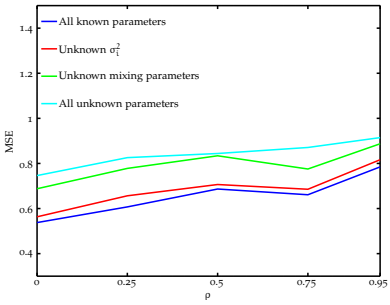


(c) x_1 for $H_1 = 0.5$, $H_2 = 0.5$.

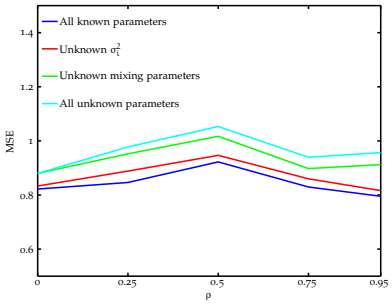


(d) x_2 for $H_1 = 0.5$, $H_2 = 0.5$.

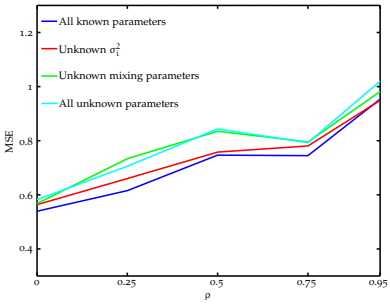
Figure 47: Filtering MSE for latent fGp and OfGp, $H_1 = 0.5$, $H_2 = 0.5$.



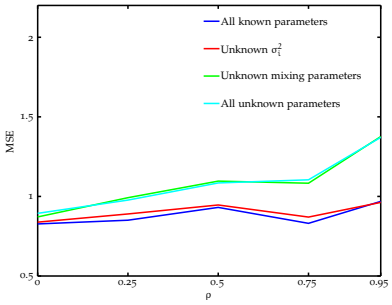
(a) u_1 for $H_1 = 0.7$, $H_2 = 0.7$.



(b) u_2 for $H_1 = 0.7$, $H_2 = 0.7$.

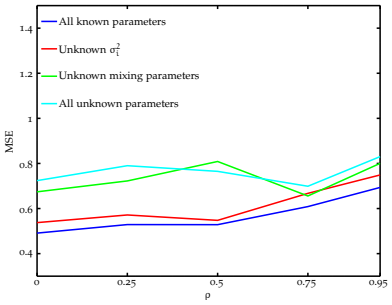


(c) x_1 for $H_1 = 0.7$, $H_2 = 0.7$.

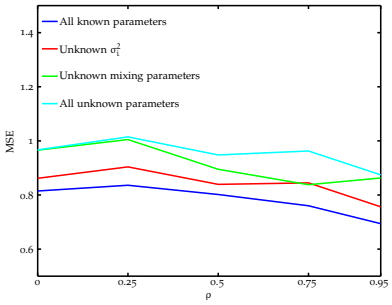


(d) x_2 for $H_1 = 0.7$, $H_2 = 0.7$.

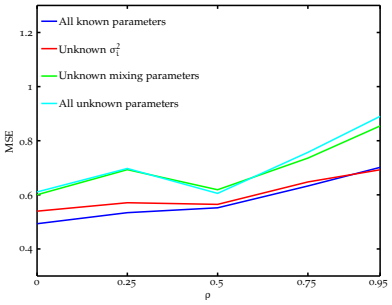
Figure 48: Filtering MSE for latent fGp and OfGp, $H_1 = 0.7$, $H_2 = 0.7$.



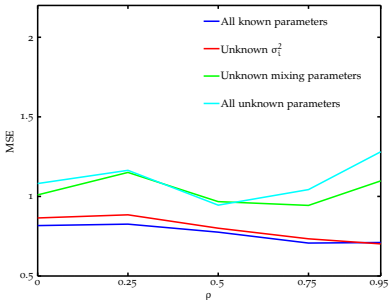
(a) u_1 for $H_1 = 0.8$, $H_2 = 0.8$.



(b) u_2 for $H_1 = 0.8$, $H_2 = 0.8$.

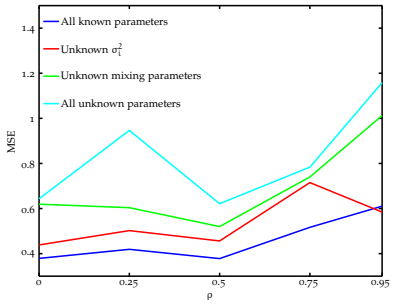


(c) x_1 for $H_1 = 0.8$, $H_2 = 0.8$.

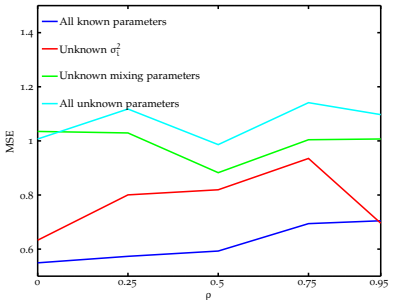


(d) x_2 for $H_1 = 0.8$, $H_2 = 0.8$.

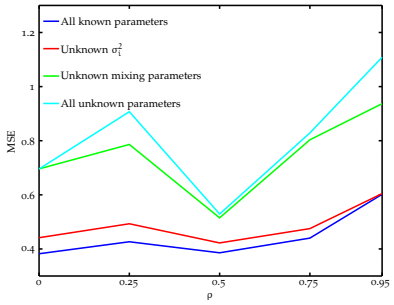
Figure 49: Filtering MSE for latent fGp and OfGp, $H_1 = 0.8$, $H_2 = 0.8$.



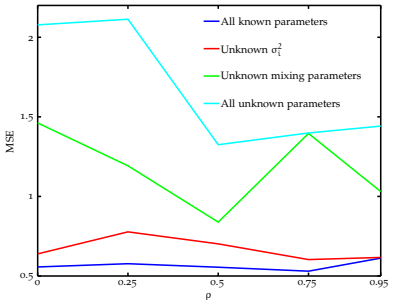
(a) u_1 for $H_1 = 0.9$, $H_2 = 0.9$.



(b) u_2 for $H_1 = 0.9$, $H_2 = 0.9$.



(c) x_1 for $H_1 = 0.9$, $H_2 = 0.9$.



(d) x_2 for $H_1 = 0.9$, $H_2 = 0.9$.

Figure 50: Filtering MSE for latent fGp and OfGp, $H_1 = 0.9$, $H_2 = 0.9$.

9

LATENT TIME-SERIES UNDER MODEL UNCERTAINTY

So far in this work, we have relied on the assumption that the underlying data dynamics follow a given model: ARMA models in Chapters 3 and 4, fGp in Chapter 5 and more generic time-series in Chapters 6 and 7.

In practice however, it is difficult (if not impossible) to know the true underlying model. In this chapter, we elaborate on SMC methods for inference and prediction of latent time-series under model uncertainty.

In Section 9.1, we first study a model selection approach and later, propose an alternative that fuses the information from the considered models within the SMC method itself. We conclude with Section 9.2, where we evaluate the performance of the proposed methods.

LATENT TIME-SERIES AND MODEL UNCERTAINTY

We hereby aim at devising a method for filtering and prediction of time-varying signals under model uncertainty. That is, we study the case where a practitioner is uncertain about the specifics of the hidden dynamics and instead, considers a set of candidate models $\mathcal{M}_k$, $k = 1, 2, \cdots, K$.

We cast the problem in state-space form and make the model uncertainty explicit. We write

$$\begin{cases} x_t = g_k(x_{t-1}, \theta_g, u_t), \\ y_t = h_k(x_t, \theta_h, v_t), \end{cases} \tag{181}$$

where $t \in \mathbb{N}_0$ represents time, $x_t \in \mathbb{R}^{d_x}$ is a time-varying hidden process whose meaning is *not* model dependent, and $y_t \in \mathbb{R}^{d_y}$ is an observation process (of dimensions d_x and d_y, respectively).

The subscript k of the functions g and h refers to each candidate model $\mathcal{M}_k$. The symbols u_t and v_t are the innovation processes of the latent state and the observation process, which we assume to be independent of each other. We refer to the

parameters of the model dependent state and space functions as θ_g and θ_h, respectively.

The goal is to infer the evolution of the latent state, as new observations are acquired, when one is not certain about the model that best describes the studied system.

To deal with model uncertainty, model selection criteria have been extensively used and popularized. For general model selection problems, the most prominent ones are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [29].

Within the Monte Carlo methodology, joint Bayesian parameter estimation and model selection have been already extensively studied [4–6, 53, 67]. However, these schemes rely on Markov Chain Monte Carlo (MCMC) or population Monte Carlo sampling techniques. Consequently, they are not designed for real-time processing.

If we shift the attention to SMC methods, an early study of PFs and model selection has been provided in [36]. Most of the studies are based on considering a predefined set of candidate models or dynamically exploring potential models [62]. As an alternative, some researchers have adopted Bayesian model averaging for SMC methods. In [84], computation of the weights is performed statically over a training data set, while in [70], Markov chains are used for exploring model probabilities.

For inference of latent time-series under model uncertainty, we consider both alternatives and thus, study (1) a set of parallel SMC methods (i.e., bank of PFs) with model selection and (2) model averaging for fusing information of all considered models.

When studying time-series in practice, the ultimate goal is to provide accurate and meaningful predictions of the observations, as these are the only variables that one can actually observe in real-life. Therefore, the predictive performance of any method is critical.

In this work, the metric for evaluating the predictive performance of our method is the predictive likelihood of the observations, that is, $f(y_{t+1}|y_{1:t})$. One can derive such density by means of Bayesian principles and compute

$$f(y_{t+1}|y_{1:t}, \mathcal{M}_k) = \int f(y_{t+1}|x_{t+1}, \mathcal{M}_k) f(x_{t+1}|y_{1:t}, \mathcal{M}_k) dx_{t+1}. \quad (182)$$

In general, obtaining the closed form solution to Equation 182 is troublesome, mostly because of the nonlinearities and non-Gaussianities considered.

However, we propose to take advantage of the random measure provided by the SMC method and thus, numerically solve the otherwise analytically non-solvable integral above.

An SMC method allows for an approximate computation of such density by using a random measure to approximate the predictive density $f(x_{t+1}|y_{1:t}, \mathcal{M}_k)$ in the integral in Equation 182. This can be obtained by drawing J state prediction samples from the transition density of the considered model, i.e.,

$$f(x_{t+1}|x_{1:t}, y_{1:t}, \mathcal{M}_k) \approx f^J_{t+1}(x_{t+1}), \quad \text{where } x^{(j)}_{t+1} \sim f(x_{t+1}|x_{1:t}, y_{1:t}, \mathcal{M}_k),\ j = 1, 2, \cdots, J. \tag{183}$$

As a result, one can numerically obtain

$$f(y_{t+1}|y_{1:t}, \mathcal{M}_k) \approx \frac{1}{J} \sum_{j=1}^{J} f(y_{t+1}|x^{(j)}_{t+1}). \tag{184}$$

How to leverage such density for improved inference and prediction of latent time-series under model uncertainty is the subject of the next subsections.

Model selection: picking the best candidate

First, we explore a solution that consists of a bank of K parallel SMC filters. That is, a PF is run for each of the candidate models $\mathcal{M}_k,\ k = 1, \cdots, K$, followed by a model selection scheme.

On the one hand, each of the SMC filters must proceed according to its assumed model. That is, the practitioner sets up the kth PF in line with the underlying model assumptions, which is run independently from other PFs ($j \neq k$) in the bank.

On the other, for the model selection criteria, we consider the predictive likelihood of the observations for each of the considered models $\mathcal{M}_k$, i.e., $f(y_{t+1}|y_{1:t}, \mathcal{M}_k)$.

Within an SMC method, we approximate the predictive likelihood of the next observation with Equation 183, which is computed per state sample $x^{(m)}_t$, that is,

$$f(y_{t+1}|y_{1:t}, \mathcal{M}_k) \approx \sum_{m=1}^{M} \frac{w^{(m)}_t}{J} \sum_{j=1}^{J} f(y_{t+1}|x^{(m,j)}_{t+1}). \tag{185}$$

Based on this metric, the best model $\mathcal{M}_k^*$ that describes the observed sequence $y_{1:t+1}$ is selected by

$$\mathcal{M}_k^* = \underset{\mathcal{M}_k}{\operatorname{argmax}} \sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, \mathcal{M}_k). \tag{186}$$

In summary, we run a bank of PFs based on a set of candidate models $\mathcal{M}_k$, $k = 1, \cdots, K$, followed by selecting the best model based on Equation 186, as detailed in Table 33.

PF WITH MODEL UNCERTAINTY: MODEL SELECTION

1. At time instant t, consider the random measures

$$f^{M}_{\mathcal{M}_k}(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right).$$

for models $\mathcal{M}_k,\ k = 1, \cdots, K$.

2. Upon reception of a new observation at time $t+1$.

3. Propagate J predictive samples $x_{t+1}^{(m,j)}$ from the available random measure and compute

$$f(y_{t+1}|y_{1:t}, \mathcal{M}_k) \approx \sum_{m=1}^{M} \frac{w_t^{(m)}}{J} \sum_{j=1}^{J} f(y_{t+1}|x_{t+1}^{(m,j)}).$$

4. Pick model $\mathcal{M}_k^*$ that describes the observed data best by

$$\mathcal{M}_k^* = \underset{\mathcal{M}_k}{\operatorname{argmax}} \sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, \mathcal{M}_k).$$

5. Perform resampling for each model, if necessary

$$\overline{x}_t^{(m)} \sim f^{M}_{\mathcal{M}_k}(x_t), \quad m = 1, \cdots, M.$$

6. Propagate the state particles for each model $\mathcal{M}_k$

$$x_{t+1}^{(m)} \sim \pi(x_{t+1}|\overline{x}_t^{(m)}, y_{1:t}, \mathcal{M}_k).$$

7. Compute the non-normalized weights of the drawn particles of each model $\mathcal{M}_k$

$$\widetilde{w}_{t+1}^{(m)} \propto \frac{f\left(y_{t+1}|x_{t+1}^{(m)}, \mathcal{M}_k\right) f\left(x_{t+1}^{(m)}|\overline{x}_t^{(m)}, \mathcal{M}_k\right)}{\pi\left(x_{t+1}^{(m)}|\overline{x}_t^{(m)}, y_{1:t}, \mathcal{M}_k\right)} w_t^{(m)},$$

and normalize them to obtain a new set of random measures

$$f^{M}_{\mathcal{M}_k}(x_{t+1}) = \sum_{m=1}^{M} w_{t+1}^{(m)} \delta\left(x_{t+1} - x_{t+1}^{(m)}\right).$$

Table 33: SMC method for inference of latent time-series under model uncertainty: model selection.

Model averaging: fusing the candidates

As an alternative to plain model selection, we now consider Bayesian model averaging. That is, we sequentially fuse information from different models in agreement with their evolving Bayesian posteriors. We achieve this autonomously within the SMC method by adjusting the weighting and resampling steps.

The aim is to propose an SMC method that systematically fuses information from candidate models, whose resources are dynamically allocated according to their predictive performance.

We start by considering the sequential update of the posterior belief of each model, as more data are observed. That is, for the posterior of each model $\mathcal{M}_k$ we write

$$\begin{aligned} p(\mathcal{M}_k|y_{1:t}) &\propto f(y_t|y_{1:t-1}, \mathcal{M}_k) p(\mathcal{M}_k|y_{1:t-1}) \\ &\propto \prod_{i=0}^{t-1} f(y_{i+1}|y_{1:i}, \mathcal{M}_k) p(\mathcal{M}_k|y_{1:i}), \end{aligned} \tag{187}$$

where $p(\mathcal{M}_k|y_{1:0}) = p(\mathcal{M}_k)$ is the prior assumed for each model.

In Section 9.1.1, we suggested to use this posterior to perform model selection. On the contrary, we now exploit the total probability theorem to sequentially fuse the information from all the models.

Given a set of densities $f(x_t|y_{1:t}, \mathcal{M}_k)$ and $f(y_{t+1}|y_{1:t}, \mathcal{M}_k)$ and a set of model posterior probabilities $p(\mathcal{M}_k|y_{1:t})$, the total probability theorem states that

$$\begin{aligned} f(x_t|y_{1:t}) &= \sum_{k=1}^{K} f(x_t|y_{1:t}, \mathcal{M}_k) p(\mathcal{M}_k|y_{1:t}), \\ f(y_{t+1}|y_{1:t}) &= \sum_{k=1}^{K} f(y_{t+1}|y_{1:t}, \mathcal{M}_k) p(\mathcal{M}_k|y_{1:t}). \end{aligned} \tag{188}$$

This allows for direct computation of both the filtering, $f(x_t|y_{1:t})$, and predictive, $f(y_{t+1}|y_{1:t})$, densities of interest, where the specific models are marginalized.

Nonetheless, the densities in Equation 188 can not be obtained in closed form in the problem of interest here, i.e., inference of latent time-series with nonlinear observations. Thus, one must study how to apply the above densities within an SMC sampling method.

Model averaging and SMC: dynamically fusing information

The goal here is to devise an SMC method that provides an overall random measure $f_t^M(x_t)$ approximation to the density of interest

$$f(x_t|y_{1:t}) \approx f_t^M(x_t) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right), \tag{189}$$

where a set of K candidate models are considered $\mathcal{M}_k$, $k = 1, \cdots, K$.

To overcome the underlying model uncertainty, we fuse the information from the different posteriors as suggested in Equation 188.

Within the SMC methodology, each model dependent density is approximated by a (model dependent) random measure

$$f(x_t|y_{1:t}, \mathcal{M}_k) \approx f_{\mathcal{M}_{k,t}}^{M_{k,t}}(x_t) = \sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} \delta\left(x_t - x_t^{(m_k)}\right), \tag{190}$$

with $m_k = 1, \cdots, M_{k,t}$; where the subscript t indicates that the number of particles used by model $\mathcal{M}_k$ may vary with time. Note that the model weights $w_t^{(m_k)}$ must, for all time instants t, satisfy $\sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} = 1$.

We require that the overall approximation of $f(x_t|y_{1:t})$ in Equation 189 is obtained from the random measures of all the models, as in Equation 190. Mathematically,

$$\begin{aligned} f(x_t|y_{1:t}) &\approx \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right) \\ &= \sum_{k=1}^{K} \left[p(\mathcal{M}_k|y_{1:t}) \left(\sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} \delta\left(x_t - x_t^{(m_k)}\right) \right) \right] \\ &= \sum_{k=1}^{K} \left(\sum_{m_k=1}^{M_{k,t}} p(\mathcal{M}_k|y_{1:t}) w_t^{(m_k)} \delta\left(x_t - x_t^{(m_k)}\right) \right), \end{aligned} \tag{191}$$

which is fulfilled if

- the total number of samples adds up, i.e.,

$$M = \sum_{k=1}^{K} M_{k,t}\ ; \tag{192}$$

- the weights are properly combined, i.e.,

$$w_t^{(m)} = p(\mathcal{M}_k|y_{1:t})w_t^{(m_k)}, \tag{193}$$

with $m = \sum_{l=1}^{k-1} M_{l,t} + m_k,\ k = 1, \cdots, K$.

As explained in Section 2.4, for the SMC to perform satisfactorily, one needs to resample particles according to their weights. After resampling, one obtains an equally weighted resampled random measure

$$f(x_t|y_{1:t}) \approx \sum_{m=1}^{M} w_t^{(m)} \delta\left(x_t - x_t^{(m)}\right) \approx \sum_{m=1}^{M} \frac{1}{M} \delta\left(x_t - \overline{x}_t^{(m)}\right). \tag{194}$$

To guarantee convergence and unbiasedness of an SMC method [43], (a) the expected number of resampled particles must be proportional to their weights and (b) the total number of particles must remain constant (although this can be relaxed).

We draw an equivalence between particles (models) and their weights (model posterior probabilities) and establish the following *model resampling* equivalence:

- Resampling from the overall random measure $f_t^M(x_t)$ with its weights $w_t^{(m)} = p(\mathcal{M}_k|y_{1:t})w_t^{(m_k)}$, is equivalent to
- assigning $M_{k,t} = Mp(\mathcal{M}_k|y_{1:t})$ samples to each model $\mathcal{M}_k$ followed by resampling particles from each model random measure $f_{\mathcal{M}_{k,t}}^{M_{k,t}}(x_t)$ based on their weights $w_t^{(m_k)}$, $m_k = 1, \cdots, M_{k,t}$.

The above scheme guarantees that the overall resampled random measure is equivalent to the mixture of per-model resampled random measures, i.e.,

$$\begin{aligned} f(x_t|y_{1:t}) &\approx \sum_{m=1}^{M} \frac{1}{M}\delta\left(x_t - \overline{x}_t^{(m)}\right) \\ &= \sum_{k=1}^{K}\left[p(\mathcal{M}_k|y_{1:t})\left(\sum_{m_k=1}^{M_{k,t}} \frac{1}{M_{k,t}}\delta\left(x_t - \overline{x}_t^{(m_k)}\right)\right)\right] \\ &= \sum_{k=1}^{K}\left(\sum_{m_k=1}^{M_{k,t}} \frac{p(\mathcal{M}_k|y_{1:t})}{M_{k,t}}\delta\left(x_t - \overline{x}_t^{(m_k)}\right)\right), \end{aligned} \tag{195}$$

only if

$$\begin{cases} M = \sum_{k=1}^{K} M_{k,t}, \\ \frac{1}{M} = \frac{p(\mathcal{M}_k|y_{1:t})}{M_{k,t}}, \rightarrow M_{k,t} = Mp(\mathcal{M}_k|y_{1:t}). \end{cases}$$

Thus, the weight of each resampled particle per model is $\frac{p(\mathcal{M}_k|y_{1:t})}{M_{k,t}}$, which is equal for all the particles within each model, but requires $M_{k,t} = Mp(\mathcal{M}_k|y_{1:t})$ to have equally weighted overall particles.

In the above equations, to avoid unnecessary diversion, we assumed that the particle allocations result in integers. This, of course, is not the case in general. There are, however, well understood approaches for handling this [69].

Furthermore, it might be of interest to avoid complete depletion of models. That is, it might be undesirable that no resources are assigned to a particular model. The straightforward extension for such cases consists on determining the number of particles to assign based on

$$M_{k,t+1} = M_{k_{min}} + \left(M - K \cdot M_{k_{min}}\right) p(\mathcal{M}_k|y_{1:t+1}), \tag{196}$$

where $M_{k_{min}}$ is the minimum number of samples assigned to each model and $\sum_{k=1}^{K} M_{k,t+1} = M$ is fulfilled.

The key for the presented fusion strategy as in Equation 195 to be applicable is to compute the model posterior probability at each time instant, i.e., $p(\mathcal{M}_k|y_{1:t})$. Such probability can be sequentially updated by means of Equation 182 in general, and Equation 184 within the SMC methodology.

One can evaluate the predictive performance of each model via a two step procedure. First, one propagates the random measure of each model in time, i.e., from $f_{\mathcal{M}_{k,t}}^{M_{k,t}}(x_t)$ to $f_{\mathcal{M}_{k,t+1}}^{M_{k,t+1}}(x_{t+1})$.

Then, one evaluates the resulting predictive density by using the next available observation y_{t+1}.

All in all, the model posterior probability computation follows

$$\begin{aligned} p(\mathcal{M}_k|y_{1:t+1}) &\propto f(y_{t+1}|y_{1:t}, \mathcal{M}_k) p(\mathcal{M}_k|y_{1:t}) \\ &\approx p(\mathcal{M}_k|y_{1:t}) \sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} f(y_{t+1}|x_{t+1}^{(m_k)}, \mathcal{M}_k). \end{aligned} \tag{197}$$

Note that a normalization step is required so that we have a proper probability mass function, i.e., $\sum_{k=1}^{K} p(\mathcal{M}_k|y_{1:t+1}) = 1$.

Everything considered, we propose a dynamic SMC method for sequential processing of data under model uncertainty in Table 34. The proposed method is dynamic in that (a) it updates model posterior probabilities as more data become available and (b) it apportions SMC resources (i.e., particles) according to such probabilities amongst the candidate models.

Note that the estimates of interest are computed at every time instant t before resampling, via

$$\mathbb{E}\{g(x_t)\} = \sum_{k=1}^{K} p(\mathcal{M}_k|y_{1:t}) \sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} g\left(x_t^{(m_k)}\right). \tag{198}$$

PF WITH MODEL UNCERTAINTY: MODEL AVERAGING

1. At time instant t, consider the random measures

$$f_{\mathcal{M}_{k,t}}^{M_{k,t}}(x_t) = \sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} \delta\left(x_t - x_t^{(m_k)}\right).$$

   for models $\mathcal{M}_k$, $k = 1, \cdots, K$.

2. Upon reception of a new observation at time $t+1$.

3. Compute the model posterior probability

$$p(\mathcal{M}_k|y_{1:t+1}) \propto p(\mathcal{M}_k|y_{1:t}) \sum_{m_k=1}^{M_{k,t}} w_t^{(m_k)} f(y_{t+1}|x_{t+1}^{(m_k)}, \mathcal{M}_k).$$

4. Determine the number of particles $M_{k,t+1}$ to be assigned to each model $\mathcal{M}_k$

$$M_{k,t+1} = M_{k_{min}} + \left(M - K \cdot M_{k_{min}}\right) p(\mathcal{M}_k|y_{1:t+1}),$$

   where $M_{k_{min}}$ is the minimum number of samples assigned to each model.

5. Perform resampling $M_{k,t+1}$ times for each model.

6. Propagate the state particles for each model $\mathcal{M}_k$

$$x_{t+1}^{(m_k)} \sim \pi(x_{t+1}|\bar{x}_t^{(m_k)}, y_{1:t}, \mathcal{M}_k).$$

7. Compute the non-normalized weights of the drawn particles of each model $\mathcal{M}_k$

$$\widetilde{w}_{t+1}^{(m_k)} \propto \frac{f\left(y_{t+1}|x_{t+1}^{(m_k)}, \mathcal{M}_k\right) f\left(x_{t+1}^{(m_k)}|\bar{x}_t^{(m_k)}, \mathcal{M}_k\right)}{\pi\left(x_{t+1}^{(m_k)}|\bar{x}_t^{(m_k)}, y_{1:t}, \mathcal{M}_k\right)} w_t^{(m_k)},$$

   and normalize them to obtain a new set of random measures

$$f_{\mathcal{M}_{k,t+1}}^{M_{k,t+1}}(x_t + 1) = \sum_{m_k=1}^{M_{k,t+1}} w_t^{(m_k)} \delta\left(x_{t+1} - x_{t+1}^{(m_k)}\right).$$

Table 34: SMC method for inference of latent time-series under model uncertainty: model averaging.

EVALUATION

To illustrate the applicability of the proposed SMC method for inference of latent time-series under model uncertainty, we return to the latent long-memory processes as in Chapter 5. There, we propose a PF for inference of latent fGps, where only the known Hurst parameter case is studied.

It is of no surprise that, in practice, assuming knowledge of the memory properties of the hidden process (i.e., parameterized by H for a fGp) is unrealistic. As explained in Section 5.3, estimation of the unknown autocovariance function of the latent state is challenging, and the estimation of the Hurst parameter H is even more so.

On the one hand, because in our problem formulation as in Equation 181, we do not observe the self-similar process directly. On the other, because the approaches proposed in the literature for estimation of unknown parameters within the SMC method may (a) break the self-similarity and stationarity properties of the fGp, and (b) hinder the convergence of the SMC method. Thus, new alternatives must be explored.

As a matter of fact, we now cast the inference of latent fGps with unknown H as a model uncertainty problem, so that we can apply the methods devised in Section 9.1. That is, we consider a set of candidate models for the latent state, each with a particular autocovariance function (parameterized by H).

Consequently, by following the explanation in Section 9.1.1, we propose to run a bank of K SMC filters, each with different fGp model assumptions. In particular, the differences are determined by the specific values of the autocovariance function, which in turn, only depend on the Hurst parameters. All in all, each model $\mathcal{M}_k$ assumes a different H parameter value, i.e., $H_k \in [0.5, 1)$, $k = 1, 2, \cdots, K$.

Each of the k SMC filters proceeds as corresponds to its assumed model (as described in Table 21 for a given H_k); followed by a model selection scheme. For the model selection criteria, we consider the predictive likelihood of the observations as in Equation 186, which for the latent fGp with unknown H results in

$$H^* = \underset{H_k}{\operatorname{argmax}} \sum_{i=1}^{t} \log f(y_{i+1} | y_{1:i}, H_k). \tag{199}$$

We are now ready to implement a bank of SMC filters as described in Table 33 with different Hurst parameter assumptions,

SMC bank	True H					
K=6	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
$H_1 = 0.5$	0.75585	0.7384	0.7414	0.76453	0.7468	0.70245
$H_2 = 0.6$	0.76235	0.73161	0.71785	0.71312	0.66156	0.59087
$H_3 = 0.7$	0.78312	0.73957	0.70206	0.67159	0.57899	0.47794
$H_4 = 0.8$	0.82449	0.76833	0.70837	0.65323	0.52737	0.39924
$H_5 = 0.9$	0.9034	0.83263	0.74516	0.67107	0.50654	0.35076
$H_6 = 0.95$	0.98512	0.90572	0.80342	0.71992	0.51913	0.33772

Table 35: MSE of the state for a bank of SMC filters.

where the transition density follows Equation 118 and we select the best model as in Equation 199.

We evaluate such an SMC approach over the SV model with latent fGp,

$$\begin{cases} x_t = u_t, \\ y_t = e^{x_t/2} v_t, \end{cases} \tag{200}$$

where $v_t \sim \mathcal{N}(0, \sigma_v^2)$ and u_t is a zero-mean fGp with unknown H.

The filtering results for this experiment are summarized in Table 35, where it is clear that the most accurate inference is attained when the correct parameter value is assumed (dark shaded entries in the diagonal of the table).

For rapidly decaying autocovariance functions (i.e., $0.5 \leqslant H < 0.75$), there is a minimal MSE performance difference among the SMC filters that assume different H values in this interval. On the contrary, as the long-memory becomes more evident ($H \geqslant 0.75$), only the filters with values close to the true one provide good accuracy.

One concludes that, when there is uncertainty about the underlying model, the relevance of the memory properties is critical. That is, assuming short-memory for a long-memory process results in poor inference and vice versa.

We run a bank of SMC schemes in parallel with a model selection scheme and thus, it is of interest to evaluate whether the method identifies the correct model (in this case, the correct H value).

First, we plot the evolution of the model selection criteria, i.e., the metric in Equation 199. See, for latent fGps with different Hurst parameters, Figure 51, Figure 52, Figure 53, Figure 54, Figure 55 and Figure 56.

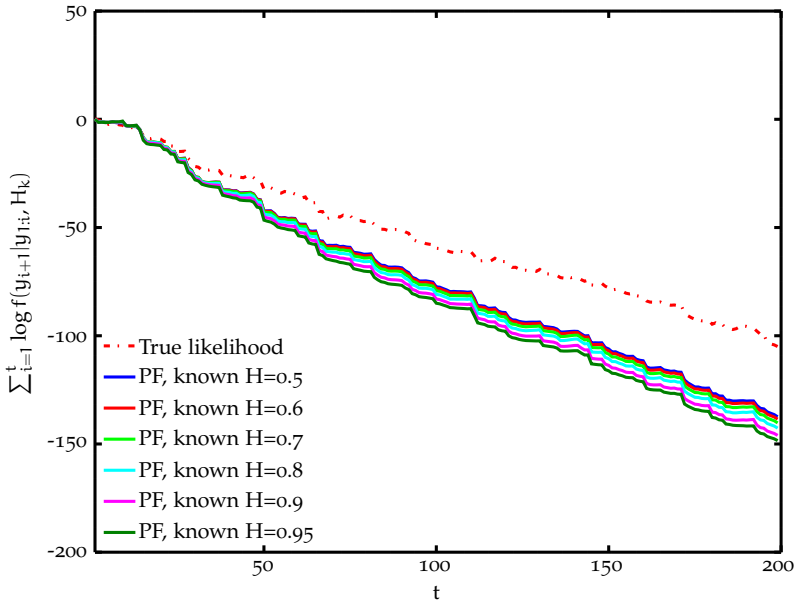


Figure 51: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.5$.

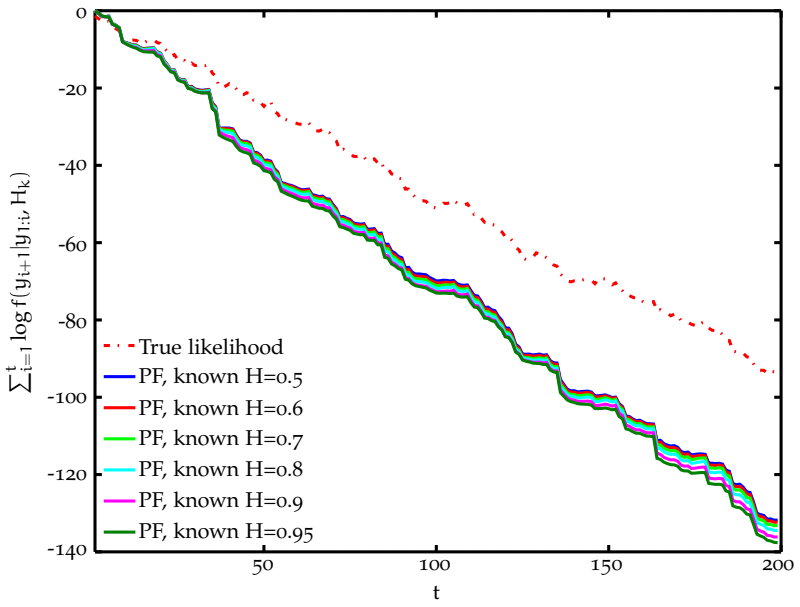


Figure 52: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.6$.

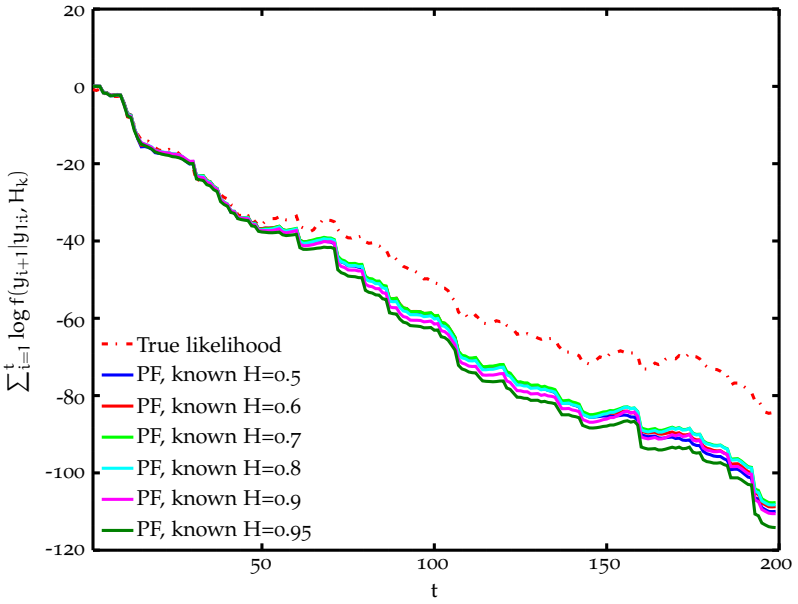


Figure 53: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.7$.

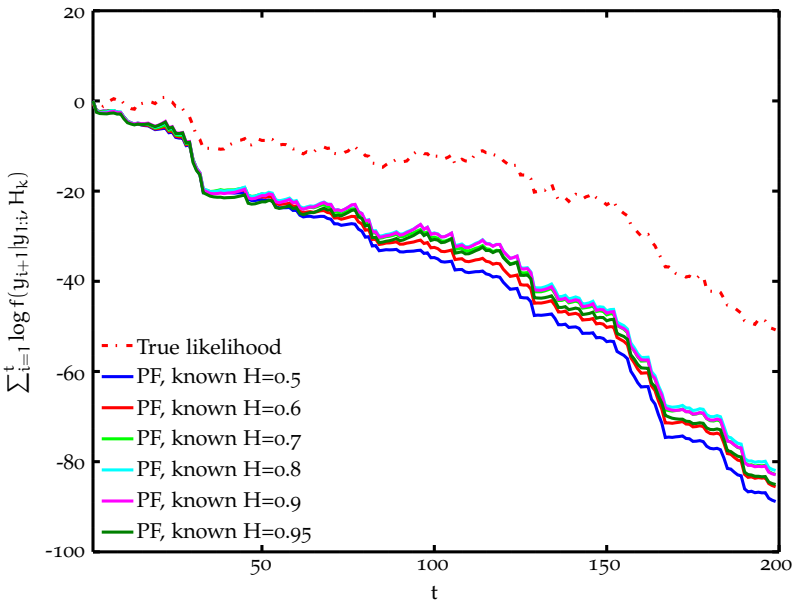


Figure 54: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.8$.

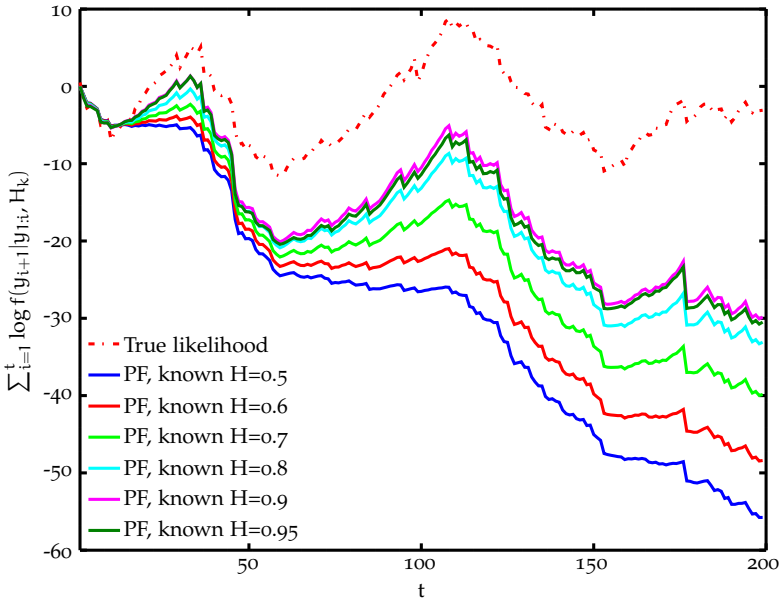


Figure 55: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.9$.

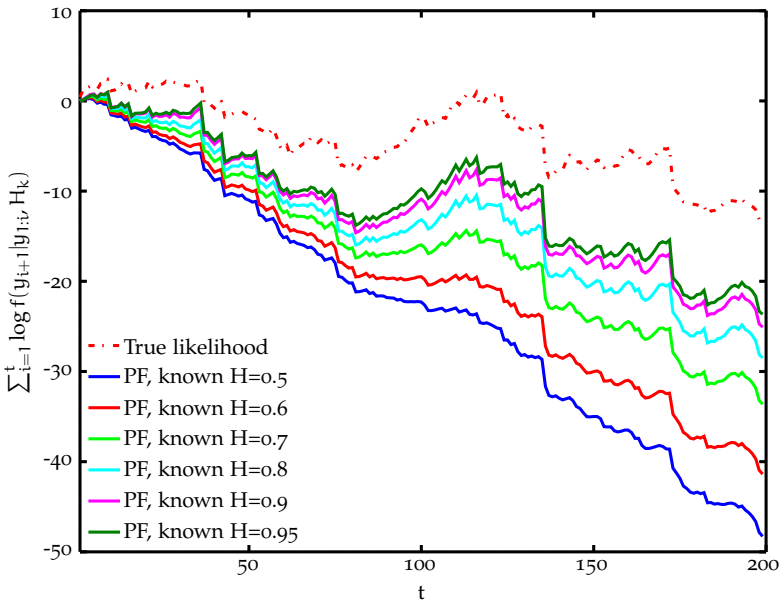


Figure 56: Cumulative log-likelihood $\sum_{i=1}^{t} \log f(y_{i+1}|y_{1:i}, H_k)$ of each SMC filter over time for fGp with $H = 0.95$.

When evaluating the likelihood of the observed time-series for different models, we conclude that only those with assumed H_k parameter close to the correct one are able to provide the best results. Besides, one observes how the difference between models becomes more evident as more data are observed. Once again, distinguishing amongst short-memory processes (i.e., $0.5 \leqslant H < 0.75$) is a much harder task than separating short- and long-memory fGps.

We can equivalently summarize the accuracy of the proposed model selection scheme by looking at the confusion matrices in Table 36 and Table 37.

We emphasize the difficulty of distinguishing the processes with rapidly decaying autocovariance functions (i.e., $0.5 \leqslant H < 0.75$). By contrast, for long-memory processes, the long-range dependence induced as $H \to 1$ helps the SMC method to identify the correct model. In other words, only when capturing the influence of the samples deep in the past on the present values, the SMC method is able to accurately predict the next observations.

Finally, the results presented in the tables in the next page (i.e., Table 36 and Table 37) confirm that the model selection accuracy improves as (1) more data are observed (i.e., as t grows) and (2) more predictive resources are used (i.e., $J > 1$). In other words, the more information we have (i.e., more data), the better the model selection is.

We elaborate on this idea by extending the model in Equation 200 to allow for multiple latent fGps to impact a set of observations. That is, we consider a general model with N_x latent processes and N_y observed time-series, where each have idiosyncratic characteristics.

More precisely, we write

$$\begin{cases} x_{1,t} = u_{1,t}, \\ x_{2,t} = u_{2,t}, \\ y_{n,t} = a_{n,0} + a_{n,1} y_{n,t-1} + x_{1,t} + e^{x_{2,t}/2} v_{n,t}, \\ \qquad n = 1, \cdots, N_y, \end{cases} \tag{201}$$

where the observed time-series have both an idiosyncratic structure (a bias term $a_{n,0}$ and an auto-regressive term $a_{n,1}$) and a common shared time-varying trend and volatility components ($x_{1,t}$ and $x_{2,t}$, respectively). The noise $v_{n,t}$ has the standard Gaussian density and $u_{i,t}$ represents a fGp.

SMC bank	True H					
K=6	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
$H_1 = 0.5$	35	47	35	30	24	19
$H_2 = 0.6$	13	18	14	13	13	6
$H_3 = 0.7$	9	9	5	13	9	8
$H_4 = 0.8$	7	5	9	7	7	6
$H_5 = 0.9$	9	5	10	9	10	13
$H_6 = 0.95$	27	16	27	28	37	48

SMC bank	True H					
K=6	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
$H_1 = 0.5$	65	42	17	3	0	2
$H_2 = 0.6$	20	28	19	15	2	2
$H_3 = 0.7$	8	22	31	18	4	2
$H_4 = 0.8$	4	4	20	42	23	7
$H_5 = 0.9$	2	3	8	14	36	18
$H_6 = 0.95$	1	1	5	8	35	69

Table 36: Model selection confusion matrix evolution over time: at $t = 10$ (top) and $t = 200$ (bottom), with $J = 1$.

SMC bank	True H					
K=6	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
$H_1 = 0.5$	40	53	44	36	28	21
$H_2 = 0.6$	13	15	5	4	10	3
$H_3 = 0.7$	5	6	6	13	8	8
$H_4 = 0.8$	9	5	11	9	5	8
$H_5 = 0.9$	9	4	5	5	9	11
$H_6 = 0.95$	24	17	29	33	40	49

SMC bank	True H					
K=6	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
$H_1 = 0.5$	71	42	16	2	1	3
$H_2 = 0.6$	11	29	17	11	2	1
$H_3 = 0.7$	9	19	34	29	1	2
$H_4 = 0.8$	2	6	21	39	21	5
$H_5 = 0.9$	3	3	10	12	38	17
$H_6 = 0.95$	4	1	2	7	37	72

Table 37: Model selection confusion matrix evolution over time: at $t = 10$ (top) and $t = 200$ (bottom), with $J = 100$.

The model in Equation 201 is an illustrative practical application with interest in finance. There, each $y_{n,t}$ would describe the price evolution of an asset (described by the CAPM model [83]) within the same market dynamics. This is captured by the trend of the market $x_{1,t}$ and its log-volatility $x_{2,t}$ [95]. The properties of the market are captured by the latent fGps [31]. Recall that, even if the described model contains only two shared variables $x_{1,t}$ and $x_{2,t}$, it can readily be generalized to a more general case (see [99]).

When the idiosyncratic parameters are known, the derivation of the observations likelihood is straightforward, i.e.,

$$\begin{aligned} f\left(y_{1:N_y,t+1}|x_{1:2,t+1}\right) &= \prod_{n=1}^{N_y} \mathcal{N}\left(y_{n,t+1}|\mu_{y_{n,t+1}}, \sigma^2_{y_{n,t+1}}\right), \\ \text{with } &\begin{cases} \mu_{y_{n,t+1}} = a_{n,0} + a_{n,1} y_{n,t} + x_{1,t+1}, \\ \sigma^2_{y_{n,t+1}} = e^{x_{2,t+1}}. \end{cases} \end{aligned} \tag{202}$$

Assuming that the distinctive parameters are known is impractical in real-problems and thus, we resort to a Bayesian analysis of the likelihood function.

Let us define the vector of unknown parameters per observation equation as $a_n = \begin{pmatrix} a_{n,0} & a_{n,1} \end{pmatrix}^\top$ and the auxiliary vector $h_{n,t} = \begin{pmatrix} 1 & y_{n,t-1} \end{pmatrix}^\top$. Recall that for the proposed SMC method, the Bayesian parameter posterior and the likelihood function evolve with the time-index t and depend, not only on the observations n, but also on each of the m particle streams.

Thus, at time instant t, we use the following parameter prior:

$$f(a^{(m)}_{n,t}|y_{n,1:t}, x^{(m)}_{1:2,1:t}) = \mathcal{N}\left(a^{(m)}_{n,t} \middle| \mu^{(m)}_{a_{n,t}}, \Sigma^{(m)}_{a_{n,t}}\right). \tag{203}$$

By rewriting the predictive density as

$$f(y_{n,t+1}|y_{n,1:t}, a^{(m)}_{n,t}, x^{(m)}_{1:2,t+1}) = \mathcal{N}\left(y_{n,t+1} \middle| h^{(m)\top}_{n,t+1} a^{(m)}_{n,t} + x^{(m)}_{1,t+1}, \; e^{x^{(m)}_{2,t+1}}\right), \tag{204}$$

we can marginalize out the unknown parameters and obtain

$$\begin{aligned}
&f\left(y_{n,t+1}|y_{n,1:t}, x^{(m)}_{1:2,t+1}\right) = \\
&= \int f(y_{n,t+1}|y_{n,1:t}, a^{(m)}_{n,t}, x^{(m)}_{1:2,t+1}) f(a^{(m)}_{n,t}|y_{n,1:t}, x^{(m)}_{1:2,1:t}) da^{(m)}_{n,t} \\
&= \mathcal{N}\left(y_{n,t+1}|\mu^{(m)}_{y_{n,t+1}}, \sigma^{2^{(m)}}_{y_{n,t+1}}\right), \qquad (205)\\
&\text{with } \begin{cases} \mu^{(m)}_{y_{n,t+1}} = h^{(m)\top}_{n,t+1} \mu^{(m)}_{a_{n,t}} + x^{(m)}_{1,t+1}, \\ \sigma^{2^{(m)}}_{y_{n,t+1}} = e^{x^{(m)}_{2,t+1}} + h^{(m)\top}_{n,t+1} \Sigma^{(m)}_{a_{n,t}} h^{(m)}_{n,t+1}. \end{cases}
\end{aligned}$$

Finally, the joint predictive density is given by

$$f\left(y_{1:N_y,t+1}|y_{1:N_y,t}, x_{1:2,t+1}\right) = \prod_{n=1}^{N_y} \mathcal{N}\left(y_{n,t+1}|\mu_{y_{n,t+1}}, \sigma^2_{y_{n,t+1}}\right).$$

Once the data at $t+1$ have been observed, we update the parameter posterior to

$$f(a^{(m)}_{n,t+1}|y_{n,1:t+1}, x^{(m)}_{1:2,1:t+1}) = \mathcal{N}\left(a^{(m)}_{n,t+1}|\mu^{(m)}_{a_{n,t+1}}, \Sigma^{(m)}_{a_{n,t+1}}\right),$$

with

$$\begin{cases} \mu^{(m)}_{a_{n,t+1}} = \Sigma^{(m)}_{a_{n,t+1}} \left(h^{(m)}_{n,t+1} e^{-x^{(m)}_{2,t+1}} \left(y_{n,t+1} - x^{(m)}_{1,t+1} \right) + \Sigma^{(m)^{-1}}_{a_{n,t}} \mu^{(m)}_{a_{n,t}} \right), \\ \Sigma^{(m)}_{a_{n,t+1}} = \left(\Sigma^{(m)^{-1}}_{a_{n,t}} + h^{(m)}_{n,t+1} e^{-x^{(m)}_{2,t+1}} h^{(m)\top}_{n,t+1} \right)^{-1}. \end{cases}$$

Thus, we can now evaluate the proposed SMC bank in this new model. The accuracy of state tracking is shown in Figure 57, where the MSE results are plotted for different instances of the SMC bank. Results are averaged over 50 realizations of 200 instants long time-series, where $M = 1000$ have been used.

First, we note that the uncertainty introduced by the unknown idiosyncratic parameters is minimal due to the implemented Bayesian estimation.

We show example estimates of the unknown idiosyncratic parameters over time in Figure 58, for a given realization.

More importantly, we emphasize the benefit of observing more data for the accuracy of the SMC bank. That is, as more time-series are observed, the MSE reduces considerably: Accuracy is improved when $N_y = 5$ as opposed to when $N_y = 1$.

As a side note, the tracking accuracy for the trend $x_{1,t}$ is better than for the log-volatility $x_{2,t}$ (which is in line with conclusions for a similar model in Section 8.3).

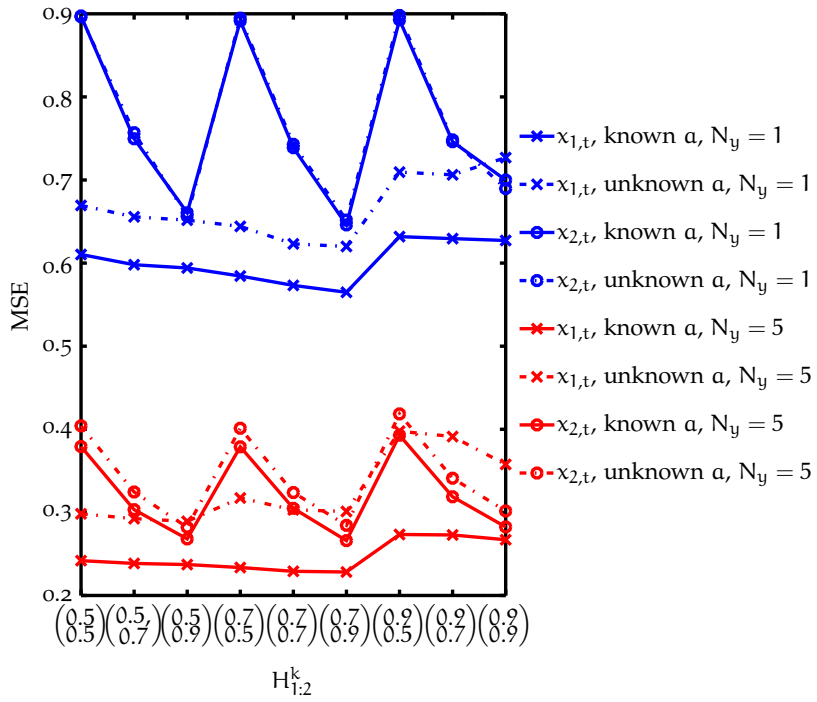
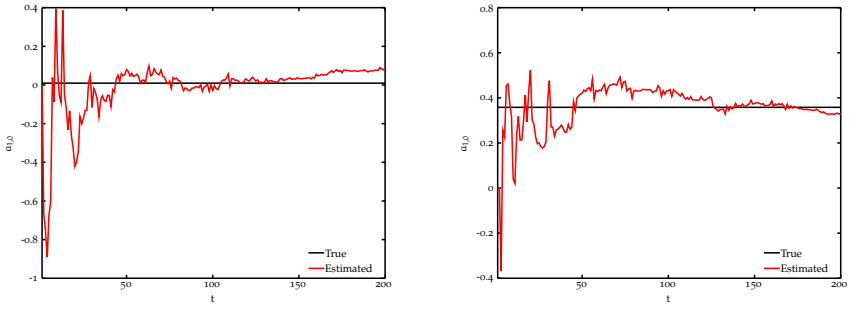


Figure 57: MSE for a bank of SMC filters with true $H_{1:2} = [0.7, 0.9]$.



(a) Estimate of the bias term $a_{1,0}$. (b) Estimate of the AR term $a_{1,1}$.

Figure 58: Idiosyncratic parameter $a_{n,0}$ and $a_{n,1}$ estimation.

Again, we observe that the SMC filter with matching $H_{1:2}$ values is able to provide the overall best performance (i.e., $H^*_{1:2} = [0.7, 0.9]$). When a mismatch occurs, the filters whose $H_{1:2}$ values are closer to the true ones provide the best results.

We now turn our attention to the predictive performance and the model selection accuracy of the bank of SMC filters.

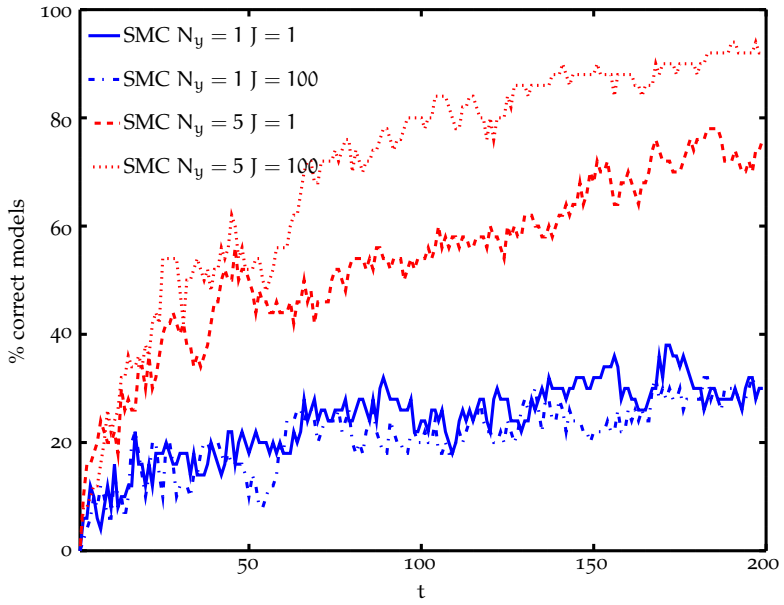


Figure 59: Model selection accuracy over time with true $H_{1:2} = [0.7, 0.9]$.

The results in Figure 59 show that the prediction of future asset returns and thus, the model selection, becomes more accurate as (1) t increases, (2) more time-series are available and (3) oversampling for prediction is implemented ($J > 1$).

The confusion matrices in Table 38 and Table 39, where correct model selection percentages are shown for $N_y = 1$ and $N_y = 5$ respectively, corroborate these statements. Once again, only those SMC filters with matching $H_{1:2}$s predict best and thus, perform best in identifying the true underlying model.

J=1	True H								
Assumed H	H = [0.5, 0.5]	H = [0.5, 0.7]	[0.5, 0.9]	H = [0.7, 0.5]	H = [0.7, 0.7]	H = [0.7, 0.9]	H = [0.9, 0.5]	H = [0.9, 0.7]	H = [0.9, 0.9]
H = [0.5, 0.5]	58 (62)	12 (20)	4 (8)	24 (16)	4 (12)	6 (14)	2 (2)	2 (6)	0 (0)
H = [0.5, 0.7]	24 (16)	24 (20)	8 (12)	10 (16)	12 (4)	14 (4)	0 (0)	6 (10)	2 (0)
H = [0.5, 0.9]	6 (8)	10 (8)	28 (34)	8 (2)	8 (8)	26 (20)	4 (6)	0 (0)	12 (10)
H = [0.7, 0.5]	8 (10)	22 (16)	10 (10)	22 (18)	26 (8)	4 (8)	6 (12)	4 (4)	0 (8)
H = [0.7, 0.7]	4 (4)	14 (16)	12 (6)	18 (28)	14 (26)	10 (10)	8 (4)	14 (10)	8 (8)
H = [0.7, 0.9]	0 (0)	8 (10)	26 (18)	6 (4)	10 (12)	30 (30)	6 (0)	10 (6)	4 (16)
H = [0.9, 0.5]	0 (0)	4 (6)	2 (0)	10 (14)	10 (10)	0 (2)	44 (48)	18 (22)	6 (4)
H = [0.9, 0.7]	0 (0)	4 (2)	2 (2)	0 (2)	14 (14)	2 (0)	26 (22)	32 (32)	16 (14)
H = [0.9, 0.9]	0 (0)	2 (2)	8 (10)	2 (0)	2 (6)	8 (12)	4 (6)	147 (10)	52 (40)

Table 38: Model selection confusion matrix for $N_y = 1$. Results are shown for $J = 1$ and $J = 100$, in parenthesis.

	True H								
Assumed H	$H = [0.5, 0.5]$	$H = [0.5, 0.7]$	$[0.5, 0.9]$	$H = [0.7, 0.5]$	$H = [0.7, 0.7]$	$H = [0.7, 0.9]$	$H = [0.9, 0.5]$	$H = [0.9, 0.7]$	$H = [0.9, 0.9]$
$H = [0.5, 0.5]$	72 (96)	12 (16)	0 (0)	10 (12)	2 (4)	0 (0)	0 (0)	0 (0)	0 (0)
$H = [0.5, 0.7]$	18 (2)	24 (62)	16 (8)	4 (2)	18 (8)	0 (0)	0 (0)	0 (0)	0 (0)
$H = [0.5, 0.9]$	0 (0)	30 (6)	52 (70)	0 (0)	0 (0)	14 (0)	0 (0)	0 (0)	0 (0)
$H = [0.7, 0.5]$	10 (2)	8 (0)	0 (0)	66 (68)	12 (14)	2 (2)	2 (0)	0 (0)	0 (0)
$H = [0.7, 0.7]$	0 (0)	26 (16)	0 (4)	16 (16)	50 (68)	8 (6)	0 (0)	2 (2)	0 (0)
$H = [0.7, 0.9]$	0 (0)	0 (0)	32 (18)	0 (0)	16 (6)	76 (92)	0 (0)	2 (0)	0 (2)
$H = [0.9, 0.5]$	0 (0)	0 (0)	0 (0)	4 (0)	0 (0)	0 (0)	76 (80)	14 (14)	0 (0)
$H = [0.9, 0.7]$	0 (0)	0 (0)	0 (0)	0 (2)	2 (0)	0 (0)	22 (20)	68 (76)	4 (6)
$H = [0.9, 0.9]$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (8)	96 (92)

Table 39: Model selection confusion matrix for $N_y = 5$. Results are shown for $J = 1$ and $J = 100$, in parenthesis.

So far, we have focused on model selection and shown that the proposed SMC method can deal with model uncertainty. However, one needs to be careful on recognizing that, in the above, the true model was always one of the candidates. In case the true model is not considered, then one will end up selecting the one that is closest to the true one, since it best describes the observed data.

We now turn our attention from plain model selection to model averaging. As explained in Section 9.1.2, the aim is to sequentially fuse information from different models in agreement with their evolving Bayesian posteriors. We achieve this autonomously within the SMC method by adjusting the weighting and resampling steps, as explained in Section 9.1.2.1.

Thus, we now evaluate the performance of the model averaging SMC proposed in Table 34 and compare it to the model selection SMC in the SV model with latent fGp with unknown Hurst parameter, i.e.,

$$\begin{cases} x_t = u_t, \\ y_t = e^{x_t/2} v_t, \end{cases} \tag{206}$$

where $v_t \sim \mathcal{N}(0, \sigma_v^2)$ and u_t is a zero-mean fGp with unknown H.

We consider a set of candidate models $\mathcal{M}_k$, $k = 1, \cdots, K$, where we assume a different $H_k \in [0.5, 1)$ value for each model, resulting in different autocovariance functions $\gamma_k(\tau)$ for the latent fGp.

An example of tracking the latent state under model uncertainty by the model averaging SMC proposed in Table 34 is shown in Figure 60.

We present statistical results of the filtering performance in Table 40. They were obtained by averaging over 100 realizations, each with time-series of length 500. The total number of

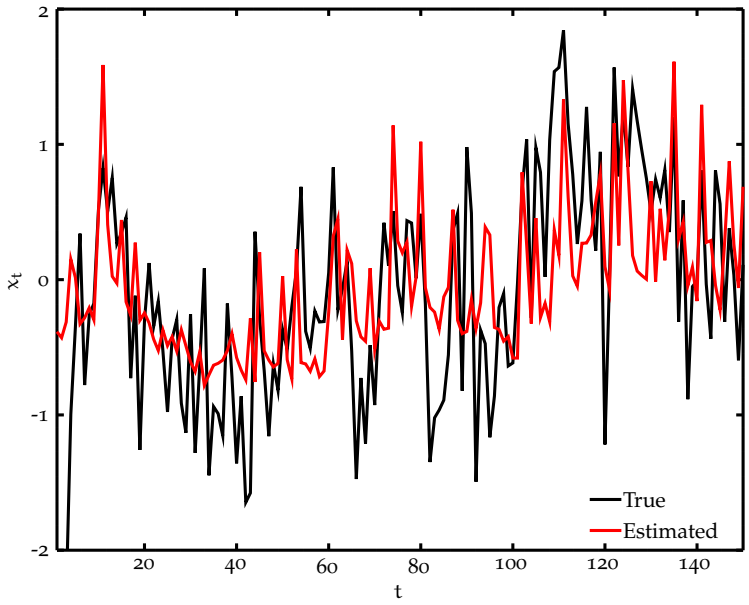


Figure 60: True state (black) and estimated (red) by proposed model averaging SMC.

particles was set to $M = 1000$. We considered $K = 5$ candidate models with parameters $H_1 = 0.5$, $H_2 = 0.7$, $H_3 = 0.8$, $H_4 = 0.9$, $H_5 = 0.95$ and $\sigma_k = 1, \forall k$.

We compared the filtering performance of the following methods:

- `Model selection SMC`: the model selection method proposed in Table 33.
- `Model averaging SMC`: a model averaging SMC, where the number of particles of each model is kept equal and constant, i.e., $M_{k,t} = 200, \forall k, \forall t$.
- `Proposed SMC method`: the SMC method proposed in Table 34, with $M_{k_{min}} = 0$.
- `True model SMC`: SMC method that knows the true underlying model (indicated in the table header).

The results in Table 40 show that the model averaging provides, overall, better performance than a plain model selection scheme.

Furthermore, the proposed dynamic SMC method is consistently more accurate than the plain averaging alternative. The reason is the implemented dynamic resampling scheme.

SMC methods	True H					
	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
`Model selection SMC`	0.75456	0.76018	0.75068	0.73097	0.69510	0.73836
`Model averaging SMC`	0.75216	0.75714	0.74734	0.72689	0.69082	0.73475
`Proposed SMC method`	0.75177	0.75265	0.74487	0.70598	0.61393	0.53713
`True model SMC`	0.74572	0.74664	0.71724	0.66040	0.53432	0.40041

Table 40: MSE of the state for different SMC methods.

By apportioning more particles to better models, the overall estimation accuracy of the SMC method is improved. We remark that the performance of the proposed method is reasonably close to the benchmark (i.e., to that of the method based on the true model).

How the method dynamically adjusts the resources (i.e., particles) over time is illustrated in Figure 61, Figure 62, Figure 63, Figure 64 and Figure 65, where the sample size (averaged over 100 realizations) assigned to each model is shown.

By studying the resource allocation evolution, one concludes that (a) the true model gets more resources as more data are processed, (b) models with extreme memory properties (i.e., $H \to 0.5$ Vs $H \to 1$) are quickly discarded, (c) for long-memory processes, $H \to 1$, short-memory processes are quickly discarded, while close-by candidates are more slowly distinguished.

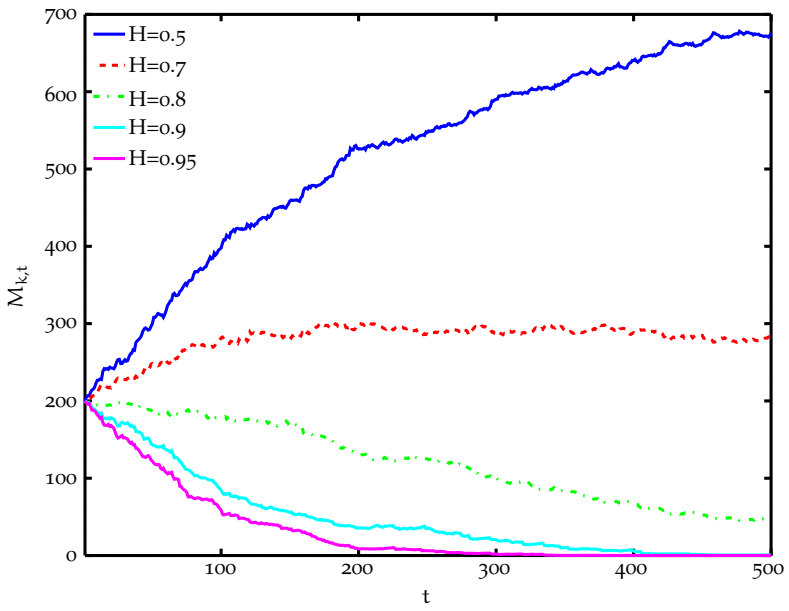


Figure 61: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.5$.

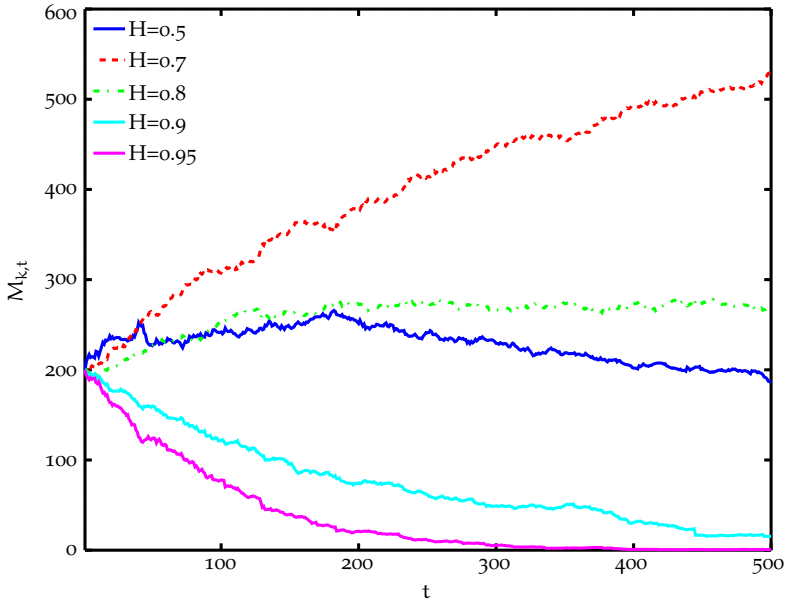


Figure 62: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.7$.

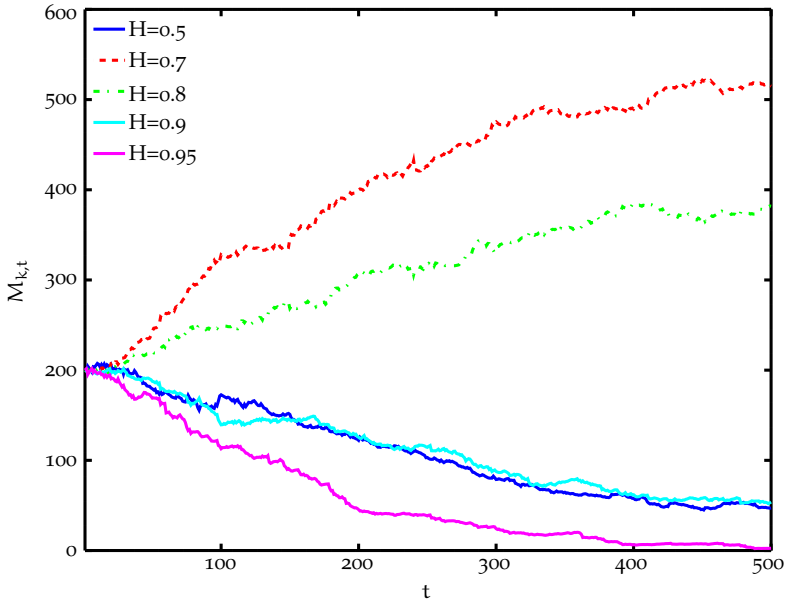


Figure 63: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.8$.

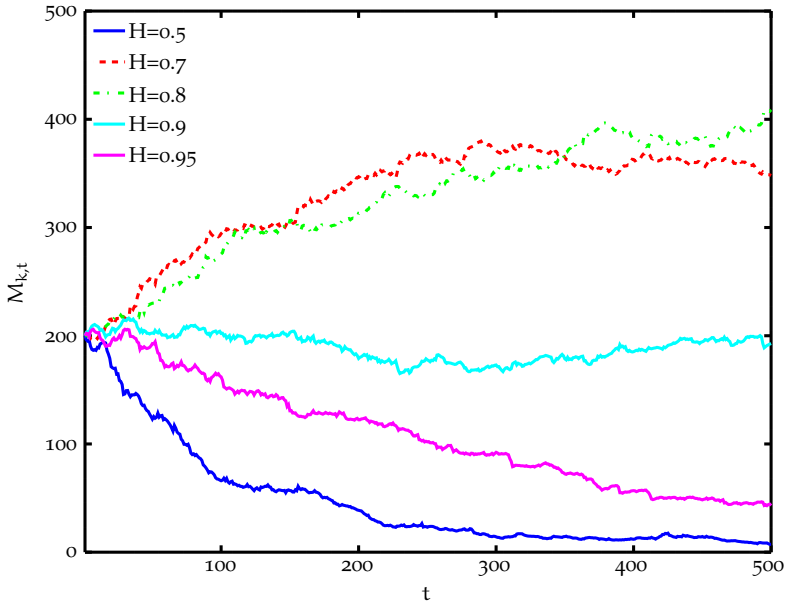


Figure 64: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.9$.

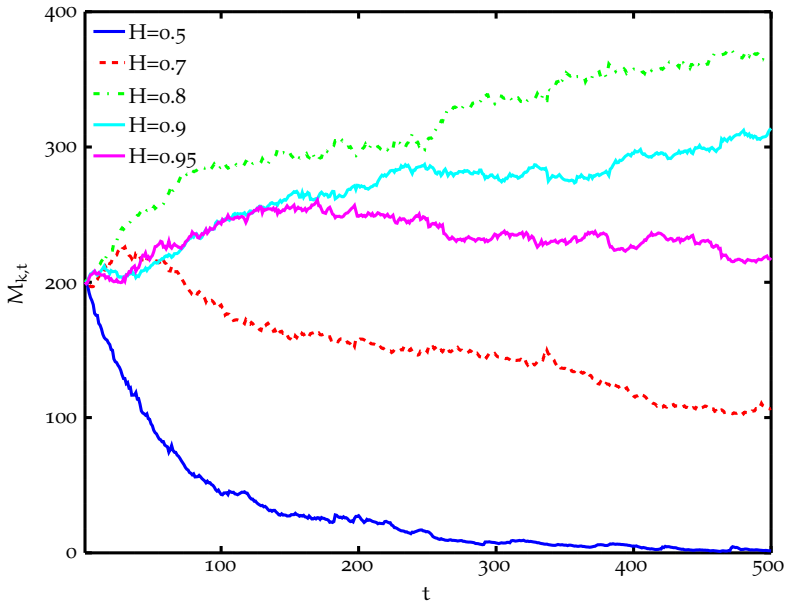


Figure 65: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.95$.

We now emphasize one of the benefits of the proposed dynamic model averaging technique: The ability to accurately estimate the underlying latent process, when the true model is not amongst the candidates (see true $H = 0.6$ in Table 40).

Even if none of the assumed models is the correct one, the performance of the proposed SMC method is the best amongst the alternatives and quite close to the benchmark (`True model SMC`). The explanation comes from the proposed dynamic averaging scheme. When the true model is not available, then the proposed SMC assigns resources to those closest to the true one and, due to model averaging, an improved performance is obtained.

This behavior is illustrated in Figure 66, where a balance between the models closest to the true one ($H = 0.6$) is achieved.

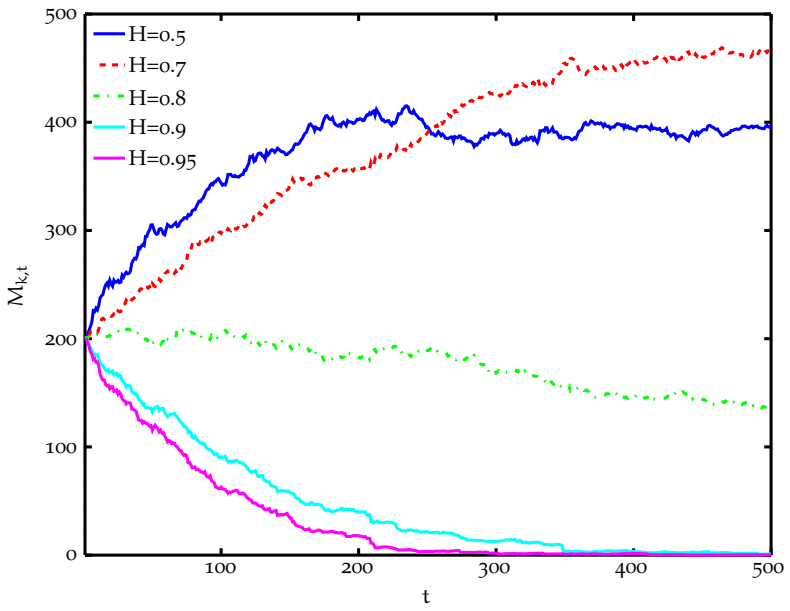


Figure 66: Number of particles $M_{k,t}$ assigned to each SMC filter with true $H = 0.6$.

Another benefit of our proposed dynamic SMC is illustrated in Figure 67, where the evolution of the filtering error is plotted over time. There, we observe how as time evolves, the proposed SMC improves its filtering accuracy, as it allocates resources to those models with good predictive performance, while discarding incorrect ones (see Figure 66). On the long run, the accuracy of the proposed SMC method would match that of the true one. On the contrary, a plain model averaging method

weights equally all models and, thus, it can not improve its performance over time.

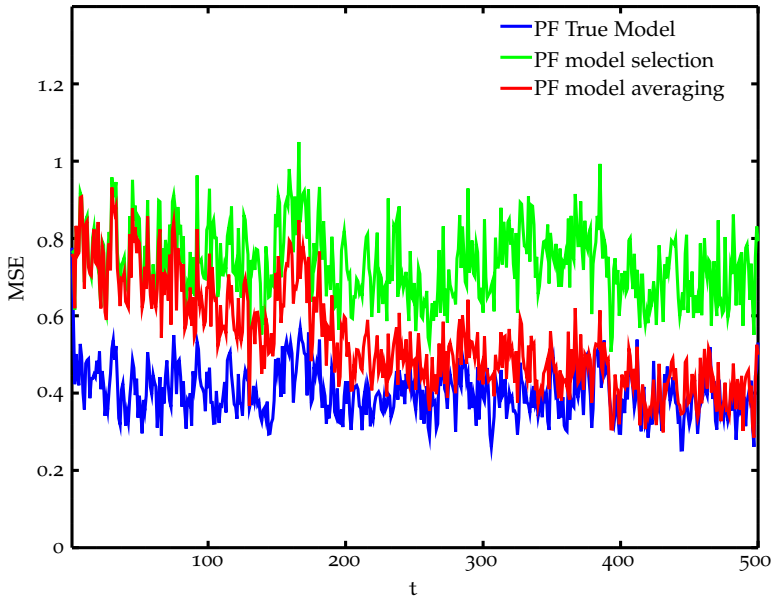


Figure 67: Filtering MSE evolution over time, for fGp with $H = 0.95$.

Finally, we evaluate the filtering performance if a minimum sample size is guaranteed per model, i.e., $M_{k_{min}} > 0$. As shown in Table 41, for the studied cases, it does not make sense to guarantee a minimum amount of resources per model, as it restricts the flexibility of the proposed method to assign resources to the best models. In a sense, the reasoning is the same as in the above paragraph, since a model averaging SMC method is the limiting case with $M_{k_{min}} = M/K$. Once again, we note that the difference for short-memory processes (i.e., $0.5 \leqslant H < 0.75$) is less remarkable than for the long-memory processes, as $H \to 1$.

SMC methods	True H					
	$H = 0.5$	$H = 0.6$	$H = 0.7$	$H = 0.8$	$H = 0.9$	$H = 0.95$
Proposed SMC method $M_{k_{min}} = 0$	0.75177	0.75265	0.74487	0.70598	0.61393	0.53713
Proposed SMC method $M_{k_{min}} = 20$	0.75063	0.75670	0.75000	0.73306	0.69647	0.72918

Table 41: State MSE for the proposed SMC method, different $M_{k_{min}}$.

Part III

CONCLUSIONS AND FUTURE WORK

Concluding remarks on the presented work and the contributions of this work are summarized in this chapter. Directions for future work are also outlined.

10 CONCLUSIONS AND FUTURE WORK

The goal of the presented work is to provide methods for accurate inference and prediction of latent time-series, in the most challenging scenarios. Consequently, we have resorted to SMC methods, so that generic state-space models with nonlinearities and non-Gaussianities are accommodated.

Throughout the dissertation, different time-series have been researched. We have thoroughly studied the different statistical features of each of the considered models. By means of Bayesian analysis, we have derived the joint and transition densities of interest, for different parameter and modeling assumptions.

As a result, a set of SMC methods have been proposed, which target inference of latent time-series of different nature and with various parameterization assumptions. Furthermore, more general solutions have been proposed for WSS and correlated time-series. Finally, inference of time-series under model uncertainty has been investigated too.

The contributions of this work are summarized in Section 10.1 and we conclude by outlining further research efforts in Section 10.2.

CONTRIBUTIONS

In Chapter 3 and Chapter 4, we addressed the problem of recursive estimation of latent ARMA(p, q) processes. Inference of latent ARMA processes is a much more challenging task than the estimation of plain AR processes. The complications are due to, amongst others, the nonlinearities induced by the MA part. Even when the ARMA process is directly observed, the derivation of the exact parameter densities is intractable. The problem is certainly much more difficult when the process is not observed, as in the work presented here.

In Chapter 3, we have provided a rigorous analysis of the ARMA model, with emphasis on the statistical properties of the process, the derivation of the key densities, and computation of its sufficient statistics (in matrix and recursive forms).

Based on such findings, we have proposed a set of novel SMC methods for inference of latent ARMA(p, q) processes with different parameter knowledge assumptions. Variants for the cases of known/unknown ARMA parameters a and b, and known/unknown driving noise variance σ_u^2 have been presented. The solutions when the parameters are unknown are based on the concept of Rao-Blackwellization, which allows for superior estimation accuracy.

We extended the analysis of time-series by considering other memory properties in Chapter 5. There, we have addressed time-series for which dependencies amongst time instants goes further into the past. A detailed description of long-memory processes is provided in Section 5.1, with emphasis on two classes of models: the fGp and FARIMA models.

The main contribution in Chapter 5 is the proposed SMC method for inference of latent fGps. These self-similar processes are of interest due to their parsimonious parameterization of diverse memory properties: from uncorrelated innovations when $H = 0.5$, to long-memory when $H \to 1$. Evaluation results show that, given the Hurst parameter, the proposed SMC successfully tracks different fGps with high accuracy.

The alternative FARIMA model, and extensions to a more general framework with ARMA(p, q) processes driven by correlated innovations, are the focus of Chapter 6. The first contribution there is the Bayesian analysis of such generic models, which leads into the derivation of their joint and transition densities. Thus, we provide the key densities for the analysis of any time-series that is, in general, described by an ARMA(p, q) driven by correlated innovations. Note that any of the previous models are sub-cases of this more general one: If innovations are uncorrelated, we have an ARMA(p, q) model; and if no ARMA filtering is applied, then we are left with a fGp.

The second contribution in Chapter 6 is the set of proposed SMC methods for inference of ARMA(p, q) processes driven by correlated innovations. Different parameter knowledge assumptions have been made and the resulting SMC variants have been shown to provide accurate state estimation for all the studied cases.

Chapter 7 addresses a more general problem on inference of latent time-series, where assumptions on the underlying model and its parameters are kept to a minimum: wide-sense stationarity is assumed. There, we only require that the time-series is described by a Gaussian density with a constant first

moment (i.e., the mean), and second order statistics that are functions of the time-lag τ (i.e., the covariance matrix). By relaxing the assumptions and resorting to the marginalization (i.e., Rao-Blackwellization) of the sufficient statistics, we derive a generic transition density: One that describes any Gaussian WSS process.

Therefore, the contribution in Chapter 7 is a generic SMC method for inference of latent Gaussian WSS time-series, that can be applied to a myriad of scenarios: latent ARMA(p, q) and FARIMA(p, d, q) models with unknown parameters and model orders, latent fGps with unknown H, etc. The accuracy and flexibility of the proposed method are shown via extensive evaluations.

We broaden the scope of our work in Chapter 8 by investigating correlation amongst multiple time-series. That is, we consider inference of multivariate latent time-series. We devise a hierarchical model, where the correlated time-series are modeled as a mixture of independent processes embedded in noise. We provide a Bayesian analysis of the system by marginalization of the unknown parameters. As a result, we propose an SMC method for inference of correlated time-series, where the assumptions on the underlying model parameters are relaxed.

The dissertation concludes with a chapter devoted to the study of latent time-series under model uncertainty. Because in practice it is difficult to know the true underlying dynamics of the data, deriving solutions for such problem is of critical importance. Our contributions in Chapter 9 are two-fold.

First, we propose an SMC-based method that combines a bank of filters with model selection. That is, an independent SMC filter is run for each of a set of candidate models; followed by a model selection scheme to determine the best model, based on their predictive power.

Second, we contribute with a novel SMC method that sequentially fuses information from different models in a dynamic fashion. We achieve this autonomously within the proposed SMC method by adjusting the weighting and resampling steps. The method is dynamic in that it updates its model posterior probabilities as more data become available, and that it apportions resources (i.e., particles) according to such probabilities amongst the candidate models. The proposed dynamically adjustable SMC method is shown to have improved filtering accuracy.

In summary, the contribution of this dissertation is a set of SMC methods for sequential inference and prediction of latent time-series. Some of the proposed solutions target an specific set of modeling or parameter knowledge assumptions, while others are less restrictive and are suitable for inference of more general time-series.

These methods are readily applicable to challenging problems of practical interest. The key to decide which solution to implement in practice will be determined by the amount of knowledge available to the practitioner.

FUTURE WORK

There are several directions of research that stem from the work presented in this dissertation.

First, there are some aspects of the presented SMC methods that can be further investigated. Specifically, the memory properties of the studied time-series and their impact on the performance of the proposed methods is of interest.

The key is to scrutinize the memory-properties of the considered processes by means of their autocovariance functions $\gamma(\tau)$. It is clear that one can discard past samples after a certain lag, once the autocovariance function becomes negligible, as they provide no information for present values. That is, if $\gamma(\tau) \to 0$, $\tau > \tau_{max}$, then there is basically no information loss in discarding lags bigger than τ_{max}; i.e., there is no information about the present farther into the past. We have also shown that the $-\alpha_t = \widetilde{\Sigma}_t^{-1}\widetilde{\gamma}_t$ term provides critical information on this matter too. Therefore, an interesting line of research is to analytically determine which τ_{max} allows to keep most of the relevant information in each time-series.

Because of this negligible information gain after τ_{max}, we have suggested to limit the computation of the sufficient statistics of the transition density to a certain lag, i.e., to use $x_{t-\tau_{max}:t}$ instead of $x_{1:t}$ (see Chapters 3, 4, 5 and 7), as it greatly reduces the computation burden of the proposed methods.

However, truncation causes negligible information loss only when τ_{max} is carefully selected. We have established that truncation is safe for most of the studied short-memory processes, i.e., ARMA(p, q) models and fGp with $0.5 < H \leqslant 0.75$. Extending the guidelines (either numerically or analytically) on which truncation τ_{max} to use for different time-series will be of undeniable help for practitioners.

Investigation on the memory properties of different time-series will also shed light on the convergence of the proposed SMC methods. It has already been shown [43] that, when exponential forgetting holds, one can establish uniform-in-time convergence of any PF, for functions that depend only on recent states. Thus, one can readily guarantee that the proposed SMC methods converge as long as the latent process has short-memory. Nevertheless, it remains unclear how to theoretically prove that the proposed SMC methods converge for long-memory processes.

In the introduction, we pointed out that one needs to be careful with the potential impact of path-degeneracy in any PF. As reported in [27], the Monte Carlo error of path functionals $\phi\left(x_{1:t}^{(m)}\right)$ remains bounded over time if the functionals relate to the filtering problem, i.e., $\phi\left(x_{1:t}^{(m)}\right) = x_t^{(m)}$, but explodes for the smoothing problem, i.e., $\phi\left(x_{1:t}^{(m)}\right) = x_1^{(m)}$.

In the proposed SMC methods, some of the derived statistics depend on the full path of the state and thus, questions might be raised on this matter. Once again, the key for the performance of the proposed methods is their memory property. That is, when dealing with short-memory processes, the dependence of the functionals on past samples decays quickly $\phi\left(x_{1:t}^{(m)}\right) \approx \phi\left(x_{t-\tau_{max}:t}^{(m)}\right)$ and thus, the impact of path-degeneracy is contained. However, a rigorous analysis of the relationship between the memory properties of the time-series, the used path functionals and their impact on the proposed SMC methods is deemed relevant.

Convergence and performance issues for SMC methods become even more critical when considering estimation of high-dimensional states. Either because joint state and parameters are considered (as in Chapter 4 or Chapter 6), or because multivariate time-series are studied (as in Chapter 8), the dimensionality of the state grows.

The limitations of SMC techniques in high-dimensional problems have been noted by many authors [11, 33, 82]. The curse-of-dimensionality in sampling methods refers to the inability to efficiently search through high-dimensional spaces. Thus, effective alternatives to deal with such curse (e.g., multiple particle filtering [30, 38], Rao-Blackwellization [89], and others) must be researched.

Finally, the implementation of the proposed SMC methods in real applications must be carried out. The work of this dissertation has been on the derivation of novel SMC methods for inference of latent time-series and its evaluation done in simulated scenarios. This has allowed for a rigorous analysis of the estimation accuracy of the methods, due to the knowledge of the underlying model and the access to the true latent state.

However, the ultimate goal is to apply the proposed methods to real-life scenarios and data. Time-series from different fields, such as biomedicine, economics, finance and others (e.g., Google Trends), should be examined.

There, one does not have access to the latent states and, even more, has no certainty on the specifics of the hidden dynamics. Thus, selecting which proposed SMC method to implement will be of critical importance.

Lastly, one must explore alternative performance metrics for evaluation of the applied methods. The proposed SMC methods provide an approximation to the predictive density, which can readily be used for the analysis of any time-series (and their evaluation).

Part IV

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Gabriel Agamennoni and Eduardo M. Nebot. "Robust Estimation in Non-Linear State-Space Models With State-Dependent Noise." In: *Signal Processing, IEEE Transactions on* 62.8 (Apr. 2014), pp. 2165–2175.

[2] Pierre-Olivier Amblard and Jean-François Coeurjolly. "Identification of the Multivariate Fractional Brownian Motion." In: *Signal Processing, IEEE Transactions on* 59.11 (Nov. 2011), pp. 5152–5168.

[3] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Dover Books on Electrical Engineering. Dover Publications, 2012.

[4] C. Andrieu, A. Doucet, W.J. Fitzgerald, and J.-M. Pérez. *Bayesian Computational Approaches to Model Selection*. 2000.

[5] Christophe Andrieu, Petar M. Djurić, and Arnaud Doucet. "Model selection by MCMC computation." In: *Signal Processing* 81.1 (2001), pp. 19–37.

[6] Christophe Andrieu, Nando de Freitas, and Arnaud Doucet. "Sequential MCMC for Bayesian model selection." In: *Higher-Order Statistics, 1999. Proceedings of the IEEE Signal Processing Workshop on*. 1999, pp. 130–134.

[7] M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking." In: *Signal Processing, IEEE Transactions on* 50.2 (Feb. 2002), pp. 174–188.

[8] Richard T. Baillie. "Long memory processes and fractional integration in econometrics." In: *Journal of Econometrics* 73.1 (1996), pp. 5 –59.

[9] Olivier Basdevant. "On Applications of State-Space Modelling in Macroeconomics." In: *SSRN Electronic Journal* (2003).

[10] Erhan Bayraktar, H. Vincent Poor, and K. Ronnie Sircar. "Efficient estimation of the Hurst parameter in high frequency financial data with seasonalities using wavelets." In: *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*. Mar. 2003, pp. 309–316.

[11] Thomas Bengtsson, Peter Bickel, and Bo Li. "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems." In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Ed. by Deborah Nolan and Terry Speed. Vol. 2. Collections. Institute of Mathematical Statistics, 2008, pp. 316–334.

[12] Jan Beran. *Statistics for Long-Memory Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[13] José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009.

[14] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[15] G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976.

[16] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. 2nd. Springer Series in Statistics. Springer, 1991.

[17] Carmen Broto and Esther Ruiz. "Estimation methods for stochastic volatility models: a survey." In: *Journal of Economic Surveys* (2004), pp. 613–649.

[18] Laurent E. Calvet and Adlai J. Fisher, eds. *Multifractal Volatility: Theory, Forecasting, and Pricing*. Academic Press, 2008.

[19] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[20] Bradely P. Carlin and Thomas A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2010.

[21] Bradley P. Carlin and Thomas A. Louis. *Bayes and Empricial Bayes Methods for Data Analysis*. Chapman and Hall/CRC, 2000.

[22] Carlos M. Carvalho and Hedibert F. Lopes. "Simulation-based sequential analysis of Markov switching stochastic volatility models." In: *Computational Statistics & Data Analysis* 51.9 (2007), pp. 4526 –4542.

[23] George Casella and Christian P. Robert. "Rao-Blackwellisation of sampling schemes." In: *Biometrika* 83.1 (1996), pp. 81–94.

[24] Yacine Chakhchoukh. "A New Robust Estimation Method for ARMA Models." In: *Signal Processing, IEEE Transactions on* 58.7 (July 2010), pp. 3512–3522.

[25] Patrick Cheridito. "Arbitrage in fractional Brownian motion models." In: *Finance and Stochastics* 7.4 (2003), pp. 533–553.

[26] Siddhartha Chib and Edward Greenberg. "Bayes inference in regression models with ARMA(p, q) errors." In: *Journal of Econometrics* 64.12 (1994), pp. 183 –206.

[27] Nicolas Chopin. "Central Limit Theorem for Sequential Monte Carlo Methods and Its Application to Bayesian Inference." In: *The Annals of Statistics* 32.6 (2004), pp. 2385–2411.

[28] Nicolas Chopin, Alessandra Iacobucci, Jean-Michel Marin, Kerrie Mengersen, Christian P. Robert, Robin Ryder, and Christian Schäfer. *On Particle Learning*. Archived online at http://arxiv.org/abs/1006.0554. 2010.

[29] Gerda Claeskensand and Nils L. Hjort. *Model Selection and Model Averaging*. Vol. 330. Cambridge University Press Cambridge, 2008.

[30] Pau Closas and Mónica F. Bugallo. "Improving Accuracy by Iterated Multiple Particle Filtering." In: *IEEE Signal Processing Letters* 19.8 (Aug. 2012), pp. 531–534.

[31] Rama Cont. "Long range dependence in financial markets." In: *Fractals in Engineering*. Ed. by Jacques Lévy-Véhel and Evelyne Lutton. Springer London, 2005, pp. 159–179.

[32] Drew Creal. "A Survey of Sequential Monte Carlo Methods for Economics and Finance." In: *Econometric Reviews* 31.3 (2012), pp. 245–296.

[33] Fred Daum and Jim Huang. "Curse of dimensionality and particle filters." In: *2003 IEEE Aerospace Conference Proceedings*. Vol. 4. Mar. 2003, pp. 1979–1993.

[34] Gustavo Didier, Hannes Helgason, and Patrice Abry. "Demixing multivariate-operator self-similar processes." In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. Apr. 2015, pp. 3671–3675.

[35] Gustavo Didier and Vladas Pipiras. "Integral representations and properties of operator fractional Brownian motions." In: *Bernoulli* 17.1 (2011), pp. 1–33.

[36] Petar M. Djurić. "Monitoring and selection of dynamic models by Monte Carlo sampling." In: *Higher-Order Statistics, 1999. Proceedings of the IEEE Signal Processing Workshop on*. 1999, pp. 191–194.

[37] Petar M. Djurić and Mónica F. Bugallo. "Particle Filtering." In: *Adaptive Signal Processing*. Ed. by Simon Haykin. John Wiley & Sons, Inc., 2010. Chap. 5, pp. 271–331.

[38] Petar M. Djurić and Mónica F. Bugallo. "Multiple particle filtering with improved efficiency and performance." In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2015, pp. 4110–4114.

[39] Petar M. Djurić, Mónica F. Bugallo, and Joaquín Míguez. "Density assisted particle filters for state and parameter estimation." In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*. Vol. 2. May 2004, pp. ii –701–704.

[40] Petar M. Djurić, Mahsiul Khan, and Dougles E. Johnston. "Particle Filtering of Stochastic Volatility Modeled With Leverage." In: *Selected Topics in Signal Processing, IEEE Journal of* 6.4 (2012), pp. 327–336.

[41] Petar M. Djurić, Jayesh H. Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F. Bugallo, and Joaquín Míguez. "Particle Filtering." In: *IEEE Signal Processing Magazine* 20(5) (Sept. 2003), pp. 19–38.

[42] Randal Douc and Olivier Cappé. "Comparison of resampling schemes for particle filtering." In: *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.* Sept. 2005, pp. 64–69.

[43] Arnaud Doucet, Nando De Freitas, and Neil Gordon, eds. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[44] Arnaud Doucet, Neil J. Gordon, and Vikram Krishnamurthy. "Particle filters for state estimation of jump Markov linear systems." In: *Signal Processing, IEEE Transactions on* 49.3 (Mar. 2001), pp. 613–624.

[45] Arnaud Doucet and Vladislav B. Tadic. "Parameter Estimation in General State-Space Models using Particle Methods." In: *Annals of the Institute of Statistical Mathematics* 55 (2003), pp. 409–422.

[46] James Durbin and Siem Jan Koopman. *Time Series Analysis by State-Space Methods*. Oxford Statistical Science Series. Oxford University Press, 2001.

[47] James Durbin and Siem Jan Koopman. *Time Series Analysis by State-Space Methods: Second Edition*. 2nd ed. Oxford Statistical Science Series. Oxford University Press, 2012.

[48] Seymour Geisser. "Bayesian-Estimation in multivariate analysis." English. In: *Annals of American Statistics* 36.1 (1965), pp. 150–159.

[49] Neil J. Gordon, D.J. Salmond, and A.F.M. Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." In: *Radar and Signal Processing, IEEE Proceedings* 140.2 (Apr. 1993), pp. 107 –113.

[50] Andrew. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter,* ed. by Cambridge University Press. Cambridge, 1989.

[51] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, 1996.

[52] Jeroen D. Hol, Thomas B. Schön, and Fredrik Gustafsson. "On Resampling Algorithms for Particle Filters." In: *IEEE Nonlinear Statistical Signal Processing Workshop, 2006*. Sept. 2006, pp. 79–82.

[53] Mingyi Hong, Mónica F. Bugallo, and Petar M. Djurić. "Joint Model Selection and Parameter Estimation by Population Monte Carlo Simulation." In: *Selected Topics in Signal Processing, IEEE Journal of* 4.3 (June 2010), pp. 526–539.

[54] Harold Edwin Hurst. "Long-term storage capacity of reservoirs." In: *Transactions of the American Society of Civil Engineers* 116 (1951), pp. 770–808.

[55] Edward L. Ionides, C. Bretó, and A. A. King. "Inference for nonlinear dynamical systems." In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18438–18443.

[56] Edward L. Ionides, Kathy S. Fang, R. Rivkah Isseroff, and George F. Oster. "Stochastic models for cell motion and taxis." English. In: *Journal of Mathematical Biology* 48.1 (2004), pp. 23–37.

[57] Kazufumi Ito and Kaiqi Xiong. "Gaussian filters for nonlinear filtering problems." In: *Automatic Control, IEEE Transactions on* 45.5 (May 2000), pp. 910–927.

[58] Eric Jacquier, Nicholas G. Polson, and Peter E. Rossi. "Bayesian Analysis of Stochastic Volatility Models." In: *Journal of Business & Economic Statistics* 12.4 (1994), pp. 371–89.

[59] Harold Jeffreys. "An Invariant Form for the Prior Probability in Estimation Problems." In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186.1007 (1946), pp. 453–461.

[60] Douglas E Johnston, Iñigo Urteaga, and Petar M. Djurić. "Replication and optimization of hedge fund risk factor exposures." In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, pp. 8712–8716.

[61] Simon J. Julier and Jeffrey K. Uhlmann. "Unscented filtering and nonlinear estimation." In: *Proceedings of the IEEE* 92.3 (Mar. 2004), pp. 401–422.

[62] V. Kadirkamanathan, M.H. Jaward, S.G. Fabri, and M. Kadirkamanathan. "Particle filters for recursive model selection in linear and nonlinear system identification." In: *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*. Vol. 3. 2000, pp. 2391–2396.

[63] Rudolph Emil Kalman. "A New Approach to Linear Filtering and Prediction Problems." In: *Transactions of the ASME–Journal of Basic Engineering* 82.Series D (1960), pp. 35–45.

[64] Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski, and Nicolas Chopin. "On Particle Methods for Parameter Estimation in State-Space Models." In: *Statistical Science* 30.3 (Aug. 2015), pp. 328–351.

[65] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation theory*. Ed. by Alan V. Oppenheim. Vol. 1. Fundamentals of Statistical Signal Processing. Prentice-Hall PTR, 1993.

[66] Genshiro Kitagawa. "Non-Gaussian State-Space Modeling of Nonstationary Time Series." English. In: *Journal of the American Statistical Association* 82.400 (1987), pp. 1032–1041.

[67] D.S. Lee and N.K.K. Chia. "A particle algorithm for sequential Bayesian parameter estimation and model selection." In: *Signal Processing, IEEE Transactions on* 50.2 (Feb. 2002), pp. 326–336.

[68] Peter Jan van Leeuwen. "Particle Filtering in Geophysical Systems." In: *Monthly Weather Review* 12.137 (2009), pp. 4089–4114.

[69] Tiancheng Li, Miodrag Bolić, and Petar M. Djurić. "Resampling Methods for Particle Filtering: Classification, implementation, and strategies." In: *Signal Processing Magazine, IEEE* 32.3 (May 2015), pp. 70–86.

[70] Bin Liu. "Instantaneous Frequency Tracking under Model Uncertainty via Dynamic Model Averaging and Particle Filtering." In: *Wireless Communications, IEEE Transactions on* 10.6 (June 2011), pp. 1810–1819.

[71] Jane Liu and Mike West. "Combined Parameter and State Estimation in Simulation Based Filtering." In: Springer, 2001. Chap. 10 in "Sequential Monte Carlo Methods in Practice", pp. 197–224.

[72] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001, p. 346.

[73] Hedibert F. Lopes and Ruey S. Tsay. "Particle filters and Bayesian inference in financial econometrics." In: *Journal of Forecasting* 30.1 (2011), pp. 168–209.

[74] Rudolph Van Der Merwe and Eric Wan. "Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models." In: *In Proceedings of the Workshop on Advances in Machine Learning*. 2003.

[75] Steffen Michalek, Mirko Wagner, and Jens Timmer. "A new approximate likelihood estimator for ARMA-filtered hidden Markov models." In: *Signal Processing, IEEE Transactions on* 48.6 (June 2000), pp. 1537–1547.

[76] John F. Monahan. "Fully Bayesian analysis of ARMA time series models." In: *Journal of Econometrics* 21 (Apr. 1983), pp. 307–331.

[77] Kevin P. Murphy. *Conjugate Bayesian analysis of the Gaussian distribution*. Tech. rep. University of British Columbia, 2007.

[78] Christopher Nemeth, Paul Fearnhead, and Lyudmila Mihaylova. "Sequential Monte Carlo Methods for State and Parameter Estimation in Abruptly Changing Environments." In: *Signal Processing, IEEE Transactions on* 62.5 (Mar. 2014), pp. 1245–1255.

[79] Tohru Ozaki. *Time Series Modeling of Neuroscience data*. Ed. by CRC Press. Chapman & Hall/CRC Interdisciplinary Statistics, 2012.

[80] W. Palma. *Long-Memory Time Series: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2007.

[81] Yohan Petetin and François Desbouvries. "Bayesian Conditional Monte Carlo Algorithms for Nonlinear Time-Series State Estimation." In: *Signal Processing, IEEE Transactions on* 63.14 (July 2015), pp. 3586–3598.

[82] Paul Bui Quang, Christian Musso, and François Le Gland. "An insight into the issue of dimensionality in particle filtering." In: *Proceedings of the 13th International Conference on Information Fusion*. ISIF. Edinburgh, United Kingdom, July 2010.

[83] Svetlozar T. Rachev, John S. J. Hsu, Biliana S. Bagasheva, and Frank J. Fabozzi. *Bayesian Methods in Finance*. Frank J. Fabozzi Series. Wiley, 2008.

[84] Joerg Rings, Jasper A. Vrugt, Gerrit Schoups, Johan A. Huisman, and Harry Vereecken. "Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments." In: *Water Resources Research* 48.5 (2012). W05520.

[85] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Ed. by Artech Print. Artech House, 2004.

[86] Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. 2nd ed. 2001. Reprint. Springer Texts in Statistics. Springer New York, 2008.

[87] Michael Roth. *On the multivariate t distribution*. Other academic LiTH-ISY-R, 3059. Sweden: Linköping University, Department of Electrical Engineering, 2012.

[88] Saikat Saha and Fredrik Gustafsson. "Particle Filtering With Dependent Noise Processes." In: *Signal Processing, IEEE Transactions on* 60.9 (Sept. 2012), pp. 4497–4508.

[89] Saikat Saha, Emre Özkan, Fredrik Gustafsson, and Václav Šmídl. "Marginalized particle filters for Bayesian estimation of Gaussian noise parameters." In: *13th Conference on Information Fusion (FUSION), 2010*. July 2010, pp. 1–8.

[90] Saikat Saha, Pranab Kumar Mandal, Arunabha Bagchi, Yvo Boers, and Johannes N. Driessen. "Particle Based Smoothed Marginal MAP Estimation for General State Space Models." In: *Signal Processing, IEEE Transactions on* 61.2 (Jan. 2013), pp. 264–273.

[91] Daniel Jonathan Scansaroli. "Stochastic Modeling with Temporally Dependent Gaussian Processes: Applications to Financial Engineering, Pricing and Risk Management." PhD thesis. Lehigh University Lehigh Preserve, 2012.

[92] Thomas Schön, Fredrik Gustafsson, and Per-Johan Nordlund. "Marginalized particle filters for mixed linear/nonlinear state-space models." In: *Signal Processing, IEEE Transactions on* 53.7 (July 2005), pp. 2279–2289.

[93] Paul Shaman. "An Approximate Inverse for the Covariance Matrix of Moving Average and Autoregressive Processes." English. In: *The Annals of Statistics* 3.2 (1975), pp. 532–538.

[94] Paul Shaman. "Approximations for Stationary Covariance Matrices and Their Inverses with Application to ARMA Models." English. In: *The Annals of Statistics* 4.2 (1976), pp. 292–301.

[95] Neil Shephard, ed. *Stochastic Volatility: Selected Readings*. Advanced texts in econometrics. Oxford University Press, 2005.

[96] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. 3rd. Springer, 2010.

[97] Geir Storvik. "Particle filters for state-space models with the presence of unknown static parameters." In: *Signal Processing, IEEE Transactions on* 50.2 (Feb. 2002), pp. 281–289.

[98] Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger. "Estimators for Long-Range Dependence: An Empirical Study." In: *Fractals* 3 (1995), pp. 785–798.

[99] Iñigo Urteaga, Mónica F. Bugallo, and Petar M Djurić. "Filtering of nonlinear time-series coupled by fractional Gaussian processes." In: *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*. 2015.

[100] Darryl Veitch and Patrice Abry. "A wavelet-based joint estimator of the parameters of long-range dependence." In: *Information Theory, IEEE Transactions on* 45.3 (Apr. 1999), pp. 878–897.

[101] A. M. Walker. "On the Asymptotic Behaviour of Posterior Distributions." English. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 31.1 (1969), pp. 80–88.

[102] Eric A. Wan and Rudolph van der Merwe. "The unscented Kalman filter for nonlinear estimation." In: *The IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000 (AS-SPCC 2000)*. 2000, pp. 153–158.

[103] Greg Welch and Gary Bishop. *An Introduction to the Kalman Filter*. Tech. rep. Department of Computer Science, University of North Carloina at Chapel Hill, 2006.

[104] Peter Whittle. *Hypothesis Testing in Time Series Analysis*. Almquist and Wicksell, 1951.

[105] Onno Zoeter, Alexander Ypma, and T. Heskes. "Improved unscented Kalman smoothing for stock volatility estimation." In: *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*. Sept. 2004, pp. 143–152.

[106] Andrew C. arvey, Esther Ruiz, and Neil Shephard. "Multivariate Stochastic Variance Models." In: *The Review of Economic Studies* 61.2 (1994), pp. 247–264.

Part V

APPENDIX

A

SIR AND OPTIMAL PFS: COMPARISON

We hereby study the difference in performance of the Sampling Importance Resampling (SIR) and the optimal PFs. To do so, we consider the following state-space model

$$\begin{cases} x_t = \alpha x_{t=1} + u_t \,, \\ y_t = x_t + v_t \,, \end{cases} \tag{207}$$

where the innovation processes follow

$$\begin{cases} u_t \sim \mathcal{N}\left(u_t|0, \sigma_u^2\right) \,, \\ v_t = \alpha\mathcal{N}\left(v_t|\mu_1, \sigma_1^2\right) + (1-\alpha)\mathcal{N}\left(v_t|\mu_2, \sigma_2^2\right) \,. \end{cases} \tag{208}$$

SIR PF

First, we provide the details of the SIR method:

- The proposal density is the transition density

$$\pi(x_t|x_{t-1}^{(m)}, y_t) = f(x_t|x_{t-1}^{(m)}). \tag{209}$$

- The resulting weight computation equation follows

$$\begin{aligned} \tilde{w}_t^{(m)} &\propto w_{t-1}\frac{f(y_t|x_t^{(m)})f(x_t^{(m)}|x_{t-1}^{(m)})}{\pi(x_t^{(m)}|x_{t-1}^{(m)}, y_t)} = w_{t-1}\frac{f(y_t|x_t^{(m)})f(x_t^{(m)}|x_{t-1}^{(m)})}{f(x_t^{(m)}|x_{t-1}^{(m)})} \\ &\propto w_{t-1}f(y_t|x_t^{(m)}). \end{aligned} \tag{210}$$

- And we perform resampling at every time instant

$$\tilde{w}_t^{(m)} \propto f(y_t|x_t^{(m)}). \tag{211}$$

The densities to be used for the model in Equation 208 are

$$\begin{cases} f(x_t|x_{t-1}) = \mathcal{N}\left(x_t|\alpha x_{t-1}, \sigma_u^2\right) \,, \\ f(y_t|x_t) = \alpha\mathcal{N}\left(y_t - x_t|\mu_1, \sigma_1^2\right) + (1-\alpha)\mathcal{N}\left(y_t - x_t|\mu_2, \sigma_2^2\right) \,. \end{cases} \tag{212}$$

OPTIMAL PF

The details of the optimal PF are:

- The proposal density is optimal

$$\pi(x_t|x_{t-1}^{(m)}, y_t) = f(x_t|x_{t-1}^{(m)}, y_t) = \frac{f(x_t, y_t|x_{t-1}^{(m)})}{f(y_t|x_{t-1}^{(m)})} = \frac{f(y_t|x_t, x_{t-1}^{(m)})f(x_t|x_{t-1}^{(m)})}{f(y_t|x_{t-1}^{(m)})}$$
$$= \frac{f(y_t|x_t)f(x_t|x_{t-1}^{(m)})}{f(y_t|x_{t-1}^{(m)})}. \tag{213}$$

- The resulting weight computation follows

$$\tilde{w}_t^{(m)} \propto w_{t-1}\frac{f(y_t|x_t^{(m)})f(x_t^{(m)}|x_{t-1}^{(m)})}{\pi(x_t^{(m)}|x_{t-1}^{(m)}, y_t)} = w_{t-1}\frac{f(y_t|x_t^{(m)})f(x_t^{(m)}|x_{t-1}^{(m)})}{\frac{f(y_t|x_t^{(m)})f(x_t^{(m)}|x_{t-1}^{(m)})}{f(y_t|x_{t-1}^{(m)})}}$$
$$\propto w_{t-1}f(y_t|x_{t-1}^{(m)}). \tag{214}$$

- And we, again, resample at every time instant

$$\tilde{w}_t^{(m)} \propto f(y_t|x_{t-1}^{(m)}). \tag{215}$$

We derive the required densities for the state-space model in Equation 207

$$\begin{cases} f(x_t|x_{t-1}, y_t) = \frac{f(y_t|x_t)f(x_t|x_{t-1})}{f(y_t|x_{t-1})}, \\ f(y_t|x_{t-1}) = \int f(y_t|x_t)f(x_t|x_{t-1})dx_{t-1}. \end{cases} \tag{216}$$

We note that because the observation equation follows a Gaussian mixture model, then

$$f(y_t|x_t)f(x_t|x_{t-1}) = \alpha\mathcal{N}\left(y_t - x_t|\mu_1, \sigma_1^2\right)\mathcal{N}\left(x_t|ax_{t-1}, \sigma_u^2\right)$$
$$+ (1-\alpha)\mathcal{N}\left(y_t - x_t|\mu_2, \sigma_2^2\right)\mathcal{N}\left(x_t|ax_{t-1}, \sigma_u^2\right). \tag{217}$$

By using the Gaussian posterior and marginal distributions [14], we conclude

$$\begin{cases} f(x_t|x_{t-1}) = \mathcal{N}\left(x_t|ax_{t-1}, \sigma_u^2\right), \\ f(y_t|x_t, x_{t-1}) = f(y_t|x_t) = \alpha\mathcal{N}\left(y_t|\mu_1 + x_t, \sigma_1^2\right) \\ \qquad\qquad +(1-\alpha)\mathcal{N}\left(y_t|\mu_2 + x_t, \sigma_2^2\right), \\ f(y_t|x_{t-1}) = \alpha\mathcal{N}\left(y_t|ax_{t-1} + \mu_1, \sigma_u^2 + \sigma_1^2\right) \\ \qquad\qquad +(1-\alpha)\mathcal{N}\left(y_t|ax_{t-1} + \mu_2, \sigma_u^2 + \sigma_2^2\right), \\ f(x_t|x_{t-1}, y_t) = \alpha\mathcal{N}\left(x_t|\frac{\sigma_1^2 ax_{t-1} + \sigma_u^2(y_t-\mu_1)}{\sigma_u^2+\sigma_1^2}, \frac{\sigma_u^2\sigma_1^2}{\sigma_u^2+\sigma_1^2}\right) \\ \qquad\qquad +(1-\alpha)\mathcal{N}\left(x_t|\frac{\sigma_2^2 ax_{t-1} + \sigma_u^2(y_t-\mu_2)}{\sigma_u^2+\sigma_2^2}, \frac{\sigma_u^2\sigma_2^2}{\sigma_u^2+\sigma_2^2}\right). \end{cases} \tag{218}$$

RESULTS

In the following, we provide illustrative results (averaged over 100 realizations) for the model in Equation 207 and Equation 208, with parameters for the observation equation set to $\alpha = 0.7, \mu_1 = -1.7, \sigma_1^2 = 2, \mu_2 = 1, \sigma_2^2 = 3$. The AR parameter a and state noise variance σ_u^2 are set as specified below.

$a = 0.9$, $\sigma_u = 0.5$

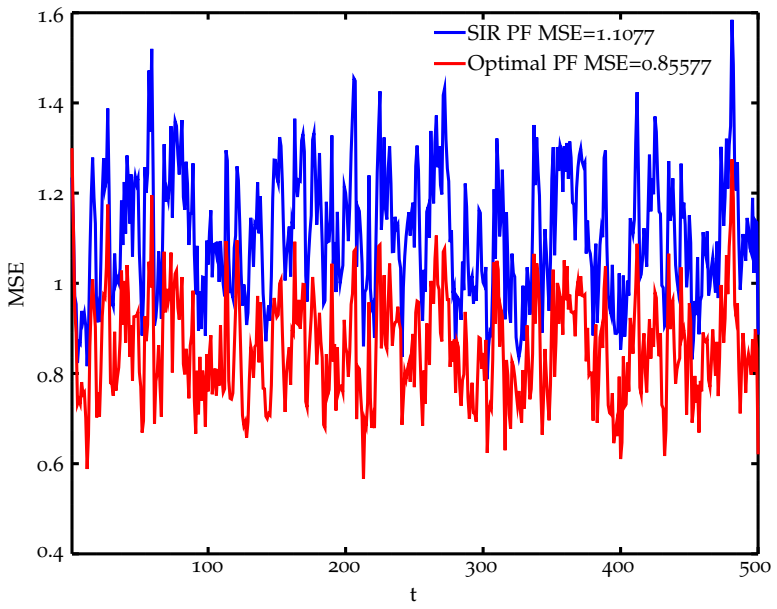


Figure 68: Averaged state filtering MSE, $a = 0.9$, $\sigma_u = 0.5$.

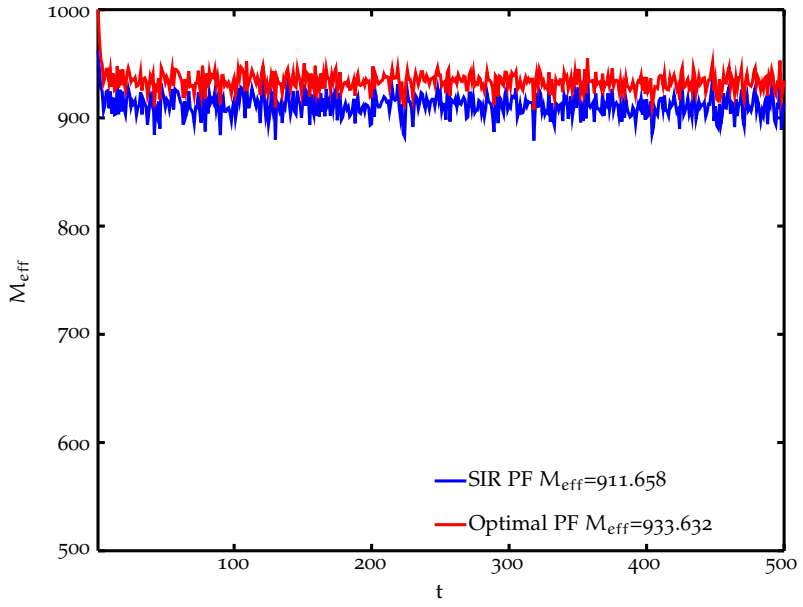


Figure 69: Effective particle size, $a = 0.9$, $\sigma_u = 0.5$.

$a = 0.9$, $\sigma_u = 1.5$

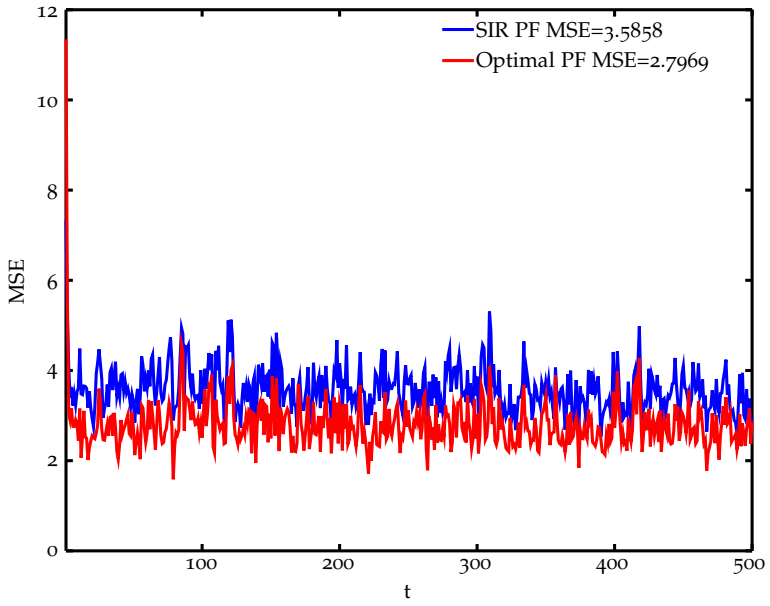


Figure 70: Averaged state filtering MSE, $a = 0.9$, $\sigma_u = 1.5$.

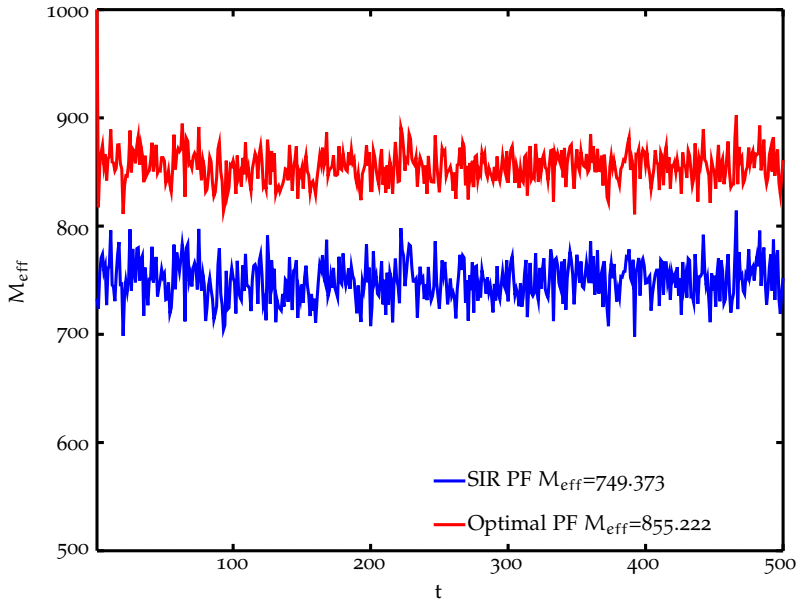


Figure 71: Effective particle size, $a = 0.9$, $\sigma_u = 1.5$.

$a = 0.9$, $\sigma_u = 2$

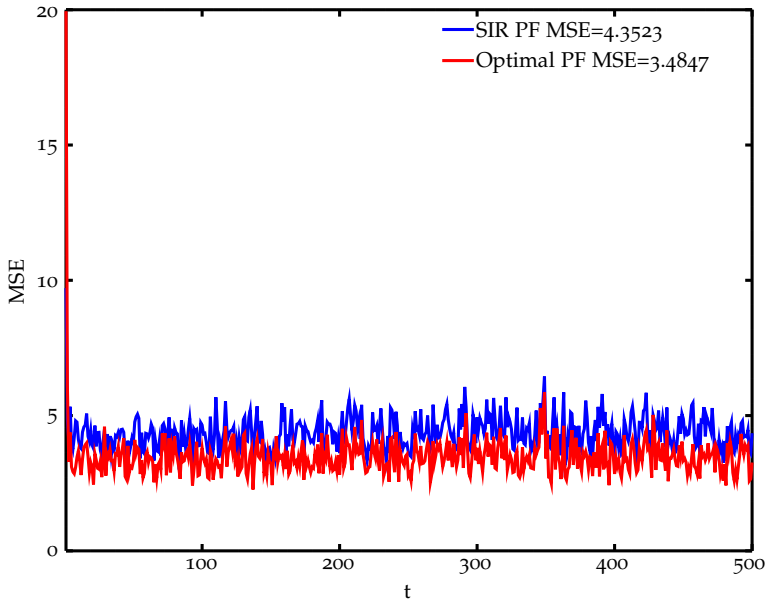


Figure 72: Averaged state filtering MSE, $a = 0.9$, $\sigma_u = 2$.

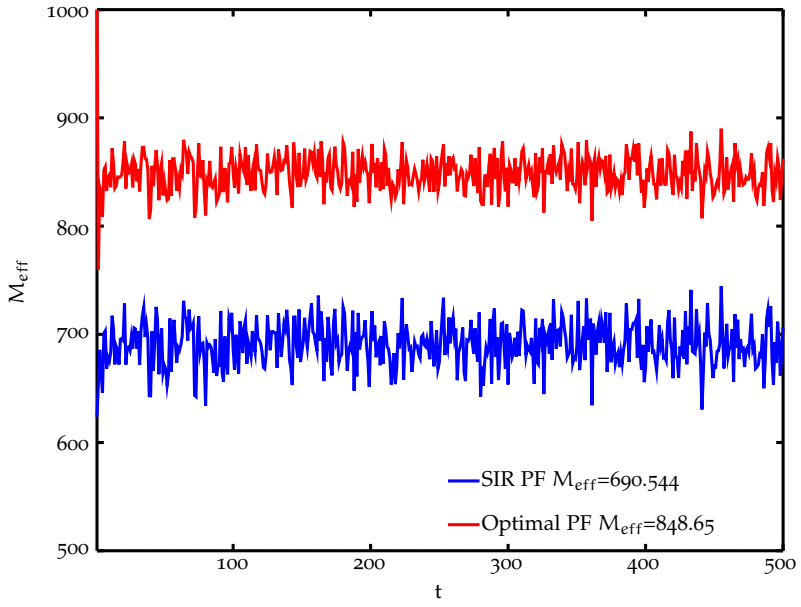


Figure 73: Effective particle size, $a = 0.9$, $\sigma_u = 2$.

$a = 0.65$, $\sigma_u = 0.5$

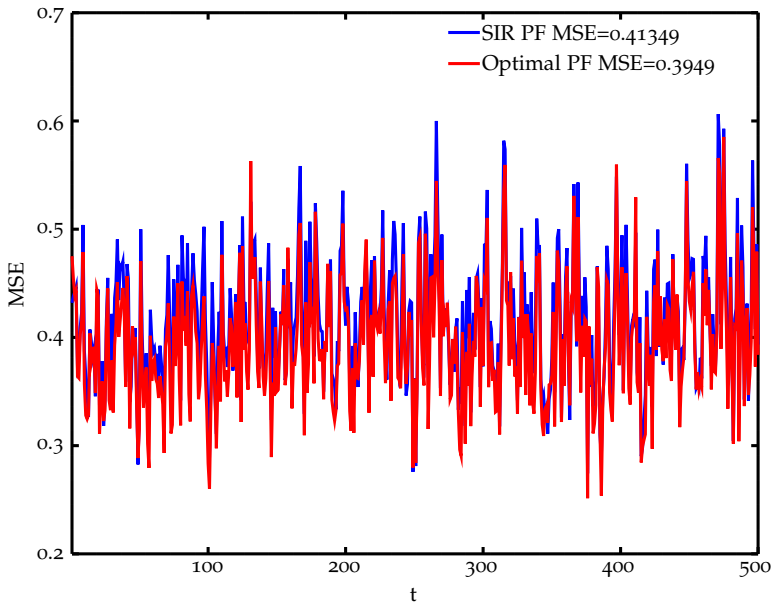


Figure 74: Averaged state filtering MSE, $a = 0.65$, $\sigma_u = 0.5$.

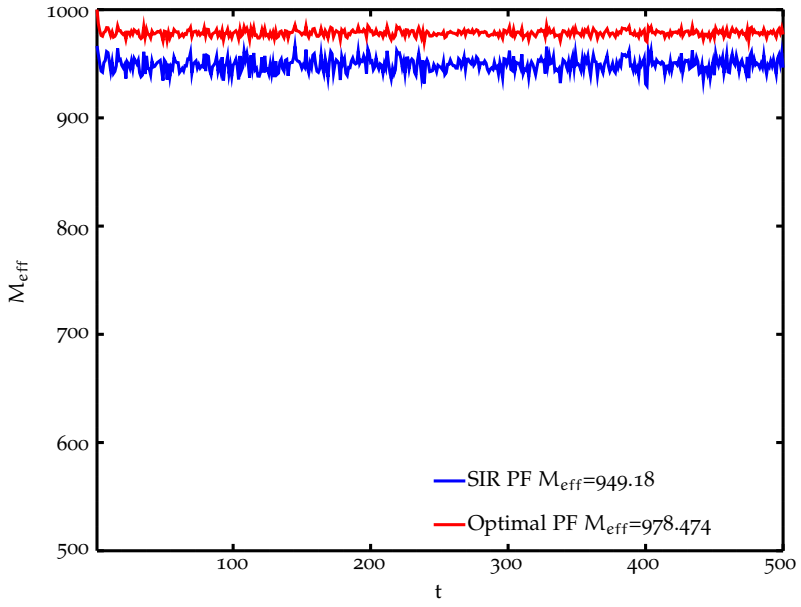


Figure 75: Effective particle size, $a = 0.65$, $\sigma_u = 0.5$.

$a = 0.65$, $\sigma_u = 1.5$

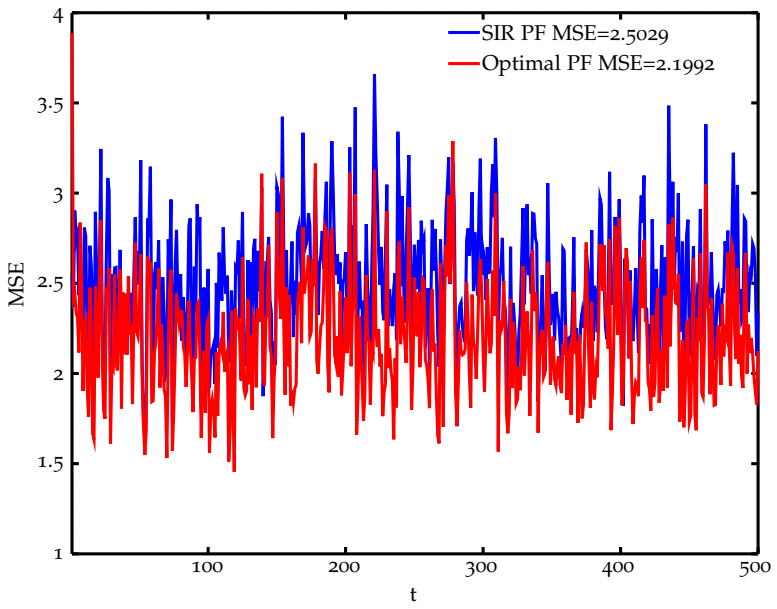


Figure 76: Averaged state filtering MSE, $a = 0.65$, $\sigma_u = 1.5$.

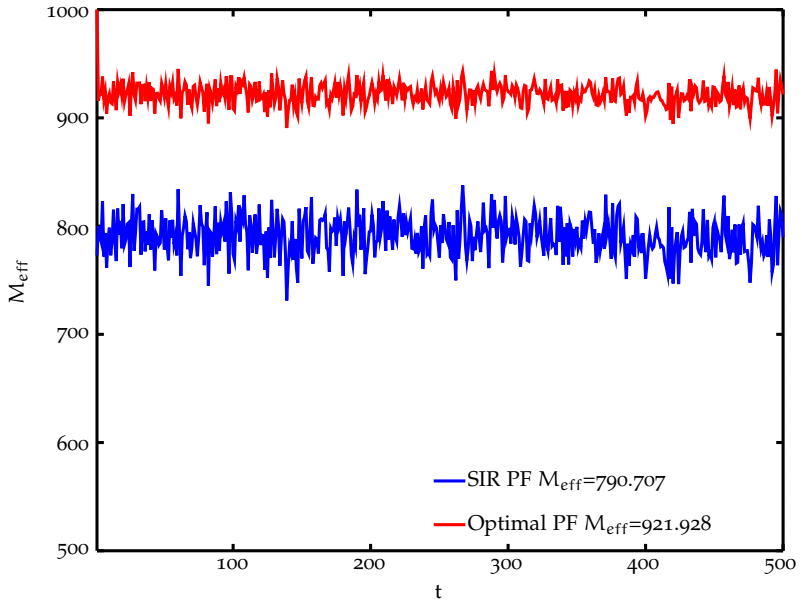


Figure 77: Effective particle size, $a = 0.65$, $\sigma_u = 1.5$.

$a = 0.65$, $\sigma_u = 2$

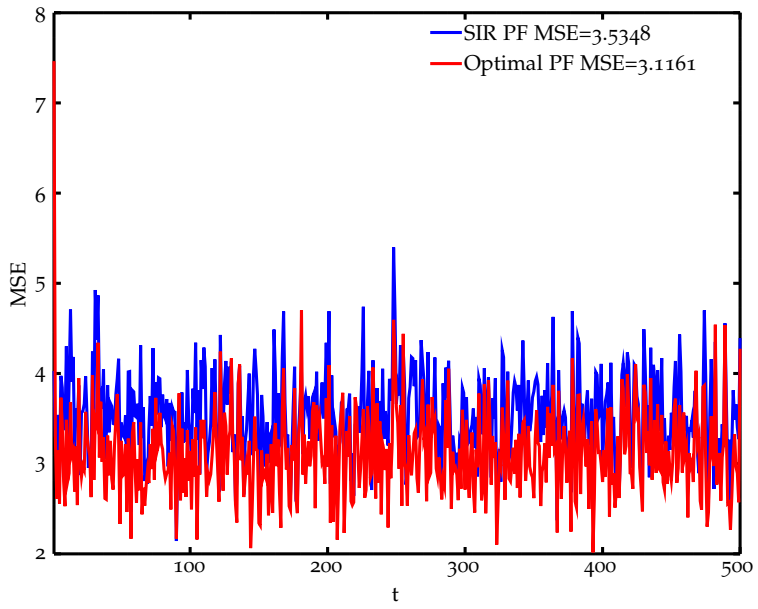


Figure 78: Averaged state filtering MSE, $a = 0.65$, $\sigma_u = 2$.

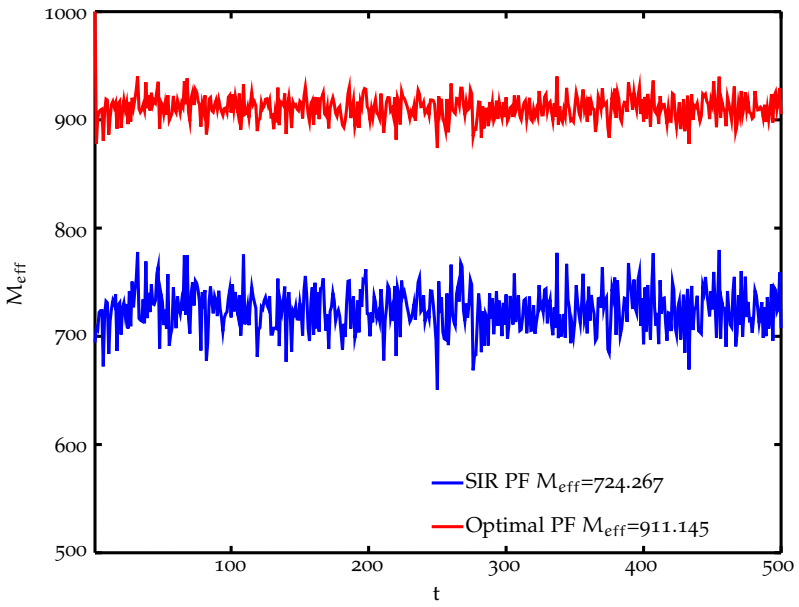


Figure 79: Effective particle size, $a = 0.65$, $\sigma_u = 2$.

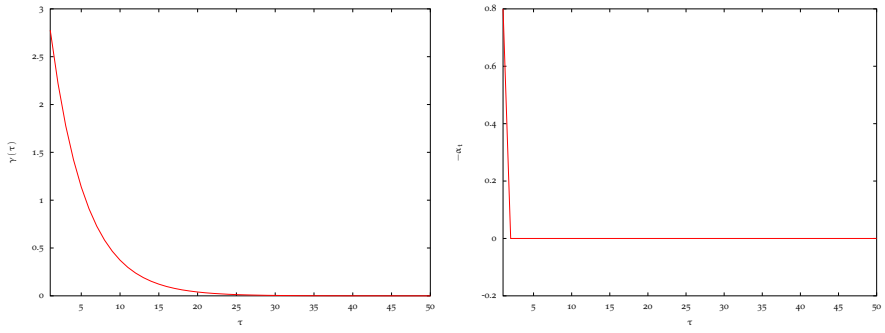
CONCLUSIONS

As expected, the performance of the optimal PF is consistently better than the SIR method. As illustrated by the effective particle size plots, the optimal PF is able to propagate particles towards regions of high probability, while it is more difficult for the SIR method to so. However, the difference in MSE highly depends on each model. That is, the more informative the transition density is, the better the SIR method performs.

B

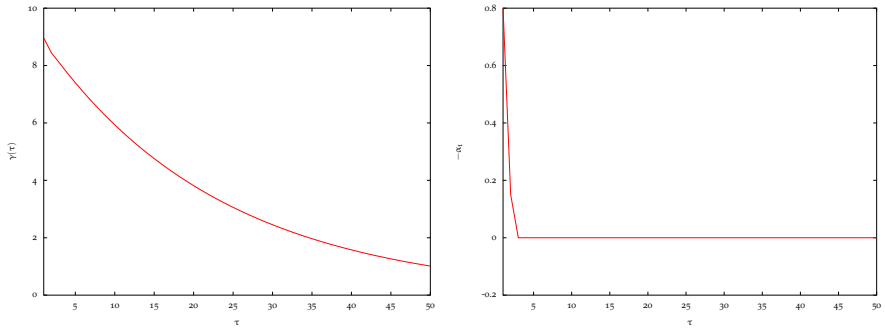
ARMA(p, q) PROCESSES: $\gamma(\tau)$ AND α_t

In this appendix, the autocovariance $\gamma(\tau)$ and $-\alpha_t = \widetilde{\Sigma}_t^{-1}\widetilde{\gamma}_t$ functions for different ARMA(p, q) models are illustrated.



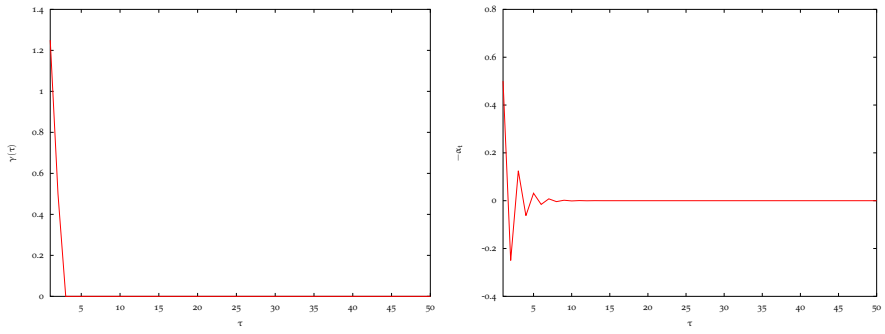
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 80: $\gamma(\tau)$ and $-\alpha_t$ for AR(1), $a = 0.8$.



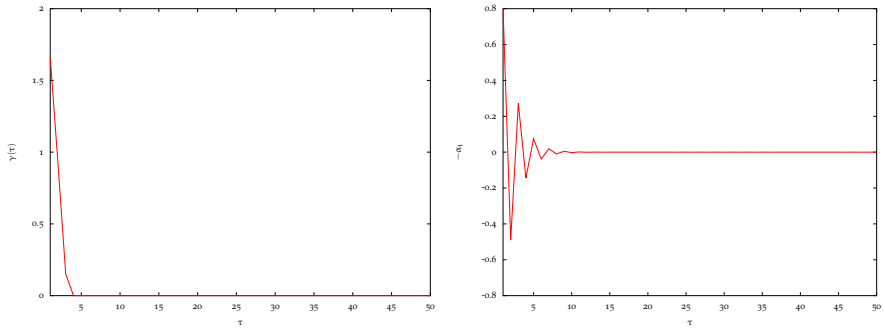
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 81: $\gamma(\tau)$ and $-\alpha_t$ for AR(2), $a_1 = 0.8$, $a_2 = 0.15$.



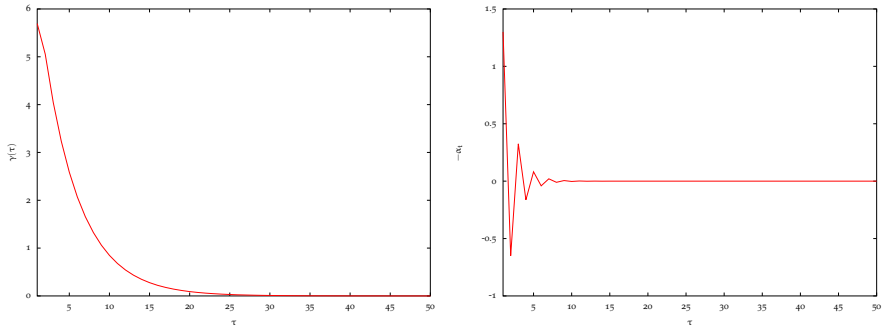
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 82: $\gamma(\tau)$ and $-\alpha_t$ for MA(1), $b = 0.5$.



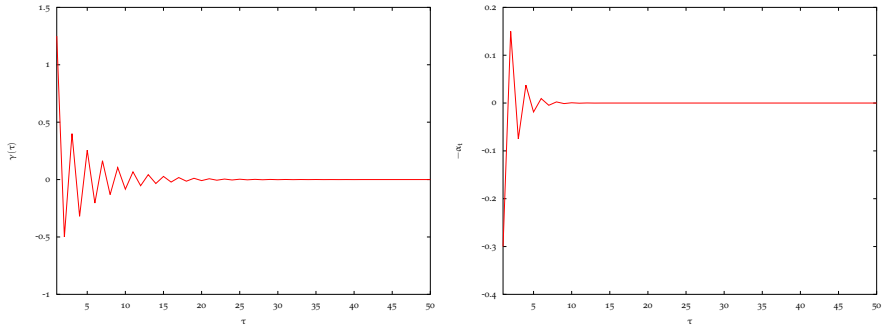
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 83: $\gamma(\tau)$ and $-\alpha_t$ for MA(2), $b_1 = 0.8$, $b_2 = 0.15$.



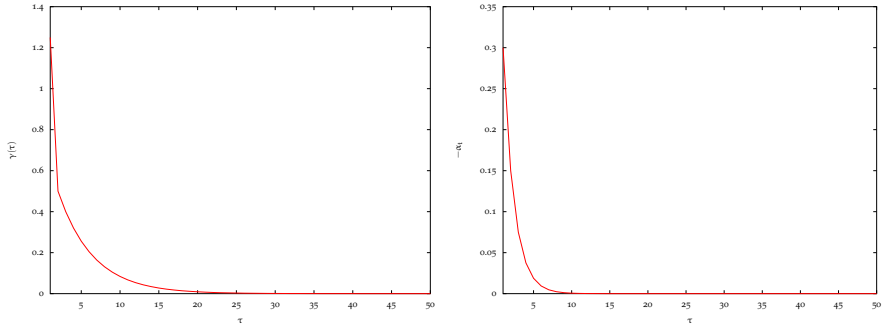
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 84: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(1, 1), $a_1 = 0.8$, $b_1 = 0.5$.



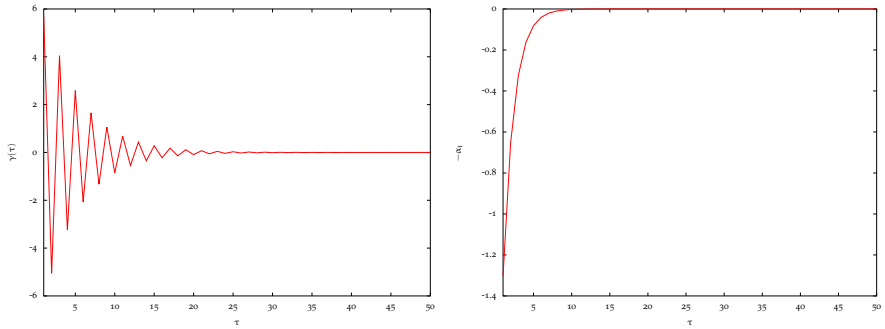
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 85: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,1)$, $a_1 = -0.8$, $b_1 = 0.5$.



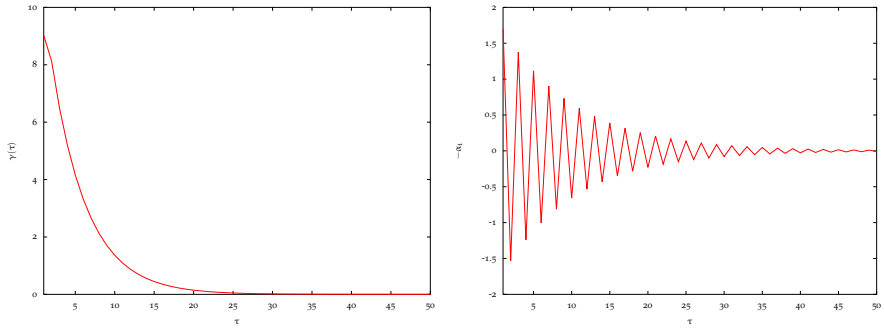
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 86: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,1)$, $a_1 = 0.8$, $b_1 = -0.5$.



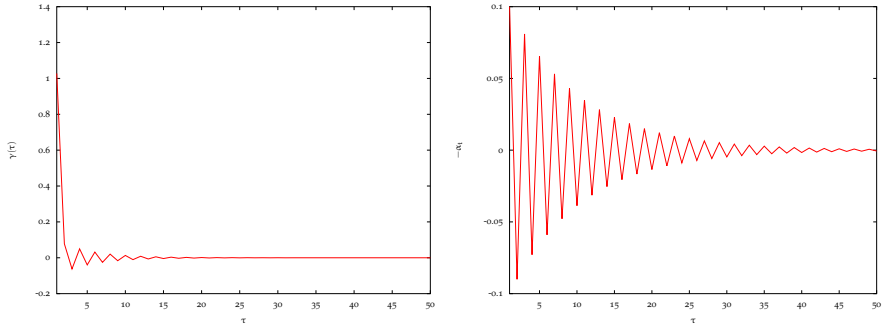
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 87: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,1)$, $a_1 = -0.8$, $b_1 = -0.5$.



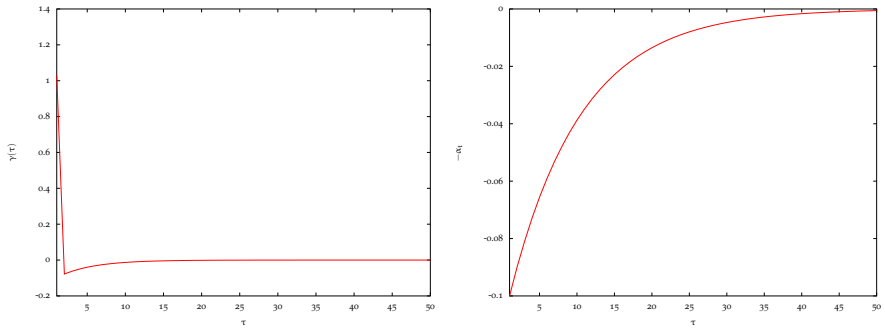
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 88: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(1, 1), $a_1 = 0.8$, $b_1 = 0.9$.



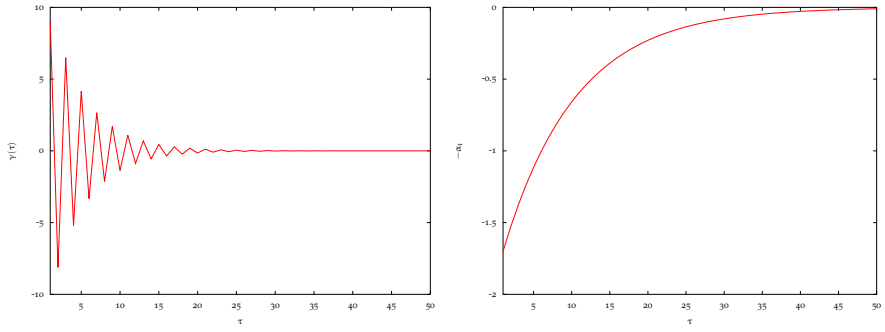
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 89: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(1, 1), $a_1 = -0.8$, $b_1 = 0.9$.



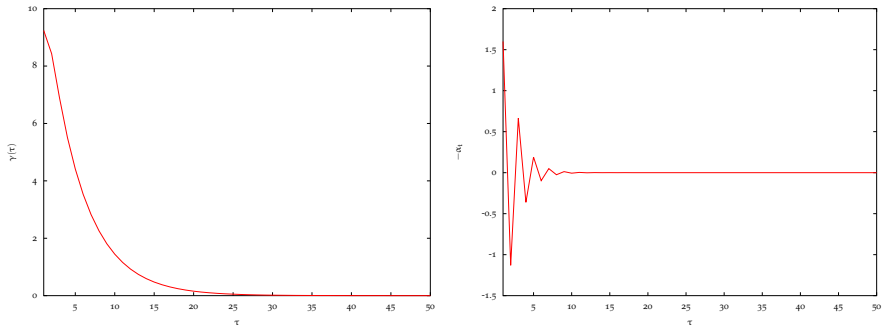
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 90: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(1, 1), $a_1 = 0.8$, $b_1 = -0.9$.



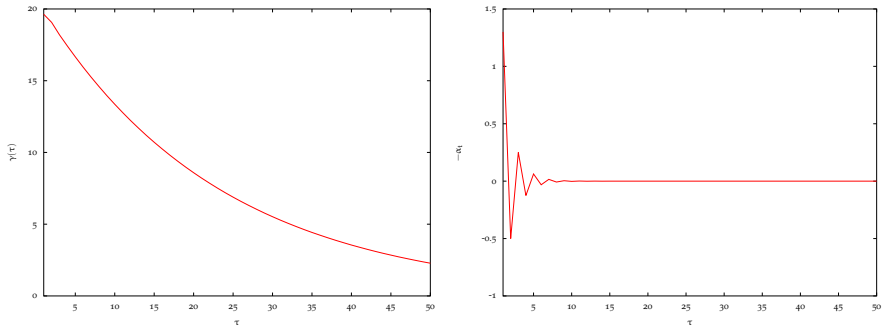
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 91: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,1)$, $a_1 = -0.8$, $b_1 = -0.9$.



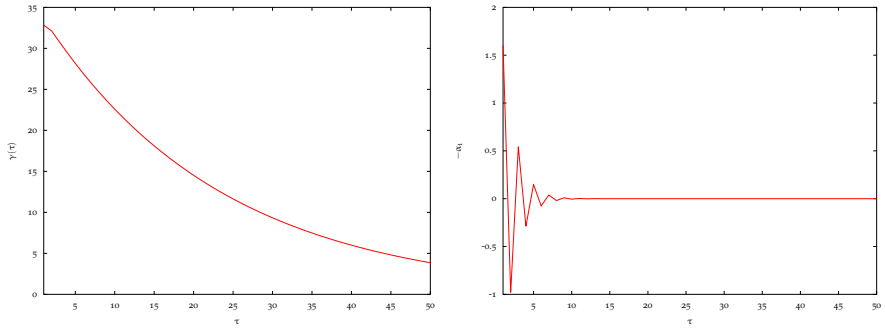
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 92: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,2)$, $a_1 = 0.8$, $b_1 = 0.8$, $b_2 = 0.15$.



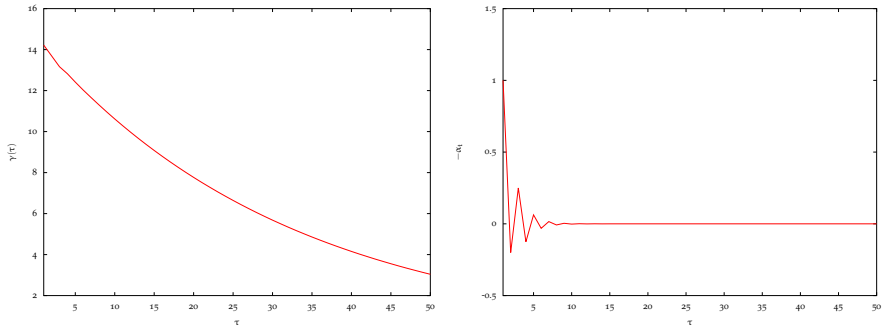
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 93: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(2,1)$, $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$.



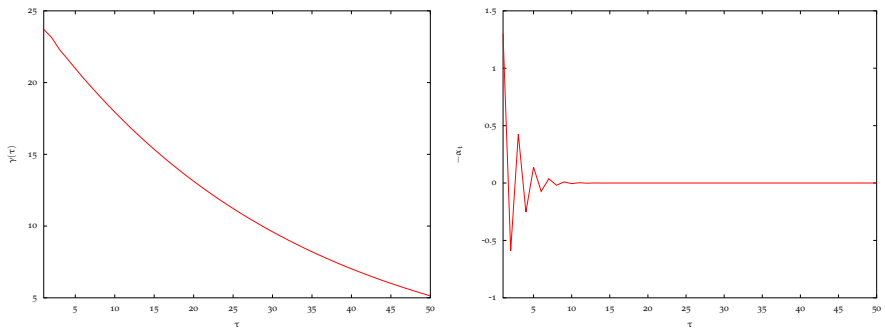
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 94: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(2, 2), $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.8$, $b_2 = 0.15$.



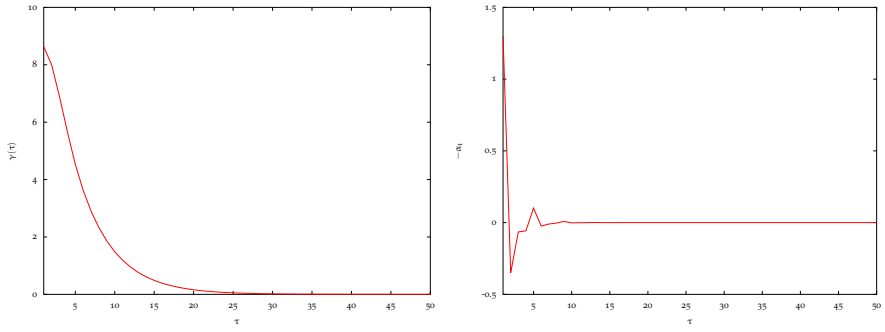
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 95: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(3, 1), $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.5$.



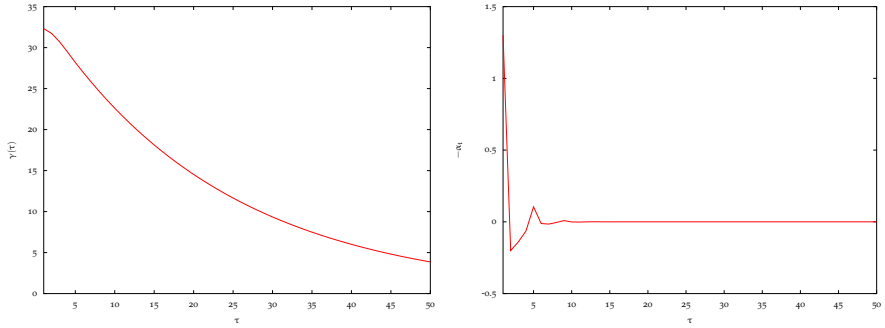
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 96: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(3, 2), $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.8$, $b_2 = 0.15$.



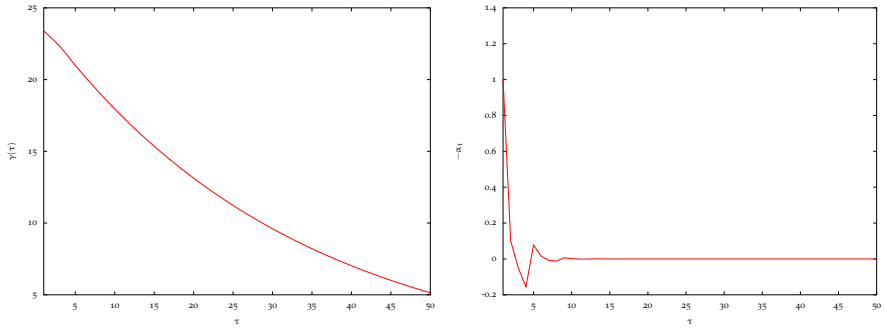
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 97: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(1,3)$, $a_1 = 0.8$, $b_1 = 0.5$, $b_2 = 0.3$, $b_3 = 0.15$.



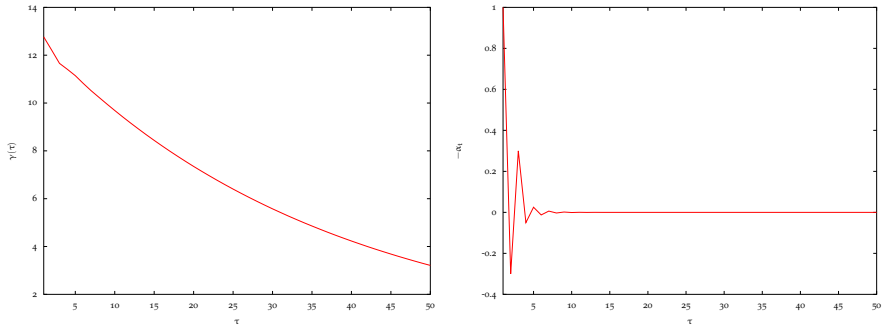
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 98: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(2,3)$, $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$, $b_2 = 0.3$, $b_3 = 0.15$.



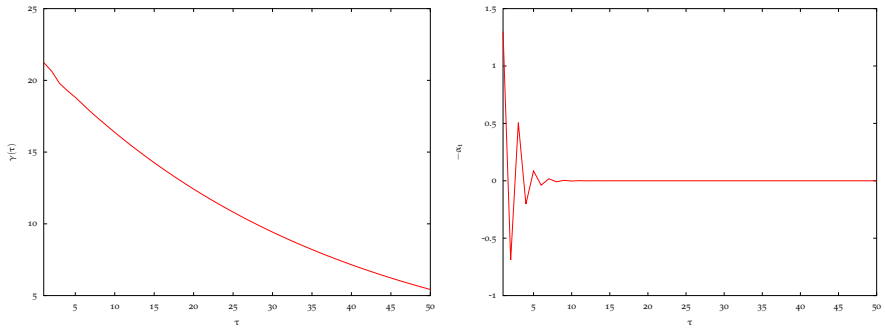
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 99: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(3,3)$, $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.5$, $b_2 = 0.3$, $b_3 = 0.15$.



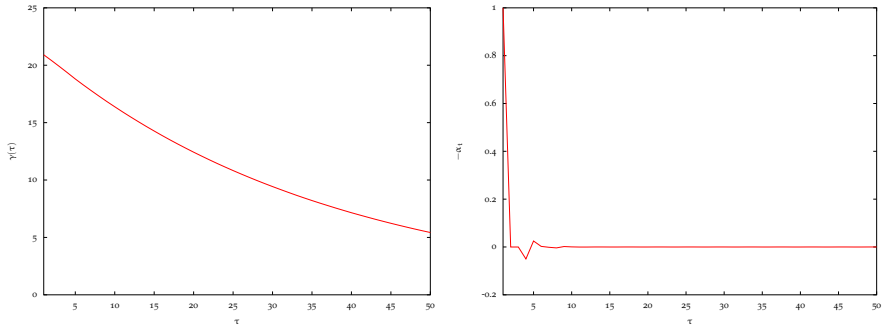
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 100: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(4,1)$, $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.5$.



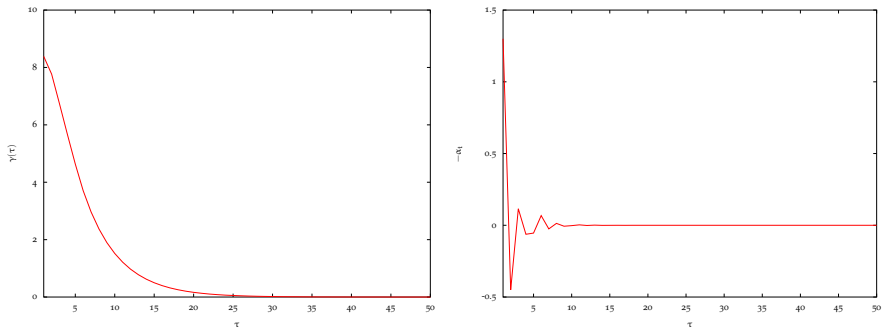
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 101: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(4,2)$, $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.8$, $b_2 = 0.15$.



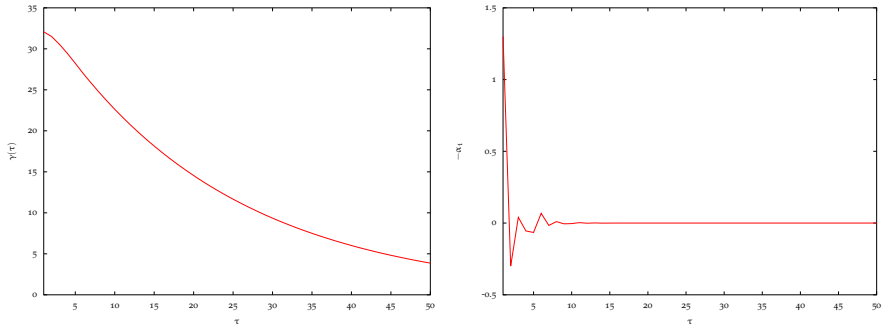
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 102: $\gamma(\tau)$ and $-\alpha_t$ for ARMA$(4,3)$, $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.5$, $b_2 = 0.3$, $b_3 = 0.15$.



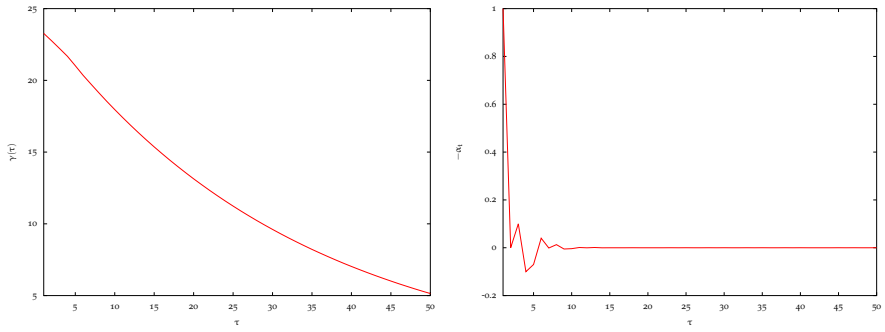
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 103: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(1,4), $a_1 = 0.8$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$.



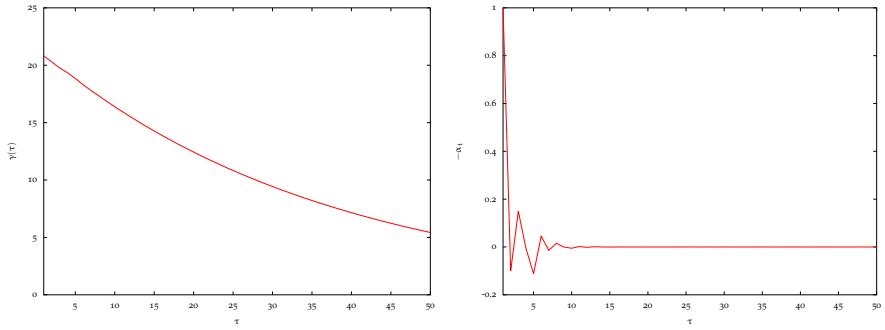
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 104: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(2,4), $a_1 = 0.8$, $a_2 = 0.15$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$.



(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 105: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(3,4), $a_1 = 0.5$, $a_2 = 0.3$, $a_3 = 0.15$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$.



(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 106: $\gamma(\tau)$ and $-\alpha_t$ for ARMA(4,4), $a_1 = 0.5$, $a_2 = 0.2$, $a_3 = 0.15$, $a_4 = 0.1$, $b_1 = 0.5$, $b_2 = 0.2$, $b_3 = 0.15$, $b_4 = 0.1$.

C

APPROXIMATED KALMAN FILTER

In this appendix, we study how to apply a KF-based solution to the stochastic ARMA log-volatility model. The observations follow a SV model, whose log-volatility is described by a latent ARMA time-series.

Mathematically, it is represented by a state-space model of the form

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j}, \\ y_t = e^{\frac{x_t}{2}} v_t, \end{cases} \tag{219}$$

with $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$.

For the latent stationary ARMA state, it has already been established in Chapter 3 that the transition density can be readily derived, where knowledge of the ARMA parameters suffices for determination of the sufficient statistics.

The transition density is Gaussian and depends linearly on previous states. Thus, the Kalman equations for updating the state can be derived. However, the nonlinear observation equation requires further consideration.

In order to deal with nonlinearities in the state-space formulation of the KF, the EKF has been proposed [3]. It consists of linearizing the problem around point estimates of the current mean and covariance. To do so, it resorts to the computation of the partial derivatives of the measurement function, i.e., $h(x_t, v_t) = e^{\frac{x_t}{2}} v_t$ in our case.

If the linearization of the SV model is performed following the guidelines of the EKF precisely, the resulting Kalman gain is null:

$$\begin{aligned} K_t &= P_t^- H_t^\top \left(H_t P_t^- h_t^\top + V_t R_t V_t^\top \right)^{-1} = 0 \\ \text{because } H_t &= \frac{\delta h}{\delta x_t}(\hat{x}_t^-, 0) = \frac{1}{2} e^{\frac{x_t}{2}} 0 = 0 \end{aligned} \tag{220}$$

Consequently, direct implementation of the EKF for the SV model is not successful, as the estimates are never updated with information from the observables. In such models, the EKF simply updates its state estimates by following the stationary transition density. Thus, it does not successfully track the state,

but converges to the expected stationary mean (i.e., $\hat{x}_t \to 0$ as $t \to \infty$) [105].

An alternative approach is now suggested, where the observation equation is rewritten so that an additive noise form is obtained.

Let us transform the observation equation

$$y_t = e^{\frac{x_t}{2}} v_t \;\to\; y_t^2 = e^{x_t} v_t^2 \;\to\; \log(y_t^2) = x_t + \log(v_t^2), \quad (221)$$

and derive the density of the transformed noise:

1. If v_t is a standard Gaussian random variable, then $w_t = v_t^2$ follows a Chi-squared distribution with one degree of freedom:

$$\text{If } v_t \sim \mathcal{N}(0,1), \text{ then } w_t = v_t^2 \sim \chi_1^2,$$
$$w_t \sim f_{w_t}(w_t) = \chi_1^2 = \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} w_t^{-\frac{1}{2}} e^{-\frac{w_t}{2}}.$$

2. For the nonlinear transformation $z_t = \log(w_t)$, the resulting density is derived in two complementary forms:
   - By using the general PDF transformation rule:

$$\text{If, } z_t = \log(w_t), \text{ then } w_t = e^{z_t} \text{ and } \frac{\delta z_t}{\delta w_t} = \frac{1}{w_t}$$
$$\text{Leading to } z_t \sim f_{z_t}(z_t) = \left|\frac{1}{\frac{\delta z_t}{\delta w_t}}\right| f_{w_t}(e^{z_t}) = e^{z_t} f_{w_t}(e^{z_t})$$
$$f_{z_t}(z_t) = e^{z_t} \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} e^{-\frac{z_t}{2}} e^{-\frac{e^{z_t}}{2}} = \frac{1}{\sqrt{(2\pi)}} e^{\left(\frac{z_t}{2} - \frac{e^{z_t}}{2}\right)}.$$

- By using the general CDF transformation rule:

$$F_{Z_t}(z_t) = P(Z_t \leqslant z_t) = P(\log(w_t) \leqslant z_t) = P(w_t \leqslant e^{z_t}) = F_{W_t}(e^{z_t})$$

Since $F_{W_t}(w_t) = \frac{1}{\Gamma(\frac{1}{2})}\gamma\left(\frac{1}{2}, \frac{w_t}{2}\right)$

where $\gamma(s,x)$ is the lower imcomplete Gamma function

$$\gamma(s,x) = \int_0^x t^{s-1}\, e^{-t}\, dt$$

Because $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\gamma\left(\frac{1}{2}, x\right) = \sqrt{\pi}\,\mathrm{erf}\left(\sqrt{x}\right)$,

$$F_{W_t}(w_t) = \mathrm{erf}\left(\sqrt{\frac{w_t}{2}}\right).$$

Then $F_{Z_t}(z_t) = F_{W_t}(e^{z_t}) = \mathrm{erf}\left(\sqrt{\frac{e^{z_t}}{2}}\right) = \mathrm{erf}\left(\frac{1}{\sqrt{2}} e^{\frac{z_t}{2}}\right)$ and

$$f_{z_t}(z_t) = \frac{\delta F_{Z_t}(z_t)}{\delta z_t} = \frac{\delta \mathrm{erf}\left(\sqrt{\frac{e^{z_t}}{2}}\right)}{\delta z_t} \frac{\delta \frac{1}{\sqrt{2}} e^{\frac{z_t}{2}}}{\delta z_t}$$
$$= \frac{2}{\sqrt{\pi}} e^{-\frac{e^{z_t}}{2}} \cdot \frac{e^{\frac{z_t}{2}}}{2\sqrt{2}} = \frac{1}{\sqrt{2\pi}} e^{\frac{z_t}{2} - \frac{e^{z_t}}{2}}.$$

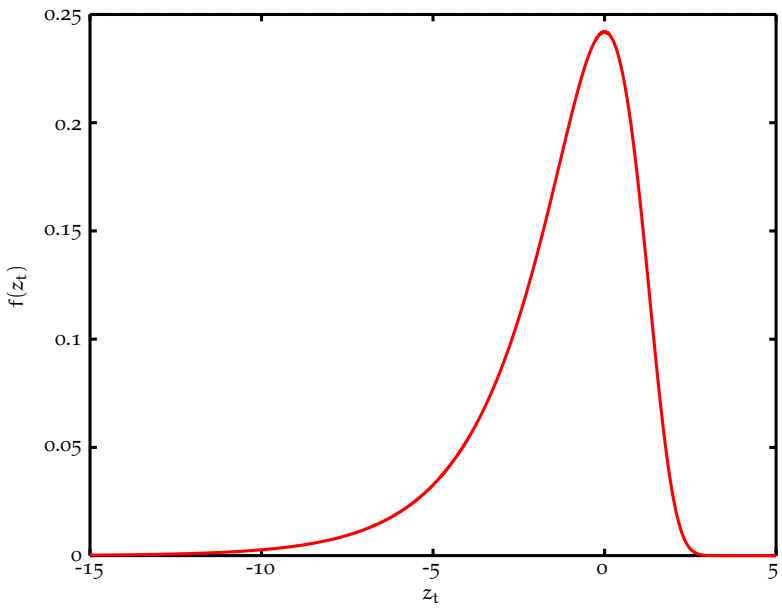
The transformed random variable z_t thus follows a non-Gaussian density $f_{z_t}(z_t) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(z_t - e^{z_t})}$ (illustrated in Figure 107) with the following mean and variance:

- Mean: $\mathbb{E}\{z_t\} = -1.2703628454615$
- Variance: $\mathrm{Var}\{z_t\} = 4.9348022005446$

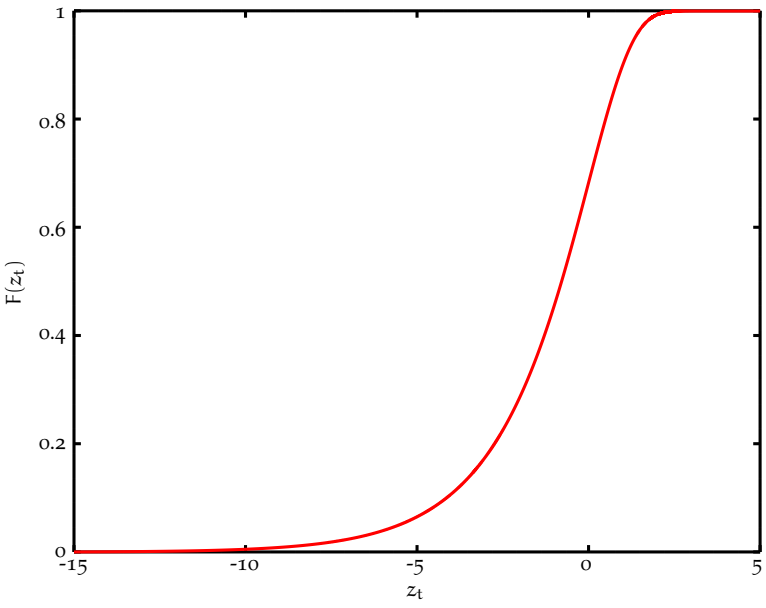
All in all, an approximated KF is derived for the modified log-volatility space model

$$\begin{cases} x_t = \sum_{i=1}^{p} a_i x_{t-i} + u_t + \sum_{j=1}^{q} b_j u_{t-j}, \\ \log(y_t^2) = x_t + z_t, \qquad \text{where } z_t \sim f_{z_t}(z_t) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(z_t - e^{z_t})}, \end{cases} \tag{222}$$

under the Gaussian approximation to the transformed noise, i.e., $z_t \approx \mathcal{N}(\mathbb{E}\{z_t\}, \mathrm{Var}\{z_t\})$. The suggested algorithm is described in Table 42.



(a) PDF of z_t.



(b) CDF of z_t.

Figure 107: PDF (top) and CDF (bottom) of variable $z_t = \log(w_t)$.

APPROXKF FOR THE MODIFIED SV MODEL

1. Project the state ahead

$$\hat{x}_t^- = \mu_{x_t|x_{t-\tau:t-1}} \hat{x}_{t-1:t-\tau}. \tag{223}$$

2. Project the error covariance ahead

$$P_t^- = \mu_{x_t|x_{t-\tau:t-1}} P_{t-1} \mu_{x_t|x_{t-\tau:t-1}}^\top + \Sigma_{x_t|x_{t-\tau:t-1}}. \tag{224}$$

3. Compute the Kalman gain

$$K_t = P_t^- \left(P_t^- + \mathrm{Var}\{z_t\}\right)^{-1}. \tag{225}$$

4. Update the state estimate with the transformed measurement $\log(y_t^2)$

$$\hat{x}_t = \hat{x}_t^- + Kt \left(\log(y_t^2) - \hat{x}_t^- - \mathbb{E}\{z_t\}\right). \tag{226}$$

5. Update the state error covariance

$$P_t = (1 - K_t) P_t^-. \tag{227}$$

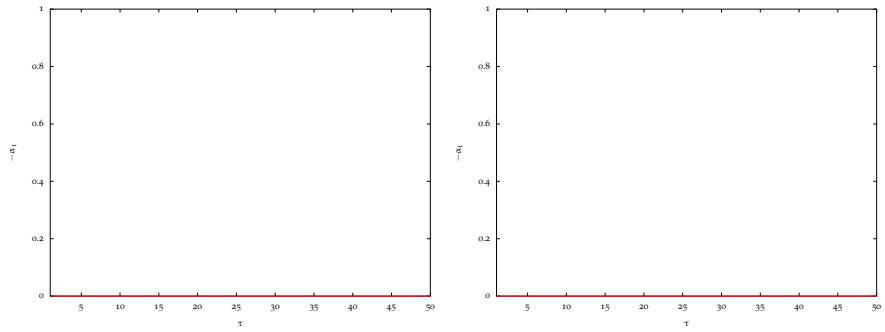
Table 42: Approximated Kalman Filter for the modified stochastic log-volatility model.

D

FGP AND FARIMA(0, d, 0) MODELS: $\gamma(\tau)$ AND α_t

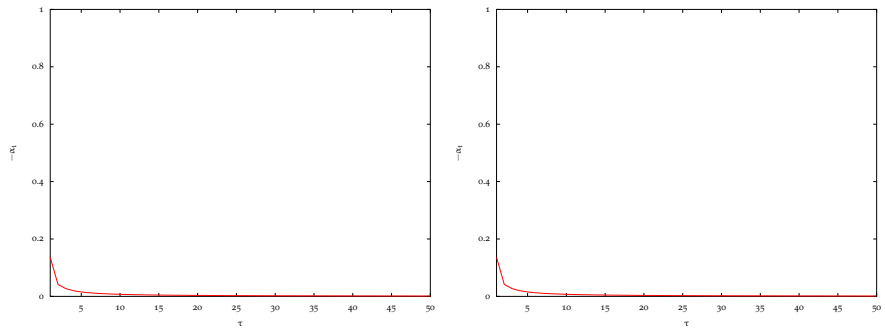
FGP MODELS: $\gamma(\tau)$ AND α_t

In this section, the autocovariance $\gamma(\tau)$ and $-\alpha_t = \widetilde{C}_t^{-1}\widetilde{C}_{t+1}^{\top}$ functions for different fGp models are illustrated.



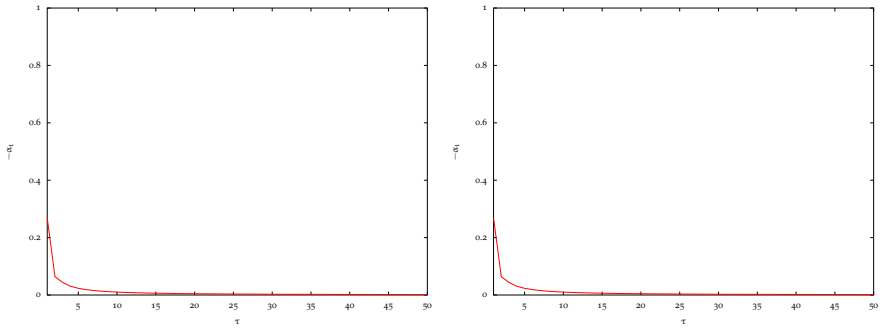
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 108: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.5$.



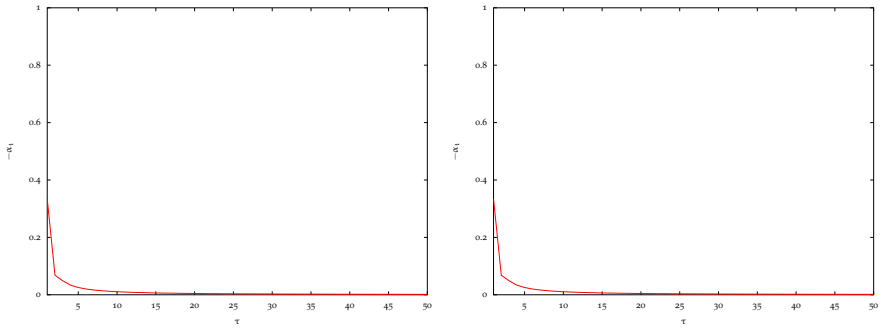
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 109: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.6$.



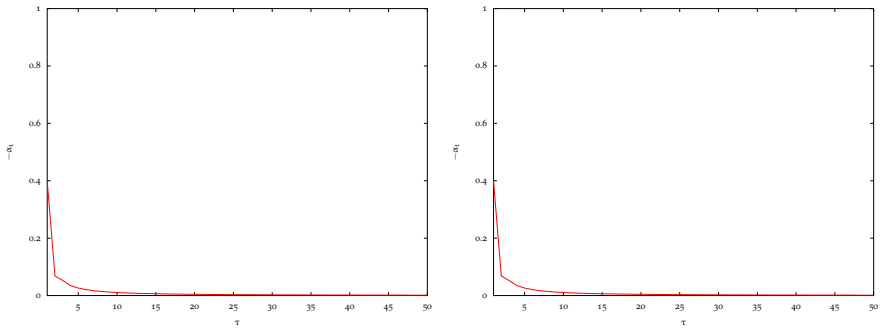
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 110: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.7$.



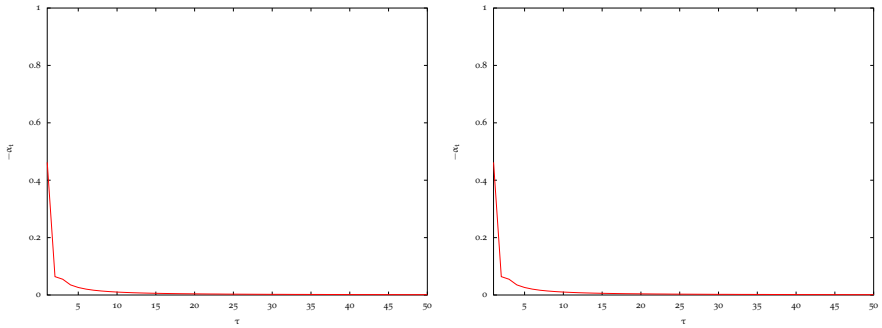
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 111: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.75$.



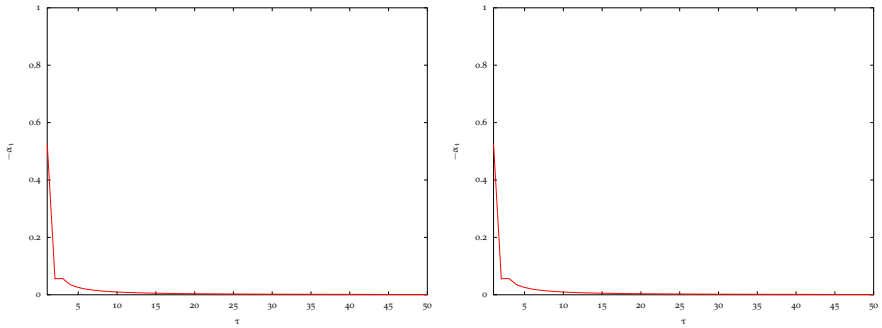
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 112: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.8$.



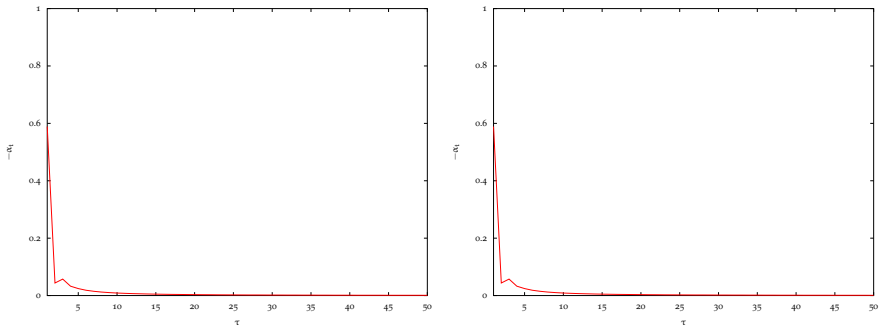
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 113: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.85$.



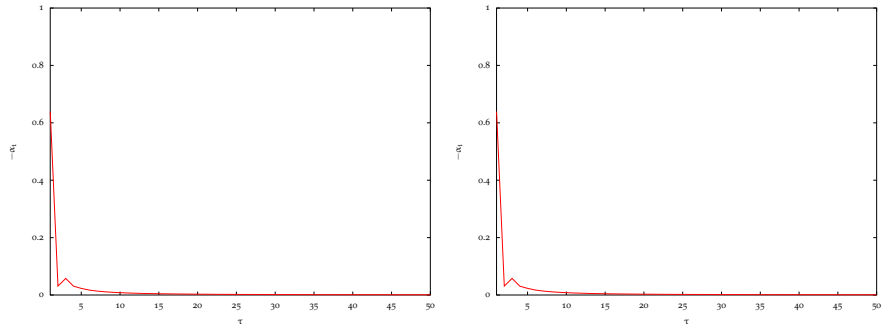
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 114: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.9$.



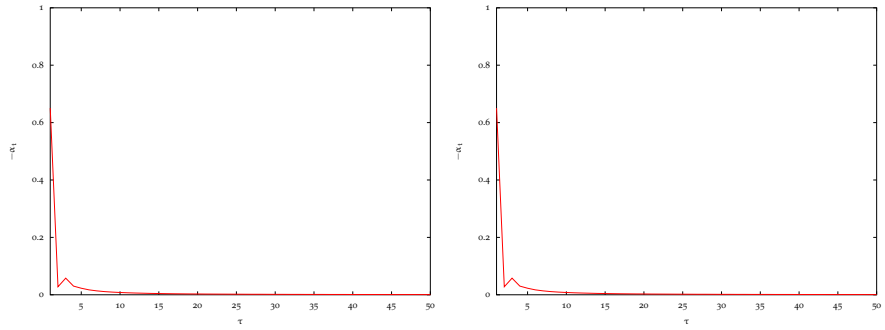
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 115: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.95$.



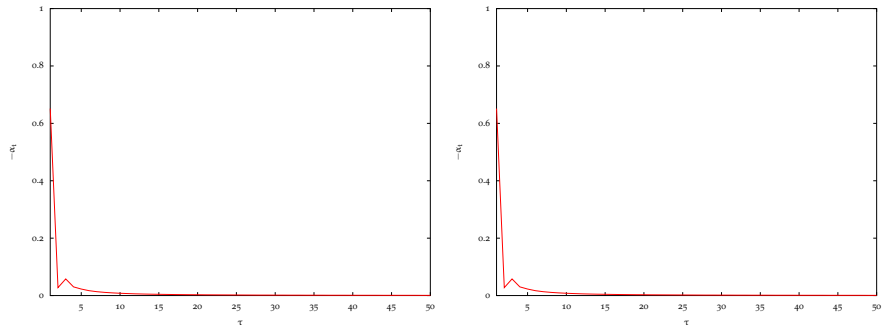
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 116: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.99$.



(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 117: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.999$.

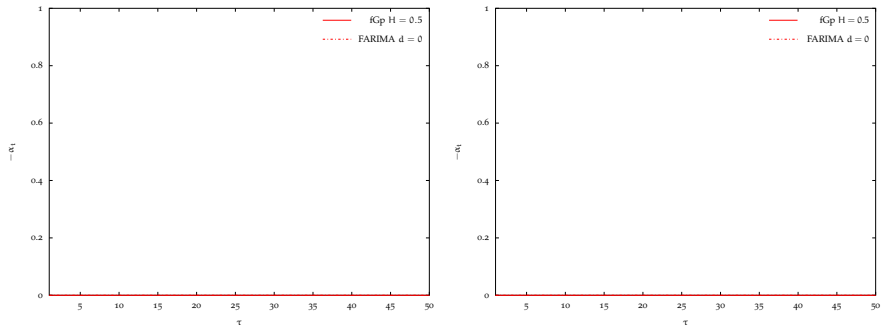


(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 118: $\gamma(\tau)$ and $-\alpha_t$ for fGp, $H = 0.9999$.

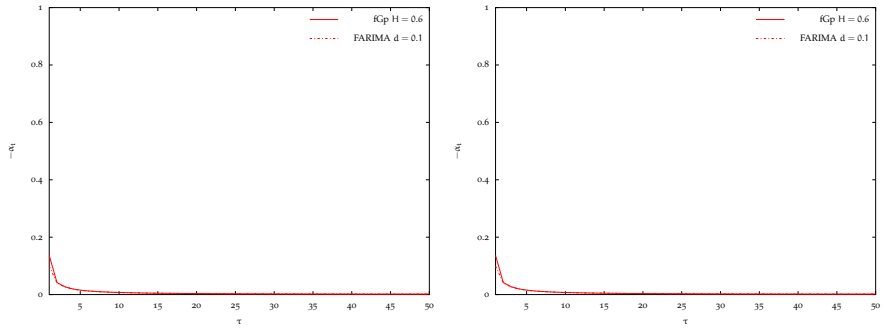
FGP AND FARIMA(0, d, 0) MODEL COMPARISON: $\gamma(\tau)$ AND α_t

In this appendix, the autocovariance $\gamma(\tau)$ and $-\alpha_t = \widetilde{\Sigma}_t^{-1}\widetilde{\gamma}_t$ functions for different fGp and FARIMA(0, d, 0) models are compared.



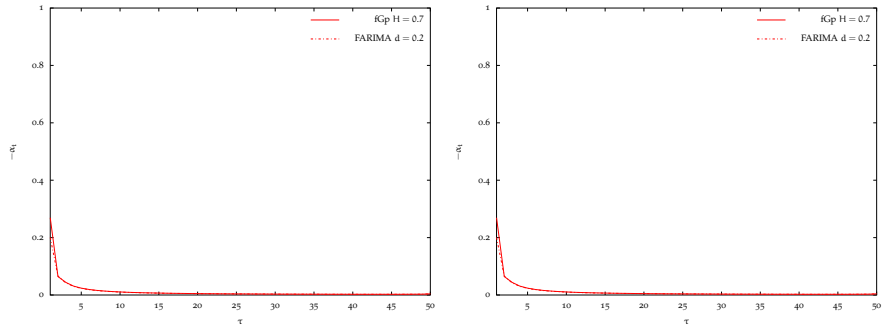
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 119: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA(0, d, 0), H = 0.5.



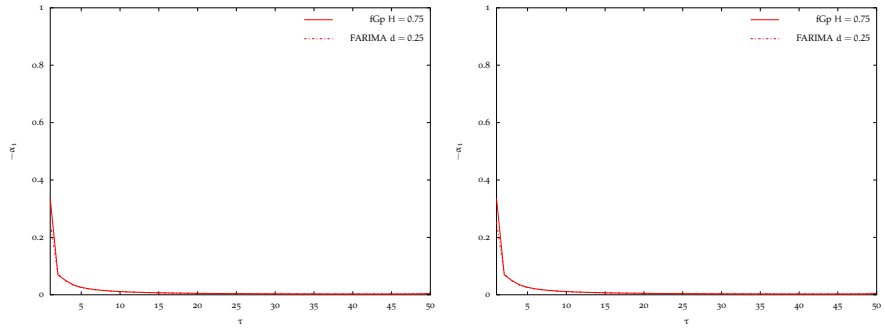
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 120: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA(0, d, 0), H = 0.6.



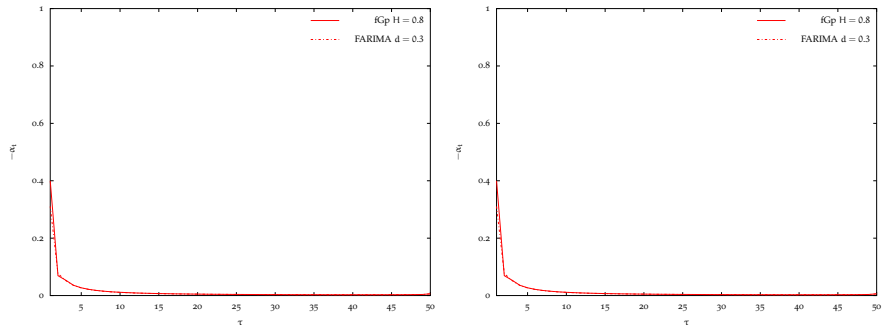
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 121: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.7$.



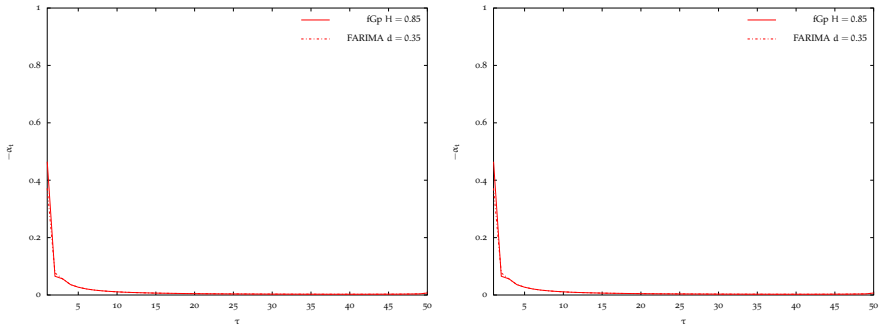
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 122: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.75$.



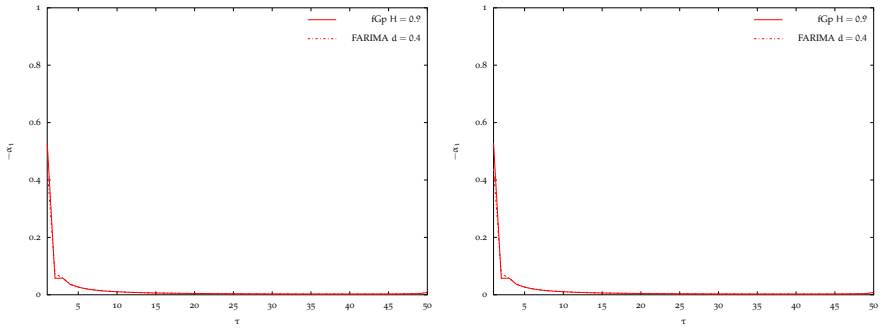
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 123: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.8$.



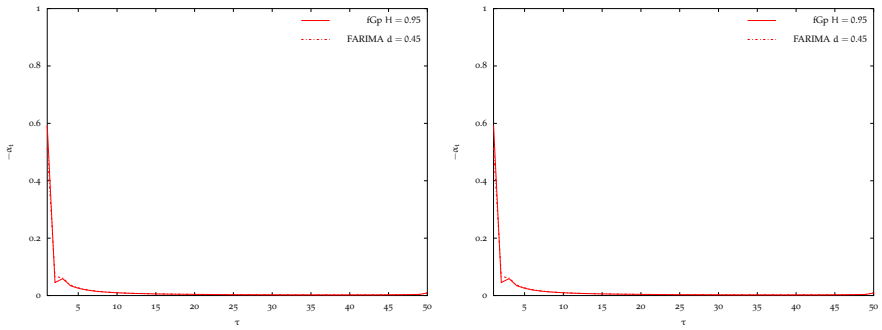
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 124: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.85$.



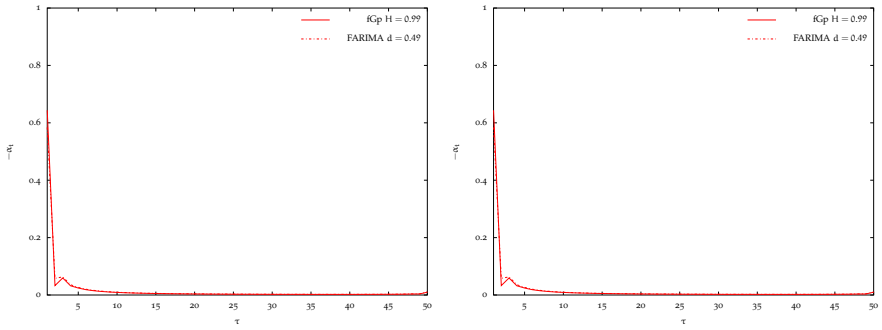
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 125: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.9$.



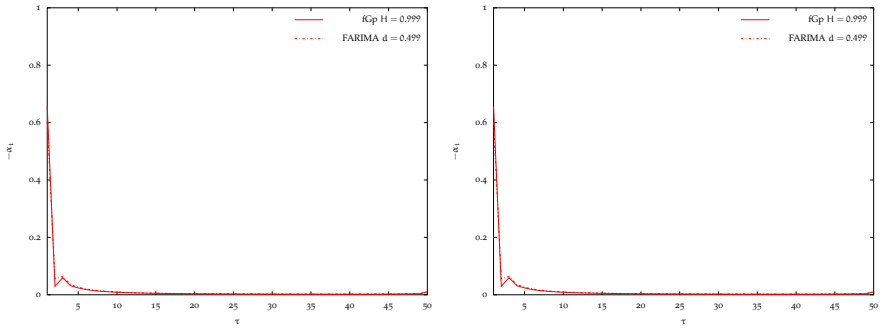
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 126: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.95$.



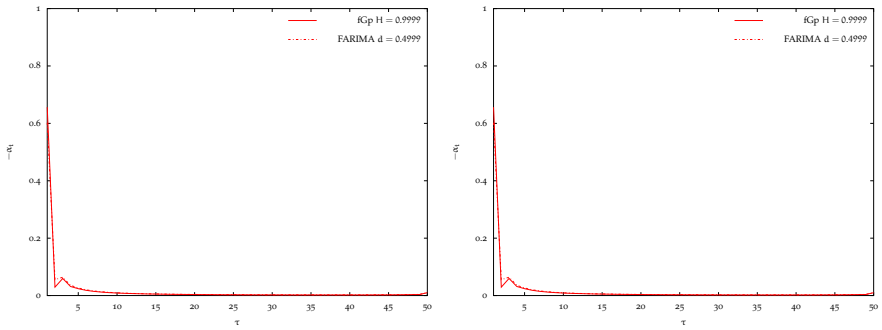
(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 127: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.99$.



(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 128: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.999$.



(a) $\gamma(\tau)$ (b) $-\alpha_t$

Figure 129: $\gamma(\tau)$ and $-\alpha_t$ for fGp and FARIMA$(0, d, 0)$, $H = 0.9999$.

E

DERIVATION OF THE MARGINALIZED GAUSSIAN DISTRIBUTIONS

Let $x \in \mathbb{R}^d$ be generated according to $x \sim f(x|\mu, \Sigma) = \mathcal{N}(x|\mu, \Sigma)$, with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. We are interested in integrating out the parameters μ and Σ, i.e.,

$$f(x) = \int\int f(x|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma, \tag{228}$$

where $f(\mu, \Sigma)$ is the prior of the unknown μ and Σ. We use conjugate priors for easiness of the analytical derivations.

UNKNOWN MEAN AND COVARIANCE

We start with the conjugate prior of the covariance matrix, which is the inverse Wishart distribution $\mathrm{IW}_d(\Sigma|\nu_w, \Lambda)$, where $\nu_w > d-1$ represents degrees of freedom and $\Lambda \in \mathbb{R}^{d\times d}$ is a scale matrix. Its probability density function is given by

$$f(\Sigma) = \mathrm{IW}_d(\Sigma|\nu_w, \Lambda) = \frac{1}{Z_{\mathrm{IW}}} |\Sigma|^{-\frac{\nu_w+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\Lambda\}}, \tag{229}$$

where

$$Z_{\mathrm{IW}} = 2^{\frac{\nu_w d}{2}} |\Lambda|^{-\frac{\nu_w}{2}} \Gamma_d\left(\frac{\nu_w}{2}\right). \tag{230}$$

Let the prior of μ given Σ be Gaussian with hyper-parameters η and κ, i.e.,

$$f(\mu|\eta, \Sigma) = \mathcal{N}(\mu|\eta, \frac{\Sigma}{\kappa}) = (2\pi)^{-\frac{d}{2}} \left|\frac{\Sigma}{\kappa}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mu-\eta)^\top \left(\frac{\Sigma}{\kappa}\right)^{-1}(\mu-\eta)}. \tag{231}$$

Then, from $f(\mu, \Sigma) = f(\mu|\Sigma)f(\Sigma)$, we deduce that the joint conjugate prior is a normal-inverse-Wishart distribution, i.e.,

$$\begin{aligned} f(\mu, \Sigma|\eta, \kappa, \Lambda, \nu_w) &= \mathrm{NIW}(\mu, \Sigma|\eta, \kappa, \nu_w, \Lambda) \\ &= \frac{1}{Z_{\mathrm{NIW}}} |\Sigma|^{-\frac{\nu_w+d+2}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\Lambda\} - \frac{\kappa}{2}(\mu-\eta)^\top \Sigma^{-1}(\mu-\eta)}. \end{aligned} \tag{232}$$

We find the normalizing constant Z_{NIW} by computing

$$Z_{\text{NIW}} = \int\int |\Sigma|^{-\frac{\nu_w+d+2}{2}} e^{-\frac{1}{2}\text{tr}\{\Sigma^{-1}\Lambda\}-\frac{\kappa}{2}(\mu-\eta)^\top\Sigma^{-1}(\mu-\eta)} d\mu d\Sigma. \tag{233}$$

First, we separate the integrals for μ and Σ

$$\begin{aligned} Z_{\text{NIW}} = &\int |\Sigma|^{-\frac{\nu_w+d+1}{2}} e^{-\frac{1}{2}\text{tr}\{\Sigma^{-1}\Lambda\}} \left(\frac{\kappa}{2\pi}\right)^{-\frac{d}{2}} \\ &\times \left(\int \left(\frac{\kappa}{2\pi}\right)^{\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{\kappa}{2}(\mu-\eta)^\top\Sigma^{-1}(\mu-\eta)} d\mu\right) d\Sigma. \end{aligned} \tag{234}$$

The inner integral with respect to μ integrates to one and so

$$\begin{aligned} Z_{\text{NIW}} &= \left(\frac{\kappa}{2\pi}\right)^{-\frac{d}{2}} \int |\Sigma|^{-\frac{\nu_w+d+1}{2}} e^{-\frac{1}{2}\text{tr}\{\Sigma^{-1}\Lambda\}} d\Sigma \\ &= \left(\frac{\kappa}{2\pi}\right)^{-\frac{d}{2}} Z_{\text{IW}}. \end{aligned} \tag{235}$$

Thus,

$$Z_{\text{NIW}} = \left(\frac{\kappa}{2\pi}\right)^{-\frac{d}{2}} 2^{\frac{\nu_w d}{2}} |\Lambda|^{-\frac{\nu_w}{2}} \Gamma_d\left(\frac{\nu_w}{2}\right). \tag{236}$$

With this prior, we can integrate out the parameters of the data distribution and find its marginal as follows:

$$\begin{aligned} f(x) &= \int\int f(x|\mu,\Sigma) f(\mu,\Sigma|\eta,\kappa,\Lambda,\nu_w) d\mu d\Sigma \\ &= \int\int \mathcal{N}(x|\mu,\Sigma)\text{NIW}(\mu,\Sigma|\eta,\kappa,\nu_w,\Lambda) d\mu d\Sigma \\ &= \int\int \mathcal{N}(x|\mu,\Sigma)\mathcal{N}\left(\mu\middle|\eta,\frac{\Sigma}{\kappa}\right)\text{IW}(\Sigma|\nu_w,\Lambda) d\mu d\Sigma \\ &= \int \text{IW}(\Sigma|\nu_w,\Lambda)\left(\int \mathcal{N}(x|\mu,\Sigma)\mathcal{N}\left(\mu\middle|\eta,\frac{\Sigma}{\kappa}\right) d\mu\right) d\Sigma. \end{aligned} \tag{237}$$

We can readily solve the inner integral and obtain the resulting marginal Gaussian

$$\int \mathcal{N}(x|\mu,\Sigma)\mathcal{N}(\mu|\eta,\frac{\Sigma}{\kappa}) d\mu = \mathcal{N}\left(x\middle|\eta,\Sigma\left(\frac{1+\kappa}{\kappa}\right)\right). \tag{238}$$

Finally,

$$
\begin{aligned}
f(x) &= \int \left(\int \mathcal{N}(x|\mu, \Sigma) \mathcal{N}\left(\mu \middle| \eta, \frac{\Sigma}{\kappa}\right) d\mu \right) \mathrm{IW}(\Sigma|\nu_w, \Lambda) d\Sigma \\
&= \int \mathcal{N}\left(x \middle| \eta, \Sigma\left(\frac{1+\kappa}{\kappa}\right)\right) \mathrm{IW}(\Sigma|\nu_w, \Lambda) d\Sigma \\
&\propto \int \left(\frac{1+\kappa}{\kappa}\right)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\eta)^\top \Sigma^{-1}\left(\frac{\kappa}{1+\kappa}\right)(x-\eta)} \\
&\quad \times |\Sigma|^{-\frac{\nu_w+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\Lambda\}} d\Sigma \\
&\propto \int |\Sigma|^{-\frac{\nu_w+d+2}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\left(\Lambda + \frac{\kappa}{1+\kappa}(x-\eta)(x-\eta)^\top\right)\}} d\Sigma \qquad (239) \\
&\propto \int \mathrm{IW}_d\left(\Sigma \middle| \nu_w + 1, \Lambda + \frac{\kappa}{1+\kappa}(x-\eta)(x-\eta)^\top\right) d\Sigma \\
&\propto \left|\Lambda + \frac{\kappa}{1+\kappa}(x-\eta)(x-\eta)^\top\right|^{-\frac{\nu_w+1}{2}} \\
&\propto \left|1 + (x-\eta)^\top \left(\frac{1+\kappa}{\kappa}\Lambda\right)^{-1} (x-\eta)\right|^{-\frac{\nu_w+1}{2}},
\end{aligned}
$$

where we have used Sylvester's determinant theorem, which states that for any invertible $d \times d$ matrix we have $|X + AB| = |X||I_n + BX^{-1}A|$.

Recall that if $x \in \mathbb{R}^d$ has a multivariate t-distribution, then

$$
f(x) \propto \left(1 + \frac{1}{\nu}(x-\mu)^\top R^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}}, \qquad (240)
$$

where ν represents the degrees of freedom and d is the dimension of x. Thus, by comparison of (239) and (240), we conclude that the joint marginal density is the following multivariate t-distribution:

$$
f(x) = \mathcal{T}_{\nu_w - d + 1}\left(x \middle| \eta, \frac{(1+\kappa)\Lambda}{\kappa(\nu_w - d + 1)}\right). \qquad (241)
$$

KNOWN MEAN, UNKNOWN COVARIANCE

The marginal density when the covariance matrix Σ is unknown and the mean of the process μ is known is derived in a similar fashion.

In this case, we only have a prior for the covariance, since the prior for the mean is a delta function at the known mean value η. Thus, we are left with $\Sigma \sim f(\Sigma|\nu_w, \Lambda) = \mathrm{IW}_d(\Sigma|\nu_w, \Lambda)$ and $x \sim f(x|\eta, \Sigma) = \mathcal{N}(x|\eta, \Sigma)$.

We can integrate out the covariance of the distribution and obtain its marginal following the same approach as in Equation 239, i.e.,

$$\begin{aligned}
f(x) &= \int f(x|\eta,\Sigma) f(\Sigma|\Lambda,\nu_w) d\Sigma \\
&= \int \mathcal{N}(x|\eta,\Sigma)\, \mathrm{IW}_d(\Sigma|\nu_w,\Lambda) d\Sigma \\
&\propto \int |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\eta)^\top \Sigma^{-1}(x-\eta)} \\
&\quad \times |\Sigma|^{-\frac{\nu_w+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}\Lambda\}} d\Sigma \\
&\propto \int |\Sigma|^{-\frac{\nu_w+d+2}{2}} e^{-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}(\Lambda+(x-\eta)(x-\eta)^\top)\}} d\Sigma \\
&\propto \int \mathrm{IW}_d\left(\Sigma|\nu_w+1, \Lambda+(x-\eta)(x-\eta)^\top\right) d\Sigma \\
&\propto \left|\Lambda+(x-\eta)(x-\eta)^\top\right|^{-\frac{\nu_w+1}{2}} \\
&\propto \left|1+(x-\eta)^\top\Lambda^{-1}(x-\eta)\right|^{-\frac{\nu_w+1}{2}}.
\end{aligned} \tag{242}$$

Thus, the data marginal when the mean is known is given by

$$f(x) = \mathcal{T}_{\nu_w-d+1}\left(x\middle|\eta, \frac{\Lambda}{(\nu_w-d+1)}\right). \tag{243}$$