University of Verona

Department of Economics

Graduate School of Economics and Management

PhD Program in Economics and Finance

XXVIII Cycle (2012)

A Journey Into State-Space Models

S.S.D: **SECS-S/01 - Statistica**

Coordinator: **Prof. Giam Pietro Cipriani**

Signature_____

Supervisor: **Prof. Marco Minozzo**

Signature_____

PhD Candidate: **Alain Julio Mbebi**

Signature_____

A Journey Into State-Space Models - Alain Julio Mbebi

PhD thesis

Verona, 28 March 2017

ISBN:

Abstract of the Dissertation

A Journey Into State-Space Models

by: Alain Julio Mbebi

Supervisor: Prof. Marco Minozzo

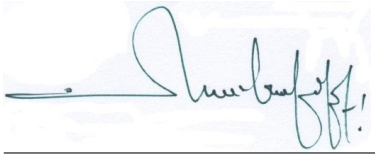
This thesis is concerned with the modelling of time series driven by unobservable processes using state space models. New models and methodologies are proposed and applied on a variety of real life examples arising from finance and economics. The dissertation is comprised of six chapters. The first chapter motivates the thesis, provides the objectives and discusses the outline of the dissertation contents. In the second chapter, we define the concept of state space modelling, review some popular filtering procedures and recall some important definitions, properties and mathematical concepts that will be used in the subsequent chapters. In Chapter three, we propose a new state-space model that accounts for asymmetry, relaxing the assumption of normality and exploiting the close skew-normal distribution which is more flexible and extends the Gaussian distribution. By allowing a stationary autoregressive structure in the state equation, and a close skew-normal distributed measurement error, we also construct a skewed version of the well known Kalman filter. Then in Chapter four, we adapt the robust filtering methodology of Calvet, Czellar and Ronchetti (2015, "Robust Filtering", *Journal of the American Statistical Association*) to build a robust filter with Student-t observation density that provides accurate state inference accounting for outliers and misspecification; this for both finite and infinite state-space models. In the fifth chapter, we provide the foundations for the construction of stochastic volatility models with close skew-normal errors in the observation equation. The summary of the thesis, future works and possible extensions appear in Chapter six.

Sommario

Questa tesi riguarda la modellizzazione di serie storiche generate da processi latenti, utilizzando modelli "state-space". Vengono proposti nuovi modelli e metodologie per poi applicarli ad una varietà di casi tipici presenti in finanza ed economia. La tesi è suddivisa in sei capitoli. Il primo capitolo presenta le motivazioni della ricerca, i suoi obiettivi e la presentazione dei contenuti. Il secondo capitolo approfondisce il concetto di modelli "state-space", riporta e discute le procedure di filtraggio più comuni, e chiarisce alcune definizioni, proprietà e concetti matematici che verranno usati nei capitoli successivi. Nel Capitolo 3 viene proposto un nuovo modello "state-space" per tener conto delle asimmetrie ("skewness") nelle osservazioni, nel quale l'assunzione di normalità non è più necessaria. La distribuzione normale viene, infatti, sostituita con la distribuzione "close skew-normal" che è più flessibile ed include la distribuzione normale. Imponendo una struttura auto-regressiva all'equazione di stato e un errore di misura distribuito secondo una "close skew-normal", si costruisce una versione "skewed" del noto filtro di Kalman. Quindi, nel Capitolo 4 si considera la metodologia di filtraggio robusta proposta da Calvet, Czellar and Ronchetti (2015, "Robust Filtering", *Journal of the American Statistical Association*) con una distribuzione t di Student per ottenere previsioni accurate che tengono conto di valori anomali e di errori di specificazione, sia per i modelli "finite state-space" sia "infinite state-space". Il Capitolo 5 presenta i fondamenti per la costruzione di modelli a volatilità stocastica con errori "close skew-normal" nelle osservazioni. Infine, il Capitolo 6 riassume il contributo della tesi e discute possibili future estensioni della ricerca.

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Alain Julio Mbebi, 24 March 2017

Acknowledgements

First and foremost, I would like to thank God, who has blessed both my family and me in so many ways. His guidance, unconditional love, and abounding compassion have kept me going during the tough times.

Second I would like to express my deepest gratitude to my advisor, Prof. Marco Minozzo, for all the guidance, help and encouragement, for teaching me how to think and write academically, for introducing me to the world of skew-normal distributions and our weekly meetings during the last 3 years, for his invaluable insights and suggestions, which gave rise to many interesting ideas and fruitful results.

I have also had the honor of working with Prof. Mohsen Pourahmadi in the department of statistics, Texas A&M University during my visit to college station, and through so many emails. His teachings in time series analysis and advices have provided me with invaluable guidance on my work. Additionally, I am deeply grateful to the entire statistics department for welcoming me and providing the best work environment ever. Xwang and Raanju, seeing you working so hard motivated me to make further efforts on my own and for that, I would like to say thank you. I also thank the entire finance and control group at EM-Lyon business school in France for making my stay in Lyon easier and productive. Especially, I would like to express my deepest gratitude to Prof. Veronika Czellar for accepting me as visiting scholar, the comments and the very helpful advices, and for always being available whenever I needed her.

I express my profound gratitude to all my professors at the graduate school of economics and management (GSEM) for sharing their knowledge with me inside and outside the classrooms. Especially to Prof. Monica Billio and to Prof. Nunzio Cappuccio. I also thank my PhD coordinator Prof. Giam Pietro Cipriani for his professionalism and my internal PhD committee Prof. Luigi Grossi and Prof. Mauro Mussini for their useful comments that gave rise to a better version of this thesis.

I also must thank all of my friends and colleagues in the economics department of the university of Verona, especially Vincenzo, Tania, Umut, Yasna, Riccardo, Zazi and the two Giovanni for

their never-ending support and encouragement. Apart from my fellow researchers, I also thank the technical staff and the PhD secretary for their valuable help. To my best friends outside of the department, namely Ksenia Dorofeeva, Thierry Zungang, Rosine Zebaze, Georges Kum and Georges Eloundou, I would like to thank you for your friendship throughout the years. Although your help has been unrelated to statistics, your advices, encouragements, and moral support have been uplifting.

Special thanks are due to Linda W. Godwin for making college station in Texas, a second home for me and showing me how a friend can stick closer than a brother. To Prof. Zoran Nikoloski, Prof. Bernd Schroers, Dr. Andres Maldonado and to Don Silvio Arieri for their endless support and encouragements.

To my family, I would like to thank them for their support and love during my adventure far away from home. I would especially like to acknowledge Bleriot and Brigitte Elah, Midel Ewane, Salomon Ehode and my mother in law Regine Ngono who have always supported my pursuits, however incomprehensible they were to them. My siblings, Jean, Elvis, Glawdys, Dany and Ida have each been there for me in incredible ways.

Last but not least, my heartfelt thanks to my wife Jeanne Mbebi and my son Gregory Mbebi for all the love, patience, and support. Without them, I would not have been able to put even half the effort into this thesis; actually, I may not have finished it at all. Their sunshine, joy and smile have been the last dose of adrenaline I needed to make it this far.

Dedication

To the loving memory of my parents, Alexit Ayaman and Lydienne Mboule, and my best friend Pascaline Mane

Contents

List of Figures

List of Tables

1. General Introduction

1.1 Motivation and objectives

In state-space models and especially in the Kalman filtering literature, it has been extensively assumed that, the variables of interest or the uncertainties we are modelling are normally distributed. In this quest, researchers will usually try to fit models into data by matching the first and the second moments or co-moments, to the expense of skewness which is often neglected. When this doesn't work, some transformations are applied on the data and on the model to achieve respectively normality and linearity.

This widely used normality assumption can be explained by the following reasons: First, a Gaussian process is completely determined by its mean and covariance functions. Thus, for model fitting, one only has to specify the first two moments. Second, the easiness to solve the prediction problem. In fact, it is well known that the best predictor of a Gaussian process at an unobserved location is simply a linear function of the observed values. Last but not least, the Kalman filter which is built on the normal and linear assumptions is available on several computing software and ready to be implemented.

However, in the context of sophisticated random phenomena and especially in the financial and economics sectors, where data usually have fat tails and exhibits skewness, symmetric distributions like the Gaussian will no longer be accurate options while modelling these kinds of data. Hence, the need of quite flexible distributions and models is required. Provided these arguments, the following natural questions may be asked. How valid is the assumption of normality? Which consequences can we face when fitting data into a Normal distribution? Given the clear evidence of skewness at different scale of macroeconomics data, how can the state-space model be modified in order to account for skewness? Since it is well known that when fitting data with a two parameters distribution and its n-parameters counterpart ($n > 2$), the latter usually provides a better fit. Is there not any other distributions that can be an alternative option while maintaining the nice properties of the Normal distribution? Moreover, instead of transforming the data to handle

outliers and misspecification, couldn't we build a model immunized against them? Shading light to these questions will constitute the corner stone of this thesis.

1.2 Thesis outline

In the second Chapter, we set a general background of the thesis by introducing some key mathematical concepts and definitions that will constantly appear in the text, this followed by a brief overview of the general state-space model and the description of some associated filtering mechanisms along with some examples.

One of the key concept we rely on in this thesis is asymmetry. Recently, some attempts to incorporate skewness in the ssm and to build the so-called skewed Kalman filter have been made. Unfortunately, some of these contributions suffer from important issues such as the skewness vanishing after several iterations and the poor characterization of the filtering densities. In Chapter 3, after revisiting these contributions and proving our statement, we study and propose a new variant of Kalman filter which overcomes the above mentioned drawbacks, and accounts for asymmetry. We then develop procedures and algorithms for prediction, filtering and estimation using closed skew-normal distributions (csn), whose Gaussian distributions are special cases. Precisely, by allowing a stationary autoregressive structure in the state equation, and a csn distributed measurement error we develop a robust modelling approach for high-dimensional multimodal data.

In Chapter 4, for both finite and infinite state-space models, we build a robust filter with Student-t observation density and provide accurate state inference accounting for outliers and misspecification. We then use simulation to compare the performance of the proposed filter with 3 other filters, namely the Gaussian filter, the robust Gaussian filter of (Calvet et al., 2015) and the Student-t filter. We further apply our theoretical results on the unobserved component model with stochastic volatility (UCSV) of (Stock and Watson, 2007).

Chapter 5 builds the foundation to new types of stochastic volatility models that takes into consideration skewness. The development of this chapter starts by using moment generating

functions of a csn random variable to find the necessary zero-mean and unit-variance conditions and construct a stochastic volatility model with csn errors in the returns equation.

2. Background Settings

In this chapter, we set a general background of the thesis by introducing some key mathematical concepts and definitions that will constantly appear in the text, this will be followed by a brief overview of the general state-space model and the description of some associated filtering mechanisms.

2.1 Introduction

The crucial need for studying, analysing and monitoring sequential information arising in several areas of engineering and science, and from various types of problems has been one of the most challenging issues over the past two centuries. In order to handle these concerns, time series analysis has become a key tool to deal with data that are usually a time series, generated by a dynamical system, or a sequence generated by a univariate spatial process such as biological sequences. In this quest and by relying on statistical modelling techniques, some of the main goals of time series analysis are to understand and reveal the dynamic driving the observed time series and to forecast future events. Thus, the requirement of an appropriate time series model that takes into account the essential feature of the observed data.

Often, two main types of analysis are considered, the offline analysis, which corresponds to the case where all the data have already been collected and the online analysis, where the data arrive in real-time and are dealt with as they become available. As already mentioned, in time series analysis and especially in the online case, one common task is the prediction of future observations, conditional of all available observations up to the time point t of interest which will be denoted by $y_{1:t} = y^t = (y_1, \cdots, y_t)$. Throughout this thesis, the notations $y_{1:t}$ and y^t refer to the same thing and will therefore be used interchangeably when there is no ambiguity regarding the starting value of t. Moreover, we only deal with discrete-time valued processes, hence t is always an integer.

With the fact that the future state of the system can generally be characterized by uncertainty,

the computation of the best guess is usually required and of course one would like to know how confident we are regarding this guess. This can be achieved by computing over future observations (in term of time horizon $h > 0$), the probability $f(y_{t+h}|y_{1:t})$. Not always, but it may happen that the researcher can control the system under investigation, this is the case for some application in engineering where some inputs are incorporated in the model. In this specific case, the predictions of future outcomes of the system are a function of the inputs as well. For instance, if we let $u_{1:t}$ be the past inputs and $u_{t+1:t+h}$ the h-periods ahead inputs, then the prediction can be computed with the following probability $f(y_{t+h}|u_{1:t+h}, y_{1:t})$.

In what can be called classical time series analysis, predictions are computed with linear models such as the autoregressive integrated moving average (ARIMA) model and the autoregressive moving average with exogenous terms (ARMAX) models among others, see (Hamilton, 1994) for detailed explanations. These classical approaches, however present some drawbacks. For example, in order to make predictions about the future, one would like a model where there is no restriction on how far we can go back in the past to gain inside information, which is not the case for the above mentioned models where the prediction of the future must be based on a finite time horizon into the past. Other challenges we face are the difficulty to incorporate prior knowledge into the model and handling multivariate variables.

One possible way of overcoming these drawbacks is the use of dynamic (linear or nonlinear) models that views the process one would like to analyse as driven by another variable, which this time is unobservable. In this framework and as it will be our case, what is often under consideration are partially observed dynamic systems driven by probability density functions, with one or more latent processes changing and interacting over time where only part of them or their linear transformations are observed. This methodology has been extensively applied to several real life examples arising from financial, engineering and biological sectors among others.

For example, the mortality curve in insurance that describes how the mortality rate as a function of age changes over time and the implied volatility as a function of time to maturity and the option's strike price in the financial sector. As for the example of observed process driven by another latent process, one can consider for instance the observed interest rate curve that can be driven by the unobserved curvature and level processes.

Furthermore, noises are incorporated in the observed and unobserved variables to take into consideration the potentially corrupted and misrepresented nature of the processes of interest. Provided that such information are observed or collected over time, that their behaviour follows a given dynamic and that one has the appropriate model's representation, analysing and forecasting these data can become very convenient. This convenience is mainly due to the utilization of what is also known as the state-space model (ssm) representation and its related filtering procedure. That is, the Kalman filter when under linear and Normal assumptions and the particle filter otherwise.

In state-space modelling, it is generally assumed that the observations are generated from an underlying hidden state of the system that evolves in time, and as a probable function of the inputs. In the online framework, the goal is to infer the hidden state given the observations up to the time period of interest. In other terms, if the latent state at time t is designated by x_t, then the goal will be to compute the belief state of the system defined as $f(x_t|y^t, u^t)$. The belief state is a very important notion since it can be viewed as a sufficient statistic for prediction, (Åström, 1965). Meaning that, there is no need to carry around the previous information, and the Bayes' rule can be used to recursively update the belief state.

With their ability to deal in a simple manner with univariate and multivariate variables and to incorporate prior knowledge into the model, it has been proved that in many aspects, ssm are better options than the classical time series tools, (West and Harrison, 1997; Harvey, 1990; Aoki, 2013; Durbin and Koopman, 2012). For instance, it is usually the case that we want to estimate some variables (hidden) but that cannot be measured. ssm offer the possibility to embed them in the model, thus generating models much closer to the real nature of the phenomenon under scrutiny (Pearl, 2000).

2.2 Notations and definitions

First of all, let us point out that, all definitions and properties in this section are borrowed from (Grimmett and Stirzaker, 2001; Roussas, 2003; Cappé et al., 2005). We start by recalling a definition of a random variable (r.v). Simply speaking, one can define a r.v as a random number which is an outcome of a random experiment. Note however that, if the random experiment is

characterized by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then a random variable can be formally defined as a measurable function $f : \Omega \to \mathbb{R}$. Here Ω called the sample space, is the set of all possible outcomes of the experiment, $\mathcal{F}$, is σ-field or a collection of subsets of Ω and the notion of measurability stands for the fact that for every Borel set $\mathcal{B} \subset \mathcal{F}$, it is true that $f^{-1}(\mathcal{B}) \in \mathcal{F}$.

Definition 2.2.1 (Stochastic process). *Let T be an arbitrary set that is sometimes called the index set and let $(\Omega, \mathcal{F}, \mathbb{P})$ represent a probability space. Then any collection of random variables $X = \{X_t : t \in T\}$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a stochastic process indexed in T .*

Definition 2.2.2 (Realization). *For a given outcome $\omega \in \Omega$, any collection $t \mapsto X_t(\omega)$ that is defined on the index set T and takes values in $\mathbb{R}$ is known as realization or sample path of the stochastic process X at ω.*

This means that, for any outcome $\omega \in \Omega$, we have a corresponding trajectory or realization of the stochastic process. With respect to the nature of T, several types of processes can be encountered. For example, if $T \in \left\{\mathbb{R}^d, \mathbb{Z}^d, \mathbb{N}^d\right\}$ for $d \geq 2$, then the process will be called random field. If instead as in this thesis, $T \in \{\mathbb{N}, \mathbb{Z}\}$, we talk about processes in discrete time. One should also recall that, the expressions random function, process, random process some times refer to stochastic process and so will be the case here.

Definition 2.2.3 (Kernel). *Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be two measurable spaces. We say that a function $Q : X \times \mathcal{Y} \to [0, \infty]$ is an unnormalized transition kernel from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ if the following holds*

(i) for all $x \in X, Q(x, .)$ is a positive measure on $(Y, \mathcal{Y})$,

(ii) for all $A \in \mathcal{Y}$, the function $x \mapsto Q(x, A)$ is measurable.

If $Q(x, Y) = 1$ for all $x \in X$, then Q is called a transition kernel, or simply a kernel. If $X = Y$ and $Q(x, X) = 1$ for all $x \in X$, then Q becomes a special case referred to as a Markov transition kernel on $(X, \mathcal{X})$.

Definition 2.2.4 (Stationary process). *A stochastic process $\{X_k\}$ is said to be stationary if its finite-dimensional distributions are translation invariant, meaning that, if for all $k, n \geq 1$ and all $n_1, \cdots, n_k$, the distribution of the random vector $(X_{n_1+n}, \cdots, X_{n_k+n})$ does not depend on n.*

2.3 Stochastic simulation and Markov chain Monte Carlo

A Markov chain can be defined as a random process or a chain of random events having the property that the future state depends only upon the current state of the process and not on the entire past history of the process. One known example of a Markov chain that can help better understand this concept is the famous "drunkard's walk". Imagine a drunk person who can move only right or left on the road and with equal probability. One can see that, the drunk's next position after moving right or left will depend only on his current position and not on any of his previous positions.

A simple question based on the above example could be the following: Assuming that the road on which the drunker is walking on is constituted by small squares, among those squares, which one are more visited than the others? A possible approach to answer this question is to define the length of the experiment and put the drunk at the same position repeatedly after each run and count the number of times he lands on each square. By doing so, we will be able to know which squares he visited the most. The type of experiment described above is known in the literature as Monte Carlo Simulation.

This term was set as the angular stone in simulation techniques in the $1940s$ by physicists working at that time on the nuclear weapon project in the Los Alamos Laboratory. The initial spark came from Stan Ulam, who while recovering from a surgery, wanted to compute the probability of a game of solitaire being successful. Instead of working out this complicated combinatorial problem analytically, he suggested to simply play out a certain (many) numbers of this game and record the outcome of each. The complicated system could then be approximated with a statistical sample.

The insights of the rapidly growing literature of the five last decades around MCMC have been

very helpful to solve some difficult problems in various areas such as Bayesian inference, computational finance and especially when multi-dimensional integrals calculation are involved. MCMC simulation is a well-known methodology for producing samples from a recognised posterior distribution for hidden variables, where the distribution is very complex, that is, it is not evident how to sample from it (Smith and Roberts, 1993). The aim of this section is to review and illustrate the use of Monte Carlo simulations and then briefly describe some MCMC methods.

2.3.1 Monte Carlo simulation

It happens very often to deal with complex problem in applied sciences. This is where numerical solutions are required, because most of those problems cannot be solved analytically. Notice that, not all numerical methods are always efficient, especially when it comes to high dimensional computations. For example, in numerical integration, the quadrature method require that the approximation of the integral is done by partitioning the integration domain into a set of discrete volumes. Thus, obtain the integral by summing the values of the weighted function. Nevertheless, as this result is simply an estimate, it can be proved that the magnitude of the error increases with respect to the dimension of the integral. Instead, Monte Carlo method can be used to sort out this problem of dimensionality.

The use of Monte Carlo methods requires to first put the quantity we would like to compute into the form of an expected value. Specifically, suppose we wish to estimate the expectation of $g(X)$ with respect to a probability distribution function f. Let us denote this expectation by $\mu = \mathbf{E}_f g(X)$. The problem is that μ is generally not tractable analytically, that is the sum or the integral that needs to be computed in order to achieve this is very complicated. A Monte Carlo approach which will provide an estimate of μ is based on the following steps: Simulate N pseudo-random values $X_1, X_2, \cdots, X_N$ iid from f , then simply take the average of $g(X_1), g(X_2), \cdots, g(X_N)$ to estimate μ. Thus,

$$\widehat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} g(X_i) \tag{2.3.1}$$

This is known as the *Monte Carlo approximation of* μ (Geyer, 2011). It can be shown that as

the sample size gets large, the estimate $\widehat{\mu}_N$ converges to the real value μ.

More formally, let us consider a function $h(x)$ that can be decomposed as a product of a probability density function $p(x)$ and another function $f(x)$ defined on (a, b). Now suppose we want to evaluate

$$J = \int_a^b h(x)dx \tag{2.3.2}$$

From the above assumptions, the integral J can be defined as the expectation of $f(x)$ with respect to the density $p(x)$. That is

$$J = \int_a^b f(x)p(x)dx = \mathbb{E}_p(x)[f(x)] \tag{2.3.3}$$

Moreover, suppose that by any mean we are able to generate a large set of random variables $x_1, x_2, \cdots, x_n$ from $p(x)$, then J can be approximated by:

$$J \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i) \tag{2.3.4}$$

This quantity is referred to as the *Monte Carlo integration*. This method is often used in Bayesian analysis to estimate posterior or marginal distribution. For example, let us consider the following integral,

$$I(y) = \int f(y|x)p(x)dx$$

From the above mathematical formulation, one can approximate $I(y)$ by

$$\widehat{I}(y) \simeq \frac{1}{n}\sum_{i=1}^{n} f(y|x_i) \tag{2.3.5}$$

Where x_i are drawn from $p(x)$. Following the setting of (Geyer, 2011) we end up with the

estimated Monte Carlo standard error defined by:

$$SE^2[\widehat{I}(y)] = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}(f(y|x_i) - \widehat{I}(y))^2\right). \tag{2.3.6}$$

Nevertheless, some related questions on this topic need to be clarified. When and how does the convergence in Monte Carlo methods take place? How enough is the approximation precise? The following theorems from stochastic analysis (Evans, 2012) will provide their explanation.

Theorem 2.3.1 (Strong Law of Large Numbers). *Let $X_1, X_2, \cdots$ be a sequence of i.i.d and integrable random variables having the same expected value μ and the same variance σ^2. Then*

$$P\left(\lim_{n\to+\infty}\frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right) = 1 \tag{2.3.7}$$

Note that, this is useless if there is no way of evaluating the quantity

$$\epsilon_n = \mathbb{E}(X_1) - \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right),$$

meaning that, the random variable X_1 has to be integrable.

Theorem 2.3.2 (Central Limit Theorem). *Let $X_1, X_2, \cdots$ be a sequence of i.i.d random variables having the same expected value $\mu < \infty$ and the same variance $\sigma^2 < \infty$ for $i = 1, 2, \cdots$. Then for all $-\infty < a < b < +\infty$ we have:*

$$\lim_{n\to+\infty} P\left(\frac{\sigma}{\sqrt{n}}a \leq \epsilon_n \leq \frac{\sigma}{\sqrt{n}}b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b \exp(-\frac{x^2}{2})dx \tag{2.3.8}$$

That is $\frac{\sqrt{n}}{\sigma}\epsilon_n$ converges in distribution to the reduced centred Gaussian distribution. For practical implementation, we often forget the limit step and replace ϵ_n by a centred Gaussian distribution having $\frac{\sigma^2}{n}$ as variance.

This shows how important is to know the magnitude of σ^2 since the error is strongly connected with σ.

Remark 2.3.1. *It is important to notice that, the Central Limit Theorem (CLT) plays a key role here, in the sense that it never permits to the user to bound the error, this simply because the support of the Gaussian is the entire* $\mathbb{R}$ *set. Usually the error in the Monte Carlo setting is characterized by the standard deviation of* ϵ_n *which is* $\frac{\sigma}{\sqrt{n}}$*. This means that, the error decreases with respect to the variance, and it is the reason why numerous methods for convergence improvement in the Monte Carlo framework focus on reducing the variance (Lapeyre et al., 2003; Glasserman, 2010). The convergence rate in this method is defined by* $\frac{1}{\sqrt{n}}$ *which holds almost surely in all dimensions.*

2.3.1.1 Example in finance

We are now applying Monte Carlo methods to compute the expected present value of a payoff of a call option. Recalling that the payoff of the call option (for the option holder) is given by

$$g(S, K) = [S(T) - K]^+ := \max(S(T) - K, 0)$$

where S is the price of the underlying asset at the maturity time T and K is the strike price. As in (Elliott and Kopp, 2006), it can be shown that the fair price of an European contingent claim is simply the discounted expected value of its payoff at the maturity. Under the risk-neutral measure $\mathbb{Q}$, the option's actual value will be given by

$$E_{\mathbb{Q}}[\exp(-rT)[(S(T) - K)^+]$$

where r is a constant interest rate and $\exp(-rT)$ is the discount factor. The following Black-Scholes (Black and Scholes, 1973) model characterize the dynamic of the stock price.

$$\frac{dS(t)}{S(t)} = rdt + \sigma dW(t) \tag{2.3.9}$$

Here σ and W are respectively the volatility of the stock price and the standard Brownian motion. A further assumption is that the rate of return on the stock is set to be equal to the interest rate.

The solution of the stochastic differential equation (2.3.9) at maturity time T is

$$\begin{aligned} S(T) &= S(0)\exp\left((r-\frac{1}{2}\sigma^2)T+\sigma W(T)\right) \\ &\simeq S(0)\exp\left((r-\frac{1}{2}\sigma^2)T+\sigma\sqrt{(T)}\phi(0,1)\right), \end{aligned}$$

with $S(0)$ the initial price of the stock and $\phi(0,1)$, the standard Normal distribution.

Proof. Starting from Equation (2.3.9) we can write,

$$\int_0^t \left[\frac{dS(z)}{S(z)}\right] \mathrm{dz} = \mathrm{rt} + \sigma \mathrm{W(t)}, \tag{2.3.10}$$

with $W(0)=0$ by convention. □

Applying Itô's formula on $U(t,x)=\log x$ leads to

$$\begin{aligned} d(\log(S(t))) &= \frac{1}{S(t)}d(S(t)) - \frac{1}{2S(t)^2}\left(\sigma^2 S(t)^2 dt\right), \\ &= \frac{dS(t)}{S(t)} - \frac{1}{2}\sigma^2 dt. \end{aligned}$$

Thus,

$$\frac{dS(t)}{S(t)} = d(\log(S(t))) + \frac{1}{2}\sigma^2 dt. \tag{2.3.11}$$

Using Equations (2.3.11) and (2.3.10), we obtain

$$\log(\frac{S(t)}{S(0)}) + \frac{1}{2}\sigma^2 t = rt + \sigma W(t),$$

then

$$\log S(t) = (rt - \frac{1}{2}\sigma^2 t) + \sigma W(t) + \log S(0)$$

that is,

$$S(0) = S(0)\exp\left[(rt-\frac{1}{2}\sigma^2 t) + \sigma W(t)\right];$$

Recalling that, for a standard Brownian motion W and for all $s < t$

$$W(t) - W(s) \sim \phi(0, t-s),$$

we have

$$W(t) - W(0) \sim \phi(0, t).$$

Since $W(0) = 0$ by hypothesis,

$$W(t) \sim \phi(0, t).$$

Setting $t = T$, we complete the proof and write

$$S(t) = S(0) \exp\left[(rt - \frac{1}{2}\sigma^2 t) + \sigma\sqrt{(T)}\phi(0,1)\right]. \tag{2.3.12}$$

Equation (2.3.9) tells us that the stock price is lognormally distributed, therefore,

$$\mathbb{E}_Q\left[\exp\left(-rT[S(T)-K]^+\right)\right] = \frac{1}{\sqrt{2\pi T}}\int_{\mathbb{R}}[S(0)\exp(\sigma y - \sigma^2\frac{T}{2})\exp(-rT)K]^+\exp(-\frac{Y^2}{2T})dy.$$

Now, drawing a sequence of independent Normal distribution $\phi_1, \phi_2, \cdots, \phi_n$, we can approximate

$$\mathbb{E}_Q\left[\exp(-rt)[S(T)-K]^+\right]$$

by

$$\frac{1}{n}\sum_{i=1}^{n}\exp(-rt)[S_T^i - K]^+,$$

where

$$S_T^i = S(0)\exp\left((r-\frac{1}{2})T + \sigma\sqrt{T}\phi_i\right)_{i=1,\cdots,n}.$$

Figure 2.1 represents the plot of the prices obtained by the Monte Carlo method (in red) compared with the true price (in blue) which is 27.66. We set $S(0) = 100$, the maturity date is 2, the strike price $K = 80$, the volatility $\sigma = 0.1$ and the interest rate $r = 0.05$.

In the top panel of Figure (2.1), the number of iteration is 300, we can see that the Monte Carlo price starts fluctuating a lot at the beginning of the iterations before showing a slow convergence

to the true price when the number of iterations approaches 250. Since we could not make a clear conclusion concerning the obtained price, we increased the number of iterations to 1000 and obtained the bottom panel of Figure (2.1), where we can see that the price obtained with the Monte Carlo method has the same behaviour as in Figure (2.1) when the number of iteration is ranging between 1 and almost 300, whereas the Monte Carlo price starts converging to the true price when the number of iteration is above 300.

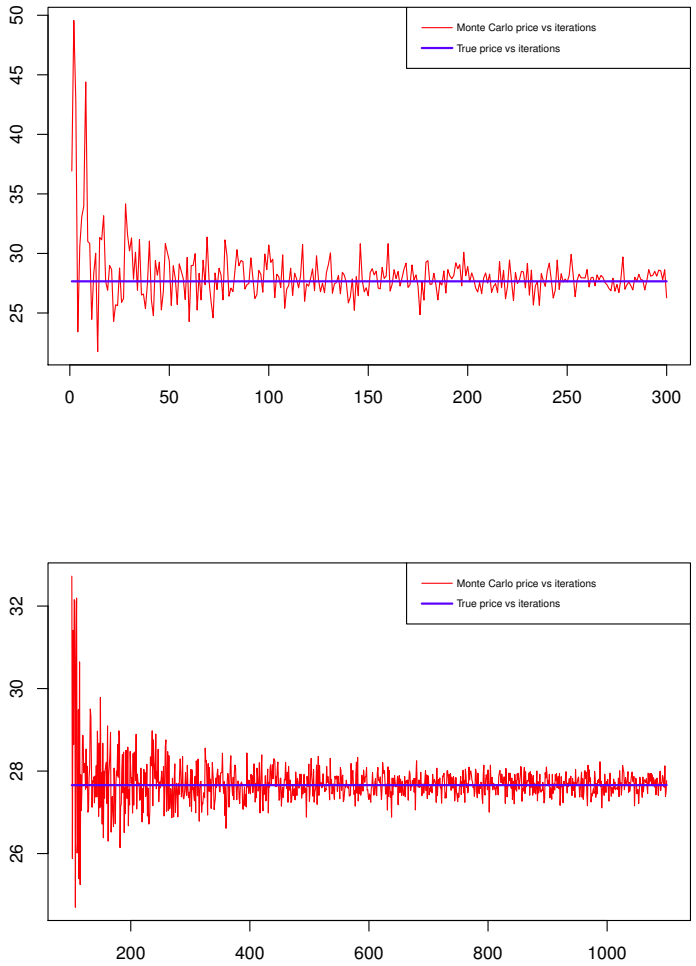


Figure 2.1: Convergence of Monte Carlo methods for the evaluation of an European call option.

2.3.1.2 Markov chain Monte Carlo methods

Before starting our investigation of the theory of MCMC, a quick tour on Markov chains is required. Let us denote by X_t the value taken by a random variable at time t, and let us call the set of possible values of X_t the *state-space*. If the transition probabilities between different values in the state-space depend only upon the current state of the random variable, then it is said to be a *Markov process*. That is, the conditional probability of X_n given the past variables depends only upon X_{n-1}.

These random variables can be seen as evolving over time, with a transition probability depending on the current state of the chain. It is possible for a Markov chain to have a stationary distribution. Intuitively, that is, if in the starting state the chain has a stationary distribution, then in the next state, the distribution of chain will still be stationary. In the case of MCMC, we want the stationary distribution to be the posterior. However, in order for the stationary distribution to be unique, no matter where the chain starts, the following properties (Robert and Casella, 2010) must hold.

Property 2.3.1 (Irreducibility). *First of all, we say that state i communicate with state j (they are accessible from each other), if there exists $n > 0$ such that, the probability of moving from i to j is not zero. Then a Markov chain is said to be irreducible if all states communicate between them, that is, the chain has only one communication class (class of equivalence). Clearly the communication relation is an equivalence relation over the state-space.*

In other words,

$$\exists n \quad \text{s.t} \quad P(X_n = j | X_0 = i) > 0 \tag{2.3.13}$$

Property 2.3.2 (Aperiodicity). *The period of the state i denoted d_i, is the biggest integer dividing all $n \geq 1$ such that, the probability of moving from state i to itself in n steps is strictly positive. When $d_i = 1$, the state is **aperiodic**, and when all the states are aperiodic, so is the chain.*

With these two properties, the Markov Chain is ensured to have a stationary distribution. This

is very important when Markov chains are used as a simulation tools. In practise, the stationary is very useful, in the sense that, the distribution of X_n converges to the stationary distribution as n gets larger.

Remark 2.3.2. *When the state-space is finite and written $x_1, x_2, \cdots, x_n$, we can link the probability of transition to a matrix formulation as follows:*
$P(X_n = x_j | X_{n-1} = x_i) = p_{ij}$, *with* $i = 1, \cdots, n$ *and* $j = 1, \cdots, n$. *Reversely this does not hold when the state-space is countably infinite which is the case of major part of Markov chain in the MCMC framework. Now let us denote by $\pi_j(n)$ the probability that the chain is in the state j at the step n and $\pi(n)$ the row vector of the state-space probabilities at step n. That is,* $\pi_j(n) = P(X_n = x_j)$.

The following known in the literature as the **Chapman-Kolomogrov equation**, usually describes the chain's evolution.

$$\pi_i(n+1) = P(X_{n+1} = x_i) = \sum_k p_{ki}\pi_k(n) \tag{2.3.14}$$

This is simply the sum over the probability of being in a particular state at the current step and the transition probability from that state into the state x_i (Dagpunar, 2007). An other key concept into the theory of Markov chains is the *deductibility* which holds when there exists a positive integer $s_{ij}^{m_{ij}} \quad \forall i, j$ where $s_{ij}^m = P(X_{n+m} = x_j | X_n = x_i)$ characterize the probability that the process is in state j given that it started in state i n steps ago. That is, it is always possible to go from one state to any other no mater how many steps it takes. Therefore, the above **Chapman-Kolomogrov equation** can be written into a matrix form as $\pi(n+1) = \pi(n)S$. s_{ij}^m is the $ij\ th$ element of S (Geyer, 2011) and (Walsh, 2004). Finally, more information about other concepts such as reversibility which plays an important role when it comes to simplify the asymptotic variance estimation and the central limit theorem (CLT) for the Markov chain can be seen in (Geyer, 2011) and (Rubinstein and Kroese, 2011).

2.3.1.3 Metropolis-Hastings algorithm

Monte Carlo methods are used for analysing Bayesian distributions in high dimension. They are helpful for either generating samples like $\theta_1, \theta_2, \cdots$ from a given probability distribution $p(\theta)$, or estimating expectations of a function under the same probability distribution, or both. They are generally preferred when the number of parameters to be estimated is large enough as well as when we are in high dimension. When combined with Markov chain, the aim become to construct a Markov chain whose equilibrium distribution is the target distribution of interest(MacKay, 2003).

The probability distribution $p(\theta)$ also called the target density, is complex enough by assumption. That is, computing expectations from it by deterministic methods is not feasible, and this is where Monte Carlo methods are helpful. The accuracy of the Monte Carlo estimate depends only on the variance and not on the dimensionality of the space sampled (MacKay, 2003), (Geweke, 1991). Since there is a possibility for the normalizing constant to be unknown or known but in high dimensional space, sampling from $p(\theta)$ can become a very painful task. This is because there is no trivial method that can help to sample from $p(\theta)$ without listing most of the possible states and obtaining an accurate estimates.

When applying Monte Carlo integration techniques, the common problem encountered is how to draw samples from some complex probability distribution $p(\theta)$ (Metropolis and Ulam, 1949), (Hastings, 1970) and (Metropolis et al., 1953). Solving such a problem has always been one of the main task of MCMC methods.

The general idea in the MCMC setting is to build a transition kernel of an ergodic[1] Markov chain with the desired invariant distribution, and then simulate the chain for many steps, so that it reaches the equilibrium. The states that have been sampled after the convergence of the chain will then have the same distribution as our distribution of interest (target).

The Metropolis-Hastings algorithm (MH) is often used when dealing with high-dimensional problems. This method consists of defining the probability of transition from a state θ_i to a state θ_{i+1}. Thus it is a Markov process in which a sequence of $\theta_1, \theta_2, \cdots$ is generated. The probability

[1]Sufficient condition for the existence of the stationary distribution $\pi(x)$ independent of the initial probability at the starting state that is, a Markov chain satisfy the stationarity, irreducibility, and aperiodicity conditions

distribution of each trial θ_{i+1} appears to depend only on the distribution of θ_i. Since we are looking for a sample which is independent from the starting condition, it is very important to run the chain for a sufficient long time horizon to avoid the dependence for successive samples.

The MH algorithm can be defined as an adaptation of a random walk based on the **acceptance rejection rule** to converge to the target distribution. One of the requirement of this algorithm is that, the proposal distribution has to be given in a very specific way, it is then recommended to use prior knowledge to achieve this.

To see how the MH algorithm works, let us assume that the variable that we want to sample is θ and the target distribution is $p(\theta)$. Furthermore, we suppose that the current state of the Markov chain is θ^n, the potential candidate is $\theta^{'}$ after n iterations and the proposal distribution is $q(\theta^{'}|\theta^n)$ which depends only upon the current state θ^n from Markov chain's definition. The next step is to generate a candidate from the proposal, then compute $p(\theta^n)$ and $p(\theta^{'})$. After this, accept the candidate with probability $\min\{1, \alpha\}$. That is, the new state after the update is $\theta^{'}$ with probability α or θ^n with probability $1-\alpha$ (Geyer, 2011). Where α knowing as the **Hasting ratio** is defined by:

$$\alpha = \frac{p(\theta^{'})q(\theta^{'}|\theta^n)}{p(\theta^n)q(\theta^n|\theta^{'})} \tag{2.3.15}$$

Finally repeat the process several times, until the convergence of the generated Markov chain is achieved. The Metropolis-Hastings algorithm can be summarise as follow:

1. Initialize $\theta^{(0)}$
2. Set $i \leftarrow 1$
3. Simulate a candidate $\theta^{'} \sim q(.|\theta^{i-1})$
4. Compute the quantity

$$\alpha = \min\left\{1, \frac{p(\theta^{'})}{p(\theta^{i-1})}\frac{q(\theta^{i-1}|\theta^{'})}{q(\theta^{'}|\theta^{i-1})}\right\}$$

5. Accept θ' with probability α, that is,

$$\theta^i = \begin{cases} \theta', & \text{with probability } \alpha. \\ \theta^{i-1}, & \text{otherwise.} \end{cases} \tag{2.3.16}$$

6. Increment i, $i \leftarrow i+1$ and return to step 3.

Moreover, when the proposed distribution $q(\theta', \theta)$ is symmetric, meaning that, $q(\theta', \theta) = q(\theta', \theta)$, we talk about the Metropolis algorithm and one just have to replace α in the step 4 of the MH algorithm by:

$$\alpha = \min\left\{1, \frac{p(\theta')}{p(\theta^{i-1})}\right\},$$

and the Metropolis algorithm will follow the steps below:

1. Initialize $\theta^{(0)}$
2. Set $i \leftarrow 1$
3. Simulate a candidate $\theta' \sim q(.|\theta^{i-1})$
4. Compute the quantity

$$\alpha = \min\left\{1, \frac{p(\theta')}{p(\theta^{i-1})}\right\}$$

5. Accept θ' with probability α, that is,

$$\theta^i = \begin{cases} \theta', & \text{with probability } \alpha. \\ \theta^{i-1}, & \text{otherwise.} \end{cases} \tag{2.3.17}$$

   In practise, an uniform distribution over $[0, 1]$ can be used.

6. Increment i, $i \leftarrow i+1$ and return to step 3.

Remark 2.3.3. *Given a "good" burn-in period, the chain should move towards the stationary distribution (or in some cases, the distribution we want to simulate). The burn-in period depends on the initial value chosen and the behaviour of the proposal q. The most used proposal is the n-dimensional Gaussian density centered at the current point and having Σ as the covariance matrix. In this case, the* ***Hasting ratio*** *becomes $\alpha = \frac{p(\theta')}{p(\theta^n)}$. On one side, the sampler will move very slowly and sample only a small local region around the starting position if the variances are too small. On the other side if the variances are very big the sampler will propose points which are far away from the high density regions of the distribution leading to high rejection rates (Gelman et al., 2011). Therefore an important consideration need to be given when setting the elements in the covariance matrix.*

In the following we are applying[2] the Metropolis algorithm on the standard Normal distribution with proposal, the uniform distribution. Recall that the probability density function of the standard Normal distribution can be written as

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \tag{2.3.18}$$

We can see in figure (2.2) below that, as the number of iterations get larger, the Metropolis sampler converges to the true pdf.

2.3.1.4 The Gibbs sampler algorithm

Known as a special case of Metropolis-Hastings algorithm, the Gibbs Sampler algorithm helps to generate samples from a joint probability distribution, under the condition that the joint distribution is unknown but the conditional distribution with respect to each parameter is known (Gelfand and Smith, 1990). The samples generated can then be used for some statistical estimation. One should notice that for a well accurate estimation, the samples generated must be as many as possible and the assumption on the complete information about the conditional distributions makes this method sometimes useless in practise (Martinez and Martinez, 2001).

[2]The R code used to implement this algorithm has been inspired by the slides "An introduction to Bayesian statistics and MCMC algoritm" of Alessandra Guglielmi.

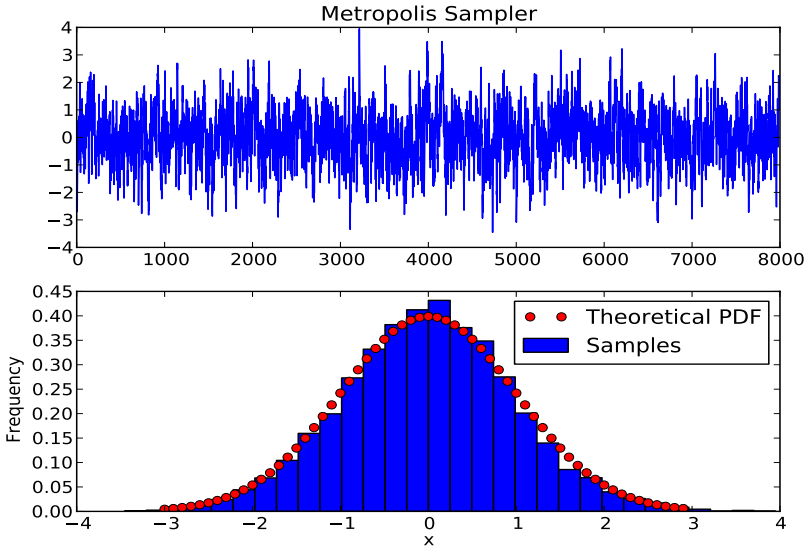


Figure 2.2: Convergence of Metropolis sampler when approximating the standard Normal distribution using the uniform distribution as proposal and with 10000 iterations

.

The acceptance probability is one, that is the Gibbs sampler always accepts the candidate point.

In practise this is how it works. Let (x, y) be a bivariate parameter, and assume that we want to evaluate the joint distribution $p(x, y)$ or the marginal densities $p(y)$ and $p(x)$. The idea behind the sampler is that, we can easily consider the conditional $p(x|y)$ or $p(y|x)$ instead of obtaining the joint $p(x, y)$. The sampler begin with one initial value y_0 for y and after generating a random variable from $p(x|y = y_0)$ we obtain x_0. The next step is to use x_0 to draw the new y_1 from $p(y|x = x_0)$. That is

$$x_i \sim p(x|y = y_{i-1}) \quad \text{and} \quad y_i \sim p(y|x = x_{i-1}) \tag{2.3.19}$$

After m iterations, we end up with a sequence of m inputs where each one represents a vector with two elements (x_i, y_j) where $1 < i, j < m$.

2.4 State-space models and the filtering problem

Modern time series analysis has been greatly impacted during the last century with the advent of the Kalman filter, its several extensions and generalizations such as the Gaussian quadrature Kalman filter (Ito and Xiong, 2000), the extended Kalman filter (Jazwinski, 1970; West et al., 1985), the unscented Kalman filter (Julier and Uhlmann, 1997; Van Der Merwe et al., 2000) and the Gaussian sum filter (Alspach and Sorenson, 1972). This impact provided a new and sophisticated tool to study and estimate intricate dynamics and helped to shade light to the big family of dynamic models by drawing a great attention from practitioners and researchers from a broad range of fields, see (Harvey, 1990; West et al., 1985; West and Harrison, 1997) among others.

Represented in a state-space form, the usually stressful study of some complex dynamic systems will become very flexible. Moreover, the state-space structure provides the ability to handle a broad range of linear and many non-linear time series models, such as the autoregressive integrated moving average (ARIMA) models, the unobserved component models and the regression models with changing coefficients, to list just few. A state-space model consists of two main parts. That is, a hidden state x_t and the observation variable y_t, sometimes identified as state and observation equations.

The main idea driving the state-space representation of a given system is that behind the observed time series, we have the underlying unobservable process which evolves with time and reflects the structure of the system. The state equation will then characterize the dynamics of the state variables while the observation equation will link the observed processes to the latent ones.

In some of its representations, the state process can incorporate seasonality, trend, regression parts with an error term and cycle. These type of models linking the time observed variable to several and different components which are themselves often modelled as individual random walks, are regrouped into the subfamily of structural time series models.

In order to be properly characterized, all ssm should specify the state transition function $f(x_t|x_{t-1}, u_t)$, the prior $f(x_0)$ and the observation function $f(y_t|x_t, u_t)$, where for the purpose of our intended application and without any loss of generality we will omit the control variable u_t. We then

end up with a model specified by the state transition function $f(x_t|x_{t-1})$ and the observation function $f(y_t|x_t)$. It is further supposed that we are under first-order Markov assumption, that is, $f(x_t|x_{1:t-1}) = f(x_t|x_{t-1})$. Note that whenever this is not the case, there is a workaround which consist of augmenting the state-space and make it first-order Markov. Simply speaking, supposed we have a third-order Markov system, then one can redefine the new state as $\tilde{x_t} = (x_t, x_{t-1}, x_{t-2})$. In the same way, the observations are assumed to be conditionally first-order Markov. That is,

$$f(y_t|y^{t-1}, x_t) = f(y_t|x_t, y_{t-1}),$$

which trough out this thesis and only for simplicity purposes, will be defined by the following assumption

$$f(y_t|y_{t-1}, x_t) = f(y_t|x_t).$$

In order to be able to model infinitely long data, we also assume that the model is time-invariant and in the case where parameters are time-varying, they will be treated as supplementary random processes and added to the model.

It is also quite usual that the model admits unknown parameters and their impact on the system is not the least. In fact, the interaction of the state process with the observation and the covariance structure of the error terms depend on them. This is why for accuracy and efficiency, the parameters together with the state process have to be estimated from the observations. One common procedure from which one can obtain maximum likelihood estimates (MLE) of the parameters is to employ the Kalman filter (KF) (Kalman, 1960, 1963) which is a recursive mechanism that estimates the hidden components at time index t, given the available information up to the same t.

Despite the fact that the KF was originally designed for works in physics and engineering fields, the finance and economics communities were facing issues that could have been handled by the KF. Still, its usage has to wait until early 1980 with the seminal contribution of (Harvey, 1981). To list just few, other contributions of KF and ssm in finance and economics are (Meinhold and Singpurwalla, 1983) who provided the statistical understanding of the KF by using a Bayesian

formulation and some well-known results in multivariate statistics and illustrated it for quality control. (Clark, 1987), who by using Kalman filtering and smoothing techniques, successfully decomposed quarterly data on industrial production and deflated gross national product in the United States from 1947 through 1985 into independent nonstationary trend and stationary cycle components. (Stock and Watson, 1988), who used a dynamic factor analysis to implicitly defines a variable that can be thought of as the overall state of the economy. By using data from 1959-1987, they estimated the unobserved process and provided a formal rationalization for the traditional methodology used to develop the coincident index. For formalism, let us represent by Y_t an $n \times 1$ observable variable evolving with time and driven by a $k \times 1$ hidden state variable X_t satisfying the following Assumptions (2.4.1) and (2.4.2).

Assumption 2.4.1. *X_t is a Markov chain*

Assumption 2.4.2. *Conditionally on X_t, the Y_t's are independent and Y_t depends on X_t only.*

Consequently for $t > 0$, a ssm can be entirely characterized by the initial distribution $f(x_0)$ and the conditional densities $f(x_t|x_{t-1})$ and $f(y_t|x_t)$ leading to the joint density in Equation (2.4.1) below

$$f(x^t, y^t) = f(x_0) \prod_{i=1}^{t} f(x_i|x_{i-1}) f(y_i|x_i), \tag{2.4.1}$$

from which one can obtain by marginalization or conditioning any other distributions he may be interested in.

2.4.1 Linear Gaussian state-space models and the Kalman filter

Also called dynamic linear model, the linear Gaussian ssm can be obtained from the above setting by considering a Normal prior distribution $x_0 \sim N_n(m_0, C_0)$ and the pair of observation Equation (2.4.2) and the state Equation (2.4.3) below

$$Y_t = F_t X_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, Q_t), \tag{2.4.2}$$

$$X_t = G_t X_{t-1} + \eta_t, \qquad \eta_t \sim N(0, H_t). \tag{2.4.3}$$

Here ε_t and η_t are assu two independent Normal random vectors with mean zero and known positive definite covariance matrices Q_t and H_t respectively and independent of x_0. Whereas F_t and G_t are two known system matrices of respective dimensions $n \times n$ and $k \times n$. With $Y_t|X_t \sim N(F_tX_t, Q_t)$ and $X_t|X_{t-1} \sim N(G_tX_{t-1}, H_t)$ it can be easily proved that the linear Gaussian ssm fulfils Assumptions (2.4.1) and (2.4.2). Moreover, the error terms in the observation and state equations assumed to be serially uncorrelated and for all lags, they are uncorrelated with each other. Once a dynamic system is written in state-space form, the goal of the researcher is to infer on the latent variables or to make predictions of the next observations based on the available data up to time t, which can be achieve in the linear Gaussian case via the KF as we will explain later. The estimation of the state can be done by simply computing the conditional densities $f(X_l|Y^t)$, with t the time period of interest, the following cases are generally encountered.[3]

Filtering: The inference regarding the state process up to time t is the concern, $l = t$ and it is assumed that the information arrive sequentially in time.

Smoothing: Sometimes called backward analysis, it is concerned with estimating the state process given the observation and $l < t$.

Prediction: This is similar to smoothing except that, we are looking forward, that is $l > t$.

It is important to point out that, although these concept are dealing with the state process, they can be reverted by marginalizing the state and turn the interest to the observation. For example, after computing $f(X_{t+l}|Y^t)$ and by marginalizing X_{t+l}, one can indeed obtain $f(Y_{t+l}|Y^t)$ if interested in. Before moving to the next section, it is important to recall the following propositions from (Petris et al., 2009), that summarizes the filtering, smoothing and forecasting recursions for a general ssm satisfying Assumptions (2.4.1) and (2.4.2).

Proposition 2.4.1 (Filtering recursions). *The one-step-ahead predictive density for the states can be obtained from the filtering density* $f(X_{t-1}|Y^{t-1})$ *as*

$$f(X_t|Y^{t-1}) = \int f(X_t|X_{t-1})f(X_{t-1}|Y^{t-1})dX_{t-1}. \tag{2.4.4}$$

[3]Most of the results in this part will be given without any proof and are borrowed from (Petris et al., 2009).

By marginalization, one can derive the one-step-ahead predictive density for the observations from Equation (2.4.4) *as*

$$f(Y_t|Y^{t-1}) = \int f(Y_t|X_t)f(X_t|Y^{t-1})dX_t. \tag{2.4.5}$$

Finally, using the Bayes' rule, the filtering density is obtained from Equations (2.4.4) *and* (2.4.5) *as*

$$f(X_t|Y^t) = \frac{f(Y_t|X_t)f(X_t|Y^{t-1})}{f(Y_t|Y^{t-1})}. \tag{2.4.6}$$

Proposition 2.4.2 (Smoothing recursions). *Given $t < l$ and Y^l, the backward transition probabilities of the collection of all state processes up to time l are defined by*

$$f(X_t|X_{t+1}, Y^l) = \frac{f(X_{t+1}|X_t)f(X_t|Y^t)}{f(X_{t+1}|Y^t)}. \tag{2.4.7}$$

Starting from $f(X_l|Y^l)$, the smoothing distributions of X_t conditionally on Y^l can be derived from the backward recursion in t described by Equation (2.4.8) *below*

$$f(X_t|Y^h) = f(X_t|Y^t) \int \frac{f(X_{t+1}|X_t)f(X_t|Y^t)}{f(X_{t+1}|Y^t)} f(X_{t+1}|Y^l)dX_{t+1}. \tag{2.4.8}$$

Proposition 2.4.3 (Forecasting recursion). *For $l > 0$, the l-steps-ahead forecast distribution of the state is defined by*

$$f(X_{t+l}|Y^t) = \int f(X_{t+l}|X_{t+l-1})f(X_{t+l-1}|Y^t)dX_{t+l-1}. \tag{2.4.9}$$

By marginalization, l-steps-ahead forecast distribution of the observation can be obtained from Equation (2.4.9) *as*

$$f(Y_{t+l}|Y^t) = \int f(Y_{t+l}|X_{t+l})f(X_{t+l}|Y^t)dX_{t+l}. \tag{2.4.10}$$

In the linear Gaussian case, the filtering and smoothing recursions above become the famous Kalman filter and smoother and are available in more clear, simple and elegant forms. This is

due to the fact that the Normal distribution is entirely characterized by its first two moments and the integrals intervening in the recursions will be analytically tractable. We omit to recall them here as they can easily be derived and we refer the reader to (Petris et al., 2009) for proofs and extensive explanations.

2.4.2 Particle filtering techniques

We recall that, one of the most important goal when dealing with data analysis and especially when real-world examples are considered, is the estimation of some unknown quantities provided that we have at our disposal some observations. In most cases, prior information regarding the phenomenon under investigation is available. Given the availability of this information, prior distributions for the unknown quantities and likelihood functions linking these quantities to the data can be formulated.

In this process, the Bayes' theorem is used to obtain the posterior distribution on which, the inference of the unknown quantities will be based. Most of the time, the data become available sequentially in time, giving rise to the concern of how to perform online inference. That is, inferring on the unknown quantities as the observations become available. For example, the volatility of some financial instruments can be estimated online using stock market data.

Among the wide range of approaches used to solve the filtering problem in state-space models, particle filtering techniques have a well established reputation. If one were to consider particle filters as one side of a given coin, then the other side will certainly be the Monte Carlo techniques whose existence date back to the 1950s (Hammersley and Morton, 1954). Due to the degeneracy problems and the lack of computational power at the time, these methods were often overlooked. With the introduction of the bootstrap filter of (Gordon et al., 1993) and more efficient resampling schemes, the scientific production in this area and related ones has been compelling. No surprise that particle filters have been routinely applied to a broad range of fields, such as target tracking (Ristic et al., 2004), economics (Kim et al., 1998; Johannes et al., 2009), neuroscience (Salimpour and Soltanian-Zadeh, 2009), biochemical networks (Djuric and Bugallo, 2009) and signal processing (Arulampalam et al., 2002) to list a few.

It is worth noticing that, before the vulgarization and the extensive use of particle filtering techniques, the Kalman filter which is optimal under the normality and the linearity assumptions, was the standard solution to the filtering problem. When these two assumptions are violated, one of its variants, namely, the extended Kalman filter and the unscented Kalman filter can be used. However, this versatility presents some drawbacks and fails to give efficient estimates when highly non linear and/or non-Gaussian situations are considered.

In order to account for the limitations presented by the Kalman filter and its variants, particle filtering methods were then introduced. Unlike other previously mentioned filtering solutions, particle filters can be applied under more general settings, making this approach very flexible and with the constant development of computing power, the future of particle filtering is even more promising.

This algorithm is a very powerful tool for non standard state-space models and it is even more attractive when dealing with online problems. Nowadays the storage limit that we usually face in data analysis is a big concern and in this respect, it is important to point out that sequential techniques are viable remedies to this issue. In fact, they offer the advantage of not having to store all the data. For detailed explanations regarding particle filters and their applications, we refer the reader to (Doucet et al., 2000, 2001), (Ristic et al., 2004), (Pollock, 2010), (Doucet and Johansen, 2009) and references hereafter.

The key concept in particle filtering is to sequentially update a given distribution using importance sampling algorithms in conjunction with the Bayesian methodology. More precisely, the marginal density of the unobserved process is approximated provided that we have at our disposal a set of available information, a measurement model, the initial estimates of the state probability density function and a non-linear state process model. Importance sampling is then used at each time index to approximate the distribution with a set of discrete values, also known as particles, each particle associated to a corresponding weight. One of the property of this Monte Carlo method is that, the particle filter representation of the posterior probability density function will converge to the true one as the sample size increases.

With the introduction of the bootstrap filter by (Gordon et al., 1993), particle filtering techniques became the appropriate tool for estimating general state-space models. Although this can be

considered as the seminal paper for online estimation, its construction was mainly motivated by the great works on importance sampling of (Müller, 1991) and (Smith and Gelfand, 1992). This is one of the reason why the bootstrap filter is sometimes defined as a mixture of sequential importance sampling and resampling procedures. Let us also point out that, all filtering algorithms having the same basis are regrouped under the label of Sequential Monte Carlo methods (SMC). Importance sampling is therefore a key concept that is worth being defined as we now do.

2.4.2.1 Importance sampling (IS)

Let us consider a state-space model in its general form defined by the observation equation $f(y_{t+1}|y_t)$ and the state equation $f(x_{t+1}|x_t)$, where for simplicity and without any loss of generality, all parameters are assumed to be known. Furthermore, we also assume that the observed process y_t and the latent process x_t are continuous and have discrete values at any time index of interest. The interest here lies in computing and efficiently estimate the filtering density $f(x_t|y^t)$ where y^t represents all available observations up to time t. One approach of studying the filtering density when the later is not analytically tractable is by using importance sampling (IS).

Importance sampling that can be defined as the art of choosing a good distribution from which random variables can be simulated, forms the basis of SMC used in particle filtering to provide a solution to the recursion problem. Monte Carlo methods are a kind of stochastic integration employed to approximate expectations by using the law of large numbers. That is, if we were to compute the following integral,

$$I = \int_a^b h(y)dy = \int_a^b w(y)f(y)dy = \mathbb{E}(w(Y)), \tag{2.4.11}$$

where $f(y) = \dfrac{1}{b-a}$ is the probability density function of the random variable Y following the uniform distribution $U(a,b)$ and $w(y) = h(y)(b-a)$. Then by the law of large number (LLN),

if N iid samples from $U(a,b)$ are considered, then the integral I can be estimated by,

$$\widehat{I} = \frac{1}{N}\sum_{i=1}^{N} w(Y_i) \to \mathbb{E}(w(Y)) = I \tag{2.4.12}$$

This technique works great if we are able to sample from the desired distribution also known as the target distribution. Overcoming the inability to sample from the target distribution, motivated the introduction of IS. In this case, the sampling is made from an other distribution known as the proposal distribution and the integral is re-weighted using importance weights in order for the true distribution to be targeted.

More precisely, let us assume one would like to compute the integral in Equation (2.4.13)

$$I = \int h(y)f(y)dy, \tag{2.4.13}$$

where f is the probability density function associated to the random variable Y and h is some function. Furthermore, if we put ourself into the situation where it is difficult to draw samples from the density f, then IS can be used to compute I by specifying a different probability density function q as the proposal density as we now do.

$$I = \int h(y)f(y)dy = \int h(y)\frac{f(y)}{q(y)}q(y)dy = \int \frac{h(y)f(y)}{q(y)}q(y)dy, \tag{2.4.14}$$

that is,

$$I = \mathbb{E}_f[h(Y)] = \int \frac{h(y)f(y)}{q(y)}q(y)dy = \mathbb{E}_q\left[\frac{h(Y)f(Y)}{q(Y)}\right]. \tag{2.4.15}$$

Given iid samples $y_i, i = \{1, \cdots, N\}$ from $q(y)$, the integral I can be estimated as

$$\widehat{I} = \frac{1}{N}\sum_{i=1}^{N}\frac{h(y_i)f(y_i)}{q(y_i)} \to \mathbb{E}_q\left[\frac{h(Y)f(Y)}{q(Y)}\right] = I. \tag{2.4.16}$$

In this process, one has to be careful in selecting the proposal as the standard error of $\widehat{I}$ can

be infinite when this selection is not done appropriately. One way of choosing g is by avoiding the ratio f/g to be large. That is, the density g should have similar shape to f but with ticker tail. Other important things to have in mind when selecting the proposal, are the ability to sample from g, otherwise we will return to the initial issue, the variance of $\widehat{I}$ is minimized when $g(y) \propto |f(y)|$ and g and f should have the same support (Pollock, 2010). This approach can be very useful, especially in a Bayesian framework where a probability distribution is only known up to a normalizing constant, IS can be used to provide a good approximation of the density and reduces the computation time as there is no need to compute the normalizing constant.

2.4.2.2 Importance sampling in Bayesian framework

In Bayesian inference, the computation of posterior expectations of the form

$$\mathbb{E}\left[g(x)|y^t\right] = \int g(x)f(x|y^t)dx \tag{2.4.17}$$

is often needed. To achieve this, one can use Monte Carlo techniques to draw samples x^i from $f(x|y^t)$ and estimate the expectation as

$$\mathbb{E}\left[g(x)|y^t\right] \approx \frac{1}{N}\sum_{i=1}^{N} g(x^{(i)}), \tag{2.4.18}$$

where $f(x)$ is simply the posterior distribution. When it is impossible to draw samples from $f(x|y^t)$ as it is often the case, IS is used to draw samples x^i from an important distribution $\pi(x|y^t)$ and computes the weights $w^{(i)}$ in such a way that,

$$\mathbb{E}\left[g(x)|y^t\right] \approx \sum_{i=1}^{N} w^{(i)}g(x^{(i)}). \tag{2.4.19}$$

Note that the derivation of IS will be based on the identity

$$\begin{aligned}\mathbb{E}\left[g(x)|y^t\right] &= \int g(x)f(x|y^t)dx \\ &= \int \left[g(x)\frac{f(x|y^t)}{\pi(x|y^t)}\right]\pi(x|y^t)dx,\end{aligned} \tag{2.4.20}$$

from which Monte Carlo approximations are computed as

$$\mathbb{E}\left[g(x)|y^t\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{f(x^{(i)}|y^t)}{\pi(x^{(i)}|y^t)}g(x^{(i)}), \tag{2.4.21}$$

where the importance weights $w^{(i)}$ are defined by the ratio

$$\frac{1}{N}\sum_{i=1}^{N}\frac{f(x^{(i)}|y^t)}{\pi(x^{(i)}|y^t)}. \tag{2.4.22}$$

Except in some simple cases, we often encounter the issue of computing the normalizing constant of $f(x^{(i)}|y^t)$. The way around this is to define the unnormalized importance weights

$$\tilde{w}^{(i)} = \frac{f(y^t|x^{(i)})f(x^{(i)})}{\pi(x^{(i)}|y^t)}, \tag{2.4.23}$$

then obtain their normalized version as $w^{(i)} = \dfrac{\tilde{w}^{(i)}}{\sum_j \tilde{w}^{(j)}}$ and finally the posterior expectation of $g(x)$ and the posterior pdf are respectively approximated by Equations (2.4.24) and (2.4.25) below.

$$\mathbb{E}\left[g(x)|y^t\right] \approx \sum_{i=1}^{N} w^{(i)}g(x^{(i)}), \tag{2.4.24}$$

$$f(x|y^t) \approx \sum_{i=1}^{N} w^{(i)}\delta(x - x^{(i)}), \tag{2.4.25}$$

where δ is the Dirac delta function. As stated before, IS is the basis to several sampling algorithms. In what follows, we briefly recall some of its extensions.

2.4.2.3 Sequential importance sampling (SIS)

When dealing with particle filtering techniques, our interest often lies in the marginal or joint distribution of the unobserved variables provided that, we have available all observations up

to the time index t of interest. However, and as already mentioned, the normalizing constant $f(y_t|y^{t-1})$ can be intractable. This intractability prohibits direct calculation and therefore IS shall come at the rescue. We want to be able to update sequentially the posterior distribution at time t without letting occur any modification on the previously simulated states x^{t-1}.

Now, let us assume the existence of an importance function $q(x^t|y^t)$, from which samples can be drawn easily and that our density of interest $f(x^t|y^t)$ has the same support as $q(x^t|y^t)$, that is,

$$f(x^t|y^t) > 0 \Rightarrow q(x^t|y^t) > 0.$$

Moreover and as in (Ristic et al., 2004), let us assume that the importance function is chosen in such a way that it can be recursively updated in time when ever the next information becomes available. Meaning that, the propagation of the current estimate to the next time period by keeping the same past simulated trajectories $x^i, i = \{1, \cdots, t-1\}$ is guarantee. In order words, we suppose the following to be true,

$$q(x^t|y^t) = q(x^{t-1}|y^{t-1})q(x_t|x^{t-1}, y^t).$$

What is known as the recursive problem will make the joint density available as follows

$$\begin{aligned} f(x_t|y^t) &= f(y_t|x_t) \int f(x_{t-1}|y^{t-1}) \frac{f(x_t|x_{t-1})}{f(y_t|y^{t-1})} dx_{t-1}, \\ &= f(y_t|x_t) \int f(x_{t-1}|y^{t-1}) \frac{f(x_t|x_{t-1}) q(x^{t-1}|y^{t-1}) q(x_t|x^{t-1}, y^t)}{f(y_t|y^{t-1}) q(x^{t-1}|y^{t-1}) q(x_t|x^{t-1}, yt)} dx_{t-1}, \\ &= f(x_t|x_{t-1}) \int \frac{f(x_{t-1}|y^{t-1}) q(x^{t-1}|y^{t-1})}{q(x^{t-1}|y^{t-1})} \frac{f(x_t|x_{t-1}) q(x_t|x^{t-1}, y^t)}{f(y_t|y^{t-1}) q(x_t|x^{t-1}, y^t)} dx_{t-1}. \end{aligned} \tag{2.4.26}$$

Provided the existence of samples $x_{t-1}^{(i)}$ drawn from the density $f(x_{t-1}|y^{t-1})$ with corresponding weights $w_{t-1}^{(i)}$, an approximation of the marginal density $f(x_t|y^t)$ can be obtained via IS with normalized weights $w_t^{(i)}$ defined by

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}},$$

where $\tilde{w}_t^{(i)}$ represent the unnormalized weights and are defined as

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{f(y_t|x_t)f(x_t|x_{t-1})}{q(x_t|x^{t-1}, y^t)}, \quad \text{for} \quad i = 1, \cdots, N.$$

At time point t of interest, the filtering density will then be approximated by the weighted set of particles $\left\{w_t^{(i)}, x_t^{(i)}\right\}_t^N$. With the fact that the transitional and the likelihood functions are available and that we are able to draw samples from the importance function, the remaining components we need to estimate the marginal density are generating initial particles and computing iteratively the importance weights. This process describes the sequential importance sampling algorithm of (Kong et al., 1994) as illustrated in Figure (2.3) and recalled in Algorithm (1) below.

Even though it is not possible, one will ideally wish to have as posterior distribution the importance density function. For the type of importance function presented in the SIS algorithm, it can be proved as in (Kong et al., 1994) that the variance of the importance weights increases at every time step. Meaning that we will almost surely converge to single non-zero weight $w^{(i)} = 1$, while other weights being negligible after few iterations . As a consequence, the accuracy of the estimate will become a major concern and a large amount of computational effort will be required when updating particles with nearly zero weights. This phenomenon is referred as the degeneracy problem which is a big issue in particle filtering. Fortunately, this can be handled by increasing the number of particles used (sometimes impractical) or by a combination of a resampling step in the SIS algorithm and a good choice of the importance density (Gordon et al., 1993; Doucet and Johansen, 2009).

2.4.2.4 Sequential importance resampling (SIR)

Originally introduced by (Gordon et al., 1993), the SIR as summarized in Algorithm (2) and illustrated in Figure (2.4) can be obtain from the above SIS algorithm by simply adding some steps in it. The key being resampling, which provides the tools to obtain approximation from the target distribution by resampling N particles from the IS approximation (Douc et al., 2014). Each particle is chosen with a probability proportional to its corresponding weight. In order words, particles with smallest weights will have high probability to be removed, by doing so, we will end

up with several copies of particles having bigger weight.

There are several techniques for implementing resampling, the stratified resampling which is optimal when the variance is considered and the adaptive resampling among others. In the adaptive case, the resampling is done when the number of effective samples $n_{eff} \approx \frac{1}{\sum_{i=1}^{N}(w_t^{(i)})^2}$ is too small, (for instance $N/10$). Although resampling reduces considerably the degeneracy issue, we have to keep in mind that it is done at the expense of additional issues such as, the interaction occurring between simulated particles that create their statistical dependence and the fact that particles obtained may not be diversified. As illustrated by Figures (2.3) and (2.4), the following summarize the SIS and SIR algorithms.

Algorithm 1: Sequential Importance Sampling

Initialization: At time index $t = 0$

for $i \leftarrow 1$ **to** N **do**

  Draw samples $x_0^{(i)}$ from the prior distribution $f(x_0)$;

  Compute the unnormalized weights as:

$$\tilde{w}_0^{(i)} = f(y_0|x_0^{(i)})$$

  Compute the normalized weights as:

$$w_0^{(i)} = \frac{\tilde{w}_0^{(i)}}{\sum_{j=1}^{N} \tilde{w}_0^{(j)}}$$

end

Iteration :

for $t \leftarrow 1$ **to** T **do**

  for $i \leftarrow 1$ **to** N **do**

    Draw samples $x_t^{(i)}$ from $q(x_t|x_{t-1}^{(i)}, y_{0:t})$ and set $x_{0:t}^{(i)} = (x_{0:t-1}^{(i)}, x_t^i)$;

    Compute the unnormalized weights as:

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-1}^{(i)} \frac{f(y_t|x_t^{(i)}) f(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{0:t})}$$

    Compute the normalized weights as:

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}$$

  end

end

Return : $\left\{x_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N}, \; t = 0, \cdots, T$

Algorithm 2: Sequential Importance Resampling

Initialization: At time index $t = 0$

for $i \leftarrow 1$ **to** N **do**

  Draw samples $x_0^{(i)}$ from the prior distribution $f(x_0)$;

  Compute the unnormalized weights as:

$$\tilde{w}_0^{(i)} = f(y_0|x_0^{(i)})$$

  Compute the normalized weights as:

$$w_0^{(i)} = \frac{\tilde{w}_0^{(i)}}{\sum_{j=1}^{N} \tilde{w}_0^{(j)}}$$

end

Iteration :

for $t \leftarrow 1$ **to** T **do**

  for $i \leftarrow 1$ **to** N **do**

    Resample $\tilde{x}_{t-1}^{(i)}$ from $\left\{x_{t-1}^{(i)}\right\}_{i=1}^{N}$ with probability $\left\{w_{t-1}^{(i)}\right\}_{i=1}^{N}$ and set $\left\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}_{i=1}^{N} = \left\{\tilde{x}_{t-1}^{(i)}, \frac{1}{N}\right\}_{i=1}^{N}$;

    Compute the unnormalized weights as:

$$\tilde{w}_t^{(i)} = \frac{f(y_t|x_t^{(i)})p(x_t^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y^t)}$$

    Compute the normalized weights as:

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}$$

  end

end

Return : $\left\{x_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N}$, $t = 0, \cdots, T$

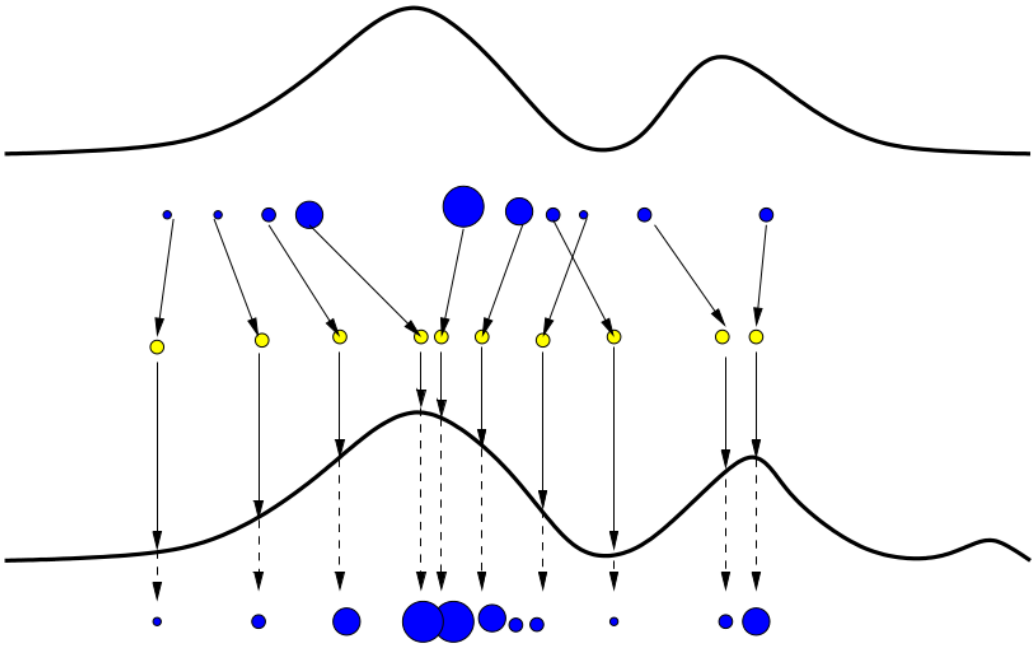


Figure 2.3: SIS illustration.

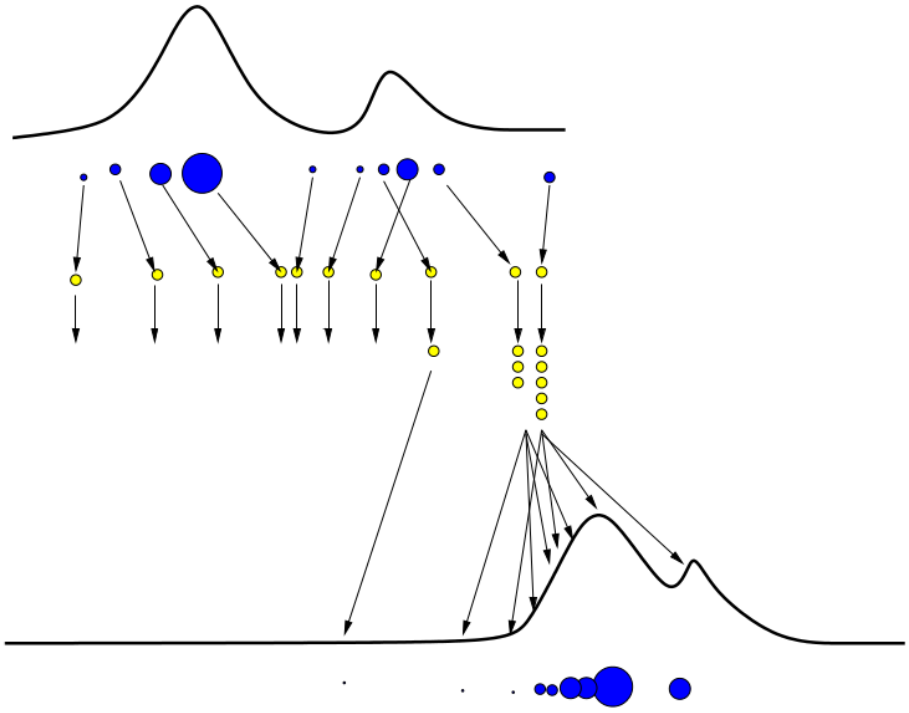


Figure 2.4: SIR illustration.

3. Kalman filter with asymmetric distributions

Mathematical models have been proposed and developed to model time observations dynamics over the past two centuries, achieving remarkable gains. Among them, we have the state-space models which can be seen as a subclass of graphical probabilistic models describing from a probabilistic point of view, the dependence between the state (latent) and the measurement (observed) variables. Till these days, the most studied filtering solution in the ssm framework is certainly the Kalman filter. In order to take into account more general applications and asymmetric distributions, some attempts to incorporate skewed distributions in the ssm and to build the so-called skewed Kalman filter have been made recently in the literature. Unfortunately, some of these contributions present considerable issues such as the skewness vanishing after several iterations, a not efficient characterization of the filtering densities and some errors. In this chapter, we study and propose a skewed Kalman filter which overcomes the above mentioned issues. We develop procedures and algorithms for prediction, filtering and estimation based on the closed skew-normal distributions, of which Gaussian are simply special cases.

3.1 Introduction

During the last two centuries, several mathematical models have been proposed and developed for the modelling of temporal observations achieving important and remarkable results. However, this is not and can not be considered as the end of the road. This because, achieving efficient and accurate estimates of the key parameters, as well as adequate forecast, remains with these models, a challenging task, especially in the cases of non-Gaussianity.

The widely used and so-called ssm that are the basis of this research, fall within the class of the above mentioned models. Koller and Friedman (2009) defined the ssm as a class of proba-

bilistic graphical models which describes the dependence between the state variable (latent) and measured variable (observable).

The expression *State-Space* was originally used in engineering and was introduced by Rudol Kalman, an electrical engineer Kalman (1960) who at that time worked at the research institute for advanced studies in Baltimore, Maryland. It should be noted that to date, the most studied and well defined discrete filtering algorithm for ssm is known under the name of Kalman filter.

An important feature of the Kalman filter is that predictions of future values can be obtained with an efficient recursive algorithm that finds application in various fields such as physics, chemistry, biology, engineering sciences and economics.

This model, however presents some lack of efficiency, especially in the financial sector, but not limited to where there is evidence that observations do not follow a Gaussian distribution. The classic Kalman filter which is built on the assumption of Gaussian observations, does not allow to take into account characteristics such as asymmetry and kurtosis.

This chapter aims to construct and study a similar model but more general, with well-defined statistical properties, which overcomes the limitations present in the classical Gaussian model. To achieve this, we will use the closed skew-normal (csn) distributions of which the Gaussian distribution is simply a special case. The model to be developed will also have to overcome the limitations of current skewed Kalman filters in the literature.

In the next section, we review some key properties and theorems of the csn distributions that will be of great importance here. Section 3 revisits the skewed Kalman filter of Naveau et al. (2005) and show that their filtering density does not follow a csn distribution which makes their filter incorrectly characterize. The description of the approach we intend to use, the presentation of an example along with the parameters estimation of the model are done in Section 4. Some concluding remarks appear in Section 5.

3.2 Some important properties of csn distributions

Among the numerous candidates of asymmetric distributions that could be incorporated in our model, we have the class of multivariate skew-normal distributions, which was originally proposed by Azzalini and Dalla Valle (1996) as an extension of the class of Normal distributions. Many studies have proven the appealing ease and convenience, when modelling the presence of skewness with these distributions.

Let us recall the fact that, we would like to mimic as much as possible, the standard Kalman filter and keep almost the same properties if not all of them. This implies that, the choice of the distribution in the above mentioned family is an important step in our study. More precisely, we would like the chosen distribution, to be closed under summation, marginalization and conditioning. Meaning for instance that, the sum or the conditional distribution of random variables, belonging to the same subset of distributions, will remain in that same subset.

From the most recent developments in the literature and based on the above mentioned facts, the perfect candidate in our opinion is certainly the csn. Moreover, it is of great importance to notice that, the presence of additional parameters to be estimated and the possibility of the independence structure between the csn and the Normal distribution in the multivariate setting, make the csn even a more efficient choice.

In addition, we can observe that, in order to obtain the right or left skewness in the csn distribution, one can take the mean of the Gaussian density component and move the mass to the right or to the left respectively. By performing this operation, we keep the tail comparable to the Normal case while gaining flexibility, asymmetry and an easier parameters estimation compare to other asymmetric distributions. Finally, because of the multivariate structure of the csn distributions, the issue of the cross-correlation among the unobservable variables become easier to deal with.

In what follows and after recalling the definition of the csn distribution, we will provide some important properties of the csn distribution that constitute the cornerstone in the implementation of our model. Although we are using the notation from Flecher et al. (2009), González-Farías et al. (2004b) and González-Farías et al. (2004a), we will mainly refer to theorems and properties from the later one and lemmas from Kim et al. (2014a) .

Definition 3.2.1 (Multivariate close skew-normal pdf). *A random vector X has a multivariate csn distribution according to González-Farías et al. (2004a) if its probability density function is given by:*

$$f(x) = \frac{1}{\Phi_{m_x}(0; \nu, \Delta + \Gamma\Sigma\Gamma^T)} \phi_{n_x}(x; \mu, \Sigma)\Phi_{m_x}(\Gamma(x-\mu); \nu, \Delta); \quad x \in \mathbb{R}^{n_x}, \tag{3.2.1}$$

where $\mu \in \mathbb{R}^{n_x}$ is the location vector, $\nu \in \mathbb{R}^{m_x}$, is the additional parameter allowing the closure under conditioning, $\Delta \in \mathbb{R}^{m_x \times m_x}$ is a covariance matrix insuring the closure under marginalization, $\Sigma \in \mathbb{R}^{n_x \times n_x}$ is the scale parameter and also a covariance matrix, $\Gamma \in \mathbb{R}^{m_x \times n_x}$ regulate the skewness, $\phi_{n_x}(x; \mu, \Sigma)$ and $\Phi_{m_x}(x; \mu, \Sigma)$ are respectively the Normal probability density function (pdf) and cumulative density function (cdf) with mean vector μ and covariance matrix Σ.

Additionally, the presence of the Normal cdf $\Phi_{m_x}(x; \mu, \Sigma)$ governs the closure properties for the joint and the sum of independent csn random vectors when $m_x \geq 1$. One can check without any difficulties that, when $\Gamma = 0$ the csn distribution reduces to the Gaussian. The notation $X \sim \text{csn}_{n_x, m_x}(\mu, \Sigma, \Gamma, \nu, \Delta)$ is often used to denote that the random variable X has a pdf as in Equation 3.2.1.

To illustrate some properties of the csn distributions, let us consider a univariate random variable z with the following benchmark distribution $z \sim \text{csn}_{1,1}(\mu, \sigma, \gamma, \nu, \delta)$. By setting $\mu = 5, \gamma = 3$ and $\nu = 0$ while playing around with other parameters. As shown in Figure 3.1 below, we can see that inducing high level of skewness in the model can be achieved by increase σ or decrease δ while keeping other parameters unchanged. On the other hand, by decreasing the value of ν while keeping all the other parameters fixed and with a high value of σ, we can see that bigger ν induce more skewness. Therefore, the effect of ν on the skewness is as much important. In the left panel, we fixed $\mu = 5, \gamma = 3$ and $\nu = 0$ for all cases. The density in orange corresponds to $(\sigma = 8, \delta = 1.5)$, the blue density is the case where $(\sigma = 4, \delta = 1)$, the green one is when $(\sigma = 3, \delta = 0.5)$, and the red density shows how the csn reduces to the Normal when $\gamma = 0$. In the right panel, all parameters except of ν are constant with values $\mu = 5, \sigma = 6, \gamma = 1$ and $\delta = 1.5$. When $\nu = 0, \nu = -3$ and $\nu = -12$, we have respectively the orange, blue and green density.

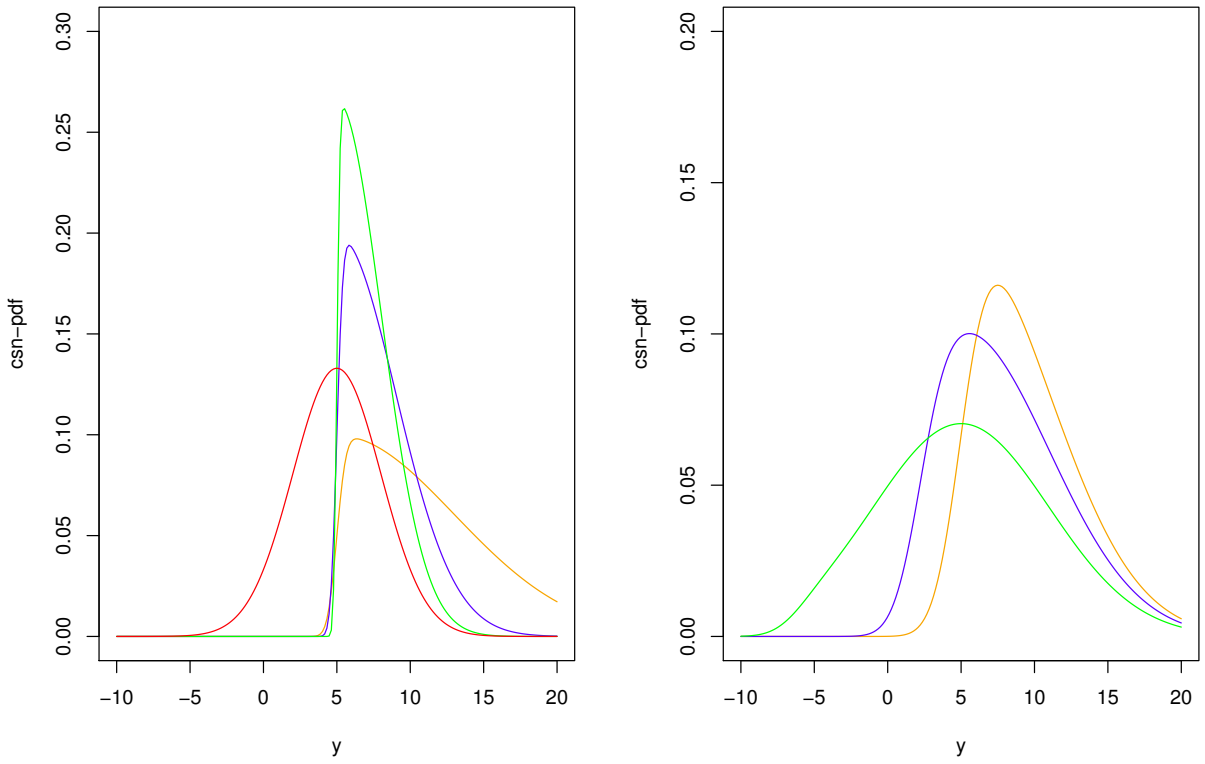


Figure 3.1: Densities of the $\text{csn}_{1,1}(\mu, \sigma, \gamma, \nu, \delta)$ distributions.

The following Theorems 3.2.1 and 3.2.2 characterize respectively, the closure properties of the joint distribution and the sum of independent csn random vectors.

Theorem 3.2.1. *If* $y_1, \cdots, y_n$ *are independent random vectors with* $y_i \sim csn_{p_i,q_i}(\mu_i, \Sigma_i, D_i, \nu_i, \Delta_i)$ *then the joint distribution of* $y_1, \cdots, y_n$ *is* $Y = (y_1', \cdots, y_n')' \sim csn_{p^\dagger,q^\dagger}(\mu^\dagger, \Sigma^\dagger, D^\dagger, \nu^\dagger, \Delta^\dagger)$,

$$\textit{where} \quad p^\dagger = \sum_{i=1}^{n} p_i, \quad q^\dagger = \sum_{i=1}^{n} q_i, \quad \mu^\dagger = (\mu_1', \cdots, \mu_n')', \quad \Sigma^\dagger = \bigoplus_{i=1}^{n} \Sigma_i,$$

$$\textit{and} \quad D^\dagger = \bigoplus_{i=1}^{n} D_i, \quad \nu^\dagger = (\nu_1', \cdots, \nu_n')', \quad \Delta^\dagger = \bigoplus_{i=1}^{n} \Delta_i.$$

Theorem 3.2.2. *If* $y_1, \cdots, y_n$ *are independent random vectors with* $y_i \sim csn_{p,q_i}(\mu_i, \Sigma_i, D_i, \nu_i, \Delta_i)$

then the distribution of the random vector $Y = \sum_{i=1}^{n} y_i$ *is* $Y \sim csn_{p,q^\star}(\mu^\star, \Sigma^\star, D^\star, \nu^\star, \Delta^\star)$,

where $q^\star = \sum_{i=1}^{n} q_i, \quad \mu^\star = \sum_{i=1}^{n} \mu_i, \quad , \Sigma^\star = \sum_{i=1}^{n} \Sigma_i, \quad D^\dagger = (\Sigma_1 D_1^{'}, \cdots, \Sigma_n D_n^{'})^{'}, \quad \nu^\star = (\nu_1^{'}, \cdots, \nu_n^{'})^{'},$

and

$$\Delta^\star = \Delta^\dagger + D^\dagger \Sigma^\dagger D^{\dagger'} - \left[\bigoplus_{i=1}^{n}(D_i\Sigma_i)\right]\left(\sum_{i=1}^{n}\Sigma_i\right)^{-1}\left[\bigoplus_{i=1}^{n}(\Sigma_i D_i^{'})\right]$$

$\Delta^\dagger, D^\dagger$ *and* $\Sigma^\dagger$ *are defined as in Theorem 3.2.1.*

The following lemma given by Naveau et al. (2005) and used in Kim et al. (2014a) is another important tool that we will need to describe the observations process as the sum of independent Gaussian and csn processes.

Lemma 3.2.1. *Let* $y \sim csn_{n_y,m_y}(\mu, \Sigma, \Gamma, \nu, \Delta)$ *and* $z \sim N_{n_y}(\psi, \Omega)$ *independent of* y*, then the process* $y+z \sim csn_{n_y,m_y}(\mu_{y+z}, \Sigma_{y+z}, \Gamma_{y+z}, \nu, \Delta_{y+z})$*, where* $\mu_{y+z} = \mu+\psi, \Sigma_{y+z} = \Sigma+\Omega, \Gamma_{y+z} = \Gamma\Sigma\Sigma_{y+z}^{-1}$ *and* $\Delta_{y+z} = \Delta + (\Gamma - \Gamma_{y+z})\Sigma\Gamma'$.

3.3 The skewed Kalman filter of Naveau et al. (2005)

A look into the literature of time series modelling shows a considerable part of probabilistic models that derived from stochastic dynamic linear models, sometimes referred as ssm. The interest in the use of ssm for time series modelling has grown tremendously in the recent years. See for example West et al. (1985), Gamerman and Migon (1993), Migon et al. (2005), Durbin and Koopman (2012) and references hereafter, among others.

In the ssm framework, an observed time series is seen as the result or output of a dynamic system perturbed by random fluctuations. While providing significant flexibilities, analysis with ssm often presents issues related to estimation and forecasting. These issues can be addressed by using a recursive algorithm that calculates the conditional density of the future observable variable given

the latent variable and the information available up to the time period of interest. This is basically the reason why the Bayesian approach is particularly suitable for this purpose.

In the literature, some algorithms used to estimate ssm suffer from computational issues, due to the use of all past observations, to estimate the current or future state of the system. On the contrary, the Kalman filter, uses recursively the last prediction made and corrected with the new measurement to obtain predictions a-priori and a-posteriori of the state of the system. This iteration of the forecast and the correction is primarily based on Bayes' formula.

As a matter of fact, difficulties arise in the case of data which have characteristics that can not be traced back to a Normal distribution. In this case, how can the Kalman filter be extended in order to overcome the limitations induced from the assumption of normality and at the same time, without losing the well behaving characteristics of the standard filter? One possible answer to this question can be given by the use of a more general statistical distribution that includes the Gaussian distribution as a special case. In this chapter, we intend to use of the *closed skew-normal* distributions (csn) which in addition of having many of the properties of the Gaussian distribution allow for greater flexibility regarding the skewness of the observations.

As mentioned early, in the recent past years, some skewed Kalman filter have been proposed in the literature. But it is still unclear how the marginal conditional densities of the state variables given the observations was characterized. The problem is that their filtering procedure is entirely based on that assumption. In what follows, we will have a look into it and see how valid their assumption is.

In the skewed Kalman filter of Naveau et al. (2005), at time index $t = 1, 2, \ldots$, Equations 3.3.1 respectively 3.3.2 below, characterize a vector of observations Y_t as a function of the unobserved states X_t respectively, the state of the system which follows an autoregressive process.

$$\begin{cases} Y_t = F_t X_t + \varepsilon_t & (3.3.1) \\ X_t = G_t X_{t-1} + \eta_t, & (3.3.2) \end{cases}$$

where F_t, G_t are scalar matrices and ε_t, η_t additive noises. For clarifications, Equations 3.3.1 and 3.3.2 correspond respectively to Equations 3 and 4 in the above mentioned paper.

In order to simplify the computations, we will be restraining ourselves to the scalar (1-dimension) form of Equations 3.3.1 and 3.3.2, which are then rewritten as $y_t = f_t x_t + \varepsilon_t$ and $x_t = g_t x_{t-1} + \eta_t$ respectively.

From their Lemma 2, one way to obtain the closure property under summation, is to consider both error terms (ε_t and η_t) to be normally and independently distributed. Thus, for the skewed Kalman filter in their Paragraph 3.1 we chose, $\varepsilon_t \sim \mathcal{N}(\mu_\varepsilon; \sigma_\varepsilon^2)$ and $\eta_t \sim \mathcal{N}(\mu_\eta; \sigma_\eta^2)$ independent and identically distributed (iid). Furthermore, for simplicity and without any loss of generality, we allow parameters to take the following values: $\mu_\varepsilon = 0, \quad \mu_\eta = 0, \quad \psi_0 = 0, \quad \eta_0 = 0, \quad \Omega_0 = \omega_0^2, \quad D_0 = \alpha \quad$ and $\quad \Delta_0 = 1$. In this specific part, we will be computing some marginal and conditional densities with appropriate parameters, the aim being to ease our analysis. Based on our assumptions, the initial distribution of the state equation can be written as follows.

$$\begin{aligned} x_0 &\sim \mathsf{csn}_{1,1}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0), \\ &\sim \mathsf{csn}_{1,1}(0, \omega_0^2, \alpha, 0, 1). \end{aligned}$$

Thus, the probability density function (pdf) of x_0 becomes

$$\begin{aligned} f(x_0) &= \frac{1}{\Phi(0; \nu_0, \Delta_0 + D_0 \Omega_0 D_0^T)} \phi(x_0; \psi_0, \Omega_0) \Phi(D_0(x_0 - \psi_0); \nu_0, \Delta_0), \\ &= \frac{1}{\Phi(0; 0, 1 + \alpha^2 \omega_0^2)} \phi(x_0; 0, \omega_0^2) \Phi(\alpha x_0; 0, 1) \end{aligned}$$

and from the definition of the general multivariate skew-normal density (GMSN) Gupta et al. (2004a), we have $X_0 \sim GMSN_{1,1}(0, \omega_0^2, \alpha_{,0}, 1) = SN_{1,1}(0, \omega_0^2, \alpha_{,0}, 1) = 2\phi(x_0; \omega_0^2)\Phi(\alpha x_0 \omega_0^{-1})$. Their proposition 3 on pages 385-386 then becomes,

Proposition 3.3.1. *Suppose* $x_0 \sim csn_{1,1}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0)$, *if* $\varepsilon_t \sim \mathcal{N}(\mu_\varepsilon; \sigma_\varepsilon^2)$ *iid and* $\eta_t \sim \mathcal{N}(\mu_\eta; \sigma_\eta^2)$ *iid, then* $x_1 \sim csn_{1,1}(\psi_1, \Omega_1, D_1, \nu_1, \Delta_1)$ *and* $y_1 \sim csn_{1,1}(\mu_1, \Gamma_1, E_1, \gamma_1, \Theta_1)$.

Where

$$\begin{aligned} \psi_1 &= g_1 \psi_0 + \mu_\eta = 0, \\ \Omega_1 &= g_1 \Omega_0 g_1^T = g_1^2 \omega_0^2 + \sigma_\eta^2, \end{aligned}$$

$$D_1 = D_0\Omega_0 g_1^T\Omega_1^{-1} = \frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2},$$

$$\nu_1 = \nu_0 = 0,$$

$$\begin{aligned}\Delta_1 &= \Delta_0 + (D_0 - D_1 g_1)\Omega_1 D_0^T = 1 + \left(\alpha - \frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2} g_1\right)\left(g_1^2\omega_0^2+\sigma_\eta^2\right)\alpha \\ &= 1 + \frac{\alpha(g_1^2\omega_0^2+\sigma_\eta^2) - g_1^2\omega_0^2 + \sigma_\eta^2}{g_1^2\omega_0^2+\sigma_\eta^2} = 1 + \alpha\frac{\left[g_1^2\omega_0^2+\sigma_\eta^2 - g_1^2\omega_0^2\right]}{g_1^2\omega_0^2+\sigma_\eta^2}(g_1^2\omega_0^2+\sigma_\eta^2)\alpha = 1+\alpha^2\sigma_\eta^2.\end{aligned}$$

The density of x_1 is then,

$$f(x_1) = \frac{1}{\Phi(0;\nu_1,\Delta_1 + D_1\Omega_1 D_1^T)}\phi(x_1;\psi_1,\Omega_1)\Phi(D_1(x_1-\psi_1);\nu_1,\Delta_1), \tag{3.3.3}$$

which can be rewritten as

$$f(x_1) = \frac{\phi\left(x_1;0,(g_1^2\omega_0^2+\sigma_\eta^2)\right)\Phi\left(\left(\frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2}x_1\right);0,(1+\alpha^2\sigma_\eta^2)\right)}{\Phi\left(0;0,(1+\alpha^2\sigma_\eta^2)+\left(\frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2}\right)^2(g_1^2\omega_0^2+\sigma_\eta^2)\right)}. \tag{3.3.4}$$

When $t=1$, Equation 3.3.1 becomes $y_1 = f_1 x_1 + \varepsilon_1$. This implies that, $y_1|x_1 \sim \mathcal{N}(f_1 x_1, \sigma_\varepsilon^2)$. That is,

$$f(y_1|x_1) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y_1-f_1x_1)^2\right\} = \phi(y_1;f_1x_1,\sigma_\varepsilon^2). \tag{3.3.5}$$

Now, let us recall that if y_1 and x_1 are realizations of two random variables, then from the Bayes' rule their joint density is defined by $f(x_1,y_1) = f(y_1|x_1)f(x_1)$ and the following holds,

$$f(x_1,y_1) = \phi(y_1;f_1x_1,\sigma_\varepsilon^2)\frac{\phi\left(x_1;0,(g_1^2\omega_0^2+\sigma_\eta^2)\right)\Phi\left(\left(\frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2}x_1\right);0,(1+\alpha^2\sigma_\eta^2)\right)}{\Phi\left(0;0,(1+\alpha^2\sigma_\eta^2)+\left(\frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2+\sigma_\eta^2}\right)^2(g_1^2\omega_0^2+\sigma_\eta^2)\right)}. \tag{3.3.6}$$

From their Proposition 3 again, we also have that $y_1 \sim \text{csn}_{1,1}(\mu_1, \Gamma_1, E_1, \gamma_1, \Theta_1)$, or equivalently

$$f(y_1) = \frac{1}{\Phi(0; \gamma_1, \Theta_1 + E_1\Gamma_1 E_1^T)} \phi(y_1; \mu_1, \Gamma_1)\Phi(E_1(y_1 - \mu_1); \gamma_1, \Theta_1),$$

with

$$\begin{aligned}
\mu_1 &= f_1\psi_1 + \mu_\varepsilon = 0, \\
\Gamma_1 &= f_1\Omega_1 f_1^T = f_1^2(g_1^2\omega_0^2 + \sigma_\eta^2 + \sigma_\varepsilon^2), \\
E_1 &= D_1\Omega_1 f_1^T\Gamma_1^{-1} = \frac{\alpha\omega_0^2 g_1 f_1}{f_1^2(g_1^2\omega_0^2 + \sigma_\eta^2) + \sigma_\varepsilon^2}, \\
\gamma_1 &= \nu_1 = 0, \\
\Theta_1 &= \Delta_1 + (D_1 - E_1 f_1)\Omega_1 D_1^T = \left(1 + \alpha^2\sigma_\eta^2\right) + \frac{\left(\alpha\omega_0^2 g_1\right)^2}{\left(g_1^2\omega_0^2 + \sigma_\eta^2\right)\left(f_1^2(g_1^2\omega_0^2 + \sigma_\eta^2) + \sigma_\varepsilon^2\right)},
\end{aligned}$$

and the density of y_1 is given by

$$f(y_1) = \frac{\phi(y_1; 0, \Gamma_1)\Phi(E_1 y_1; 0, \Theta_1)}{\Phi(0; 0, \Theta_1 + E_1\Gamma_1 E_1^T)} \tag{3.3.7}$$

From the Bayes' rule, we have $f(x_1|y_1) = \dfrac{f(x_1, y_1)}{f(y_1)} = \dfrac{f(y_1|x_1)f(x_1)}{f(y_1)}$, which implies,

$$f(x_1|y_1) = \phi(y_1; f_1 x_1, \sigma_\varepsilon^2)\frac{\phi(x_1; 0, \Omega_1)\Phi(D_1 x_1; 0, \Delta_1)}{\Phi(0; 0, \Delta_1 + D_1\Omega_1 D_1^T)}\frac{\Phi(0; 0, \Theta_1 + E_1\Gamma_1 E_1^T)}{\phi(y_1; 0, \Gamma_1)\Phi(E_1 y_1; 0, \Theta_1)}. \tag{3.3.8}$$

The fact that the product of the two Normal densities $\phi(y_1; f_1 x_1, \sigma_\varepsilon^2)$ and $\phi(x_1; 0, \Omega_1)$ of the dependent variables y_1 and x_1 respectively is not a Normal pdf, contradicts the definition of a csn pdf. Therefore, the density in Equation 3.3.8 is clearly not csn. This is a contradiction with respect to their Proposition 7 and proves our statement regarding the wrong characterization of their filtering density. Another way to see this is by using Proposition 7 on Page 391 of the same paper, and by applying our 1-dimensional setting, we have the following. Suppose that the initial state vector x_0 of the system composed by Equations 3.3.1 and 3.3.2 is such that

$x_0 \sim \text{csn}_{1,1}(\psi_0, \Omega_0, D_0, \nu_0, \Delta_0)$ and that $\varepsilon_t \sim \mathcal{N}(\mu_\varepsilon; \sigma_\varepsilon^2)$ iid and $\eta_t \sim \mathcal{N}(\mu_\eta; \sigma_\eta^2)$ iid.

Then, without loss of generality, we can assume that at $t = 0$, $y_0 = 0$ and therefore $x_0|y_0 = x_0$ holds. From their Proposition 7, we obtain the following: $\widehat{\psi}_0 = 0, \quad \widehat{\Omega}_0 = \omega_0^2, \quad \widehat{D}_0 = \alpha, \quad \widehat{\nu}_0 = 0, \quad \widehat{\Delta}_0 = 1$. We then obtain the parameters of the posterior (filtering) distribution of x_1 defined by

$$(x_1|y_1) \sim \text{csn}_{1,1}(\widehat{\psi}_1, \widehat{\Omega}_1, \widehat{D}_1, \widehat{\nu}_1, \widehat{\Delta}_1), \tag{3.3.9}$$

where

$$\begin{aligned}
\widehat{\psi}_1 &= g_1\widehat{\psi}_0 + \mu_\eta + \widetilde{\Omega}_1 f_1(\sigma_\varepsilon^2 + f_1^2\widetilde{\Omega}_1)^{-1}\left\{y_1 - f_1(g_1\widehat{\psi}_0 + \mu_\eta) - \mu_\varepsilon\right\} = \frac{f_1 y_1(g_1^2\omega_0^2 + \sigma_\eta^2)}{\sigma_\varepsilon^2 + f_1^2(g_1^2\omega_0^2 + \sigma_\eta^2)}, \\
\widetilde{\Omega}_1 &= g_1\widehat{\Omega}_0 g_1^T + \sigma_\eta = g_1^2\omega_0^2 + \sigma_\eta^2, \\
\widehat{\Omega}_1 &= \widetilde{\Omega}_1 - \widetilde{\Omega}_1 f_1^T(\sigma_\varepsilon + f_t\widetilde{\Omega}_1 f_1^T)^{-1} f_1\widetilde{\Omega}_1 = (g_1^2\omega_0^2 + \sigma_\eta^2)\left\{1 - \frac{f_1^2}{\sigma_\varepsilon^2 + f_1^2(g_1^2\omega_0^2 + \sigma_\eta^2)}\right\}, \\
\widehat{D}_1 &= \widehat{D}_0\widehat{\Omega}_0 g_1^T\widetilde{\Omega}_1^{-1} = \frac{\alpha\omega_0^2 g_1}{g_1^2\omega_0^2 + \sigma_\eta^2}, \\
\widehat{\nu}_1 &= \widehat{\nu}_0 = 0, \\
\widehat{\Delta}_1 &= \widehat{\Delta}_0 + (\widehat{D}_0 - \widehat{D}_1 g_1)\widehat{\Omega}_0\widehat{D}_0 = 1 + \frac{\alpha^2\omega_0^2\sigma_\eta^2}{g_1^2\omega_0^2 + \sigma_\eta^2}.
\end{aligned}$$

Comparing the two filtering densities $f(x_1|y_1)$ from Equations 3.3.8 and 3.3.9 will show the problem. Perhaps a different assumption on these densities would have provided results with better performance.

3.4 The closed skew-normal state-space model

From now on, we work under the following state-space model

$$\begin{cases} Y_t = H_t X_t + \varepsilon_t & (3.4.1) \\ X_t = \Phi_t X_{t-1} + \eta_t, & (3.4.2) \end{cases}$$

where $t \in \{1, \ldots, N\}$ represents the time index, $Y_t \in \mathbb{R}^{n_y}$ the vector of observations, $X_t \in \mathbb{R}^{n_x}$ the unobservable state vector, $H_t \in \mathbb{R}^{n_x \times n_y}$ and $\Phi_t \in \mathbb{R}^{n_x \times n_x}$ two known matrices. $\varepsilon_t \sim \text{csn}_{n_y,m_y}(\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon)$ iid and independent of $\eta_t \sim \text{N}_{n_x}(0, \Sigma_\eta)$ iid for all t. Furthermore, let us assume that $X_t \in \mathbb{R}^{n_x}$ is a stationary vector autoregressive process of order 1 (VAR(1)) where this latter assumption is the particularity and novelty of the model we are proposing.

To simplify the computations and without any loss of generality, we consider the univariate situation (lower case letters) of all processes in Equations 3.4.1 and 3.4.2 above. Notice that, with similar arguments and after some simple computations, the multivariate generalization of our results can easily be derived.

We then assume that the error terms ε_t are independent and identically distributed as closed skew-normal (iid.csn) and η_t are independent and identically distributed as Normal (iid.N) distribution with appropriate dimension. Moreover, the η_t are independent of x_t and the series $x_1, x_2, \ldots, x_t$ is stationary and since it is an AR(1) process, we consider the fact that stationarity holds if $|\phi| < 1$.

In the case in which $|\phi| < 1$, let us set $\text{B}x_t = x_{t-1}$ with B the lag operator. We can then write $x_t = \phi \text{B} x_t + \eta_t$ and using a geometric series, the following moving average (MA) expansion holds $x_t = \eta_t + \phi\eta_{t-1} + \phi^2\eta_{t-2} + \phi^3\eta_{t-3} + \cdots$. For this MA($\infty$) process we have,

$$\begin{aligned} \mathbb{E}(x_t) &= \mathbb{E}(\eta_t + \phi\eta_{t-1} + \phi^2\eta_{t-2} + \cdots) \\ &= \mathbb{E}(\eta_t) + \phi\mathbb{E}(\eta_{t-1}) + \phi^2\mathbb{E}(\eta_{t-2}) + \cdots = 0. \end{aligned}$$

By independence of the errors and values of x_t, its variance denoted $\text{Var}(x_t)$ is obtained as

$$\text{Var}(x_t) = \text{Var}(\phi x_{t-1} + \eta_t) = \text{Var}(\phi x_{t-1}) + \text{Var}(\eta_t) = \phi^2 \text{Var}(x_{t-1}) + \sigma_\eta^2.$$

The fact that $\text{Var}(x_t) = \text{Var}(x_{t-1}) = \sigma^2$ by stationarity, leads to $\sigma^2 = \dfrac{\sigma_\eta^2}{1-\phi^2}$. The series x_t is marginally distributed as a Normal with mean 0 and variance σ^2, $\forall t = 0, \pm 1, \pm 2, \ldots$. To conclude with the characterization of x_t, let us look at the autocorrelation function (ACF).

$$\begin{aligned} x_t = \phi x_{t-1} + \eta_t \quad &\Longrightarrow \quad x_{t-h} - x_t = \phi x_{t-h} x_{t-1} + x_{t-h}\eta_t, \\ &\Longrightarrow \quad \mathbb{E}(x_{t-h}x_{t-1}) = \mathbb{E}(\phi x_{t-h}x_{t-1}) + \mathbb{E}(x_{t-h}\eta_t), \\ &\Longrightarrow \quad \gamma_k = \phi\gamma_{k-1}. \end{aligned}$$

Indicating with γ_k the kth order covariance of x_t, we can write

$$\gamma_0 = \mathsf{Var}(x_t) = \frac{\sigma_\eta^2}{1-\phi^2},$$

we obtain

$$\begin{aligned} \gamma_1 &= \phi\gamma_0, \\ \gamma_2 &= \phi\gamma_1 = \phi(\phi\gamma_0) = \phi^2\gamma_0, \\ \gamma_3 &= \phi\gamma_2 = \phi(\phi^2\gamma_0) = \phi^3\gamma_0, \\ &\vdots \\ \gamma_h &= \phi^h\gamma_0 = \phi^h\frac{\sigma_\eta^2}{1-\phi^2}. \end{aligned}$$

Hence, for $I = \{\cdots -2, -1, 0, 1, 2, \cdots\}$, the covariance and the correlation of $\{x_t\}_{t\in I}$ at two different lags are respectively given by

$$\mathsf{Cov}(x_t, x_{t-h}) = \frac{\phi^h\sigma_\eta^2}{1-\phi^2}, \quad \text{and} \quad \mathsf{Corr}(x_t, x_{t-h}) = \rho_h = \frac{\mathsf{Cov}(x_t, x_{t-h})}{\mathsf{Var}(x_t)}, \quad \text{for} \quad h = 1, 2, \ldots.$$

Since from stationarity, the product of the standard deviation (SD) of x_t at two different lags will give the variance, we can write,

$$\mathsf{SD}(x_t)\mathsf{SD}(x_{t-h}) = \mathsf{SD}^2(x_t) = \mathsf{Var}(x_t),$$

which implies,

$$\rho_h = \mathsf{Corr}(x_t, x_{t-h}) = \phi^h\frac{\mathsf{Var}(x_t)}{\mathsf{Var}(x_t)} = \phi^h.$$

Now, for all t let us recall the measurement Equation 3.4.1. We are interested in defining the distribution of $\{y_t\}_{t=0,1,2,\dots}$. Before going to the computations, a detour into their proper characterisation is required. We know that it is always possible to derive the Normal distribution from the csn by letting the parameter governing the skewness or the skewness dimension to be equal to zero. From Lemma 3.2.1 and since $\varepsilon_t \sim \text{csn}_{1,1}(\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon)$, y_t which is a sum of a csn and a Gaussian processes will be distributed as a csn with appropriate dimension and parameters. Having made this clear, we can now compute some simple examples (one dimension) in order to easily generalize the result in the multivariate setting. For $n_x = m_y = n_y = 1$ and $t = 0$, we have

$$y_0 = (x_0 + \varepsilon_0) \sim \text{csn}_{1,1}(\mu_{y_0}, \Sigma_{y_0}, \Gamma_{y_0}, \nu_{y_0}, \Delta_{y_0}),$$

where $\mu_{y_0} = \mu_\varepsilon$, $\Sigma_{y_0} = \Sigma_\varepsilon + \sigma^2$, $\Gamma_{y_0} = \Gamma_\varepsilon \Sigma_\varepsilon (\Sigma_\varepsilon + \sigma^2)^{-1}$, $\nu_{y_0} = \nu_\varepsilon$ and $\Delta_{y_0} = \Delta_\varepsilon + \dfrac{\Gamma_\varepsilon^2 \Sigma_\varepsilon \sigma^2}{\Sigma_\varepsilon + \sigma^2}$

We know that, x_t has a stationary distribution $(x_t \sim N(0, \sigma^2))$ and $\varepsilon_t \sim \text{csn}_{1,1}(\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon)$ iid. Then for all t, y_t has a stationary distribution as well, and this distribution is $\text{csn}_{1,1}(\mu_{y_0}, \Sigma_{y_0}, \Gamma_{y_0}, \nu_{y_0}, \Delta_{y_0})$. With these information, we can now state the following proposition.

Proposition 3.4.1. *Under the model composed by both Equations 3.4.1 and 3.4.2, if the noises ε_t, respectively η_t are iid csn$_{1,1}(\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon,)$ random variables, respectively a Gaussian random variable with mean 0, and variance σ_η^2 and as assumed before, the process $x_t \sim N(0, \sigma^2)$ is stationary AR(1). Then for $t = 0, 1, 2, \dots,$ the observations y_t are stationary and distributed as $y_t \sim$ csn$_{1,1}(\mu_{y_t}, \Sigma_{y_t}, D_{y_t}, \nu_{y_t}, \Delta_{y_t})$, with parameters satisfying the following relationships $\mu_{y_t} = \mu_\varepsilon$, $\Sigma_{y_t} = \Sigma_\varepsilon + \sigma^2$, $\Gamma_{y_t} = \Gamma_\varepsilon \Sigma_\varepsilon (\Sigma_\varepsilon + \sigma^2)^{-1}$, $\nu_{y_t} = \nu_\varepsilon$ and $\Delta_{y_t} = \Delta_\varepsilon + \dfrac{\Gamma_\varepsilon^2 \Sigma_\varepsilon \sigma^2}{\Sigma_\varepsilon + \sigma^2}$.*

We will now look at the finite-dimensional marginal distribution of $\{y_t\}_{t=0,1,2,\dots,n}$. By considering

a given n with values in $\{1, 2, 3, \dots, \}$ Equation 3.4.1 can be rewritten as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Let us start with $n = 2$ and then generalize the result.

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \\ &\overset{\text{d}}{=} \left[N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \phi\sigma^2 \\ \phi\sigma^2 & \sigma^2 \end{pmatrix} \right] \right] + \left[\text{csn}_{2,2} \left[\begin{pmatrix} \mu_\varepsilon \\ \mu_\varepsilon \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\varepsilon \end{pmatrix}, \begin{pmatrix} \Gamma_\varepsilon & 0 \\ 0 & \Gamma_\varepsilon \end{pmatrix}, \begin{pmatrix} \nu_\varepsilon \\ \nu_\varepsilon \end{pmatrix}, \begin{pmatrix} \Delta_\varepsilon & 0 \\ 0 & \Delta_\varepsilon \end{pmatrix} \right] \right] \end{aligned}$$

Using Lemma 3.2.1 we can then write $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \text{csn}_{2,2}(\mu_{y_{12}}, \Sigma_{y_{12}}, \Gamma_{y_{12}}, \nu_{y_{12}}, \Delta_{y_{12}})$ where,

$$\mu_{y_{12}} = \begin{pmatrix} \mu_\varepsilon \\ \mu_\varepsilon \end{pmatrix}, \quad \Sigma_{y_{12}} = \begin{bmatrix} \sigma^2 + \Sigma_\varepsilon & \phi\sigma^2 \\ \phi\sigma^2 & \sigma^2 + \Sigma_\varepsilon \end{bmatrix},$$

$$\Gamma_{y_{12}} = \frac{\Gamma_\varepsilon \Sigma_\varepsilon}{(\sigma^2 + \Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} \begin{bmatrix} \sigma^2 + \Sigma_\varepsilon & -\phi\sigma^2 \\ -\phi\sigma^2 & \sigma^2 + \Sigma_\varepsilon \end{bmatrix}, \quad \nu_{y_{12}} = \begin{bmatrix} \nu_\varepsilon \\ \nu_\varepsilon \end{bmatrix},$$

$$\Delta_{y_{12}} = \begin{bmatrix} \Delta_\varepsilon + \Gamma_\varepsilon - \dfrac{(\Gamma_\varepsilon \Sigma_\varepsilon)^2 (\sigma^2 + \Sigma_\varepsilon)}{(\sigma^2 + \Sigma_\varepsilon)^2 - (\sigma^2\phi)^2} & \dfrac{(\Gamma_\varepsilon \Sigma_\varepsilon)^2 \phi\sigma^2}{(\sigma^2 + \Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} \\ \dfrac{(\Gamma_\varepsilon \Sigma_\varepsilon)^2 \phi\sigma^2}{(\sigma^2 + \Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} & \Delta_\varepsilon + \Gamma_\varepsilon - \dfrac{(\Gamma_\varepsilon \Sigma_\varepsilon)^2 (\sigma^2 + \Sigma_\varepsilon)}{(\sigma^2 + \Sigma_\varepsilon)^2 - (\sigma^2\phi)^2} \end{bmatrix}.$$

So, for any set of two indexes i and j, the resulting vector $\begin{bmatrix} y_i \\ y_j \end{bmatrix} \sim \text{csn}_{2,2}(\mu_{y_{ij}}, \Sigma_{y_{ij}}, D_{y_{ij}}, \nu_{y_{ij}}, \Delta_{y_{ij}})$ with parameters defined as above.

From this, we can then construct and define the joint distribution for any set of indexes $\{1, 2, 3, \cdots, n\}$. The following proposition characterizes a vector $X_t = x_1, x_2, \ldots, x_n$ as a csn distribution with appropriate parameters.

Proposition 3.4.2. *Let us consider a given n with values in $\{1, 2, 3, \cdots\}$, a given set of indexes $t = \{1, 2, \ldots, n\}$, the noise η_t a Gaussian random variable with mean 0 and variance σ_η^2. As demonstrated before, all x_t defined by Equation 3.4.2 follow the stationary distribution $x_t \sim N(0, \sigma^2)$ and the distribution of X_t which is the joint distribution of all x_t can be written as*

$$X_t = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_t \end{bmatrix} \sim csn_{t,t} \left[\mathbf{0}_{t\times 1}, \sigma^2 \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{t-1} \\ \phi & 1 & \phi & \cdots & \phi^{t-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{t-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{t-1} & \phi^{t-2} & \phi^{t-3} & \cdots & 1 \end{pmatrix}, \mathbf{0}_{t\times t}, \mathbf{0}_{t\times 1}, \mathbf{1}_{t\times t} \right],$$

where $\sigma^2 = \dfrac{\sigma_\eta^2}{1-\phi^2}$, $\mathbf{0}_{i\times j}$ and $\mathbf{1}_{i\times j}$ are $i \times j$ matrices of 0 and 1, respectively.

This proposition can easily be generalised for the multivariate situation.

We now want to evaluate the distribution of the state vector conditional on the set of all available information (observations), at each time index. That is, $f(x_t|Y^t)$, where $Y^t = (y_1, y_2, \ldots, y_t)$ represents the set of all available information up to time t. These data are assumed to be conditionally independent, given the state variable at the indicated time step. The filtering problem will then consist on characterizing, the distribution of the states given all the available information. Moreover, the likelihood for the data can be defined as $f(Y^t|X_t)$ and from Equation 3.4.1, we can see that there is a linear relationship between the observations and the state variables.

As stated before, the csn distribution is an extension or generalization of the Gaussian distribution and as such, it has numerous properties similar to the Gaussian one. Another of these properties which will be the key of what follows is that, *the csn distribution is conjugate*. Meaning that, if the prior probability density function $f(X_t)$ and the likelihood $f(Y^t|X_t)$ are both distributed as a csn and if in addition, there is a linear relation in the likelihood, then the posterior pdf $f(X_t|Y^t)$

will also be distributed as a csn with appropriate dimension and parameters. The reader can refer to Karimi et al. (2010) and Rezaie et al. (2014) for more information about this. For convenience and later use, we will now present the following Proposition 3.4.3 which describes the filtering density and is simply a particular case of the above mentioned proposition.

Proposition 3.4.3. *If the prior pdf for the state variables is* $X_t \sim N_{n_x}(\mu_x, \Sigma_x)$ *and the likelihood is* $Y^t|X_t \sim csn_{n_y,m_y}(HX_t + \mu_{y|x}, \Sigma_{y|x}, \Gamma_{y|x}, \upsilon_{y|x}, \Delta_{y|x})$*, then the posterior of the state variables is* $X_t|Y_t \sim csn_{n_x,m_{x|y}}(\mu_{x|y}, \Sigma_{x|y}, \Gamma_{x|y}, \upsilon_{x|y}, \Delta_{x|y})$*. The skewness dimension parameter of the posterior pdf is* $m_{x|y} = m_x + m_y$ *and for the others we have:*

$$
\begin{aligned}
\mu_{x|y} &= \mu_x + \Sigma_x H^{'}[H\Sigma_x H^{'} + \Sigma_{y|x}]^{-1}(Y^t - H\mu_x - \mu_{y|x}),\\
\Sigma_{x|y} &= \Sigma_x - \Sigma_x H^{'}[H\Sigma_x H^{'} + \Sigma_{y|x}]^{-1}H\Sigma_x,\\
\Gamma_{x|y} &= \left[-\left[\Gamma_{y|x}\Sigma_{y|x}\right]\left[H\Sigma_x H^{'} + \Sigma_{y|x}\right]^{-1}H\Sigma_x\right]\Sigma_{x|y}^{-1},\\
\nu_{x|y} &= \left[-\nu_{y|x}\right] + \left[\Gamma_{y|x}\Sigma_{y|x}\right]\left[H\Sigma_x H^{'} + \Sigma_{y|x}\right]^{-1}(Y^t - H\mu_x - \mu_{y|x}),\\
\Delta_{x|y} &= \left[\Delta_{y|x} + \Gamma_{y|x}\Sigma_{y|x}\Gamma_{y|x}^{'}\right] - \left[\Gamma_{y|x}\Sigma_{y|x}\right]\left[H\Sigma_x H^{'} + \Sigma_{y|x}\right]^{-1}\left[\Gamma_{y|x}\Sigma_{y|x}\right]^{'} - \Gamma_{x|y}\Sigma_{x|y}\Gamma_{x|y}^{'}.
\end{aligned}
$$

When looking at $(Y^t - H\mu_x - \mu_{y|x})$, which can be considered as the innovation, we obtain as in Rezaie et al. (2014) a Gaussian and a Skewed Kalman gains defined respectively by

$$K_{Gaussian} = \Sigma_x H^{'}[H\Sigma_x H^{'} + \Sigma_{y|x}]^{-1} \quad \text{and} \tag{3.4.3}$$

$$K_{Skewed} = \left[\Gamma_{y|x}\Sigma_{y|x}\right]\left[H\Sigma_x H^{'} + \Sigma_{y|x}\right]^{-1} \tag{3.4.4}$$

The fact that $m_x = 0$ in the present situation implies $(m_{x|y} = m_y)$ and it is worth noticing that the skewness dimension of the posterior distribution is considerably reduced compared to previous approaches in the literature. This will be very valuable in handling the updating scheme and the computational cost of the parameters' estimation.

For simplicity purpose, let us assume that $\varepsilon_t \sim \text{csn}_{1,1}(0, \Sigma_\varepsilon, \Gamma_\varepsilon, 0, 1)$. Having these ingredients and by applying Proposition 3.4.3 we have the following.

For $t = 1$ our state-space model can be written as follows

$$\begin{cases} y_1 = x_1 + \varepsilon_1 \\ x_1 = \phi x_0 + \eta_1 \end{cases}$$

We know from Proposition 3.4.1 that y_1 will be distributed as a csn and so does $(y_1|x_1) \sim \text{csn}_{1,1}(x_1, \Sigma_\varepsilon, \Gamma_\varepsilon, 0, 1)$ and the posterior model $f(x_1 \mid y_1)$ with respect to the Bayesian closed skew-normal inversion (BCSNI) Karimi et al. (2010) is given by:

$$f(x_1|y_1) \sim \text{csn}_{1,1}(\mu_{x_1|y_1}, \Sigma_{x_1|y_1}, \Gamma_{x_1|y_1}, \nu_{x_1|y_1}, \Delta_{x_1|y_1}) \tag{3.4.5}$$

Where,

$$\begin{aligned} \mu_{x_1|y_1} &= \frac{\sigma^2}{\sigma^2 + \Sigma_\varepsilon} y_1 \\ \Sigma_{x_1|y_1} &= \frac{\sigma^2}{\sigma^2 + \Sigma_\varepsilon} \Sigma_\varepsilon \\ \Gamma_{x_1|y_1} &= -\Gamma_\varepsilon \\ \nu_{x_1|y_1} &= \frac{\Gamma_\varepsilon \Sigma_\varepsilon}{\sigma^2 + \Sigma_\varepsilon} y_1 \\ \Delta_{x_1|y_1} &= 1 + \frac{\sigma^2 \Gamma_\varepsilon \Sigma_\varepsilon (1 + \Gamma_\varepsilon)}{\sigma^2 + \Sigma_\varepsilon} \end{aligned}$$

Now let us look at the conditional marginal distribution $f(x_2 \mid y_1, y_2)$.

$$\begin{cases} y_1 = x_1 + \varepsilon_1 \\ y_2 = x_2 + \varepsilon_2 \end{cases}$$

Which can be rewritten as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{3.4.6}$$

We know that,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathsf{csn}_{2,0}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \phi\sigma^2 \\ \phi\sigma^2 & \sigma^2 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

and

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathsf{csn}_{2,2}\left(\begin{pmatrix} \mu_\varepsilon \\ \mu_\varepsilon \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\varepsilon \end{pmatrix}, \begin{pmatrix} D_\varepsilon & 0 \\ 0 & D_\varepsilon \end{pmatrix}, \begin{pmatrix} \nu_\varepsilon \\ \nu_\varepsilon \end{pmatrix}, \begin{pmatrix} \Delta_\varepsilon & 0 \\ 0 & \Delta_\varepsilon \end{pmatrix}\right),$$

thus,

$$\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) \sim \mathsf{csn}_{2,2}\left(\mu_{x_{12}|y_{12}}, \Sigma_{x_{12}|y_{12}}, \Gamma_{x_{12}|y_{12}}, \nu_{x_{12}|y_{12}}, \Delta_{x_{12}|y_{12}}\right),$$

with

$$\begin{aligned} \mu_{x_{12}|y_{12}} &= \mu_{x_{12}} + \Sigma_{x_{12}}\left[\Sigma_{x_{12}} + \Sigma_{\varepsilon_{12}}\right]^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \frac{1}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2}\begin{pmatrix} \sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4 & \phi\sigma^2\Sigma_\varepsilon \\ \phi\sigma^2\Sigma_\varepsilon & \sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &= \frac{1}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2}\begin{pmatrix} y_1(\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4) + y_2(\phi\sigma^2\Sigma_\varepsilon) \\ y_1(\phi\sigma^2\Sigma_\varepsilon) + y_2(\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4) \end{pmatrix}. \end{aligned} \tag{3.4.7}$$

Finally, we have that

$$\begin{aligned} \mu_{x_2|y_{12}} &= (0,1)\mu_{x_{12}|y_{12}} \\ &= \frac{1}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2}\left[y_1(\phi\sigma^2\Sigma_\varepsilon) + y_2(\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4)\right] \end{aligned} \tag{3.4.8}$$

By setting

$$\begin{aligned} A &= \frac{(\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4)(\sigma^2) + (\phi\sigma^2\Sigma_\varepsilon)(\phi\sigma^2)}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} \\ B &= \frac{(\phi\sigma^2\Sigma_\varepsilon)(\sigma^2) + (\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4)(\phi\sigma^2)}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} \\ C &= \frac{-(D_\varepsilon\Sigma_\varepsilon)(\sigma^4 + \sigma^2\Sigma_\varepsilon - \phi^2\sigma^4)}{(\sigma^2+\Sigma_\varepsilon)^2 - (\phi\sigma^2)^2} \end{aligned}$$

$$\text{and} \quad D = \frac{-(D_\varepsilon \Sigma_\varepsilon)(\phi\sigma^2\Sigma_\varepsilon)}{(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2},$$

we have

$$\begin{aligned}\Sigma_{x_{12}|y_{12}} &= \Sigma_{x_{12}} - \Sigma_{x_{12}}\left[\Sigma_{x_{12}}+\Sigma_{\varepsilon_{12}}\right]^{-1}\Sigma_{x_{12}} \\ &= \begin{pmatrix} \sigma^2-A & \phi\sigma^2-B \\ \phi\sigma^2-B & \sigma^2-A \end{pmatrix},\end{aligned}$$

which implies

$$\begin{aligned}\Sigma_{x_2|y_{12}} &= (0,1)\Sigma_{x_{12}|y_{12}}\begin{pmatrix}0\\1\end{pmatrix} \\ &= \sigma^2-A \\ &= \sigma^2 - \frac{(\sigma^4+\sigma^2\Sigma_\varepsilon-\phi^2\sigma^4)(\sigma^2)+(\phi\sigma^2\Sigma_\varepsilon)(\phi\sigma^2)}{(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2} \\ &= \frac{\sigma^2}{(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2}\left[\sigma^2\Sigma_\varepsilon(1-\phi^2)+\Sigma_\varepsilon^2\right] \qquad (3.4.9)\end{aligned}$$

$$\begin{aligned}\Gamma_{x_{12}|y_{12}} &= \left[-\Gamma_{\varepsilon_{12}}\Sigma_{\varepsilon_{12}}\left[\Sigma_{x_{12}}+\Sigma_{\varepsilon_{12}}\right]^{-1}\Sigma_{x_{12}}\right]\Sigma_{x_{12}|y_{12}}^{-1} \\ &= \frac{\Gamma_\varepsilon\Sigma_\varepsilon}{[(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2][(\sigma^2-A)^2-(\phi\sigma^2-B)^2]} \\ &\times\begin{pmatrix}(\sigma^2+\Sigma_\varepsilon)(\sigma^2-A)-\phi\sigma^2(B-\phi\sigma^2) & (\sigma^2+\Sigma_\varepsilon)(B-\phi\sigma^2)-\phi\sigma^2(\sigma^2-A)\\ (\sigma^2+\Sigma_\varepsilon)(B-\phi\sigma^2)-\phi\sigma^2(\sigma^2-A) & (\sigma^2+\Sigma_\varepsilon)(\sigma^2-A)-\phi\sigma^2(B-\phi\sigma^2)\end{pmatrix} \qquad (3.4.10)\end{aligned}$$

Yields,

$$\Gamma_{x_2|y_{12}} = \Gamma_{x_{12}|y_{12}}\Sigma_{x_{12}|y_{12}}\begin{pmatrix}0\\1\end{pmatrix}\left[(0,1)\Sigma_{x_{12}|y_{12}}\begin{pmatrix}0\\1\end{pmatrix}\right]^{-1} \qquad (3.4.11)$$

$$\begin{aligned}
\nu_{x_{12}|y_{12}} &= \nu_{x_2|y_{12}} = \Gamma_{\varepsilon_{12}}\Sigma_{\varepsilon_{12}}\left[\Sigma_{x_{12}}+\Sigma_{\varepsilon_{12}}\right]^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \begin{pmatrix} \Gamma_\varepsilon\Sigma_\varepsilon & 0 \\ 0 & \Gamma_\varepsilon\Sigma_\varepsilon \end{pmatrix}\frac{1}{(\sigma^2+\Sigma)^2-(\phi\sigma^2)^2}\begin{pmatrix} \sigma^2+\Sigma_\varepsilon & -\phi\sigma^2 \\ -\phi\sigma^2 & \sigma^2+\Sigma_\varepsilon \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
&= \frac{\Gamma_\varepsilon\Sigma_\varepsilon}{(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2}\begin{pmatrix} (\sigma^2+\Sigma_\varepsilon)y_1-\phi\sigma^2 y_2 \\ -\phi\sigma^2 y_1+(\sigma^2+\Sigma_\varepsilon)y_2 \end{pmatrix} \qquad (3.4.12)
\end{aligned}$$

Finally,

$$\begin{aligned}
\Delta_{x_{12}|y_{12}} &= \mathbf{I}_2+\Gamma_{\varepsilon_{12}}\Sigma_{\varepsilon_{12}}\Gamma_{\varepsilon_{12}}^{'}-\Gamma_{\varepsilon_{12}}\Sigma_{\varepsilon_{12}}[\Sigma_{x_{12}}+\Sigma_{\varepsilon_{12}}]^{-1}[\Gamma_{\varepsilon_{12}}\Sigma_{\varepsilon_{12}}]^{'} \\
&= \frac{1}{(\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2}\begin{pmatrix} E & F \\ F & E \end{pmatrix} \qquad (3.4.13)
\end{aligned}$$

where

$$E = ((\sigma^2+\Sigma_\varepsilon)^2-(\phi\sigma^2)^2)(1+\Gamma_\varepsilon^2\Sigma_\varepsilon)-(\Gamma_\varepsilon\Sigma_\varepsilon)^2(\sigma^2+\Sigma_\varepsilon) \quad \text{and} \quad F=(\Gamma_\varepsilon\Sigma_\varepsilon)^2\phi\sigma^2$$

and we then obtain

$$\begin{aligned}
\Delta_{x_2|y_{12}} =& \Delta_{x_{12}|y_{12}}+\Gamma_{x_{12}|y_{12}}\Sigma_{x_{12}|y_{12}}\Gamma_{x_{12}|y_{12}}^{'} \\
&-\Gamma_{x_{12}|y_{12}}\Sigma_{x_{12}|y_{12}}\begin{pmatrix} 0 \\ 1 \end{pmatrix}\left[(0,1)\Sigma_{x_{12}|y_{12}}\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right]^{-1}(0,1)\Sigma_{x_{12}|y_{12}}\Gamma_{x_{12}|y_{12}}^{'} \qquad (3.4.14)
\end{aligned}$$

We then conclude that,

$$f(x_2|y_1,y_2)\sim \mathsf{csn}_{1,2}(\mu_{x_2|y_{12}},\Sigma_{x_2|y_{12}},\Gamma_{x_2|y_{12}},\nu_{x_2|y_{12}},\Delta_{x_2|y_{12}}) \qquad (3.4.15)$$

and as the computations can show, except from μ and ν, where we there exists an updating procedure, the other parameters do not depend on data.

3.5 Model estimation: The maximum likelihood approach

In this section, the model estimation will be conducted based on data generated from a benchmark model with true parameters ($\sigma_\varepsilon = 2, \gamma_\varepsilon = 8, \sigma_\eta = \sqrt{2}$ and $\phi = 0.7$). Before proceeding, let us recall the ssm defined by the observation equation $y_t = x_t + \varepsilon_t$ and the state equation $x_t = \phi x_{t-1} + \eta_t$, where $\varepsilon_t \sim \mathsf{csn}_{1,1}(\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon)$ and $x_t \sim \mathsf{N}(0, \sigma^2)$. Moreover, if we let the parameters linked to the state process x_t and $(y_t|x_t)$ be $\Lambda := \{\sigma_\eta, \phi\}$ and $\Theta := \{\mu_\varepsilon, \Sigma_\varepsilon, \Gamma_\varepsilon, \nu_\varepsilon, \Delta_\varepsilon\}$ respectively, then we can define the parameters vector as $\psi = \{\Theta, \Lambda\}$.

Now, let $\mathbf{0}$ represents the null vector, $\mathbf{I}_n$ the $n \times n$ identity matrix, Σ_x the variance-covariance matrix of the state process, Σ_ε an $n \times n$ diagonal matrix with entries σ_ε and Γ_ε an $n \times n$ diagonal matrix with entries γ_ε. The marginal likelihood is then distributed as

$$\mathsf{Y}^n \overset{\mathsf{d}}{=} [\mathsf{N}_n(\mathbf{0}, \Sigma_x)] + [\mathsf{csn}_{n,n}(\mathbf{0}, \Sigma_\varepsilon, \Gamma_\varepsilon, \mathbf{0}, \mathbf{I}_n)] \tag{3.5.1}$$

$$\sim [\mathsf{csn}_{n,n}(\mu_Y, \Sigma_Y, \Gamma_Y, \nu_Y, \Delta_Y)]\,, \tag{3.5.2}$$

where

$$\mu_Y = \mathbf{0}, \quad \nu_Y = \mathbf{0}, \quad \Sigma_Y = \Sigma_x + \Sigma_\varepsilon, \quad \Gamma_Y = \Gamma_\varepsilon \Sigma_\varepsilon (\Sigma_x + \Sigma_\varepsilon)^{-1}, \quad \Delta_Y = \mathbf{I}_n + (\Gamma_\varepsilon - \Gamma_Y)\Sigma_\varepsilon \Gamma_\varepsilon'$$

and $\overset{\mathsf{d}}{=}$ stands for, distributed as. The marginal likelihood is then written in closed form as

$$\begin{aligned}
\mathcal{L}(\Psi; \mathsf{Y}^n) &= \frac{1}{\Phi_n(\mathbf{0}; \mathbf{0}, \Delta_Y + \Gamma_Y \Sigma_Y \Gamma_Y')} \times \phi_n(\mathsf{Y}^n; \mathbf{0}, \Sigma_Y) \times \Phi_n(\Gamma_Y \mathsf{Y}^n; \mathbf{0}, \Delta_Y) \\
&= \frac{1}{(2\pi)^{n/2} |\Sigma_Y|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathsf{Y}^n)' \Sigma_Y^{-1} (\mathsf{Y}^n) \right\} \times (2\pi)^{n/2} |\Delta_Y + \Gamma_Y \Sigma_Y \Gamma_Y')|^{1/2} \\
&\times \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} \frac{1}{(2\pi)^{n/2} |\Delta_Y|^{1/2}} \exp\left\{ -\frac{1}{2} (\Gamma_Y Z)' \Delta_Y^{-1} (\Gamma_Y Z) \right\} \mathsf{d}z_1 \cdots \mathsf{d}z_n,
\end{aligned} \tag{3.5.3}$$

where |A| is simply the determinant of the square matrix A. Taking out some constant that do not depend on parameters, we finally have that

$$-2\log\mathcal{L}(\Psi;\mathsf{Y}^n) = -\log|\Delta_Y+\Gamma_Y\Sigma_Y\Gamma_Y'| + \log|\Sigma_Y| + (\mathsf{Y}^n)'\Sigma_Y^{-1}(\mathsf{Y}^n) + \log|\Delta_Y| - 2\log\int_{-\infty}^{y_1}\cdots\int_{-\infty}^{y_n}\exp\left\{-\frac{1}{2}(\Gamma_Y Z)'\Delta_Y^{-1}(\Gamma_Y Z)\right\}\mathrm{d}z_1\cdots\mathrm{d}z_n \tag{3.5.4}$$

In many statistics problems, such as the one under investigation in this chapter, one usually encounters issue is that of numerically evaluating the n-dimensional Gaussian distribution function. To tackle this problem, several solutions have been proposed in the literature for various (relatively small) values of n, see for instance the contributions of Donnelly (1973) and Cox and Wermuth (1991) among others. As our application require big values of n, we follow the approach proposed by Genz (1992) where three successive transformations are used to transform the initial multidimensional integral into an integral over a unit hypercube with constant limits. Detailed explanations on these transformations and the algorithm can be found in the above mentioned reference where as for the purpose of this study, we implemented the same algorithm via the R software.

A first attempt on estimating the above likelihood led to several computational issues such as the singularity of covariance matrices when the sample size goes above 100, and the high time consumption required to generate data from a multivariate csn distribution. Additionally, even with relaxed assumptions on parameters such as the ones leading to the skew-normal density of Azzalini and Capitanio (1999), we end up with the same conclusion as the authors that called for an alternative estimation method rather than the maximum likelihood estimation procedure which leads to several statistical issues. Moreover, given that for large sample, the evaluation of multivariate Gaussian cdf is computationally costly, a simple reformulation of the likelihood function is most desirable.

3.5.1 Reformulation of the likelihood function

In the following, we will reformulate the likelihood function to allow for simple and fast simulation from the csn distribution which will also contribute to make feasible and reliable the maximum

likelihood estimation technique by overcoming the above mentioned drawbacks.

To proceed, we invite the reader to make a brief detour to linear algebra and especially to the following two properties of symmetric positive semi-definite matrices (PSD). It is well known that any real symmetric matrix $A \in \mathbf{R}^{n\times n}$ can always be rewritten as $A = U\Lambda U'$, with U a full rank orthogonal matrix with columns the eigenvectors of A, and Λ a diagonal matrix with entries the eigenvalues of the matrix A. Second, the eigenvalues of any PSD matrix A are all non negative. By using the first property also known as the eigendecomposition, we have that the square root of any PSD matrix A can always be computed as

$$A = U\Lambda U' = (U\Lambda^{1/2}U')(U\Lambda^{1/2}U') = SS,$$

where $S = U\Lambda^{1/2}U'$ is the square root of A. In fact if A is PSD, the existence and the uniqueness of the PSD matrix denoted $A^{1/2}$ such that $(A^{1/2})^2 = A$ can be found in appropriate graduate text book of linear algebra.

Now, let $y_1, \ldots, y_n$ be a set normally distributed random variables with expected values $\mu_1, \ldots, \mu_n$ and covariance matrix $\Sigma \in \mathbb{S}^n_{++}$, where $\mathbb{S}^n_{++}$ is the space of symmetric positive-definite matrices of dimension $n \times n$ defined by

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & \cdots & \sigma^2_{1n} \\ \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \cdots & \sigma^2_{nn} \end{pmatrix}, \quad \sigma^2_{ij} = \mathsf{Cov}(y_i, y_j) \quad \text{and} \quad \sigma^2_{ii} = \mathsf{Var}(y_i).$$

Then, the pdf of the random vector denoted $Y = [y_1, \ldots, y_n]'$ is defined as

$$f(Y) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)\right\} \tag{3.5.5}$$

This means that, if we have a pdf of the form described in Equation (3.5.5), then by a suitable

change of variables, one sees that the new random vector $Z = \Sigma^{-1/2}(Y - \mu)$ has pdf

$$f(Z) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left\{-\frac{1}{2}(Z)'(Z)\right\} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}{z_1}^2\right\} \times \ldots \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}{z_n}^2\right\} \quad (3.5.6)$$

and the components of the random vector Z are independent and normally distributed with mean 0 and variance 1. It is important to see that, the random vector Y can be recovered by simply noticing that $Y = \mu + \Sigma^{1/2}Z$ which is more convenient if one wants to generate sample from Y. Finally, by taking out some constants that do not depend on parameters and using the above properties and transformations, the log-likelihood in Equation (3.5.4) becomes without loss of generality

$$\mathcal{L}(\Psi; \mathsf{Y}^n) = -\frac{1}{2}\sum_{t=1}^{n}\left(\Sigma_Y^{-1/2}Y^n\right)[t] + \sum_{t=1}^{n}\log\left(\int_{-\infty}^{(\Delta_Y^{-1/2}\Gamma_Y Y^n)[t]} \exp\left\{-\frac{1}{2}z^2\right\} \mathrm{dz}\right), \quad (3.5.7)$$

or equivalently,

$$-2\mathcal{L}(\Psi; \mathsf{Y}^n) = \sum_{t=1}^{n}\left(\Sigma_Y^{-1/2}Y^n\right)[t] - 2\sum_{t=1}^{n}\log\Phi_1\left((\Delta_Y^{-1/2}\Gamma_Y Y^n)[t]; 0, 1\right), \quad (3.5.8)$$

where $V[t]$ represents the t^{th} component of the vector V.

Definition (3.5.1) below is the result of applying the above transformation to the definition of the csn density.

Definition 3.5.1 (Alternative representation of the multivariate close skew-normal pdf). *If a random vector X has a multivariate csn distribution according to González-Farías et al. (2004a), then under the same assumption as in Definition (3.2.1), its pdf defined by Equation (3.2.1) can equivalently be represented as*

$$f(x) = \frac{1}{\Phi_{m_x}(-(\Delta + \Gamma\Sigma\Gamma^T)^{-1/2}; \mathbf{0}, \mathbf{I}_{m_x})}\phi_{n_x}(\Sigma^{-1/2}(X - \mu); \mathbf{0}, \mathbf{I}_{n_x}) \quad (3.5.9)$$

$$\times \Phi_{m_x}(\Delta^{-1/2}(\Gamma(x-\mu)-\nu); \mathbf{0}, \mathbf{I}_{m_x}). \tag{3.5.10}$$

It is important to notice that we do not intend to substitute the original definition but instead we are offering an alternative representation for fast and more friendly computation, as the latest has a well-behaving structure allowing for factorization of multivariate Normal pdf and cdf as product of standard univariate Normal pdf and cdf respectively.

Corollary 3.5.1. *With parameters' transformation as in definition* (3.5.1) *above, the elements of the vector X will then be independent as a consequence of the diagonal structure of the covariance matrices.*

This corollary can be very useful when it comes to the parameters estimation of the csn density in general. The independence structure of the random vector it offers constitutes the key ingredient for an alternative method to estimate the parameters of the csn likelihood by using the weighted method of moment of Flecher et al. (2009), where the authors demonstrated that their method seems to outperform the mle for small sample sizes in the univariate case. It would have been interesting to compare their results with ours in the csn-ssm framework, but we leave it for future studies.

3.5.2 Estimation results

Based on simulation studies conducted with the R software on a laptop operating under Ubuntu 14.04.5 LTS, with the following properties (Processor: 8x Intel(R) Core(TM) i7-2670QM, CPU @ 2.20GHz, Memory: 4G), Tables (3.1) and (3.2) below were obtained and summarize our findings. For both tables, we used the alternative representation in Equation (3.5.9).

Parameters	σ_ε	γ_ε	σ_η	ϕ
True values	2	8	$\sqrt{2}$	0.7
Estimates	1.6	7.6	1.226555	0.5

Table 3.1: Parameter estimates of a csn likelihood via the optimx package and the "L-BFGS-B" method and sample size $N = 400$.

Parameters	σ_ε	γ_ε	σ_η	ϕ
True values	2	8	$\sqrt{2}$	0.7
Estimates	1.7	7.8	1.584982	0.6

Table 3.2: This table summarizes the parameter estimates of a csn likelihood via the optimx package and the "L-BFGS-B" method and sample size $N = 1000$.

One can see that, as the sample size becomes larger, the estimates converge to their true values. It is worth noticing that, we first ran a grid-search algorithm around the parameters and then supplied the obtained estimates as initial values for the mle routine. We did this because the plot of the likelihood function presents local maxima and depending on the starting values, relatively different estimates were obtained. To avoid the search to stuck into these local maxima, we suggest to first search the parameters regions via the grid-search method and then use the best set of obtained estimates as starting value. In Figure (3.2) below, the curve in red, blue, green and orange correspond respectively to the case where all parameters are keep fixed except for $\sigma_\varepsilon, \gamma_\varepsilon, \sigma_\eta$ and ϕ respectively.

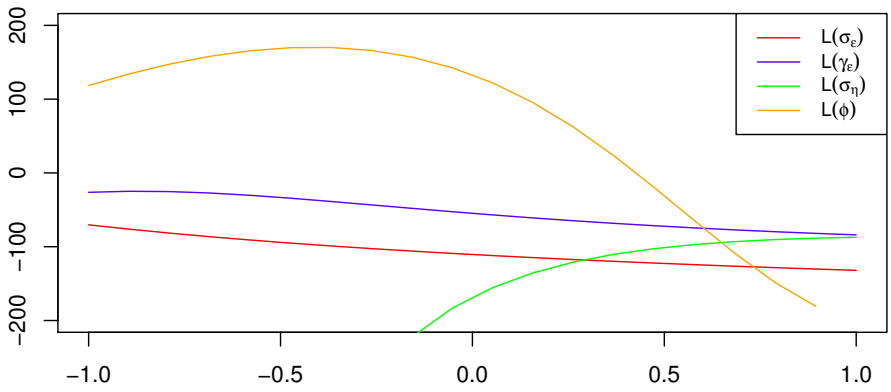


Figure 3.2: Likelihood of the closed skew-normal state-space model as a function of one parameter.

3.6 Concluding remarks

The so called skew Kalman filter (SKF) which we derived from a state-space model with closed skew-normal innovations in the measurement equation combined with a stationary autoregressive state process was introduced. Given that the Normal distribution is a special case of the csn distribution and that after a suitable transformation, one can recover the linear Gaussian ssm from our model, the proposed approach can be considered as a simple generalization of the Kalman filter.

For simulation and estimation purpose, we proposed a suitable reformulation of the csn likelihood function, making independent the observed data. The independence structure presented by these data allow for a factorization of the multivariate csn density as a product of univariate ones and leading to the parameter estimation via the maximum likelihood technique which in the present context, is proven to be computationally not costly, efficient and overcomes some of the drawbacks presented in Azzalini and Capitanio (1999).

Moreover, taking into account the fact that we wanted a model that can account for skewness, the present ssm representation is more practical as the skewness is not fading away when the sample size increase as shown by tables (3.1) and (3.2). Note that, this is not the case with some existing models. Also, compare to the existing representations in the literature such as the skew Kalman filter of Naveau et al. (2005), the skewness dimension for the observation and filtering densities is considerably reduce by at least half, this makes the estimation even faster and reliable compare to theirs.

Since by making use of BCSNI we ended with a posterior, likelihood and filtering densities within the csn family and analytically tractable, we can additionally say that the ultimate goal of mimicking as much as possible the Gaussian ssm was achieved up to a certain extend. For more general applications, some numerical challenges encountered in the current study along with other estimation techniques such as the method of weighted moment of Flecher et al. (2009) required further studies.

4. Robust Student-t state-space model

4.1 Introduction

For many problems from time-series analysis and related areas, the estimation of the state of the system evolving with time by employing available and noisy measurements provided by the system is usually required. With the same idea, this chapter makes use of discrete-time formulation of state-space model to characterize dynamical systems.

In state-space modelling framework, the attention is generally focussed on the state equation as it contains all important information needed to describe the system under consideration. For instance, in problems related to econometrics applications, this could be related to some economics indicators such as inflation and interest rates among others. Where the measurement equation represents noisy observations linked to the state vector. Compare to standard and traditional approaches in time-series analysis, one shall notice that this techniques offers considerable insights and can appear to be more convenient when it comes to handle nonlinear and/or non-Gaussian variables and multivariate observations.

A key property for situations evolving with time, is the ability to reformulate many of their recursive techniques as general solution of prediction, smoothing and filtering problems; making the flexibility, the suitability as well as the usefulness of state-space models not arguable any more (Härdle et al., 2000). In the second half of the 19C and especially with the seminal paper by (Kalman, 1960), these models back then started to offer new perspectives to tackle many unsolved/non-well solved problems in several fields. The interesting future which was glimpsed in these models and the need to make them more reliable, led to the study of their statistical properties which has received a considerable attention in the recent years; with contributions such as (Diderrich, 1985; Harrison and Stevens, 1976; West et al., 1985; Gamerman and Migon, 1993; Migon et al., 2005; Durbin and Koopman, 2012), and references hereafter in this chapter.

An inside look into a box containing mathematical algorithms which combine simplicity and optimality, and that can solve state-space models will show the Kalman filter in a prominent position. But it is unfortunate that this prominence holds only under specific assumptions such as the linearity of the model and the normality of the error terms driving the model.

An important feature of the Kalman and related filters also known as classical solutions to the filtering problem is that, their methodologies are mostly based on the moment of second order of the underlying distribution. Despite this appealing property, the quality of the filter depends strongly on the assumptions governing the model. This means that, even a small departure from these assumptions will have a considerable impact on the filter's quality or efficiency. Depending on where the perturbation or deviation occurs, different impacts can be encountered.

For instance, let us assume that the error term in the observation equation of our favourite state-space model is contaminated and that this error and the state process are independent. As result, we will end up with significantly bias values of single observations. On a different perspective, if we assume instead that, the contamination has occurred in the state equation via its error term, even if the effect may diminish on time, depending on the magnitude and the time horizon, the value of the state at the moment when the contamination has occurred and all subsequent states will be erroneous and so will be the related observations as the state process enter the observation equation. These are sometimes referred in the literature as additive and innovation outliers respectively. For detailed readings concerning this matter, see for instance (Fox, 1972) and (Ruckdeschel, 2000).

While standard filtering solutions are easy to implement, they do not always provide viable solutions, and especially when it comes to more general and non-linear systems. An alternative solution which is the one used in this chapter is the particle filter. The specificity of this approach which consists on representing the filtering density by a system of particles, resides in the fact that the entire probability density function is estimated instead of parametrized as in the standard solutions. It can be proved that, there is always a possibility to characterised any non-linear system by using particle filter given that we have at our disposal a sufficient large number of particles.

The wide range of applicability of this powerful recursive algorithm and the need of models

immunized against misspecification and outliers, motivated the need for counter-measures such as robustness. Before continuing, it is perhaps very important to state a clear difference between the widely used concept of robustness in the state-space model literature and the one we are using.

In state-space modelling, *Robustness* usually refers to the ability of the distribution on which the inference is based, to explain up to a certain extend, the true state and measurement equation. Whereas in the present case, we follow the idea of (Calvet et al., 2015) who succeeded to mimic the methodology of robust statistic and adapted it to sequential filtering. In this specific framework, the filter is consider to be robust if "*the relative error in the state distribution caused by misspecification is uniformly bounded by a linear function of the perturbation size*". Simply speaking, this is the ability of the model to handle the sensitivity of the filter due to the presence of little misspecification of the underlying model, to outliers in the observation process and to the possible occurrence of some contaminations at the instant just before the time period of interest.

The non-robustness of the Kalman filter and as explained in (Meinhold and Singpurwalla, 1989) can also be due to the fact that, the function describing the mean of the state process is not bounded and its variance is independent of the observations. Consequently, the inference of the state will greatly be impacted if an outlier occurs in the observation.

In the paper of (Calvet et al., 2015) that inspired us, a mechanism describing how to robustify the entire filtering density along with the robust particle filter algorithm has been provided and the Gaussian case implemented. In this chapter, we are concerned with following their steps and construct a robust state-space model allowing for Student-t error in the measurement equation. The goal being to take into consideration fat tails behaviour and therefore, a more general model.

The remaining of the chapter is organized as follows. In section 2, after recalling some key definitions and propositions, we will construct the robust Student-t filtering density. In section 3, we apply the theoretical results to the linear Gaussian model and to the unobserved component model with Stochastic volatility (UCSV) of (Stock and Watson, 2007). Then, in order to locate our filter with existing ones, we conduct some diagnostics checking and compare the performance of our robust Student-t filter with three other filters. Namely, the standard Gaussian, the standard Student-t and the robust Gaussian filter. Some concluding remarks and possible extensions appear

in section 4.

4.2 On the robustification of a filtering density

In its simple form and given a time index $t \in \mathbb{N}$, a state-space model establishes a temporal relationship between a sequence of available information (observations) $Y_t = y_1, y_2, \ldots, y_t$ and a set of latent variables $x_1, x_2, \ldots, x_t$ via the following two mechanisms.

$$y_t = x_t + \varepsilon_t \quad \text{and} \tag{4.2.1}$$

$$x_t = x_{t-1} + \eta_t. \tag{4.2.2}$$

Equations (4.2.1) and (4.2.2) are usually referred as observation and state equations respectively. ε_t and η_t are their respective associated error terms which are assumed to be independently distributed. Without any loss of generality, the model under investigation in this chapter characterizes the observation and the sate equations by the conditional observation density $f(y_t|x_t, Y_{t-1})$ and a Markov process x_t in the set $\mathcal{X}$ with kernel $\rho(x_t|x_{t-1})$ respectively.

Our concern is estimating the quantity $g(x_t|Y_t)$, also known as the filtering density. Given the observation density $f(y_t|x_t, Y_{t-1})$ and with the use of Bayes' rule, the filtering density can then be rewritten as $g(x_t|y_t, Y_{t-1}) \propto f(y_t|x_t, Y_{t-1})g(x_t|Y_{t-1})$. Before proceeding and in order to make things clear, it is worth making a brief detour into the paper by (Calvet et al., 2015). The goal being to recall some crucial assumptions, definitions and propositions from which the construction of our robust filtering density will be organized.

4.2.1 Background settings from (Calvet et al., 2015)

Let $\eta \in \mathcal{J}$ represent the perturbation size where $\mathcal{J}$ is a non-degenerate interval of the real line containing zero, $f(\cdot|x_t, Y_{t-1})$ the non-contaminated density and $f_{cont}(\cdot|x_t, Y_{t-1}, \eta)$ the contaminated probability density function (pdf). By assuming that at any given time t, the noisy

measurement y_t sampled from the contaminated pdf is made available, it can be proved that,

$$\lim_{\eta \to 0} f_{cont}(\cdot|x_t, Y_{t-1}, \eta) \to f(\cdot|x_t, Y_{t-1}). \tag{4.2.3}$$

That is, the two densities become closer and closer as the perturbation size tends to zero. From the above mentioned contaminated pdf and using the Bayes' rule, the contaminated filtering density is obtained as

$$g_{cont}(x_t|y_t, Y_{t-1}, \eta) \propto f_{cont}(y_t|x_t, Y_{t-1}, \eta) g_{cont}(x_t|Y_{t-1}, \eta) \tag{4.2.4}$$

It is important to notice that, before the time period t of interest, if no outliers have occurred, then in conjunction with the relation in Equation (4.2.3) one can easily prove that the contaminated filtering density becomes the non-contaminated one. Since $g_{cont}(x_t|y_t, Y_{t-1}, \eta)$ and $g(x_t|y_t, Y_{t-1})$ are available, simple analysis can be conducted in order to measure the impact of the contamination and to quantify the quality of the estimation when studying the filtering density. Thanks to (Calvet et al., 2015), this can be done with the following Equation (4.2.5), that computes the relative error between the non-contaminated and the true (contaminated) filtering densities.

$$|\log g(x_t|y_t, Y_{t-1}) - \log g_{cont}(x_t|y_t, Y_{t-1}, \eta)| \tag{4.2.5}$$

For convenience and later use, we now recall the following.

Definition 4.2.1 (Robustness of the filtering density). *The filtering density is said to be robust with respect to a family of contaminations f_{cont} if there exists a constant $c_1 \in \mathbb{R}_+$ such that*

$$|\log g(x_t|y_t, Y_{t-1}) - \log g_{cont}(x_t|y_t, Y_{t-1}, \eta)| \leq c_1|\eta| \tag{4.2.6}$$

for all $x_t \in \mathcal{X}, y_t \in \mathbb{R}^p, Y_{t-1} \in \mathbb{R}^{(t-1)p}$ and $\eta \in \mathcal{J}$.

In order words, a filter is said to be robust whenever any linear function of perturbation size is bounded below by Equation (4.2.5).

Definition 4.2.2 (Conditional means). *If $f(y_t|Y_{t-1}) = \int_{\mathcal{X}} f(y_t|x_t, Y_{t-1}) g(x_t|, Y_{t-1}) dx_t$, for every*

x_t *and* $t \geq 1$,

$$\mu_t = \int_{\mathbb{R}^p} y f(y|Y_{t-1}) dy, \tag{4.2.7}$$

$$\mu(x_t) = \int_{\mathbb{R}^p} y f(y|x_t, Y_{t-1}) dy. \tag{4.2.8}$$

Let us consider the observation y_t^* sampled from the non-contaminated density $f(\cdot|x_t, Y_{t-1})$, η the parameter driving the perturbation and $u_t \in \mathbb{R}^p$ a disturbance. Then, $y_t = y_t^* + \eta u_t$ characterizes the contamination dynamic. Every time the term disturbance is used, the reader should refer to one of the following.

Definition 4.2.3 (Point-mass disturbance). *Assuming that the disturbance is characterized by* $u_t = y_t^* - \mu_t$, *with* μ_t *defined as in Equation* (4.2.7) *leads to the following contaminated observation* $y_t = y_t^* + \eta(y_t^* - u_t)$, *which provides the contaminated observation density as*

$$f_{cont}(y_t|x_t, Y_{t-1}; \eta) = (1+\eta)^{-p} f[(1+\eta)^{-1}(y_t + \eta u_t)|x_t, Y_{t-1}], \tag{4.2.9}$$

with $\eta \in [\underline{\eta}, +\infty)$ *and the negative constant* $\underline{\eta} > -1$.

Definition 4.2.4 (Continuous disturbance). *If instead, disturbances* u_t *with conditional density* $\xi(\cdot|y_t^*, Y_{t-1})$ *are considered. Then, for all* $x_t \in \mathcal{X}, y_t \in \mathbb{R}^p, Y_{t-1} \in \mathbb{R}^{(t-1)p}$ *and* $\eta \in \mathbb{R}$, *one obtains the following contaminated observation density*

$$f_{cont}(y_t|x_t, Y_{t-1}; \eta) = \int_{\mathbb{R}^p} f(y_t - \eta u_t|x_t, Y_{t-1}) \xi(u_t|y_t - \eta u_t, Y_{t-1}) du_t, \tag{4.2.10}$$

It is important to point out that, the existence of Equation (4.2.10), is subject to Assumption (4.2.1) below. Whereas more details on how (4.2.10) is derivation can be found in the online Appendix of (Calvet et al., 2015).

Assumption 4.2.1. *The conditional pdf of the continuous disturbance* u_t *holds for* $\mathbb{E}(u_t|Y_{t-1}) = 0$ *and* $\xi(u_t|y_t^*, Y_{t-1}) = 0$ *given that* $\|u_t\| > \|y_t^* - \mathbb{E}(y_t^*|Y_{t-1})\|$ *is satisfied. Moreover, it should*

exist a nonnegative constant c_2 *such that,*

$$\left| u_t' \frac{\partial \log \xi}{\partial y_t^*}(u_t|y_t^*, Y_{t-1}) \right| \leq c_2 \tag{4.2.11}$$

for all $u_t \in \mathbb{R}^p, y_t \in \mathbb{R}^p$ *and* $Y_{t-1} \in \mathbb{R}^{(t-1)p}$.

We now recall the following assumption that provides the framework under which Proposition (4.2.1) below holds.

Assumption 4.2.2. *For every time index* t*, observation* $y_t \in \mathbb{R}^p$*, state process* x_t *and past observations* Y_{t-1}*, the observation density* $f(y_t|x_t, Y_{t-1})$ *is strictly positive and twice continuously differentiable with respect to* y_t.

Proposition 4.2.1 (Sufficient condition for robustness). *Assume that there exists* $c \in \mathbb{R}_+$ *such that*

$$\left\| \frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t} \right\| \|y_t - \mu_t\| \leq c, \tag{4.2.12}$$

for all x_t, y_t *and* Y_{t-1}*. Then the filter is robust to point-mass disturbances and to continuous disturbances that satisfy Assumption* (4.2.1).

One can easily see the link between Proposition (4.2.1), inequality in Definition (4.2.1) and the two types of above mentioned disturbances. In fact, Inequality (4.2.6) will be valid in the case of point-mass disturbances if $c_1 = 2c/(1+\underline{\eta})$ and for all $\eta \in [\underline{\eta}, +\infty)$, where as for continuous disturbances, if $c_1 = 2(c + c_2)$ and all $\eta \in \mathbb{R}$.

Given Assumption (4.2.2) and as we will see in the next section, the construction of a robust filter requires solving Inequality (4.2.12), provided the validity of the additional Assumption (4.2.3) below.

Assumption 4.2.3 (Critical region). *For all* $x_t, Y_{t-1}, z \in \mathbb{R}^p$ *and every* $c \in \mathbb{R}_+$*, the critical region*

$$\left\{ y \in \mathbb{R}^p \ s.t. \ \left\| \frac{\partial \log f(y|x_t, Y_{t-1})}{\partial y} \right\| \|y - \mathbb{E}(y_t|Y_{t-1})\| = c \right\}, \tag{4.2.13}$$

intersects the segment $[\mu(x_t), z]$ *finitely many times.*

Proposition 4.2.2 (Robustified observation density). *Let us consider the following function*

$$G(y) = h_{\frac{c}{\|y-\mu_t\|}}\left[\frac{\partial \log f(y|x_t, Y_{t-1})}{\partial y}\right], \tag{4.2.14}$$

where $h_\tau(z) = z\min(1; \tau/\|z\|)$ *is the multivariate Huber function and* $c \in \mathbb{R}_+$ *is a tuning constant. Then under Assumptions* (4.2.2) *and* (4.2.3)*, the function*

$$\tilde{f}(y_t|x_t, Y_{t-1}) = f[\mu(x_t)|x_t, Y_{t-1}] \exp\left(\int_0^1 G[y_t(s)]'[y_t - \mu(x_t)]ds\right) \tag{4.2.15}$$

belongs to $\mathcal{C}^1(\mathbb{R}^p)$ *and satisfies the sufficient condition for robustness for every* $y_t \in \mathbb{R}^p$*.*

4.2.2 Robustification of a univariate Student-t density

With all these ingredients, we simply need to place ourselves into a state-space model framework and consider the observation equation from a Student-t distribution with ν degrees of freedom. That is,

$$f(y_t|x_t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1 + \frac{(y_t - x_t)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}. \tag{4.2.16}$$

We only have to replace accordingly the observation density in (4.2.12) with the Student-t and solve it for equality in order to find the critical roots. Given the position of $\mu(x_t)$, the value of c and with respect to the roots, a different value of the observation density will be provided. All this will then leads to a specific (robust) filtering density.

As pointed out in (Calvet et al., 2015), Assumption (4.2.3) is satisfied by many models such as the Student's t-distributions. This is one of the motivation driving the use of this particular distribution in the current chapter.

To ease our computations, let us make these further settings,

$$A = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}.$$

Equation (4.2.16) can then be rewritten as $f(y_t|x_t) = A\left(1 + \frac{(y_t - x_t)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$, which implies that

$$\begin{aligned} \log f(y_t|x_t) &= \log\left[A\left(1 + \frac{(y_t - x_t)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}\right] \\ &= \log A - \frac{1}{2}(\nu+1)\log\left(1 + \frac{(y_t - x_t)^2}{\nu}\right). \end{aligned} \tag{4.2.17}$$

Yields

$$\frac{\partial \log f(y_t|x_t)}{\partial y_t} = -(\nu+1)\frac{y_t - x_t}{\nu + (y_t - x_t)^2} \tag{4.2.18}$$

In order to find the critical roots, we need to check when the sufficient condition for robustness as defined in Calvet et al. (2015), will hold as equality. As we now do, this implies solving Equation (4.2.19) below.

$$\begin{aligned} \left\|\frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t}\right\| \|y_t - \mu_t\| &= c \\ \text{ie.} \quad \left|-(\nu+1)\frac{y_t - x_t}{\nu + (y_t - x_t)^2}\right| |y_t - \mu_t| &= c \\ \text{ie.} \quad (\nu+1)|(y_t - x_t)(y_t - \mu_t)| &= c(\nu + (y_t - x_t)^2) \end{aligned} \tag{4.2.19}$$

$$\begin{cases} (\nu+1)(y_t - x_t)(y_t - \mu_t) = c(\nu + (y_t - x_t)^2) & \text{if} \quad (y_t < x_t, y_t < \mu_t) \quad \text{or} \quad (y_t > x_t, y_t > \mu_t) \\ -(\nu+1)(y_t - x_t)(y_t - \mu_t) = c(\nu + (y_t - x_t)^2) & \text{if} \quad (y_t < x_t, y_t > \mu_t) \quad \text{or} \quad (y_t > x_t, y_t < \mu_t) \end{cases}$$

which implies respectively Equations (4.2.20) and (4.2.21) below.

$$y_t^2(\nu + 1 - c) + y_t(2cx_t - (x_t + \mu_t)(\nu+1)) + x_t\mu_t(\nu+1) - cx_t^2 - c\nu = 0 \tag{4.2.20}$$

$$y_t^2(-\nu-1-c)+y_t(2cx_t+(x_t+\mu_t)(\nu+1))-x_t\mu_t(\nu+1)-cx_t^2-c\nu = 0 \quad (4.2.21)$$

In what follows, Equations (4.2.20) and (4.2.21) will be referred as case 1 and case 2 respectively.

4.2.3 Condition on c in order for case 1 to admit two distinct roots

The discriminant Δ_1 with respect to the variable y_t is defined as follows:

$$\begin{aligned}\Delta_1 &= [(2cx_t-(x_t+\mu_t)(\nu+1))]^2-4[(\nu+1-c)][x_t\mu_t(\nu+1)-cx_t^2-c\nu] \\ &= [(\nu+1)(x_t-\mu_t)]^2+4c\nu(\nu+1-c) \quad (4.2.22)\end{aligned}$$

Now as a function of $c, \Delta_1(c) > 0$ if and only if $\nu+1 > c$ and c belongs to

$$\left(\frac{4\nu(\nu+1)-\sqrt{[4\nu(\nu+1)]^2+16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu}, \frac{4\nu(\nu+1)+\sqrt{[4\nu(\nu+1)]^2+16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu}\right)$$

Since $c \in \mathbb{R}_+$ and $\frac{4\nu(\nu+1)-\sqrt{[4\nu(\nu+1)]^2+16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu} < 0$, we conclude that, case 1 has two distinct roots if

$$c \in \left(0, c_1 = \frac{4\nu(\nu+1)+\sqrt{[4\nu(\nu+1)]^2+16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu}\right) \quad (4.2.23)$$

4.2.4 Condition on c in order for case 2 to admit two distinct roots

The discriminant Δ_2 with respect to the variable y_t is defined as follows:

$$\begin{aligned}\Delta_2 &= [(2cx_t+(x_t+\mu_t)(\nu+1))]^2-4[(\nu+1+c)][x_t\mu_t(\nu+1)+cx_t^2+c\nu] \\ &= [(\nu+1)(x_t-\mu_t)]^2-4c\nu(\nu+1+c) \quad (4.2.24)\end{aligned}$$

This implies that as a function of $c, \Delta_2(c) > 0$ if and only if c belongs to the interval

$$\left(\frac{-4\nu(\nu+1) - \sqrt{[4\nu(\nu+1)]^2 + 16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu}, \frac{-4\nu(\nu+1) + \sqrt{[4\nu(\nu+1)]^2 + 16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu} \right)$$

Given the fact that $c \in \mathbb{R}_+$ and $\frac{-4\nu(\nu+1) - \sqrt{[4\nu(\nu+1)]^2 + 16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu} < 0$, we state that, case 2 has two distinct roots if

$$c \in \left(0, c_2 = \frac{-4\nu(\nu+1) + \sqrt{[4\nu(\nu+1)]^2 + 16\nu[(\nu+1)(x_t-\mu_t)]^2}}{8\nu} \right) \tag{4.2.25}$$

The reader can easily check that $c_1 > c_2$, condition that will play an important role later. Therefore, if $c > c_2$, we are in the situation where $(y_t < x_t, y_t < \mu_t)$ or $(y_t > x_t, y_t > \mu_t)$ and equation (4.2.19) has two distinct solutions y_-^* and y_+^* satisfying $y_-^* < x_t, \mu_t < y_+^*, \quad \forall (x_t, \mu_t)$ and defined by

$$y_-^* = \frac{(x_t+\mu_t)(\nu+1) - 2cx_t - \sqrt{[(\nu+1)(x_t-\mu_t)]^2 + 4c\nu(\nu+1-c)}}{2(\nu+1-c)} \tag{4.2.26}$$

and

$$y_+^* = \frac{(x_t+\mu_t)(\nu+1) - 2cx_t + \sqrt{[(\nu+1)(x_t-\mu_t)]^2 + 4c\nu(\nu+1-c)}}{2(\nu+1-c)} \tag{4.2.27}$$

Otherwise, we are in the situation where $(y_t < x_t, y_t > \mu_t)$ or $(y_t > x_t, y_t < \mu_t)$ and we have the following two additional solutions z_-^* and z_+^*

$$z_-^* = \frac{(x_t+\mu_t)(\nu+1) + 2cx_t - \sqrt{[(\nu+1)(x_t-\mu_t)]^2 - 4c\nu(\nu+1+c)}}{2(\nu+1+c)} \tag{4.2.28}$$

$$z_+^* = \frac{(x_t+\mu_t)(\nu+1) + 2cx_t + \sqrt{[(\nu+1)(x_t-\mu_t)]^2 - 4c\nu(\nu+1+c)}}{2(\nu+1+c)} \tag{4.2.29}$$

and the following will hold

$$\begin{cases} y_-^* < x_t < z_-^* < z_+^* < \mu_t < y_+^* & \text{if} \quad x_t < \mu_t \\ y_-^* < \mu_t < z_-^* < z_+^* < x_t < y_+^* & \text{if} \quad x_t > \mu_t \end{cases} \tag{4.2.30}$$

4.2.5 Robustified Student-t observation density

In this section, we intend to "Huberize" the derivative of the log-observation density computed previously, its integration will provide us with the robust Student density. Equation (4.2.14) implies

$$\begin{aligned} G(y) &= h_{\frac{c}{\|y-\mu_t\|}}\left[\frac{\partial \log f(y|x_t, Y_{t-1})}{\partial y}\right] \\ &= -(\nu+1)\frac{y_- x_t}{\nu+(y-x_t)^2}\min\left\{1; \frac{\frac{c}{|y-\mu_t|}}{\left|-(\nu+1)\frac{y_- x_t}{\nu+(y-x_t)^2}\right|}\right\} \\ &= \max\left\{-(\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}; -\frac{c(y-x_t)}{|(y-\mu_t)(y-x_t)|}\right\} \end{aligned} \tag{4.2.31}$$

We can then define the followings functions

$$G_1(y) = -(\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2} \tag{4.2.32}$$

and

$$G_2(y) = \begin{cases} -\dfrac{c}{y-\mu_t}, & \text{if } (y_t < x_t, y_t < \mu_t) \quad \text{or} \quad (y_t > x_t, y_t > \mu_t). \\ \dfrac{c}{y-\mu_t}, & \text{otherwise.} \end{cases} \tag{4.2.33}$$

The robustified density Calvet et al. (2015) is defined by

$$\tilde{f}(y_t|x_t, Y_{t-1}) = f[\mu(x_t)|x_t, Y_{t-1}]\exp\left\{\int_0^1 [y_t - \mu(x_t)]' G[\mu(x_t) + s(y_t - \mu(x_t))]ds\right\} \tag{4.2.34}$$

In order to determine which of $G_1(y)$ or $G_2(y)$ is the maximum in each interval, one can choose to evaluate the two functions at $\mu(x_t) = x_t$, then alternate the maximum in the next interval.

It is easy to check that $f(\mu(x_t)|x_t, Y_{t-1}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}$. For simplicity, let $f(y)$ represents the Student-t observation density evaluated at y.

The robustified density $\tilde{f}(y_t|x_t, Y_{t-1})$, as defined by equation (4.2.34) will be characterized as follows.

4.2.5.1 When $c > c_2$

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} D_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^* \\ f(y_t) & \text{if } y_t \in [y_-^*, y_+^*) \\ D_{2,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^* \end{cases} \tag{4.2.35}$$

Now, let $B_{1,t}(x_t)$ denotes the normalizing constant of the density in (4.2.35) and $F(y)$ the Student cdf evaluated at y, then

$$B_{1,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1})dy\right]^{-1}$$

and we can write

$$B_{1,t}^{-1}(x_t) = \frac{|y_+^* - \mu_t|f(y_+^*) - |y_-^* - \mu_t|f(y_-^*)}{c-1} + F(y_+^*) - F(y_-^*) \tag{4.2.36}$$

4.2.5.2 When $c \leq c_2$ and $x_t < \mu_t$

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} C_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^* \\ f(y_t) & \text{if } y_t \in [y_-^*, z_-^*) \\ C_{2,t}(x_t)|y_t - \mu_t|^{c} & \text{if } y_t \in [z_-^*, z_+^*) \\ C_{3,t}(x_t)f(y_t) & \text{if } y_t \in [z_+^*, y_+^*) \\ C_{4,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^* \end{cases} \tag{4.2.37}$$

and the normalizing constant $B_{2,t}(x_t)$ defined as follows

$$B_{2,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1})dy\right]^{-1}$$

with

$$\begin{aligned} B_{2,t}^{-1}(x_t) &= \frac{f(y_-^*)|y_-^* - \mu_t|}{c-1} + F(z_-^*) - F(y_-^*) + f(z_-^*)\left[\frac{|z_+^* - \mu_t|}{|z_-^* - \mu_t|}\right]^c \times \\ &\times \left\{\frac{|z_+^* - \mu_t|}{c+1} + \frac{F(y_+^*) - F(z_+^*)}{f(z_+^*)} + \frac{|y_+^* - \mu_t| f(y_+^*)}{f(z_+^*)(c-1)}\right\} \end{aligned} \tag{4.2.38}$$

4.2.5.3 When $c \leq c_2$ and $x_t > \mu_t$

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} E_{4,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^* \\ E_{3,t}(x_t) f(y_t) & \text{if } y_t \in [y_-^*, z_-^*) \\ E_{2,t}(x_t)|y_t - \mu_t|^c & \text{if } y_t \in [z_-^*, z_+^*) \\ f(y_t) & \text{if } y_t \in [z_+^*, y_+^*) \\ E_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^* \end{cases} \tag{4.2.39}$$

and the normalizing constant $B_{3,t}(x_t)$ defined by

$$B_{3,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1})dy\right]^{-1}$$

Where,

$$\begin{aligned} B_{3,t}^{-1}(x_t) &= \frac{f(y_+^*)|y_+^* - \mu_t|}{c-1} + F(y_+^*) - F(z_+^*) + \frac{f(z_+^*)}{c+1}\left\{|z_+^* - \mu_t| - \left[\frac{|z_-^* - \mu_t|}{|z_+^* - \mu_t|}\right]^c |z_-^* - \mu_t|\right\} + \\ &+ \frac{f(z_+^*)}{f(z_-^*)(|z_+^* - \mu_t||z_-^* - \mu_t|)^c}\left[F(z_-^*) - F(y_-^*) - \frac{f(y_-^*)|y_-^* - \mu_t|}{c-1}\right] \end{aligned} \tag{4.2.40}$$

Figure (4.1) shows the behaviour of the robustified Student-t observation density for various values of c. The left and the right panels correspond respectively to the case when $\mu(x_t) > \mu_t$ and $\mu(x_t) < \mu_t$. Just like in the Gaussian case presented in Calvet et al. (2015), the robustified

density converges to the original observation density as the constant c becomes larger. This means that decreasing values of the tuning constant c leads to an increase strength of the robustification.

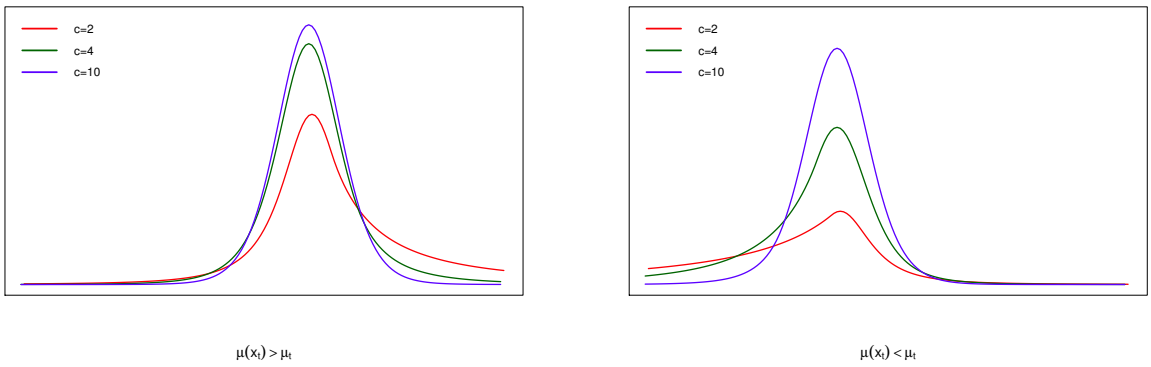


Figure 4.1: Robustified Student-t observation density.

4.3 Application and analysis

In this part, we are going to implement the theoretical results and conduct some statistical analysis. Since we would like to compare the performance of our robust filtering density with exiting models, we choose for application the the linear Gaussian ssm and the UCSV. We will then proceed with some comparison between the standard model in (Stock and Watson, 2007) (GPF), the robust Gaussian particle filter (RGPF) (Calvet et al., 2015), the standard Student-t particle filter (SSPF) and the proposed robust Student-t particle filter (RSPF). The availability of conditional and robust filtering density suggests to use the maximum likelihood approach combined with the particle filter technique for parameter estimation.

However, it is well known that when using the particle filter, the resampling step is the most costly computationally speaking. The combination of the ***Malmquist ordered statistics*** (Cappé et al., 2005) and the robust filtering density will considerably decrease the time usually needed. The MSE and the empirical rejection frequency will be computed and compared for all methods in order to state the efficiency of ours.

4.3.1 The linear Gaussian ssm

The linear Gaussian ssm considered is defined by Equation (4.3.1) below.

$$\begin{cases} y_t = ax_t + \varepsilon_t \\ x_t = \phi x_{t-1} + \eta_t, \end{cases} \tag{4.3.1}$$

where $y_t, x_t \in \mathbb{R}^p$, $\varepsilon_t \sim N(0,1)$ for the Gaussian case and $\varepsilon_t \sim t(\nu, 1)$ for the Student-t case, $\eta_t \sim N(0,1)$, and ε_t and η_t are independent. For the simulations, we use the point-mass disturbance with 5% contamination of size $\eta = 10$, we further set $a = 0.4$ and $\phi = 0.9$.

In tables (4.1) and (4.2), we report respectively the empirical rejection frequency and the the mean squared error (MSE) of the forecast observation for the Normal linear model with 500 observations. For the uncontaminated series as well as under 5% contamination, the robust Student-t model has the best performance both in terms of rejection frequency and the MSE, followed by the robust Gaussian model.

Models	No contamination	5% contamination
Gaussian	159	157
Student-t	145	143
Robust Gaussian	81	84
Robust Student-t	50	52

Table 4.1: Empirical rejection frequency of the 90% prediction band for the Normal linear model

Models	No contamination	5% contamination
Gaussian	3337.166	3172.259
Student-t	3478.01	3277.573
Robust Gaussian	2858.989	2776.012
Robust Student-t	2794.055	2764.348

Table 4.2: Mean squared error of the forecast observations for the Normal linear model

4.3.2 The UCSV model

In this section, all the analysis will be conducted under the assumption that there is no contamination and we consider the UCSV model defined as follows,

$$\begin{cases} \pi_t = \tau_t + \eta_t, & \eta_t \sim N(0, \sigma_t^{\eta}) \\ \tau_t = \tau_{t-1} + \varepsilon_t, & \varepsilon_t \sim N(0, \sigma_t^{\varepsilon}) \\ \log(\sigma_t^{\eta}) = \log(\sigma_{t-1}^{\eta}) + \nu_t^{\eta}, & \nu_t^{\eta} \sim N(0, \gamma_1) \\ \log(\sigma_t^{\varepsilon}) = \log(\sigma_{t-1}^{\varepsilon}) + \nu_t^{\varepsilon}, & \nu_t^{\varepsilon} \sim N(0, \gamma_2) \end{cases} \tag{4.3.2}$$

Figure (4.2) represents the filtered mean of the state processes $x_1(t), x_2(t), x_3(t)$ using the 4 particle filters. That is, the GPF in red, the SSPF in blue, the RGPF in green and the RSPF in orange using B=30000 particles. The left panel correspond to the case where the truncation constant $c = 5.1413$ and the degree of freedom for the Student-t and robust Student-t filtering densities $\nu = 5$. The middle panel is for $c = 10$ and $\nu = 10$, where as the right panel is for $c = 20$ and $\nu = 20$.

Figure (4.3) depicts the prediction interval using the GPF in red, the SSPF in blue, the RGPF in green and RSPF in orange with 30000 particles. The left panel correspond to the case where $c = 5.1413$ and $\nu = 5$. The middle panel is for $c = 10$ and $\nu = 10$ and the right one is for $c = 20$ and $\nu = 20$. Table (4.3) reports the proportion of time the data y_t fall outside the 90% prediction bands, using the GPF, the SSPF, the RGPF and RSPF with sample B=30000.

Figure (4.4) describes the observations forecast using the 4 particle filters with various values of the tuning constant c and with 30000 particles. For ($c = 5.1413$ $\nu = 5$), ($c = 10$ $\nu = 10$) and ($c = 20$ $\nu = 20$), we have respectively the left, the middle and the right panel. Table (4.4) reports the mean squared error for the forecasted data using the GPF, the SSPF, the RGPF and the RSPF.

Figure (4.5) represents the likelihood function using the 4 particle filters and with various values of c. Namely, the G (red), S (blue), RG (green) and the RS (orange) with 30000 particles. For ($c = 5.1413$ $\nu = 5$), ($c = 10$ $\nu = 10$) and ($c = 20$ $\nu = 20$), we have respectively the left, the

middle and the right panel. Table (4.5) provides the parameter's estimate with the likelihood method via grid-search using the GPF, the SSPF, the RGPF and the RSPF. In parentheses are provided the optimized values.

When looking at Figure (4.2), independently of the values of c and ν, one can see that for all cases in the first row which corresponds to the process $x_1(t)$, all the filtered means are in a close neighbourhood of the data. The Gaussian, the robust Gaussian and the Student-t filtered mean all together match perfectly with the data. When the tuning constant c and the degree of freedom ν increase simultaneously all the filtered means are still in the neighbourhood of the data but, robust Student-t becomes the less viable option. The Gaussian, the robust Gaussian and the Student-t filtered mean still match the data and have apparently similar performance.

In the top row of the prediction intervals presented by Figure (4.3), the first part of the hidden state $x_1(t)$ show that the prediction intervals obtained with all particle filters methods contain best the data. For all values of ν and c, the robust Student-t has the wider prediction interval. This may suggest it as the best filter as it is more probable for future observations to be contained in that interval.

As c and ν increase the Gaussian, the robust Gaussian and the Student-t prediction intervals become similar and best option compare to the robust Student-t prediction interval. We can observe that the three intervals have an apparent similar performance, this because as ν increases, the SSPF converges to the Gaussian and as c increases, the robust Gaussian converges to the Gaussian. Therefore, they will perform almost equally if we were to consider even bigger values of ν and c.

The empirical rejection frequency which we define as the failure rate or the percentage of data outside the prediction interval presented in Table (4.3) confirm these findings. The accuracy of the RSPF depends on the degree of robustness. That is, for smaller value of the tuning constant c, the RSPF outperforms the SSPF, GPF and RGPF and reaches 60% rejection frequency for mild robustification.

Additionally, we can see that for whatever values of c and ν the failure rate of the Gaussian particle filter doesn't change. This is in line with the theory as the GPF is neither a function of ν

nor c. As expected, even if their performance is not that different, the SSPF outperform the GPF and the RGPF. When c and ν increase the accuracy of these three filters becomes similar, this result is theoretically supported by the fact that the Student-t density converges to the Normal one for increasing degree of freedom.

The mean squared error of the forecast observations in Table (4.4) shows that for all four filters and for all values of c and ν, the robust Student-t is the less accurate, the forecast observations in Figure (4.4) confirms the accuracy of the RSPF for smaller values of c compare to the other three counterparts. Similar to what discussed earlier, Table (4.4) confirms the convergence of the robust Gaussian and robust Student-t when c and ν increase.

Finally, in Table (4.5) we have the parameter estimate and the optimized values in parentheses produced with the likelihood method. This estimation which is also in accordance with the theoretical results, suggests that the robust Gaussian, the robust Student-t and the Student-t filters converge respectively to the Gaussian, the Student-t and the Gaussian filters when c increases.

	$c = 5.1413, \nu = 5$	$c = 10, \nu = 10$	$c = 20, \nu = 20$
Gaussian	30%	30%	30%
Student-t	29%	30%	29%
Robust Gaussian	32%	30%	30%
Robust Student-t	16%	35%	60%

Table 4.3: Empirical rejection frequency of the 90% prediction band for the UCSV

	$c = 5.1413, \nu = 5$	$c = 10, \nu = 10$	$c = 20, \nu = 20$
Gaussian	750.426	750.426	750.426
Student-t	762.2935	756.9334	754.2368
Robust Gaussian	762.626	760.6836	755.7935
Robust Student-t	1460.192	1640.807	2359.121

Table 4.4: Mean squared error of the forecast observation for the UCSV

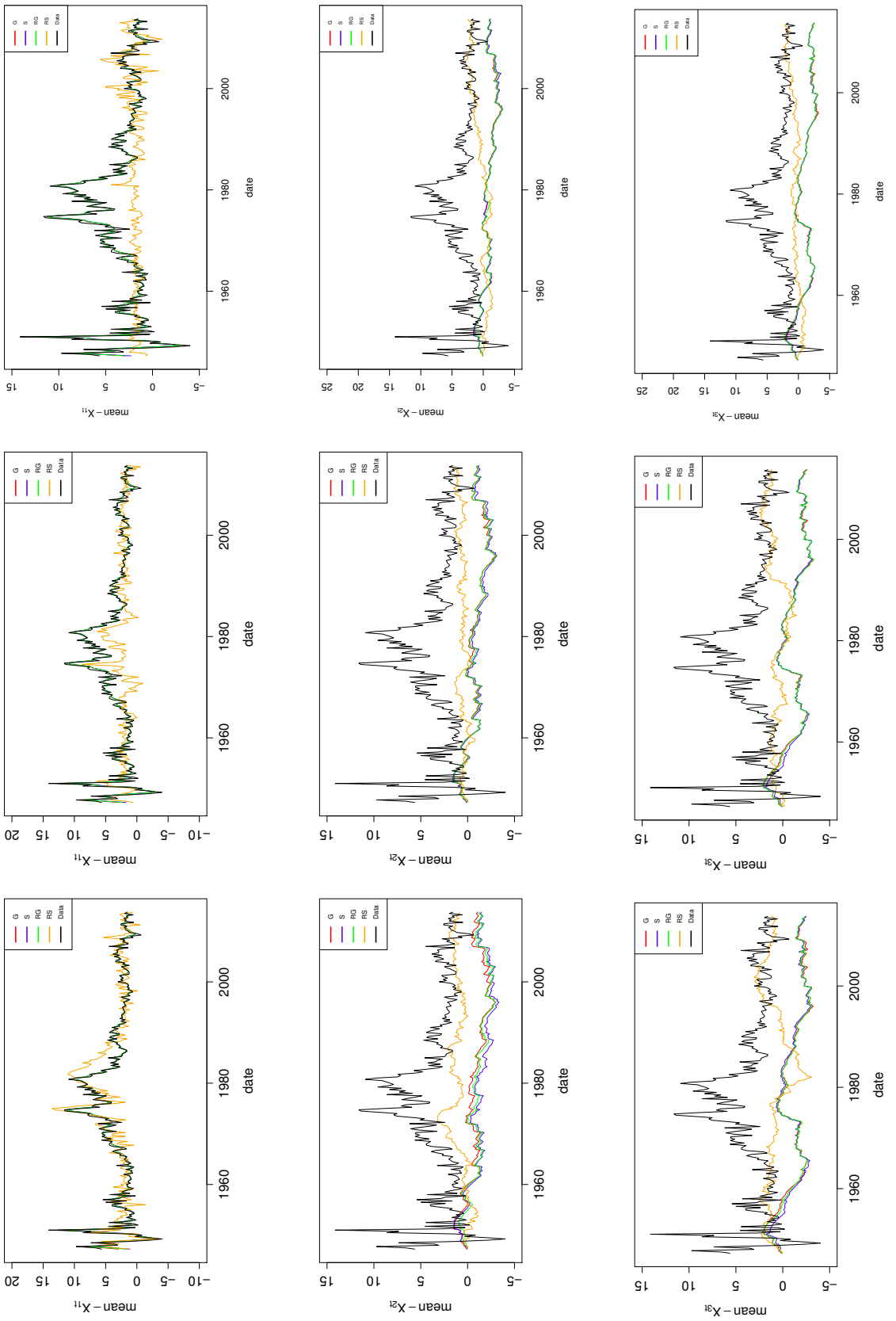


Figure 4.2: Filtered mean of the state processes $x_1(t), x_2(t), x_3(t)$ for the UCSV

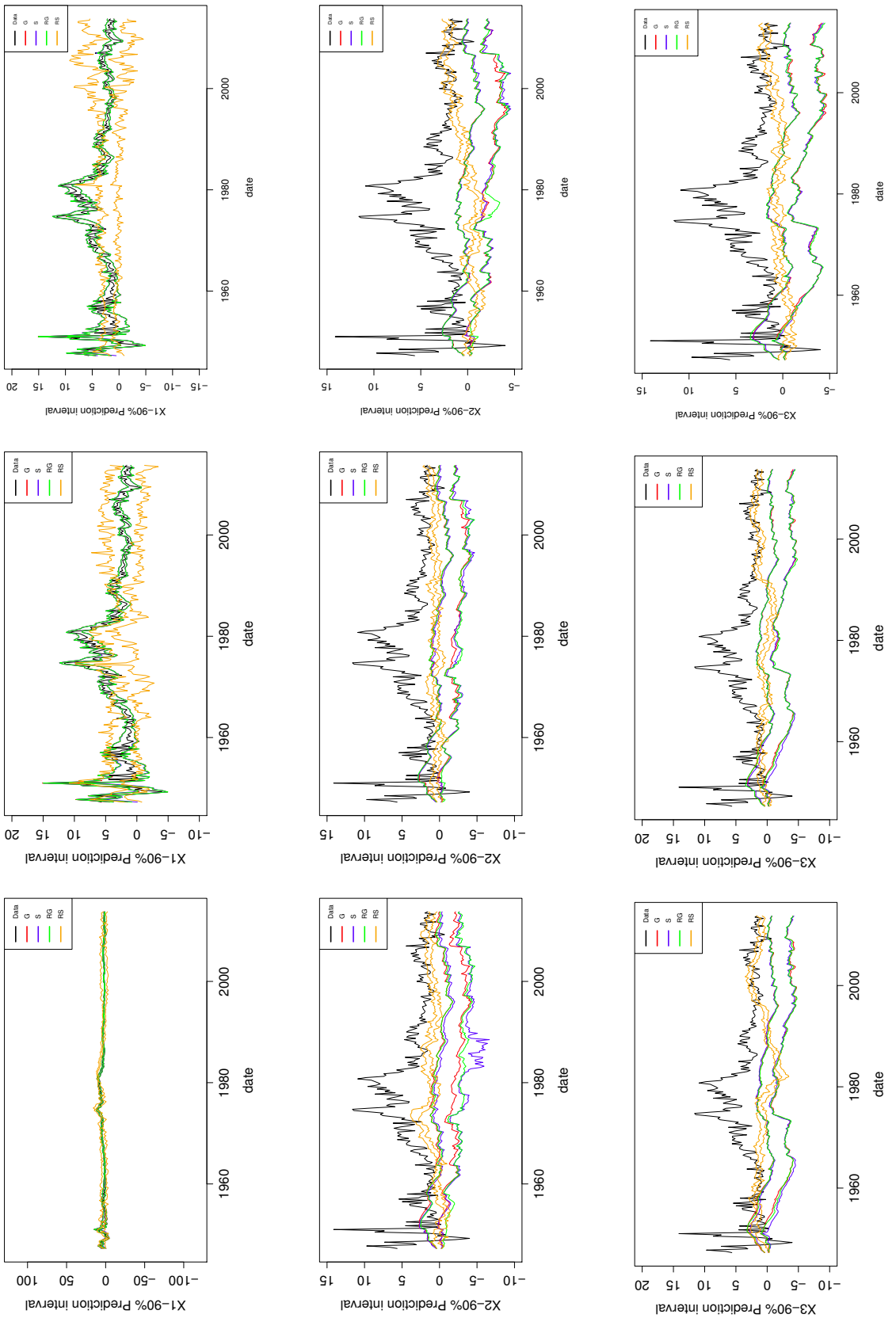


Figure 4.3: Prediction intervals using the 4 particle filters method for the UCSV

γ	$c = 5.1413, \nu = 5$	$c = 10, \nu = 10$	$c = 20, \nu = 20$
Gaussian	0.46 (-424.5036)	0.43 (-424.0915)	0.46 (-424.1451)
Student-t	0.95 (-1583.547)	0.93 (-1550.909)	0.46 (-1650.1)
Robust Gaussian	0.39 (-424.925)	0.47 (-424.4674)	0.42 (-424.243)
Robust Student-t	0.87 (-1642.791)	0.54 (-1626.151)	0.65 (26875.27)

Table 4.5: Parameter estimate via the likelihood for the UCSV

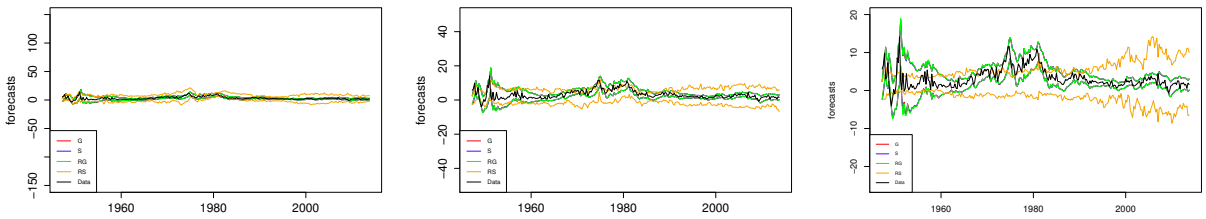


Figure 4.4: Forecast observations using the 4 particle filters for the UCSV

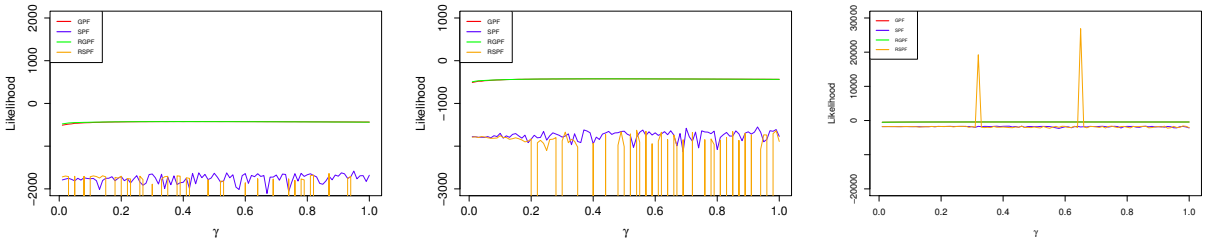


Figure 4.5: Likelihood function using the 4 particle filters for the UCSV

4.4 Concluding remarks

In this chapter, we proposed and constructed a robust Student-t state-space model which under no contamination can be used as a viable alternative to the often used linear Gaussian approach and achieve remarkably well than the robust Gaussian filter proposed by Calvet et al. (2015). We proved that the obtained robust density belongs to the Student-t family and in term of efficiency, under contamination the results obtained with the simulated model are better than in the standard Student-t case which is naturally robust. By using the robustified density, we were able to derive a recursive procedure to obtain all finite marginal and conditional distributions, and therefore the likelihood function. From simulations conducted on the Gaussian linear model, we proved that the proposed method is robust, accurate and efficient compare to the models where standard assumptions are made. By applying our theoretical results on the US inflation data from 1947 to 2013, we were able estimate the parameter in the UCSV and our model performed well.

As for the question "why has the U.S. inflation become harder to forecast?" by (Stock and Watson, 2007), it is well known that, obtaining perfect accuracy when forecasting inflation is hard to achieve. Such an imperfection can be explained by the following non exhaustive phenomenons. The presence of some external shocks that for example can be originated from the occurrence of a jump in the world oil or energy prices can have a severe impact on the world's economy. The fact that, the available measurements of inflation are usually direct or indirect consequences of important amount of pricing decisions. The error committed and sometimes the omission of some information when computing the consumer price index. The importation prices can become very volatile due to some fluctuation in the exchange rate. Recurrent policy changes by central banks on the interest rate when they realize that the inflation target will not be met on a given time horizon. One can indeed see that, most if not all of these facts are somehow related to the occurrence of outliers or misspecification. The foregoing and clear evidence suggest that for a relatively good inflation forecast, the need of mathematical models that can take into consideration the above mentioned issues is most desirable. We suggest that in the framework of state-space modelling, the use of robust filtering density and different time index for inflation data are potential way out.

Finally and as suggested by the computed MSE forecast,there is a price to pay when robustifying a filtering density and this price is even higher when robustifying a naturally robust model such as the Student-t. One then has to be willing to cope with this cost in the search of robustness. For further research and in order to simultaneously reduce this cost and achieve robustness, it can be interesting look at the construction of the robust filtering density using other robust function and study their statistical properties.

4.5 Appendix: Detailed computations for the robust filtering density

4.5.1 When $c > c_2$ and given the position of y_t, let us compute $\tilde{f}(y_t|x_t, Y_{t-1})$ and denote it by $I_i(y_t)$

It is easy to check that $f(\mu(x_t)|x_t, Y_{t-1}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}$. Moreover, let $f(y)$ represents the Student-t observation density evaluated at y. Thus,

If $y_t < y_-^*$

$$\begin{aligned}
I_1(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{y_t}^{y_-^*} G_2(y)dy - \int_{y_-^*}^{\mu(x_t)} G_1(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{y_t}^{y_-^*} \frac{c}{y-\mu_t}dy + \int_{y_-^*}^{\mu(x_t)} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{c\Big[\log|y-\mu_t|\Big]_{y_t}^{y_-^*} + \frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{y_-^*}^{\mu(x_t)}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{y_-^*-\mu_t}{y_t-\mu_t}\right|^c + \log\left|\frac{\nu+(x_t-x_t)^2}{\nu+(y_-^*-x_t)^2}\right|^{\frac{\nu+1}{2}}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{\left|\frac{y_-^*-\mu_t}{y_t-\mu_t}\right|^c \left|\frac{\nu+(x_t-x_t)^2}{\nu+(y_-^*-x_t)^2}\right|^{\frac{\nu+1}{2}}\right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{(y_-^*-x_t)^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}\right]|y_-^*-\mu_t|^c|y_t-\mu_t|^{-c} \\
&= f(y_-^*)|y_-^*-\mu_t|^c|y_t-\mu_t|^{-c} \\
&= D_{1,t}(x_t)|y_t-\mu_t|^{-c}, \quad \text{where} \quad D_{1,t}(x_t) = f(y_-^*)|y_-^*-\mu_t|^c
\end{aligned} \tag{4.5.1}$$

If $y_t \in [y_-^*, y_+^*)$

$$I_2(y_t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\}$$

$$
\begin{aligned}
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G_1(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} -(\nu+1)\frac{y_t - x_t}{\nu + (y-x_t)^2}dy\right\} \\
&= \exp\left\{\frac{-(\nu+1)}{2}\Big[\log|\nu + (y-x_t)^2|\Big]_{\mu(x_t)}^{y_t}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{\nu + (y_t - x_t)^2}{\nu + (x_t - x_t)^2}\right|^{\frac{-(\nu+1)}{2}}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{(y_t - x_t)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \\
&= f(y_t) \qquad (4.5.2)
\end{aligned}
$$

If $y_t \geq y_+^*$

$$
\begin{aligned}
I_3(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_+^*} G_1(y)dy + \int_{y_+^*}^{y_t} G_2(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{\mu(x_t)}^{y_+^*} (\nu+1)\frac{y - x_t}{\nu + (y-x_t)^2}dy - \int_{y_+^*}^{y_t} \frac{c}{y - \mu_t}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\frac{-(\nu+1)}{2}\Big[\log|\nu + (y-x_t)^2|\Big]_{\mu(x_t)}^{y_+^*} - c\Big[\log|y - \mu_t|\Big]_{y_+^*}^{y_t}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{\nu + (y_+^* - x_t)^2}{\nu + (x_t - x_t)^2}\right|^{\frac{-(\nu+1)}{2}} + \log\left|\frac{y_t - \mu_t}{y_+^* - \mu_t}\right|^{-c}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{\left|\frac{\nu + (y_+^* - x_t)^2}{\nu + (x_t - x_t)^2}\right|^{\frac{-(\nu+1)}{2}} \left|\frac{y_t - \mu_t}{y_+^* - \mu_t}\right|^{-c}\right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{(y_+^* - x_t)^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}\right] |y_+^* - \mu_t|^c |y_t - \mu_t|^{-c} \\
&= f(y_+^*)|y_+^* - \mu_t|^c |y_t - \mu_t|^{-c} \\
&= D_{2,t}(x_t)|y_t - \mu_t|^{-c}, \quad \text{where} \quad D_{2,t}(x_t) = f(y_+^*)|y_+^* - \mu_t|^c \qquad (4.5.3)
\end{aligned}
$$

Then,

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} D_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^* \\ f(y_t) & \text{if } y_t \in [y_-^*, y_+^*) \\ D_{2,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^* \end{cases} \tag{4.5.4}$$

4.5.2 Computation of the normalizing constant $B_{1,t}(x_t)$

We know that

$$B_{1,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1}) dy \right]^{-1},$$

therefore,

$$\begin{aligned} B_{1,t}^{-1}(x_t) &= \int_{-\infty}^{y_-^*} I_1(y)dy + \int_{y_-^*}^{y_+^*} I_2(y)dy + \int_{y_+^*}^{+\infty} I_3(y)dy \\ &= D_{1,t}(x_t) \int_{-\infty}^{y_-^*} |y - \mu_t|^{-c} dy + \int_{y_-^*}^{y_+^*} f(y)dy + D_{2,t}(x_t) \int_{y_+^*}^{+\infty} |y - \mu_t|^{-c} dy \\ &= D_{1,t}(x_t) \left[\frac{-1}{(c-1)|y_-^* - \mu_t|^{c-1}} \right] + F(y_+^*) - F(y_-^*) + D_{2,t}(x_t) \left[\frac{1}{(c-1)|y_+^* - \mu_t|^{c-1}} \right] \\ &= \frac{|y_+^* - \mu_t| f(y_+^*) - |y_-^* - \mu_t| f(y_-^*)}{c-1} + F(y_+^*) - F(y_-^*) \end{aligned} \tag{4.5.5}$$

4.5.3 When $c \leq c_2$ and $y_-^* < x_t < z_-^* < z_+^* < \mu_t < y_+^*$ ie. $x_t < \mu_t$ let us compute $\tilde{f}(y_t|x_t, Y_{t-1})$ and denote it by $J_i(y_t)$

If $y_t < y_-^*$

$$\begin{aligned} J_1(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G(y)dy \right\} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ -\int_{y_t}^{y_-^*} G_2(y)dy - \int_{y_-^*}^{\mu(x_t)} G_1(y)dy \right\} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{y_t}^{y_-^*} \frac{c}{y - \mu_t} dy + \int_{y_-^*}^{\mu(x_t)} (\nu+1) \frac{y - x_t}{\nu + (y - x_t)^2} dy \right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ c\Big[\log|y-\mu_t|\Big]_{y_t}^{y_-^*} + \frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{y_-^*}^{\mu(x_t)} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log\left|\frac{y_-^*-\mu_t}{y_t-\mu_t}\right|^c + \log\left|\frac{\nu+(x_t-x_t)^2}{\nu+(y_-^*-x_t)^2}\right|^{\frac{\nu+1}{2}} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{ \left|\frac{y_-^*-\mu_t}{y_t-\mu_t}\right|^c \left|\frac{\nu+(x_t-x_t)^2}{\nu+(y_-^*-x_t)^2}\right|^{\frac{\nu+1}{2}} \right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1+\frac{(y_-^*-x_t)^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}\right] |y_-^*-\mu_t|^c |y_t-\mu_t|^{-c} \\
&= f(y_-^*)|y_-^*-\mu_t|^c|y_t-\mu_t|^{-c} \\
&= C_{1,t}(x_t)|y_t-\mu_t|^{-c}, \quad \text{where} \quad C_{1,t}(x_t) = f(y_-^*)|y_-^*-\mu_t|^c
\end{aligned} \tag{4.5.6}
$$

If $y_t \in [y_-^*, z_-^*)$

$$
\begin{aligned}
J_2(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G_1(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} -(\nu+1)\frac{y_t-x_t}{\nu+(y-x_t)^2}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\frac{-(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{\mu(x_t)}^{y_t}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{\nu+(y_t-x_t)^2}{\nu+(x_t-x_t)^2}\right|^{\frac{-(\nu+1)}{2}}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1+\frac{(y_t-x_t)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)} \\
&= f(y_t)
\end{aligned} \tag{4.5.7}
$$

If $y_t \in [z_-^*, z_+^*)$

$$
\begin{aligned}
J_3(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{z_-^*} G_1(y)dy + \int_{z_-^*}^{y_t} G_2(y)dy\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ -\int_{\mu(x_t)}^{z_-^*} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy + \int_{z_-^*}^{y_t} \frac{c}{y-\mu_t}dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \frac{-(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{\mu(x_t)}^{z_-^*} + c\Big[\log|y-\mu_t|\Big]_{z_-^*}^{y_t} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log\left|\frac{\nu+(z_-^*-x_t)^2}{\nu+(x_t-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} + \log\left|\frac{y_t-\mu_t}{z_-^*-\mu_t}\right|^c \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{ \left|\frac{\nu+(z_-^*-x_t)^2}{\nu+(x_t-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} \left|\frac{y_t-\mu_t}{z_-^*-\mu_t}\right|^c \right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{(z_-^*-x_t)^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}\right]|z_-^*-\mu_t|^{-c}|y_t-\mu_t|^c \\
&= f(z_-^*)|z_-^*-\mu_t|^{-c}|y_t-\mu_t|^c \\
&= C_{2,t}(x_t)|y_t-\mu_t|^c, \quad \text{where} \quad C_{2,t}(x_t) = f(z_-^*)|z_-^*-\mu_t|^{-c} \qquad (4.5.8)
\end{aligned}
$$

If $y_t \in [z_+^*, y_+^*)$

$$
\begin{aligned}
J_4(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G(y)dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{z_-^*} G_1(y)dy + \int_{z_-^*}^{z_+^*} G_2(y)dy + \int_{z_+^*}^{y_t} G_1(y)dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ -\int_{\mu(x_t)}^{z_-^*} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy + \int_{z_-^*}^{z_+^*} \frac{c}{y-\mu_t}dy - \int_{z_+^*}^{y_t} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \frac{-(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{\mu(x_t)}^{z_-^*} \right\} \times \\
&\quad \exp\left\{ +c\Big[\log|y-\mu_t|\Big]_{z_-^*}^{z_+^*} - \frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{z_+^*}^{y_t} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log\left|\frac{\nu+(z_-^*-x_t)^2}{\nu+(x_t-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} + \log\left|\frac{z_+^*-\mu_t}{z_-^*-\mu_t}\right|^c + \log\left|\frac{\nu+(y_t-x_t)^2}{\nu+(z_+^*-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} \right\} \\
&= \left[C_{2,t}(x_t)|z_+^*-\mu_t|^c/f(z_+^*)\right] f(y_t) \\
&= C_{3,t}(x_t)f(y_t), \quad \text{where} \quad C_{3,t}(x_t) = C_{2,t}(x_t)|z_+^*-\mu_t|^c/f(z_+^*) \qquad (4.5.9)
\end{aligned}
$$

If $y_t \geq y_+^*$

$$
\begin{aligned}
J_5(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{z_-^*} G_1(y)dy + \int_{z_-^*}^{z_+^*} G_2(y)dy + \int_{z_+^*}^{y_+^*} G_1(y)dy + \int_{y_+^*}^{y_t} G_2(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{\mu(x_t)}^{z_-^*} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy\right\} \times \\
&\exp\left\{\int_{z_-^*}^{z_+^*} \frac{c}{y-\mu_t}dy - \int_{z_+^*}^{y_+^*} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy - \int_{y_+^*}^{y_t} \frac{c}{y-\mu_t}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\frac{-(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{\mu(x_t)}^{z_-^*} + c\Big[\log|y-\mu_t|\Big]_{z_-^*}^{z_+^*}\right\} \times \\
&\exp\left\{-\frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{z_+^*}^{y_+^*} - c\Big[\log|y-\mu_t|\Big]_{y_+^*}^{y_t}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{\nu+(z_-^*-x_t)^2}{\nu+(x_t-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} + \log\left|\frac{z_+^*-\mu_t}{z_-^*-\mu_t}\right|^c\right\} \times \\
&\exp\left\{\log\left|\frac{\nu+(y_+^*-x_t)^2}{\nu+(z_+^*-x_t)^2}\right|^{\frac{-(\nu+1)}{2}} + \log\left|\frac{y_t-\mu_t}{y_+^*-\mu_t}\right|^{-c}\right\} \\
&= C_{3,t}(x_t)|y_+^*-\mu_t|^c f(y_+^*)|y_t-\mu_t|^{-c} \\
&= C_{4,t}(x_t)|y_t-\mu_t|^{-c}, \quad \text{where} \quad C_{4,t}(x_t) = C_{3,t}(x_t)|y_+^*-\mu_t|^c f(y_+^*)
\end{aligned} \tag{4.5.10}
$$

That is,

$$
\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases}
C_{1,t}(x_t)|y_t-\mu_t|^{-c} & \text{if } y_t < y_-^* \\
f(y_t) & \text{if } y_t \in [y_-^*, z_-^*) \\
C_{2,t}(x_t)|y_t-\mu_t|^{c} & \text{if } y_t \in [z_-^*, z_+^*) \\
C_{3,t}(x_t)f(y_t) & \text{if } y_t \in [z_+^*, y_+^*) \\
C_{4,t}(x_t)|y_t-\mu_t|^{-c} & \text{if } y_t \geq y_+^*
\end{cases} \tag{4.5.11}
$$

4.5.4 Computation of the normalizing constant $B_{2,t}(x_t)$

$$B_{2,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1}) dy \right]^{-1}$$

This implies that,

$$\begin{aligned}
B_{2,t}^{-1}(x_t) &= \int_{-\infty}^{y_-^*} J_1(y)dy + \int_{y_-^*}^{z_-^*} J_2(y)dy + \int_{z_-^*}^{z_+^*} J_3(y)dy + \int_{z_+^*}^{y_+^*} J_4(y)dy + \int_{y_+^*}^{+\infty} J_5(y)dy \\
&= C_{1,t}(x_t) \int_{-\infty}^{y_-^*} |y - \mu_t|^{-c} dy + \int_{y_-^*}^{z_-^*} f(y)dy + \\
&\quad + C_{2,t}(x_t) \int_{z_-^*}^{z_+^*} |y - \mu_t|^{c} dy + C_{3,t}(x_t) \int_{z_+^*}^{y_+^*} f(y)dy + C_{4,t}(x_t) \int_{y_+^*}^{+\infty} |y - \mu_t|^{-c} dy \\
&= \frac{f(y_-^*)|y_-^* - \mu_t|}{c-1} + F(z_-^*) - F(y_-^*) + f(z_-^*) \left[\frac{|z_+^* - \mu_t|}{|z_-^* - \mu_t|} \right]^c \times \\
&\quad \times \left\{ \frac{|z_+^* - \mu_t|}{c+1} + \frac{F(y_+^*) - F(z_+^*)}{f(z_+^*)} + \frac{|y_+^* - \mu_t| f(y_+^*)}{f(z_+^*)(c-1)} \right\} \qquad (4.5.12)
\end{aligned}$$

4.5.5 When $c \leq c_2$ and $y_-^* < \mu_t < z_-^* < z_+^* < x_t < y_+^*$ ie. $x_t > \mu_t$ let us compute $\tilde{f}(y_t|x_t, Y_{t-1})$ and denote it by $K_i(y_t)$

If $y_t < y_-^*$

$$\begin{aligned}
K_1(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G(y)dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ -\int_{y_t}^{y_-^*} G_2(y)dy - \int_{y_-^*}^{z_-^*} G_1(y)dy - \int_{z_-^*}^{z_+^*} G_2(y)dy - \int_{z_+^*}^{\mu(x_t)} G_1(y)dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{y_t}^{y_-^*} \frac{c}{y - \mu_t} dy + \int_{y_-^*}^{z_-^*} (\nu+1) \frac{y - x_t}{\nu + (y - x_t)^2} dy \right\} \times \\
&\quad \exp\left\{ + \int_{z_-^*}^{z_+^*} \frac{c}{y - \mu_t} dy + \int_{z_+^*}^{\mu(x_t)} (\nu+1) \frac{y - x_t}{\nu + (y - x_t)^2} dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log\left| \frac{y_-^* - \mu_t}{y_t - \mu_t} \right|^c + \log\left| \frac{\nu + (z_-^* - x_t)^2}{\nu + (y_-^* - x_t)^2} \right|^{\frac{(\nu+1)}{2}} \right\} \times \\
&\quad \exp\left\{ \log\left| \frac{z_+^* - \mu_t}{z_-^* - \mu_t} \right|^c + \log\left| \frac{\nu + (x_t - x_t)^2}{\nu + (z_+^* - x_t)^2} \right|^{\frac{(\nu+1)}{2}} \right\}
\end{aligned}$$

$$= E_{3,t}(x_t)|y_-^* - \mu_t|^c f(y_-^*)|y_t - \mu_t|^{-c}$$
$$= E_{4,t}(x_t)|y_t - \mu_t|^{-c}, \quad \text{where} \quad E_{4,t}(x_t) = E_{3,t}(x_t)|y_-^* - \mu_t|^c f(y_-^*) \tag{4.5.13}$$

If $y_t \in [y_-^*, z_-^*)$

$$\begin{aligned}
K_2(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{y_t}^{z_-^*} G_1(y)dy - \int_{z_-^*}^{z_+^*} G_2(y)dy - \int_{z_+^*}^{\mu(x_t)} G_1(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{y_t}^{z_-^*} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy + \int_{z_-^*}^{z_+^*} \frac{c}{y-\mu_t}dy + \int_{z_+^*}^{\mu(x_t)} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{y_t}^{z_-^*}\right\} \times \\
&\exp\left\{+c\Big[\log|y-\mu_t|\Big]_{z_-^*}^{z_+^*} + \frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{z_+^*}^{\mu(x_t)}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{\nu+(z_-^*-x_t)^2}{\nu+(y_t-x_t)^2}\right|^{\frac{(\nu+1)}{2}} + \log\left|\frac{z_+^*-\mu_t}{z_-^*-\mu_t}\right|^c + \log\left|\frac{\nu+(x_t-x_t)^2}{\nu+(z_+^*-x_t)^2}\right|^{\frac{(\nu+1)}{2}}\right\} \\
&= \left[E_{2,t}(x_t)|z_-^* - \mu_t|^{-c}/f(z_-^*)\right] f(y_t) \\
&= E_{3,t}(x_t)f(y_t), \quad \text{where} \quad E_{3,t}(x_t) = E_{2,t}(x_t)|z_-^* - \mu_t|^{-c}/f(z_-^*)
\end{aligned} \tag{4.5.14}$$

If $y_t \in [z_-^*, z_+^*)$

$$\begin{aligned}
K_3(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\int_{\mu(x_t)}^{y_t} G(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{y_t}^{z_+^*} G_2(y)dy - \int_{z_+^*}^{\mu(x_t)} G_1(y)dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-\int_{y_t}^{z_+^*} \frac{c}{y-\mu_t}dy + \int_{z_+^*}^{\mu(x_t)} (\nu+1)\frac{y-x_t}{\nu+(y-x_t)^2}dy\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{-c\Big[\log|y-\mu_t|\Big]_{y_t}^{z_+^*} + \frac{(\nu+1)}{2}\Big[\log|\nu+(y-x_t)^2|\Big]_{z_+^*}^{\mu(x_t)}\right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{\log\left|\frac{z_+^*-\mu_t}{y_t-\mu_t}\right|^{-c} \log\left|\frac{\nu+(x_t-x_t)^2}{\nu+(z_+^*-x_t)^2}\right|^{\frac{(\nu+1)}{2}}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{ \left| \frac{z_+^* - \mu_t}{y_t - \mu_t} \right|^{-c} \left| \frac{\nu}{\nu + (z_+^* - x_t)^2} \right|^{\frac{(\nu+1)}{2}} \right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{(z_+^* - x_t)^2}{\nu} \right)^{\frac{-(\nu+1)}{2}} \right] |z_+^* - \mu_t|^{-c} |y_t - \mu_t|^c \\
&= f(z_+^*) |z_+^* - \mu_t|^{-c} |y_t - \mu_t|^c \\
&= E_{2,t}(x_t) |y_t - \mu_t|^c, \quad \text{where} \quad E_{2,t}(x_t) = f(z_+^*) |z_+^* - \mu_t|^{-c}
\end{aligned} \tag{4.5.15}
$$

If $y_t \in [z_+^*, y_+^*)$

$$
\begin{aligned}
K_4(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G(y) dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G_1(y) dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} -(\nu+1) \frac{y_t - x_t}{\nu + (y - x_t)^2} dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \frac{-(\nu+1)}{2} \Big[\log|\nu + (y - x_t)^2| \Big]_{\mu(x_t)}^{y_t} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log \left| \frac{\nu + (y_t - x_t)^2}{\nu + (x_t - x_t)^2} \right|^{\frac{-(\nu+1)}{2}} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{(y_t - x_t)^2}{\nu} \right)^{-\frac{1}{2}(\nu+1)} \\
&= f(y_t)
\end{aligned} \tag{4.5.16}
$$

If $y_t \geq y_+^*$

$$
\begin{aligned}
K_5(y_t) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_t} G(y) dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \int_{\mu(x_t)}^{y_+^*} G_1(y) dy + \int_{y_+^*}^{y_t} G_2(y) dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ - \int_{\mu(x_t)}^{y_+^*} (\nu+1) \frac{y - x_t}{\nu + (y - x_t)^2} dy - \int_{y_+^*}^{y_t} \frac{c}{y - \mu_t} dy \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \frac{-(\nu+1)}{2} \Big[\log|\nu + (y - x_t)^2| \Big]_{\mu(x_t)}^{y_+^*} - c \Big[\log|y - \mu_t| \Big]_{y_+^*}^{y_t} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \exp\left\{ \log\left| \frac{\nu + (y_+^* - x_t)^2}{\nu + (x_t - x_t)^2} \right|^{\frac{-(\nu+1)}{2}} + \log\left| \frac{y_t - \mu_t}{y_+^* - \mu_t} \right|^{-c} \right\} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left\{ \left| \frac{\nu + (y_+^* - x_t)^2}{\nu} \right|^{\frac{-(\nu+1)}{2}} \left| \frac{y_t - \mu_t}{y_+^* - \mu_t} \right|^{-c} \right\} \\
&= \left[\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{(y_+^* - x_t)^2}{\nu} \right)^{\frac{-(\nu+1)}{2}} \right] |y_+^* - \mu_t|^c |y_t - \mu_t|^{-c} \\
&= f(y_+^*)|y_+^* - \mu_t|^c |y_t - \mu_t|^{-c} \\
&= E_{1,t}(x_t)|y_t - \mu_t|^{-c}, \quad \text{where} \quad E_{1,t}(x_t) = f(y_+^*)|y_+^* - \mu_t|^c
\end{aligned} \tag{4.5.17}
$$

4.5.6 Computation of the normalizing constant $B_{3,t}(x_t)$

$$
B_{3,t}(x_t) = \left[\int_{\mathbb{R}} \tilde{f}(y|x_t, Y_{t-1}) dy \right]^{-1}
$$

This implies,

$$
\begin{aligned}
B_{3,t}^{-1}(x_t) &= \int_{-\infty}^{y_-^*} K_1(y)dy + \int_{y_-^*}^{z_-^*} K_2(y)dy + \int_{z_-^*}^{z_+^*} K_3(y)dy + \int_{z_+^*}^{y_+^*} K_4(y)dy + \int_{y_+^*}^{+\infty} K_5(y)dy \\
&= E_{4,t}(x_t) \int_{-\infty}^{y_-^*} |y - \mu_t|^{-c} dy + E_{3,t}(x_t) \int_{y_-^*}^{z_-^*} f(y)dy + \\
&\quad + E_{2,t}(x_t) \int_{z_-^*}^{z_+^*} |y - \mu_t|^c dy + \int_{z_+^*}^{y_+^*} f(y)dy + E_{1,t}(x_t) \int_{y_+^*}^{+\infty} |y - \mu_t|^{-c} dy \\
&= \frac{f(y_+^*)|y_+^* - \mu_t|}{c-1} + F(y_+^*) - F(z_+^*) + \frac{f(z_+^*)}{c+1} \left\{ |z_+^* - \mu_t| - \left[\frac{|z_-^* - \mu_t|}{|z_+^* - \mu_t|} \right]^c |z_-^* - \mu_t| \right\} + \\
&\quad + \frac{f(z_+^*)}{f(z_-^*)(|z_+^* - \mu_t||z_-^* - \mu_t|)^c} \left[F(z_-^*) - F(y_-^*) - \frac{f(y_-^*)|y_-^* - \mu_t|}{c-1} \right]
\end{aligned} \tag{4.5.18}
$$

5. Stochastic volatility models with close skew-normal errors

Markov chain Monte Carlo (MCMC) is a well-known methodology for producing samples from a recognised posterior distribution for hidden variables, where the distribution is very complex, that is, it is not evident how to sample from it. MCMC methods are often used in practise as a rescue, when the computation is not accessible with deterministic methods.These powerful stochastic calculus techniques have many applications in a wide range of area such as physics, chemistry, biology, engineering sciences and economics among others. In this chapter, we used MCMC methods to approximate parameter of discrete time stochastic volatility models with csn error in the observation equation.

5.1 Introduction

The seek to find models that can explain in better way, the dynamics of observed stock prices has been one of the main task for financial mathematicians over the last four decades. The cornerstone of these models was the Black-Scholes-Merton model postulated in early 1970's. This model built its reputation and success around the fact that, option hedging and pricing are easily done from it. As pointed out in (Black, 1975) soon after the publication of what can be consider as the turning point paper for option pricing (Black and Scholes, 1973), this ease however presents some drawbacks such as, the inconsistency of the constant volatility assumption which is not in line with the real financial markets data. This simply because, volatility can be a function of the underlying price level and thus affected by the changes in the price level.

One of the evidence of these weaknesses is probably the occurrence of market crashes that these dynamics sometimes fail to predict. The most memorable in our opinion being the one of October 1987 also known as the Black Monday. The catastrophic impacts and revelations of such

phenomenons, quickly incited many experts in the finance area to explore the possibilities of new statistical models very close to the reality, and in which volatility and co-dependence between variables is allowed to fluctuate over time rather than remaining constant. That is, models that can take into consideration more complex characteristics and stylized facts such as, the excess kurtosis and skewness exhibited by stock returns and the presence of jumps in stock prices to list just few; among these type of models, we have the so called stochastic volatility model (SV).

When it comes to modelling and predicting time varying volatility on financial markets, assessing or managing risk and pricing asset, SV models can be very useful and even considered as essential tools. In financial economics and financial mathematics, the continuous time framework is mostly used to model SV. This because, it captures in the best way the empirical features of asset markets such as derivative pricing. It allows for the computation of internally consistent model implications across all sampling and return horizon, like the very complex dynamic of volatilities and its non observable nature.

However, taking into account the fact that data are discrete time observations, equal consideration can be given to discrete time setting of SV in practise. The basic economic motivations of SV models can be derived from the mixture of distribution hypothesis (MDH) as postulated by (Clark, 1973) and stating that, asset return volatility is driven by its own stochastic process updated by an unobservable innovation. Hypothesis that is not in line with the autoregressive conditional heteroskedasticity (ARCH) setting.

Recently, there has been an increasing interest in modelling the volatility of high frequency financial data using two well known and competing approaches. The generalized autoregressive conditional heteroskedasticity (GARCH) type models (Bollerslev, 1986) and the SV models. Each of these models presenting specific features, as we now recall.

In the GARCH specification, a single error term is assumed, the variance is conditionally deterministic given past observations, the parameters explaining persistence and kurtosis are closely linked, (Carnero et al., 2004). Were as in the SV framework, the presence of two errors processes is assumed, there is an unpredictable component of the conditional variance at time-t. Moreover, parameters explaining the persistence and the kurtosis can be modelled independently and the volatility is modelled as a latent variable. SV models are close to the models often used in finan-

cial theory to represent the behaviour of financial prices, their statistical properties are easy to derive and the empirical irregularities usually observed in financial time series are better captured, (Danielsson, 1994; Kim et al., 1998). This makes the SV model more flexible in fitting data than the popular GARCH.

Although SV models are intuitively prevalent, their empirical application has been limited. Mainly because of difficulties related to their estimation and the intractability of the likelihood function, all these leading to the need of numerical methods that are known to be computationally costly, issues that can efficiently be handled nowadays. Thanks to the availability of more sophisticated simulation methods, recent and considerable increase of computers' power and the advanced development of efficient sampling techniques.

Among the above mentioned simulation methods, we have the simulated maximum likelihood, the method of simulated moments and the Markov chain Monte Carlo (MCMC) methods (Jacquier et al., 2002). The existence of these powerful tools and the need to take into account more general specificities such as, occurrence of jumps and leverage effects, make SV models suitable candidates to study financial return.

The main purpose of this chapter is to revisit discrete-time SV models, by providing some of their most important characteristics and study their implementation via MCMC based on Bayesian statistical inference. The remainder of the chapter develops as follows. The next section briefly reviews the notion of volatility as the measure of risk and presents some key specifications of GARCH type models along with an application. Section 3 reviews the stochastic volatility models, whereas a description of MCMC methods applied on SV is done in section 4. Section 5, considers the SV model with csn distributed error in the observation equation and some concluding remarks appear in section 5.

5.2 Univariate volatility models

5.2.1 Volatility as a measure of risk and stylized facts

Dealing with enormous amounts of data recorded over time is a common issue in financial markets. Prices of various financial products such as stocks are some examples. From their initial value and in a random way, the fluctuations of stock prices are not constant functions of time parameters. Since there is possibility for stock prices to fall below their initial values, volatility can simply be defined as the magnitude of price change. Thus, interpreting volatility as a measure of risk is straightforward.

Volatility which is certainly the most used measured of risk in finance, is usually treated as latent variable and can only be inferred from another observable variable. One can find trivial to say that, lower volatility induces less risk on the asset since there is less price fluctuation. Nevertheless, and as pointed out in (Danielsson, 2011), in the presence of returns with fat tail densities, there is a possibility of overlooking some extreme values of return leading to a misrepresentation of the real data.

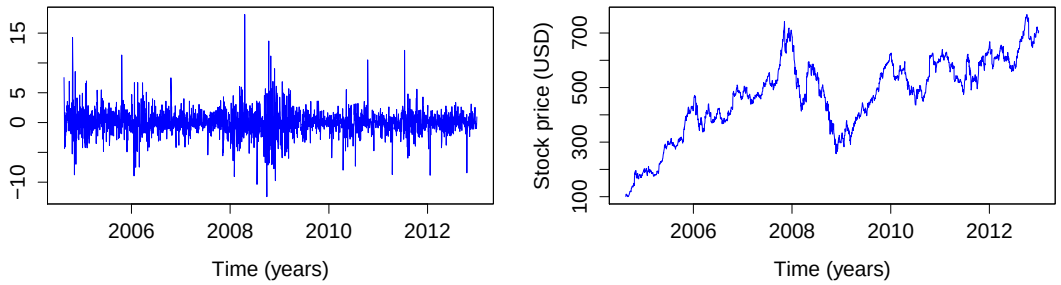
As we will be dealing with financial returns, we now recall some of their important characteristics highlighted in (Danielsson, 2011) and (Cai, 2008).

1. **Non-linearity dependence**: When returns series move in the same way over time, we say that they are linearly dependent. However this is not true with financial returns, as they have the tendency to move independently and the behaviour of each individual return depends on the conditions in the market, this is the reason why financial returns are said to be non-linearly dependent.

2. **Fat tails**: Usually, returns are assumed to be normally distributed, assumption that does not hold with real data, that exhibit fat tails.

3. **Volatility clusters**: This important notion explain the continuous cycle from periods with high volatility to period with lower volatility and can be used to forecast future volatility.

In the following, an analysis of Google stock price returns will be conducted. Figure (5.1) represents the returns and the daily closing prices of Google stock, the period chosen is from 1st January 2000 to 31st December 2012.



(a) Google returns (b) Google closing stock prices

Figure 5.1: Google returns and closing stock prices

Remark 5.2.1. *It is important to notice that, the notion of volatility is usually categorised as either unconditional or conditional. That is, σ the volatility computed over a period T of a returns series y_t with mean μ,*

$$\sigma = \sqrt{\sum_{t=1}^{T}(y_t - \mu)^2} \tag{5.2.1}$$

and the volatility σ_t computed over a specific time period and conditional on past information respectively. In the latter case which will be our concern, for each time t, there is a given σ_t.

In what follows and unless explicitly specified, every time volatility will be defined as σ_t, this will implied conditional volatility.

5.2.2 The family of GARCH(p,q) models

Introduce in (Bollerslev, 1986), the GARCH model is the generalization of the ARCH model of (Engle, 1982). These models have been widely used to study conditional volatility in time series analysis and one of their main goal is to study the statistical properties of returns at time t given

the set of available information up to time $t-1$. The term heteroskedasticity appearing in both ARCH and GARCH, characterises the fact that the variance is not constant.

Let the stochastic processes $y_t = \mu + \sigma_t \varepsilon_t$ characterize the returns. Where, μ is the mean of the returns and the market shocks ε_t are iid normally distributed with mean zero and unit variance. The normality assumption here is just to simplify the computation, in practise other distributions can be used to define these processes. We further assume that the mean is zero, implying that $y_t = \sigma_t \varepsilon_t$.

The estimation of volatility with GARCH model has been extensively done by many practitioners and seems offer realistic and meaningful insights on the data. Indeed, the model is very useful for modelling the conditional variance and for capturing the effect of volatility on stock prices. In its general form, the model is written as GARCH(p, q) and the volatility forecast at time $t+1$ which depends entirely on the set of previously available information is defined by

$$\sigma_{t+1}^2 = w + \sum_{i=1}^{p} \alpha_i y_{t+1-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t+1-j}^2, \tag{5.2.2}$$

where σ_{t+1-j}^2 represents the history estimate of the variance, y_{t+1-i}^2 the square of history return and $w > 0, \alpha_i$ and β_j the model's parameters. The most simple and used GARCH model is the GARCH (1,1) and it is characterised by the following Equations (5.2.3) and (5.2.4).

$$y_t = \sigma_t z_t \tag{5.2.3}$$

$$\sigma_t^2 = w + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{5.2.4}$$

The parameters w, α and β can be used to illustrate the effectiveness of the model to capture the effect of volatility clustering, by considering arbitrary values of these parameters in the calculation. These parameters are plugged into Equation (5.2.4) to estimate the variance σ_t^2 from the returns y_{t-1}^2 and variances σ_{t-1}^2. One of the advantages of using GARCH model is the possibility of making accurate predictions of volatility in the short-term horizon.

In order to ensure that the volatility is strictly positive some constraints are usually imposed on

α_i and β_j and as in (Nelson and Cao, 1992), for the GARCH(1,1) we require that $\alpha \geq 0$ and $\beta \geq 0$, leading to the following unconditional volatility.

$$\begin{aligned}\sigma^2 &= E(\sigma_t^2) \\ &= E(w + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2) \\ &= w + \alpha\sigma^2 + \beta\sigma^2 \\ &= \frac{w}{1-\alpha-\beta}.\end{aligned}$$

$\alpha+\beta$ measure the persistence of volatility with respect to the time and one must have $\alpha+\beta < 1$, to insure the mean reverting to the variance in the long term and the covariance stationarity. As $\alpha + \beta$ converges to 1, more the volatility is persistent over time. On the contrary, when $\alpha + \beta$ tends to 0, more fast is the converge of volatility to the variance over long time horizon. Note however that, in the conditional variance case, one can get rid of these restrictions.

5.2.3 Application to Google stock returns

In what follows, we will use the historical returns of Google stocks index to estimate the parameters of the GARCH(1,1) via quasi-maximum likelihood estimation (QLM) approach. We assume that these returns follow a normal distribution with zero mean and a variance σ_t^2. Since the normal assumption on the distribution of returns is not strict, the method can be applied on non-normal data as well. Furthermore, as we are mainly interested on the volatility, if the zero mean condition is not satisfied by the returns, then the unconditional mean will be subtracted from each observation in order to guaranty that $\mu \simeq 0$.

Now, we recall that if a random variable $Y \sim \mathcal{N}(\mu, \sigma)$, then its density function is defined by,

$$f(y, \mu; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\}. \tag{5.2.5}$$

Provided that iid-observations of size T are available, we obtain the following conditional likelihood

function

$$\begin{aligned}\mathcal{L}(\mu, \sigma_t; y) &= \prod_{t=1}^{T} f(y_t, \mu; \sigma) \\ &= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{1}{2}\frac{(y_t-\mu)^2}{\sigma_t^2}\right\},\end{aligned} \tag{5.2.6}$$

which leads to the following log-likelihood

$$\ln(\mathcal{L}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left\{\log(\sigma_t^2) + \frac{(y_t-\mu)^2}{\sigma_t^2}\right\}. \tag{5.2.7}$$

We then obtain the log-likelihood function for the GARCH(1,1) model, by plugging Equation (5.2.4) into Equation (5.2.7). Yields,

$$\ln(\mathcal{L}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{T}\left\{\log(w + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2) + \frac{y_t^2}{w + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2}\right\} \tag{5.2.8}$$

We will use R software and the GARCH(1,1) model will be fitted using Google's daily adjusted closing stock prices from 1st January 2000 to 31st December 2012, downloaded from <http://finance.yahoo.com>. The QQ-plots help us to determine the type of distribution followed by the returns. Then the Ljung-Box test for autocorrelation and the Jarque-Bera test for normality are conducted. The Ljung-Box test summarised in Table (5.1) suggests that the returns are not autocorrelated. Also, as suggested by the Jarque-Bera test in Table (5.1), the distribution of returns is not normal.

Finally, the implementation of the model is done under the hypothesis that returns follow a conditional fat tail distribution as suggested by the above performed tests. We chose the Student-t density with one degree of freedom and the fitted equation is given by

$$\sigma_t^2 = 0.03154853 + 0.04158007 y_{t-1}^2 + 0.95384487\sigma_{t-1}^2 \tag{5.2.9}$$

	Ljung-Box test	Jarque-Bera test
X-squared	0.4562	3871.821
df	2	2
p-value	0.8046	< 2.2e-16

Table 5.1: Ljung-Box and Jarque-Bera tests.

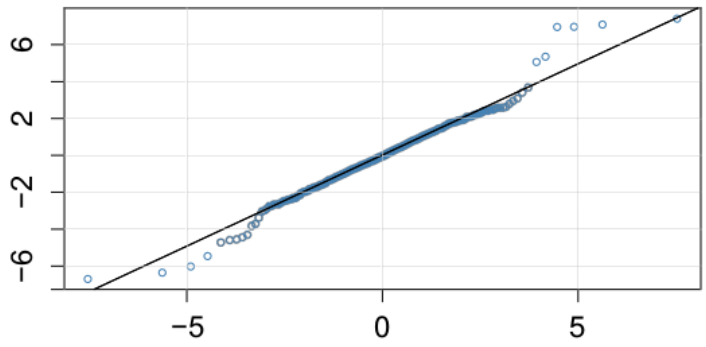


Figure 5.2: Normal QQ-plot of Google returns and density plots

5.3 Stochastic volatility models

First introduced in (Taylor, 1982), this model appears to be more flexible than the ARCH-type, as it takes into account the randomness caused by the observations and the latent volatilities. The standard SV model in the theoretical finance literature, is usually characterized in term of stochastic differential equation (SDE) given by,

$$\begin{cases} dS(t) = \sigma(t).dW_1(t) \\ d\log\sigma^2(t) = \alpha + \beta\log\sigma^2(t)dt + \eta dW_2(t). \end{cases} \tag{5.3.1}$$

In this setting, $S(t)$ is the logarithm of asset price, $\sigma^2(t)$ the volatility, $W_1(t)$ and $W_2(t)$ two Brownian motions. In empirical studies, the above continuous formulation is discretised using the Euler-Maruyama approximation. That is, $S(t+1) - S(t) = Y(t)$, $W_1(t+1) - W_1(t) = u_t$, $W_2(t+1) - W_2(t) = \nu_t$, $1 + \beta = \phi$, $\log\sigma^2(t) = h_t$ and, $\alpha(1+\phi) = \mu$.

We then obtain the so called standard SV model which is defined by Equation (5.3.2) below.

$$\begin{cases} y_t = \sigma(t)\mu_t = \exp(\frac{h_t}{2}) \\ h_{t+1} = \mu + \phi(h_t - \mu) + \eta\nu_t. \end{cases} \tag{5.3.2}$$

Where y_t represents the log-return (observations) at time t, ν is the standard deviation of the log-volatility h_t which by assumption is driven by an AR(1) process, with persistence parameter $\phi < 1$. $\nu_t \sim \text{iid}N(0, \sigma_\eta^2)$ and $u_t \sim \text{iid}N(0, 1)$ are respectively, the volatility and the return shocks. The independence condition of the two above mentioned Brownian motions will be equivalent to the zero correlation of (u_t, ν_t) for $t = 1, \cdots, T$.

Remark 5.3.1. *The condition on the persistence parameter $|\phi| < 1$ in the SV models plays an important role, in the sense that it ensures the stationary of the log-volatility process.*

I turns out as in (Hautsch and Ou, 2008) that, the unconditional distribution of h_t and the kurtosis $K(y_t)$ are respectively given by

$$h_t \sim \mathcal{N}(\mu, \sigma_h^2) \quad \text{with} \quad \sigma_h^2 = \frac{\sigma_\eta^2}{1 - \phi} \tag{5.3.3}$$

and

$$K(y_t) = \frac{E(y_t^4)}{E(y_t^2)^2} = 3\exp(\sigma_h^2) \tag{5.3.4}$$

From Equation (5.3.4) above, one can see that if $\sigma_h^2 > 0$, then $K(y_t) > 3$ provided that the condition on the persistence parameter ϕ is satisfied. The problem usually encounter with this type of model, is the parameters estimation, since a direct computation of the likelihood function is not possible, thus the need of other methods than the often used likelihood is required.

5.4 MCMC methods applied on SV models

One of the most important specificity of SV models is the non correlation of the errors components in the log-volatility and the mean equations. Due to the hard evaluation of the likelihood function

which relies essentially on the computation of a high dimensional integral, this represents a considerable issue when it comes to the estimation. This likelihood intractability can be explain by the fact that the process characterising the variance is unobservable making the likelihood function available only in the form of a very complicated multidimensional integral.

These drawbacks rise questions on the efficiency and reliability of the QML and the method of moments (MM); respectively because, the hight dimensional integral approximation is difficult when producing QML estimators and there is no way of knowing which moment we should use for the MM estimation, since the score function is not available. One possible way to handle these issues is the use of Bayesian inference techniques based on MCMC simulation algorithms introduced in (Jacquier et al., 1994). In what follows, a quick overview of these methods will be done along with a description of their implementation on SV.

5.4.1 Bayesian inference and MCMC algorithm for SV models

In statistical inference, two main approaches are often used. The maximum likelihood method as seen in the parameters estimation of the GARCH model in the previous section, and the Bayesian inference. In the latter, data are combined with prior belief in order to produce posterior distributions that will be used for the inference.

Monte Carlo methods are used for analysing Bayesian distributions in high dimension. They are helpful for either generating samples like $\theta_1, \theta_2, \cdots$ from a given probability distribution $p(\theta)$, or estimating expectations of a function under the same probability distribution, or both (MacKay, 2003). They are generally preferred when the number of parameters to be estimated is large, as well as when we are dealing with high dimensional problems. When combined with Markov chain, the aim become to construct a Markov chain whose equilibrium distribution is the target distribution of interest(MacKay, 2003).

The probability distribution $p(\theta)$ also called the target, is complex enough by assumption. That is, computing expectations from it by deterministic methods is not feasible, thus, require Monte Carlo methods. The accuracy of the Monte Carlo estimate depends only on the variance and not on the dimensionality of the space sampled (MacKay, 2003; Geweke, 1991). Since there is a

possibility for the normalizing constant to be unknown or known but in high dimensional space, sampling from $p(\theta)$ can be cumbersome. This because, there is no trivial techniques that can help to sample from $p(\theta)$ without simultaneously listing most of the possible states and obtaining accurate estimates.

In Monte Carlo integration, the common difficulty encountered is how to draw samples from some complex probability distributions. With early contributions of (Metropolis and Ulam, 1949), (Hastings, 1970) and (Metropolis et al., 1953), solving such a problem has always been one of the main task of MCMC methods.

The general purpose of MCMC settings is to build a transition kernel of an ergodic[1] Markov chain with the desired invariant distribution, and simulate the chain for many steps, so that it reaches the equilibrium. The states sampled after the convergence of the chain will then have the same distribution as the distribution of interest (target).

To estimate standard SV model and its many extensions, various MCMC methods such as the Gibbs sampler and the family of Metropolis have been extensively used in the literature as well as in practise, with key contributions of (Geman and Geman, 1984),(Gelfand and Smith, 1990) and (Jacquier et al., 2002) among others.

5.4.2 Gibbs sampler for SV models

The aim when implementing the Gibbs sampler, is to make an approximation of the parameters' posterior distributions such that, inference can be made using the fitted model by assuming that the conditional distributions of one parameter given others is available. That is, $f_i(\theta_i|\theta_{j\neq i}, y, h)$ for $i, j = 1, 2, \cdots K$ are assumed known. Now, if we assume that it is possible to make draws from each conditional distribution and set $K = 3$, then the following steps characterise the Gibbs sampler.

1- Set initial values $\theta_{2,0}$ and $\theta_{3,0}$ for θ_2 and θ_3 respectively.

[1]Sufficient condition for the existence of the stationary distribution $\pi(x)$ independent of the initial probability at the starting state that is, a Markov chain satisfy the stationarity, irreducibility, and aperiodicity conditions

2- Complete the Gibbs iteration.

(i) Draw randomly $\theta_{1,1}$ from $f_1(\theta_1|\theta_{2,0}, \theta_{3,0}, y, h)$

(ii) Draw randomly $\theta_{2,1}$ from $f_2(\theta_2|\theta_{1,1}, \theta_{3,0}, y, h)$

(iii) Draw randomly $\theta_{3,1}$ from $f_3(\theta_{1,1}|\theta_{2,1}, \theta_{2,1}, y, h)$

3- Use $\theta_{1,1}, \theta_{2,1}, \theta_{3,1}$ as the new starting values and return to step 2 to complete another Gibbs iteration that will produce new values $\theta_{1,2}, \theta_{2,2}$ and $\theta_{3,2}$. This is usually called the update step .

4- Repeat this several times, say n and obtain a sequence of n samples, $(\theta_{1,1}, \theta_{2,1}, \theta_{3,1}), \cdots (\theta_{1,n}, \theta_{2,n}, \theta_{3,n})$.

For large n and under some sufficient conditions, it is possible to show that, $(\theta_{1,n}, \theta_{2,n}, \theta_{3,n})$ converges to a random draw from the joint distribution $f(\theta_1, \theta_2, \theta_3|y, h)$. This realisation can then be used for inference as it constitute a random sample from the joint posterior distribution (Robert and Casella, 1999; Tsay, 2005). In practise and in order to achieve efficiency, one can chose n large enough and get rid of the first m draws (burn-in sample).

5.4.3 The Metropolis Hasting (MH) algorithm for SV model

The Metropolis-Hastings algorithm (MH) is often used when dealing with high-dimensional problems. This method consists of defining the transition probability from a state θ_i to a state θ_{i+1}. Thus, it is a Markov process in which a sequence of $\theta_1, \theta_2, \cdots$ is generated. The probability distribution of each trial θ_{i+1} appears to depend only on the distribution of θ_i. Since we are looking for a sample which is independent from the initial condition, it is very important to run the chain for a sufficient long time horizon to avoid the dependence of successive samples.

The MH algorithm can be defined as an adaptation of a random walk based on the **acceptance rejection rule** to converge to the target distribution. One of the requirement of this algorithm is that, the proposal distribution has to be given in a very specific way, it is then recommended to use prior knowledge to achieve this. The algorithm is efficient when the conditional posterior distribution is available at least, up to a normalisation constant (Metropolis and Ulam, 1949).

To see how the MH algorithm works, let us assume that the variable that we want to sample is θ and that the target distribution $f(\theta|y)$ contains a non easily tractable normalisation constant, making direct sampling impossible or time consuming. Furthermore, we suppose that we have a good approximation (proposal) of $f(\theta|y)$ from which we can generate random draws in a simple way. The sequence of random draws from this approximated distribution will then converge to $f(\theta|y)$ by definition.

Now let the current state of the Markov chain be θ^t, the potential candidate is θ' and the proposal distribution is $q(\theta'|\theta^t)$ which depends only on the current state θ^t. The next step is to generate a candidate from the proposal to compute $f(\theta^t)$ and $f(\theta')$. Then, accept the candidate with probability $\min\{1, \alpha\}$. That is, the new state after the update step is θ' with probability α or θ^t with probability $1-\alpha$ (Geyer, 2011). Where α knowing as the **Hasting ratio** is given by

$$\alpha = \frac{f(\theta'|y)q(\theta'|\theta^t)}{f(\theta^t|y)q(\theta^t|\theta')} \tag{5.4.1}$$

Finally repeat the process several times, until the convergence of the generated Markov chain is achieved. The following summarizes the MH algorithm for the SV models.

1) Initialize θ^0 such that $f(\theta^0|y) > 0$

2) For $t = 1, 2\cdots$, set $t \leftarrow 1$

3) Draw a candidate θ' from $q(\theta^t|\theta^{t-1})$

4) Compute the quantity

$$\alpha = \min\left\{1, \frac{f(\theta'|y)}{f(\theta^{t-1}|y)}\frac{q(\theta^{t-1}|\theta')}{q(\theta'|\theta^{t-1})}\right\}$$

5) Accept θ' with probability α, that is,

$$\theta^t = \begin{cases} \theta', & \text{with probability } \alpha. \\ \theta^{t-1}, & \text{otherwise.} \end{cases} \tag{5.4.2}$$

6) Set $t = t + 1$ and return to step 3.

Note that, when the proposed distribution $q(\theta', \theta)$ is symmetric, that is, $q(\theta'|\theta) = q(\theta|\theta')$, we talk about the Metropolis algorithm and one just have to set the formula of α in the step 4 of the MH algorithm by

$$\alpha = \min\left\{1, \frac{f(\theta'|y)}{f(\theta^{t-1}|y)}\right\}$$

and given a "good" burn-in period, the chain or sequence θ^t should move towards the distribution $f(\theta|y)$.

5.5 Close skew-normal SV model

We now discuss an extension of the SV in which the error terms in the observation equation follow a csn distribution. Specifically, the model we call SV-csn is defined by equations (5.5.1) and (5.5.2) below.

$$y_t = \exp(h_t/2)\varepsilon_t, \quad \varepsilon_t \sim \text{csn}_{1,1}(\mu_\varepsilon, \sigma_\varepsilon, \gamma_\varepsilon, \nu_\varepsilon, \delta_\varepsilon) \tag{5.5.1}$$

$$h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \tag{5.5.2}$$

Before proceeding, let us recall the definition of the moment generating function (mgf) for the csn random variable.

Definition 5.5.1 (Multivariate Close Skew Normal mgf). *A random vector X has a multivariate* $csn_{m,m}(\mu, \Sigma, \Gamma, \nu, \Delta)$ *distribution according to (González-Farías et al., 2004a), if its mgf is given by*

$$M_x(t) = \frac{\Phi_m(\Gamma\Sigma t; \nu, \Delta + \Gamma\Sigma\Gamma^T)}{\Phi_m(0; \nu, \Delta + \Gamma\Sigma\Gamma^T)} \exp(t^T\mu + \frac{1}{2}t^T\Sigma t), \quad t \in \mathbb{R}^n, \tag{5.5.3}$$

where $\Phi_m(\cdot;\cdot,\cdot)$ *represents the Gaussian cdf and* V^T *the transpose of the vector* V.

In high dimensional problem, the multivariate Normal cdf is not tractable analytically. Thus, we restrain ourself to the univariate processes $X \sim \text{csn}_{1,1}(\mu, \sigma, \gamma, \nu, \delta)$. If we further assume that $\delta + \gamma^2\sigma \neq 0$, then from Definition (5.5.1) one can easily derive the following.

$$\mathbb{E}(X) = \mu + \sqrt{\frac{2}{\pi}} \frac{\gamma\sigma}{\sqrt{\delta + \gamma^2\sigma}} \tag{5.5.4}$$

$$\text{Var}(X) = \sigma - \frac{2}{\pi} \frac{\gamma^2\sigma^2}{\delta + \gamma^2\sigma} \tag{5.5.5}$$

and

$$\mathbb{E}(X - \mathbb{E}(X))^3 = \left(2 - \frac{\pi}{2}\right) \left(\sqrt{\frac{2}{\pi}}\right)^3 \left(\frac{\gamma\sigma}{\sqrt{\delta + \gamma^2\sigma}}\right)^3 \tag{5.5.6}$$

Now, in the SV-csn model, we assume that $\mathbb{E}(\varepsilon_t) = 0$ and $\text{Var}(\varepsilon_t) = 1$ which from Equation (5.5.4) and Equation (5.5.5) imply respectively

$$\mu_\varepsilon = -\sqrt{\frac{2}{\pi}} \frac{\gamma_\varepsilon\sigma_\varepsilon}{\sqrt{\delta_\varepsilon + \gamma_\varepsilon^2\sigma_\varepsilon}}, \tag{5.5.7}$$

and

$$\sigma_\varepsilon = 1 + \mu_\varepsilon^2. \tag{5.5.8}$$

We conclude the specification of the proposed SV-csn by assuming that ε_t is independent of η at all leads and lags. We also consider an AR(1) structure for h_t with persistence parameter $|\phi| < 1$ and we initialize the log-volatility with

$$h_1 \sim N\left(\mu, \frac{\sigma_\eta^2}{1 - \phi^2}\right). \tag{5.5.9}$$

With these restriction imposed on the parameters of ε_t, assuming that all prior are independent and using the Bayes rule, we obtain the following full conditional posteriors.

$$\begin{cases} p(\sigma_\eta^2|y,h,\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\sigma_\eta^2) \\ p(\phi|y,h,\mu,\sigma_\eta^2,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\phi) \\ p(\mu|y,h,\sigma_\eta^2,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\mu) \\ p(\mu_\varepsilon|y,h,\sigma_\eta^2,\phi,\mu,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\mu_\varepsilon) \\ p(\nu_\varepsilon|y,h,\sigma_\eta^2,\phi,\mu,\mu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\nu_\varepsilon) \\ p(\delta_\varepsilon|y,h,\sigma_\eta^2,\phi,\mu,\mu_\varepsilon,\nu_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\delta_\varepsilon) \\ p(\sigma_\varepsilon|y,h,\delta_\eta^2,\phi,\mu,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\gamma_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\sigma_\varepsilon) \\ p(\gamma_\varepsilon|y,h,\delta_\eta^2,\phi,\mu,\mu_\varepsilon,\nu_\varepsilon,\sigma_\varepsilon,\delta_\varepsilon) \propto p(h|\mu,\phi,\mu_\varepsilon,\nu_\varepsilon,\delta_\varepsilon,\sigma_\varepsilon,\gamma_\varepsilon)p(\gamma_\varepsilon). \end{cases} \tag{5.5.10}$$

Finally, we assume that

$$\begin{cases} p(\sigma_\eta^2 \sim \mathcal{IG}(\alpha_1,\beta_1) \\ p(\phi) \sim N(\alpha_2,\beta_2)\mathbb{I}_{(-1,1)}(\phi) \\ p(\mu) \sim N(\alpha_3,\beta_3) \\ p(\mu_\varepsilon) \sim N(\alpha_4,\beta_4) \\ p(\nu_\varepsilon) \sim N(\alpha_5,\beta_5) \\ p(\delta_\varepsilon) \sim \mathcal{IG}(\alpha_6,\beta_6) \\ p(\sigma_\varepsilon) \sim \mathcal{IG}(\alpha_7,\beta_7) \\ p(\gamma_\varepsilon) \sim N(\alpha_8,\beta_8), \end{cases} \tag{5.5.11}$$

where $\mathcal{IG}(\cdot,\cdot)$ is the inverse Gamma distribution.

Having all this, one can then proceed with the estimation. As far as this chapter is concerned, we do not investigate this further and will be considered deeply in a near future.

5.6 Concluding remarks

In this chapter, we presented some important characteristics of GARCH type models and estimated the parameters with Google stock returns. We have briefly reviewed simulation methods based on MCMC algorithms such as Gibbs sampling and Metropolis-Hastings algorithm apply on SV models. We then provided foundation for the construction of the so called SV-csn which is an asymmetric SV allowing for csn errors in the observation equation. We defined all the full conditional posteriors. In a near future, we intend to deeply study the statistical properties of this model, estimate the parameters using Bayesian inference based on MCMC techniques and provide some comparison with existing models.

6. Summary

This thesis has presented various methodologies that can be used to achieve accurate and efficient inference in linear, Gaussian and non-Gaussian state-space models, with results that can easily be adapted to the nonlinear cases.

The contribution of this dissertation was threefold. First, we revisited some recent attempts to incorporate skewness in state-space models and especially the skewed Kalman filter by (Naveau et al., 2005). We proved that their statement regarding the close skew-normal nature of the filtering density was not correct. Depending on the assumptions made, other contributions faced challenges such as the explosion of skewness dimension and the skewness vanishing as the recursion proceeds over many time steps. As a remedy, we proposed a new state-space model that overcomes these limitations by relaxing the assumption of normality and exploiting the close skew-normal distribution which is more flexible and extends the Gaussian distribution. This has been achieved by allowing a stationary autoregressive structure in the state equation, and a csn distributed measurement error. This structure allowed us to develop a skewed version of the well known Kalman filter and provided new procedures and algorithms for prediction, filtering and estimation that can be employed to analyse multivariate time series data where the symmetry assumption can not be legitimately made. Moreover, with the proposed methodology we obtained a $\text{csn}_{1,n}$ filtering density which is a big improvement compared to the existing models in the literature where these densities were $\text{csn}_{n,2n}$ or $\text{csn}_{n,3n}$.

Second, we adapted the robust filtering methodology of (Calvet et al., 2015) to build a robust filter with Student-t observation density that provides accurate state inference accounting for outliers and misspecification, this for both finite and infinite state-space models. With simulations, we were able to compare the performance of the proposed robust Student-t filter with the Gaussian filter, the robust Gaussian filter of (Calvet et al., 2015) and the Student-t filter. Furthermore, we applied the theoretical findings on the unobserved component model with stochastic volatility of (Stock and Watson, 2007) and showed that, the Student-t filter performs well which is not surprising given that the Student-t density is naturally robust. Regarding the robust version, one can see that the price to pay when robustifying the filtering density can be hight in the sense of

MSE especially when no outliers have occurred.

Third, we laid the foundations for the construction of stochastic volatility models with csn errors in the observation equation. Even though we did not explore this question in much details, we believe that this approach should not be overlooked. The estimation of this model and the study of its statistical properties need to be done in details, as this can be a good alternative for SV models accounting for skewness.

However, for real life applications, as well as for interest in the state-space models, many other features need to be investigated. For example the curse of dimensionality in the filtering density for the skewed Kalman filter and the robustification of such model. For further research, it might be interesting to look at some techniques of dimension reduction such as the projection pursuit. This can be a way out to construct the csn variant of the well known Gaussian AR, MA and ARMA models. For the estimation part of the Kalman filter with asymmetric distributions, it can be interesting to look at the weighted method of moment of (Flecher et al., 2009) which can lead to possibly more accurate estimate when combined with our methodology. Finally, in the robustification setting, it can be interesting to check the accuracy of the robust filtering density for naturally robust densities and investigate the use of other robust function.

References

Agamennoni, G., Nieto, J. I., and Nebot, E. M. (2011). An outlier-robust kalman filter. In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1551–1558.

Alspach, D. and Sorenson, H. (1972). Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448.

Aoki, M. (2013). *State Space Modeling of Time Series*. Universitext. Springer, Berlin, Heidelberg, New York.

Arulampalam, M. S., Maskell, S., Gordon, N. J., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. Signal Processing*, 50(2):174–188.

Aruoba, S. B., Diebold, F., Nalewaik, J., Schorfheide, F., and Song, D. (2011). Improving gdp measurement: a forecast combination perspective. Working Papers 11-41, Federal Reserve Bank of Philadelphia.

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Black, F. (1975). Fact and fantasy in the use of options. *Financial Analysts Journal*, 31(4):36–72.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, 11(2):143–172.

Cabral, C. R. B., Da-Silva, C. Q., and Migon, H. S. (2014). A dynamic linear model with extended skew-normal for the initial distribution of the state parameter. *Computational Statistics & Data Analysis*, 74:64–80.

Cai, Z. (2008). Financial econometrics: Analysis of financial market data. Technical report, ECON6219.

Calvet, L. E., Czellar, V., and Ronchetti, E. (2015). Robust filtering. *Journal of the American Statistical Association*.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Carnero, M. A., Peña, D., and Ruiz, E. (2004). Persistence and kurtosis in garch and stochastic volatility models. *Journal of Financial Econometrics*, 2:319–342.

Chib, S., Omori, Y., and Asai, M. (2007). Multivariate stochastic volatility. Cirje f-series, CIRJE, Faculty of Economics, University of Tokyo.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155.

Clark, P. K. (1987). The cyclical component of us economic activity. *The Quarterly Journal of Economics*, 102(4):797–814.

Cox, D. and Wermuth, N. (1991). A simple approximation for bivariate and trivariate normal integrals. *International Statistical Review*, 59(2):263–269.

Dagpunar, J. S. (2007). *Simulation and Monte Carlo: With Applications in Finance and MCMC*. Wiley Series in Probability and Statistics. John Wiley & Sons, second edition.

Danielsson, J. (1994). Stochastic volatility in asset prices: estimation with simulated maximum likelihood. *Journal of Econometrics*, 64:375–400.

Danielsson, J. (2011). *Financial risk forecasting: the theory and practice of forecasting market risk with implementation in R and Matlab*. Wiley-Blackwell. © 2011 Wiley-Blackwell.

Diderrich, G. T. (1985). The kalman filter from the perspective of goldberger-theil estimators. *The American Statistician*, 39(3):193–198.

Djuric, P. M. and Bugallo, M. F. (2009). Estimation of stochastic rate constants and tracking of species in biochemical networks with second-order reactions. In *17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, Scotland, UK, August 24-28, 2009*, pages 2308–2311.

Domínguez-Molina, J., González-Farías, G., and Gupta, A. K. (2003). General multivariate closed skew normal distribution. Technical Report 01, Department of Mathematics and Statistics.

Donnelly, T. G. (1973). Algorithm 462: Bivariate normal distribution. *Communications of the Association for Computing Machinery*, 16(10):638.

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Doucet, A., Freitas, N., and Gordon, N. J. (2001). An introduction to sequential monte carlo methods. In Doucet, A., Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, pages 3–14. Springer New York.

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

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: fifteen years later. *The Oxford Handbook of Nonlinear Filtering*, pages 4–6.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89:603–616.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Number 38 in Oxford statistical science series. Oxford University Press.

Elliott, R. J. and Kopp, P. E. (2006). *Mathematics of Financial Markets*. Springer Finance. Springer.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Evans, L. C. (2012). *An Introduction to Stochastic Differential Equations*. American Mathematical Society.

Flecher, C., Naveau, P., and Allard, D. (2009). Estimating the closed skew-normal distribution parameters using weighted moments. *Statistics & Probability Letters*, 79(19):1977 – 1984.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 350–363.

Franke, J. and Christian M. Hafner, W. K. H. (2011). *Statistics of financial markets: An introduction*. Springer, 3rd edition.

Gamerman, D. and Migon, H. S. (1993). Dynamic hierarchical models. *Journal of the Royal Statistical Society. Series B, Methodological*, 55(3):629–642.

Gandhi, M. A. and Mili, L. (2010). Robust kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing*, 58(5):2509–2520.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Brooks, S., G. J., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149.

Geweke, J. F. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report 148, Federal Reserve Bank of Minneapolis.

Geyer, C. J. (2011). *Handbook of Markov Chain Monte Carlo*, pages 3–48. Boca Raton, FL: CRC Press.

Glasserman, P. (2010). *Monte Carlo Methods in Financial Engineering*. Applications of Mathematics. Springer.

González-Farías, G., Domínguez-Molina, J., and Gupta, A. K. (2004a). Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference*, 126(2):521–534.

González-Farías, G., Domínguez-Molina, J., and Gupta, A. K. (2004b). The closed skew-normal distribution. In Genton, M., editor, *Skew-elliptical distributions and their applications: A journey beyond normality*, chapter 2, pages 25–42. CRC Chapman and Hall, London, chapman and hall/crc edition.

Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press.

Gupta, A. K., González-Farías, G., and Domínguez-Molina, J. (2004a). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, 89(1):181 – 190.

Gupta, A. K., González-Farías, G., and Domínguez-Molina, J. (2004b). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, 89(1):181 – 190.

Hafner, C. M. and Manner, H. (2012). Dynamic stochastic copula models: estimation, inference and applications. *Journal of Applied Econometrics*, 27(2):269–295.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hammersley, J. M. and Morton, K. W. (1954). Poor man's monte carlo. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(1):23–38.

Härdle, W., Hlávka, Z., and Klinke, S. (2000). *XploRe® - Application Guide: Application Guide*. Springer Berlin Heidelberg.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):205–247.

Harvey, A. C. (1981). Finite sample prediction and overdifferencing. *Journal of Time Series Analysis*, 2(4):221–232.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:99–109.

Hautsch, N. and Ou, Y. (2008). Discrete-time stochastic volatility models and mcmc-based statistical inference. Technical Report 2008, 063, SFB 649 discussion paper.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343.

Ito, K. and Xiong, K. (2000). Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 12:pp. 371–389.

Jacquier, E., Polson, N. G., and Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 20:69–87.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*, volume 64. Academic Press.

Johannes, M. S., Polson, N. G., and Stroud, J. R. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies*, 22(7):2759–2799.

Julier, S. J. and Uhlmann, J. K. (1997). New extension of the kalman filter to nonlinear systems. In *AeroSense'97*, pages 182–193. International Society for Optics and Photonics.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45.

Kalman, R. E. (1963). Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192.

Karimi, O., Omre, H., and Mohammadzadeh, M. (2010). Bayesian closed-skew gaussian inversion of seismic avo data for elastic material properties. *Geophysics*, 75(1):R1–R11.

Kim, H.-M., Ryu, D., Mallick, B. K., and Genton, M. G. (2014a). Mixtures of skewed kalman filters. *Journal of Multivariate Analysis*, 123(0):228 – 251.

Kim, H.-M., Ryu, D., Mallick, B. K., and Genton, M. G. (2014b). Mixtures of skewed kalman filters. *Journal of Multivariate Analysis*, 123(0):228 – 251.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *Review of Economic Studies*, 65:361–393.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288.

Lapeyre, B., Pardoux, E., and SentisSentis, R. (2003). *Introduction to Monte Carlo Methods for Transport and Diffusion Equations*. Oxford University Press.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Martinez, W. L. and Martinez, A. R. (2001). *Computational Statistics Handbook with MATLAB*. CRC Press.

Meinhold, R. J. and Singpurwalla, N. D. (1983). Understanding the kalman filter. *The American Statistician*, 37(2):123–127.

Meinhold, R. J. and Singpurwalla, N. D. (1989). Robustification of kalman filter models. *Journal of the American Statistical Association*, 84(406):479–486.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341.

Migon, H. S., Gamerman, D., Lopes, H. F., and Ferreira, M. A. (2005). Dynamic models. *Handbook of Statistics, Elsevier*, 25:553–588.

Müller, P. (1991). Monte carlo integration in general dynamic models. *Contemporary Mathematics*, 115:145–163.

Naveau, P., Genton, M. G., and Shen, X. (2005). A skewed kalman filter. *Journal of Multivariate Analysis*, 94(2):382 – 400.

Nelson, D. B. and Cao, C. Q. (1992). Inequality constraints in the univariate garch model. *Journal of Business & Economic Statistics*, 10(2):229–235.

Øksendal, B. (2010). *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer New York.

Pole, A., West, M., and Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis*. CRC Press.

Pollock, M. (2010). Algorithms & computationally intensive inference reading group "introduction to particle filtering". Discussion- Notes.

Ramaprasad, B. (2010). *Stochastic Filtering With Applications In Finance:*. Number 7736 in World Scientific Books. World Scientific Publishing Co. Pte. Ltd.

Rees, D., Lancaster, D., and Finlay, R. (2014). A state-space approach to australian gdp measurement. Rba research discussion papers, Reserve Bank of Australia.

Rezaie, J., Eidsvik, J., and Mukerji, T. (2014). Value of information analysis and bayesian inversion for closed skew-normal distributions: Applications to seismic amplitude variation with offset data. *GEOPHYSICS*, 79(4):R151–R163.

Ristic, B., Arulampalam, S., and Gordon, N. (2004). *Beyond the Kalman filter : particle filters for tracking applications*. Artech House.

Robert, C. and Casella, G. (2010). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, Second Edition.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*, volume 58. Springer New York.

Roussas, G. G. (2003). *An Introduction to Probability and Statistical Inference*. Elsevier Science.

Rubinstein, R. Y. and Kroese, D. P. (2011). *Simulation and the Monte Carlo Method*. Wiley, Second Edition.

Ruckdeschel, P. (2000). *Robust kalman filtering. In "XploRe. Application Guide" (Härdle, Wolfgang and Z. Hlavka and Klinke, Sigbert. Eds)*, chapter 18, pages 483–516. Springer, New York.

Salimpour, Y. and Soltanian-Zadeh, H. (2009). Particle filtering of point processes observation with application on the modeling of visual cortex neural spiking activity. In *2009 4th International IEEE/EMBS Conference on Neural Engineering*, pages 718–721.

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88.

Smith, A. F. M. and Roberts, G. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B*, 55:3–23.

Stock, J. H. and Watson, M. W. (1988). A probability model of the coincident economic indicators. National Bureau of Economic Research Cambridge, Mass., USA.

Stock, J. H. and Watson, M. W. (2007). Why has u.s. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33.

Subbotin, A. (2009). Volatility models: from conditional heteroscedasticity to cascades at multiple horizons. *Applied Econometrics*, 15(3):94–138.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time Series Analysis: Theory and Practice*, 1:203–226.

Tsay, R. S. (2005). *Analysis of financial time series*, volume 543. John Wiley & Sons.

Van Der Merwe, R., Doucet, A., De Freitas, N., and Wan, E. (2000). The unscented particle filter. In *NIPS*, pages 584–590.

Walsh (2004). Markov chain monte carlo and gibbs sampling. Lecture Notes for EEB 581.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer, 2nd edition.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80:73–83.

Xie, L., Soh, Y. C., and de Souza, C. E. (1994). Robust kalman filtering for uncertain discrete-time systems. *IEEE Transactions on Automatic Control*, 39(6):1310–1314.