CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

MASTER THESIS

Bayesian estimation of time series with empirical likelihood

Bayesovské odhadování modelů časových řad s využitím empirické věrohodnostní funkce

Author: Juraj Panek
Supervisor: Kamil Dedecius
Year: 2017

České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská

Katedra: matematiky Akademický rok: 2014/2015

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	Juraj Panek
Studijní program:	Aplikace přírodních věd
Obor:	Aplikované matematicko-stochastické metody
Název práce (česky):	Bayesovské odhadování modelů časových řad s využitím empirické věrohodnostní funkce
Název práce (anglicky):	Bayesian estimation of time series with empirical likelihood

Pokyny pro vypracování:

1. Seznamte se s problematikou klasického (tj. nebayesovského) modelování časových řad.
2. Seznamte se s využitím metody empirické věrohodnosti v oblasti časových řad.
3. Navrhněte možné bayesovské přístupy využívající empirickou věrohodnost pro modelování časových řad.
4. Experimentálně (a případně i teoreticky) ověřte vlastnosti navržené metody (metod).
5. Shrňte výhody a nevýhody navržené metody (metod).

Doporučená literatura:

1. A. B. Owen, Empirical Likelihood, Chapman and Hall/CRC, 2001.
2. D. J. Nordman, H. Bunzel, S. N. Lahiri, "A nonstandard empirical likelihood for time series," The Annals of Statistics, Vol. 41, No. 6, 3050–3073, 2013.
3. S. Chen and I. van Keilegom, "A review on empirical likelihood methods for regression," vol. 18, no. 3, pp. 415-447, 2009.
4. K. Mengersen, P. Pudlo, and C. Robert, "Bayesian computation via empirical likelihood.," Proc. Natl. Acad. Sci. U. S. A., vol. 110, no. 4, pp. 1321-1326, Jan. 2013.
5. Y. Kitamura, "Empirical Likelihood Methods in Econometrics: Theory and Practice." Cowles Foundation. Yale University, 2006.
6. H. Harari-Kermadec, "Regenerative block empirical likelihood for Markov chains," J. Nonparametr. Stat., vol. 23, no. 3, pp. 781-802, 2011.

Jméno a pracoviště vedoucí diplomové práce:

Ing. Kamil Dedecius, Ph.D.
Ústav teorie informace a automatizace AV ČR, v.v.i., Pod Vodárenskou věží 4, 182 08 Praha

Jméno a pracoviště konzultanta:

Datum zadání diplomové práce: 28.2.2015

Datum odevzdání diplomové práce: 4.1.2016

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 5. ledna 2017

. . . . . .
vedoucí katedry

. . . . . .
děkan

I hereby declare that this thesis is based on my own work, with the only help provided by my supervisor and the referred-to literature.

In Prague on 6 January 2017 .....

Juraj Panek

Název práce: **Bayesovské odhadování modelů časových řad s využitím empirické věrohodnostní funkce**

Autor: Juraj Panek

Zaměření: Aplikované Matematicko-Stochastické Metody

Vedoucí práce: Kamil Dedecius, Ph.D., ÚTIA AV ČR

Abstrakt: Empirická věrohodnostní funkce je neparametrická metoda statistických odhadů a tvoří alternativu k parametrické věrohodnostní funkci s chí-kvadrát limitní distribcí. Současný vývoj vedl k modifikacím empirické věrohodnosti na analýzu závislých dat. V praci zkoumáme správaní se empirické věrohodnosti na závislých datech a navrhujeme metody statistických odhadů časových řad v Bayesovském prístumu pomocí empirické věrohodnosti s využitm váženýho vzorkování a metod Monte Carlo Markovových řetězců.

Klíčová slova: Empirická věrohodnostní funkce, Bayesovská statistika, časové řady, neparametrická, statistika, Monte Carlo, Markovův řetězec

Title: **Bayesian estimation of time series with empirical likelihood**

Author: Juraj Panek

Specialization: Applied mathematical and stochastic methods

Supervisor: Kamil Dedecius, Ph.D., UTIA CAS

Abstract: Empirical likelihood is a nonparametric method of statistical inference and is an alternative to the parametric likelihood function with chi-square limiting distribution. The recent development has seen the empirical likelihood framework modified to accommodate the inference on dependent data.In this work we study the behaviour of the empirical likelihood treatment of dependent data and propose methods of inference on time series. By proposing the use the empirical likelihood function in the Bayesian setting via importance sampling and Markov chain Monte Carlo methods.

Key words: Empirical likelihood, Bayesian inference, time series, nonparametric, statistics, Monte Carlo, Markov chain

Contents

Introduction

The concept of the empirical likelihood (EL) has recently enjoyed a popularity in statistics mainly because it forms a nonparametric alternative to parametric statistics whilst keeping the preferable features of its parametric counterpart. EL allows us to build parameter estimates and confidence intervals via its chi-square asymptotic properties, i.e. the EL has a chi-squared limiting distribution. Originally the EL framework was designed to treat independent and identically distributed data, but the recent development lead to modified EL structure to accommodate a dependence among data samples. We review the current state of art modifications of EL for time series and propose a new methods of Bayesian inference on time series via the framework of the EL. The rest of this work is organized as follows.

The first chapter gives a basic overview of time series. We introduce the concept of general linear process and define its the representative candidates such as the autoregressive process and the moving average process. We present the idea of stationarity and its conditions which will be referred to and used throughout this work.

The EL concept is introduced in Chapter2. Firstly we build the classical EL for independent data to allow for better understanding of this method and we formulate its asymptotic properties. Further on in this chapter we provide an overview of the recent development of EL for time series, dividing it into two parts. The framework of block-wise EL for weakly dependent time series and then the model-based EL.

Chapter 3 recalls the basic ideas of Bayesian inference such as the prior and the posterior distribution. At this point we propose two methods of sampling from the posterior distribution using the EL for time series. First sampler produces posterior weights as an importance sampler and the other uses the Markov chain Monte Carlo framework to sample directly from the posterior distribution.

Data examples of proposed algorithms are provided in Chapter 4. The chapter is divided into two parts, the first section shows the inference using the block-wise EL sampler and the latter section shows a inference on parameters of the autoregressive process via the model-based EL.

Chapter 1

Time series

This chapter is dedicated to provide basic definitions and ideas behind time series which will be referred to and utilized throughout the rest of this work. However this thesis is not meant to give the reader a detailed overview of time series and its modeling concepts. For a good review on time series, the publications of Box, Jenkins [1] or Wei [2] are recommended.

In this presentation is restricted to stationary time series as the relevant representative on which we will build the proposed methods of statistical inference in latter chapters.

Definition 1.1. *Time series is a set, sequence of observations* $X_1, X_2, \ldots$ *that is generated sequentially over time and can be thought of as a realization of underlying stochastic process* $z_1, z_2, \ldots$.

The observation times of time series $\tau_1, \tau_2, \ldots$ can either come from a discrete set, then we refer to them as *discrete* time series. Or the observations come from a continuous space, and we refer to them as *continuous*. Throughout this work, we will limit ourselves to the discrete case.

A crucial property of the underlying stochastic process is its stationarity which will be later also inherited to the concept of time series.

Definition 1.2. *Let* $z_1, z_2, \ldots$ *denote a stochastic process. We say that the stochastic process is (strictly) stationary, when its properties are not affected by the shift in time. Meaning the probability distribution of given* m *observations* $z_1, z_2, \ldots, z_m$ *is the same as for* $z_{k+1}, z_{k+2}, \ldots, z_{k+m}$ *for any given integer value* $k > 0$.

Definition 1.3. *Real valued uncorrelated random process* $\{e_t\}$ *is called the white noise with zero mean and variance* σ^2 *if*

$$E[e_t] = 0 \tag{1.1}$$

and

$$\mathrm{var}[e_t] = \sigma^2. \tag{1.2}$$

The condition requiring the process $\{e_t\}$ to be uncorrelated implies for the autocovariance function γ to be written in the form of

$$\gamma_k = E[e, e_{t+k}] = \begin{cases} \sigma^2 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0. \end{cases} \tag{1.3}$$

Definition 1.4. *Let the process $\{e_t\}$ denote a white noise. Let $\theta_0, \theta_1, \ldots$ be real numbers. Then the general linear process $\{X_t\}$ is defines as*

$$X_t = \theta_0 e_t + \theta_1 e_{t-i} + \cdots = \sum_{i=0}^{\infty} \theta_i e_{t-i}. \tag{1.4}$$

The definition of the general linear process can be interpreted in a way that the process $\{X_t\}$ is a weighted sum of previous shocks, i.e. the random white noise values e_t.

The formulation of the general linear process form Equation 1.4 is fundamental to us. It can be shown, for reference see Chapter 3 in [1], that any purely nondeterministic stationary process $\{X_t\}$ can be represented in the form from definition 1.4, when satisfying the condition $\sum_{i=0}^{\infty} \theta_i^2 \leq \infty$ and the random shocks e_t are uncorrelated with common variance.

Proposition 1.1. *The general linear process from definition 1.4 represents a stationary process if the set of coefficients $\theta_1, \theta_2, \ldots$ is absolutely summable* [1]*. Meaning that*

$$\sum_{i=0}^{\infty} |\theta_i| < \infty. \tag{1.5}$$

The stationarity property of a general linear process will be important to us in the following chapters, because it implies that the probability distribution of observing X_t, denoted by $p(X_t)$ is constant with respect to the time t and therefore can be written as $p(X)$. As a consequence the stationary process has a constant mean

$$\mu = E[X_t] = \int_{-\infty}^{\infty} X p(X) dX \tag{1.6}$$

and also a constant variance

$$\sigma^2 = E[(X_t - \mu)^2] = \int_{-\infty}^{\infty} (X - \mu)^2 p(X) dX. \tag{1.7}$$

The statistics of mean and variance can be estimated directly from the observed sample by the sample mean $\bar{X}$ (1.8) and sample variance $\hat{\sigma}^2$ (1.9)

$$\bar{X} = \frac{1}{N} \sum_{t=1}^{N} X_t \tag{1.8}$$

[1] The condition restricting the values of coefficients $\theta_1, \theta_2, \ldots$ is sometimes in literature presented in its weaker form by requiring $\sum_{i=0}^{\infty} \theta_i^2 < \infty$

$$s^2 = \frac{1}{N-1}\sum_{t=1}^{N}(X_t - \bar{X})^2 \tag{1.9}$$

where N denotes the sample size.

Generally the stationary linear process can have its mean anywhere on the process range, however we can always rearrange it so that it has a zero mean by writing

$$\widetilde{X_t} = X_t - \mu \tag{1.10}$$

for $t = 1, \ldots, N$. Throughout the rest of this chapter to simplify the notations we will assume that the stationary processes have a zero mean.

Further on in the text we will take use of the notations of the backward and forward shift operator. The backshift operator performs the time shift backwards

$$BX_t = X_{t-1} \qquad B^k X_t = X_{t-k} \tag{1.11}$$

and respectively the forward shift operator performs a forward time shift such that

$$FX_t = X_{t+1} \qquad F^k X_t = X_{t+k}. \tag{1.12}$$

The relationship between the operators can be expressed as $F = B^{-1}$.

It is important to note that under suitable conditions the linear process from definition 1.4 can be written in a form

$$X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \cdots + e_t = \sum_{i=1}^{\infty} \psi_i X_{t-i} + e_t, \tag{1.13}$$

i.e. the current value of X_t is a linear combination of its previous values plus added shock e_t. For instance consider a process

$$X_t = \psi_1 X_{t-1} + e_t \tag{1.14}$$

where all $\theta_j = 0$ for $j > 1$. Recursively substituting $k \in \mathcal{N}$ times for X_t we obtain

$$X_t = \psi_1(\psi_1 X_{t-2} + e_{t-1}) + e_t = \cdots = \psi_1 e_t - 1 + \psi_1^2 e_{t-2} + e_t + \ldots \psi_1^k e_{t-k} + e_t. \tag{1.15}$$

Under conditions $|\psi_1| < 0$ and $E[X_t] < \infty$ and by letting $k \to \infty$ we can rewrite Eq. 1.14 as

$$X_t = e_t + \psi_1 e_{t-1} + \psi_1^2 e_{t-2} + \cdots = \sum_{i=0}^{\infty} \psi_1^i e_{t-i}. \tag{1.16}$$

Therefore the process can be viewed as a linear process where the factors are related via $\psi_1^i = \theta_i$.

Generally speaking, the two presented representations of a general linear process from Equations 1.13 and 1.4 are not used in practical applications since they contain an unbounded number of parameters θ_i and ψ_i respectively. In the next text we present the most commonly used versions of linear time series models.

1.1 Autoregressive process

Autoregressive process is a special case of the linear process formulated by the Equation 1.13, where the current value (X_t) is a regression of its previous values plus the added random shock e_t.

Definition 1.5. *By autoregressive process of order $p \in \mathcal{N}$ we understand the process defined as*

$$X_t = \sum_{i=1}^{p} \psi_i X_{t-i} + e_t \tag{1.17}$$

or alternatively as

$$(1 - \psi_1 B - \psi_2 B^2 - \dots \psi_p B^p) X_t = e_t \tag{1.18}$$

and is denoted by AR(p).

The autoregressive process must meet certain conditions to be stationary. As an example consider the simplest case of AR(1). Writing

$$(1 - \psi_1 B) X_t = e_t \tag{1.19}$$

and rearranging for X_t and using the polynomial notation we get

$$X_t = \frac{1}{1 - \psi_1 B} e_t = \sum_{j=0}^{\infty} (\psi_1 B)^j e_t = \sum_{j=0}^{\infty} \psi_1^j e_{t-j}. \tag{1.20}$$

This expression can also be formulated in form of the *characteristic function* $\Psi(B)$

$$\Psi(B) = \frac{1}{1 - \psi_1 B} = \frac{1}{\psi(B)} = \sum_{j=0}^{\infty} \psi_1^j B^j \tag{1.21}$$

where the term $\psi(B)$ is called the *characteristic polynomial* of the autoregressive process. The above equation yields a condition of $|\psi_1| < 1$ for the AR(1) process to be stationary. Because the root of $\psi(B) = 1 - \psi_1 B$ is $B = \psi_1^{-1}$. This condition is equivalent to requiring that the root of the characteristic polynomial must lie outside of the unit circle.

In the case of a general autoregressive process of order p, we can write down the process in terms of its characteristic polynomial as

$$X_t = \Psi(B) a_t = \frac{1}{\psi(B)} e_t, \tag{1.22}$$

where $\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_p B^p$. Box, Jenkins [1] show via the expansion of $\psi(B)$ into partial fractions that the condition for the stationarity of a general autoregressive process of order p, is that the roots

of the characteristic polynomial of the process $\psi(B)$ must all lie outside of the unit circle.

To analyze the statistical properties of time series, it is common to evaluate the autocorrelation function. To take a closer look at the autocorrelation of a stationary autoregressive process let us multiply the defining Equation 1.17 by shifted value of X_t by k time steps

$$X_t X_{t-k} = \psi_1 X_{t-k} X_{t-1} + \psi_2 X_{t-k} X_{t-2} + \cdots + \psi_p X_{t-k} X_{t-p} + X_{t-k} e_t \quad (1.23)$$

and apply the expected value on both sides. We get

$$\gamma_k = \psi_1 \gamma_{k-1} + \psi_2 \gamma_{k-2} + \cdots + \psi_p \gamma_{k-p}, \quad (1.24)$$

where function γ_k denotes the covariance of a stationary process with time lag k, i.e. $\gamma_k = \mathrm{cov}[X_t, X_{t-k}]$. Note that the term reflecting the covariance between X_{t-k} and random shock e_t vanishes since the observation X_{t-k} is only dependent upon the random shocks up to time $t - k$.

By dividing the Equation 1.24 by the variance of the process, which is an invariant of the time t, we get a term for the autocorrelation function, denoted by ρ. The autocorrelation function of a AR(p) process satisfies the difference equation

$$\rho_k = \psi_1 \rho_{k-1} + \psi_2 \rho_{k-2} + \ldots \psi_p \rho_{k-p}. \quad (1.25)$$

To get clearer understanding of the autocorrelation function, take as an example the case of AR(1) process. After a recursive application of substitution, the autocorrelation function takes a form of $\rho_k = \psi_1^k$. Since the stationarity property requires for $|\psi| < 1$, the function is decreasing with increasing k.

The behaviour of the autocorrelation function depends on the value of ψ. For $\psi > 0$ the function decays exponentially to zero, and for $\psi < 0$ it decreases with oscillations around zero. The larger the value of ψ the slower is the approach of the autocorrelation function to zero.

The autocorrelation function is a valuable tool when analyzing time series, however one must know in an advance the order of the underlying autoregressive process. This information is usually unknown when dealing with real data. One way to exploit the order is the use of the partial autocorrelation function.

Definition 1.6. *Let ψ_{kj} denote the jth coefficient of AR(p) model with representation of order k. Given the formulation from Equation 1.25 the ψ_{kj} satisfies*

$$\rho_j = \psi_{k1} \rho_{j-1} + \cdots + \psi_{k(k-1)} \rho_{j-k+1} + \psi_{kk} \rho_{j-k} \quad (1.26)$$

for $j = 1, 2, \dots k$. After solving these equations we get values for ψ_{kk} as a function of the lag k and call it the partial autocorrelation function.

Note that for AR(p) the partial autocorrelation function ψ_{kk} becomes nonzero for $k \leq p$ and zero otherwise. This is used in determining the order p of AR(p) process by plotting the values of ψ_{kk} and looking for a lag k after which all lags are zero.

1.2 Moving average process

Definition 1.7. *Consider a special case of the general linear process from definition 1.4 where only the first $q \in \mathcal{N}$ of the weights $\theta_1, \theta_2, \dots$ are nonzero. Writing*

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}, \tag{1.27}$$

such process will be called the moving average process of order q and will be denoted as MA(q).

Similarly as with the autoregressive process when we were interested in the property of stationarity. The counterpart property of MA(q) processes is the invertibility. Meaning that we can rewrite the process in terms of the autoregressive model.

As an example consider the MA(1) process

$$X_t = (1 - \theta_1 B) e_t. \tag{1.28}$$

The invertibility condition is now $|\theta_1| < 1$. This is similar as with the AR(1) process, where the stationarity condition required the root of $\psi(B) = 1 - \psi_1 B$ to lie outside of the unit circle. For the MA(1) process the characteristic polynomial can be written as $\theta(B) = 1 - \theta_1 B$ and the invertibility condition requires its root to be outside of the unit circle.

To derive the invertibility condition for a general moving average process of order q, we rewrite it in the terms of the characteristic polynomial

$$e_t = \theta(B) X_t, \tag{1.29}$$

where $\theta(B)$ is

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0. \tag{1.30}$$

After the expansion of $\theta(B)$ into partial fractions one sees that the invertibility property is met when all roots of the characteristic polynomial lie outside of the unit circle.

The stationary property in the case of a moving average process is satisfied naturally since its characteristic polynomial forms a finite sequence, therefore there are no additional restrictions for MA(q) process to ensure stationarity.

For the general moving average process of order q the autocovariance function is

$$\gamma_k = E[(e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q})(e_{t-k} - \theta_1 e_{t-k-1} - \cdots - \theta_q e_{t-k-q})]. \quad (1.31)$$

For the variance, note that the random shocks e_t are generated by a white noise with constant variance σ_e^2 therefore they are uncorrelated and all the terms $E[e_i ej]$ vanish for $i \neq j$, writing

$$\sigma^2 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_e^2. \quad (1.32)$$

Finally the autocorrelation function is of a form

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \ldots \theta_{q-k}\theta_q}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} & k = 1, 2, \ldots, q \\ 0 & k > q. \end{cases} \quad (1.33)$$

From the form of Equation 1.33 it is apparent that the autocorrelation function for MA(q) process is zero beyond the processes order, i.e. there is a *cutoff* after lag q.

For the moving average process of order q, the expression for partial autocorrelation function is rather complicated. However there is a strong duality in the terms of the partial autocorrelation and the autocorrelation function between moving average and autoregressive process. The duality property can be summarized as follows.

1. When dealing with a stationary AR(p) process the current value X_t is defined by a weighted sum of a finite number of its previous values. It can also be represented by an infinite weighted sum of previous shock values e_i. Same applies for the moving average of order q, but in this case the necessary condition is the invertibility of the process, i.e. X_t can be represented by either a finite number of random shocks or by a infinite weighted sum of previous values of X_t. The duality property is also reflected in the condition of stationarity and invertibility respectively. Both conditions require that the roots of the characteristic polynomial of the underlying process lie outside of the unit circle on a complex plane.

2. This duality is reflected in the behaviour of the autocorrelation and partial autocorrelation function. Take for instance the autocorrelation function of MA process, which is zero after a certain lag. But since the MA process corresponds to an infinite AR process, its partial autocorrelation function decays infinitely with exponential power. This applies vice versa to the AR process with partial autocorrelation function that is zero after given lag but its autocorrelation is infinite with a exponential decay.

1.3 Mixed autoregressive-moving average process

We have discussed in previous section that the autoregressive process can be represented as the moving average process and the other way around. However this representation is not always efficient. To obtain a good explanatory parametrization of the underlying process, it is feasible to use both the autoregressive as well as the moving average model.

Definition 1.8. *Let $\psi_1, \psi_2, \dots$ and $\theta_1, \theta_2, \dots$ be real parameters. Let $e_t, e_{t-1}, \dots$ be random shocks generated by a white noise with variance σ^2. Then by*

$$X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \cdots + \psi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots \theta_q e_{t-q} \quad (1.34)$$

we will denote the autoregressive-moving average process or order (p, q) and use the abbreviation ARMA(p,q).

Alternatively we can write down the form of ARMA(p,q) process in terms of the backshift operator B

$$(1 - \psi_1 B - \psi_2 B^2 \dots, \psi_p B^p) X_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) e_t, \quad (1.35)$$

or in the notation of the characteristic polynomials $\psi(B)$ and $\theta(B)$ for AR and MA part respectively by writing

$$\psi(B) X_t = \theta(B) e_t. \quad (1.36)$$

It was already showed in Section 1.2 that the stationarity of MA(q) process is guaranteed naturally therefore the moving average part of Equation 1.34 will not interfere with the stationarity of ARMA(p,q) process. This implies that the condition for stationarity of a ARMA process is the same as for the stationarity of its autoregressive part, i.e. the roots of the characteristic polynomial $\psi(B) = 0$ lie outside of the unit circle. Similarly the autoregressive part of Equation 1.34 does not affect the invertibility of the process. The invertibility condition of ARMA(p,q) process matches the one for MA(q) process, i.e. ARMA(p,q) process is invertible if the roots of $\theta(B) = 0$ from Equation 1.36 are outside of the unit circle.

Recall now the derivation of the autocorrelation function of AR(p) process from Section 1.1. The defining equation of the process was multiplied by X_{t-k} and then the expected value was taken. We proceed the same way for ARMA(p,q) process.

$$\gamma_k = \psi_1 \gamma_{k-1} + \cdots + \psi_p \gamma_{k-p} + \gamma_{Xa}(k) - \theta_1 \gamma_{Xa}(k-1) - \cdots - \theta_q \gamma_{Xa}(k-q) \quad (1.37)$$

where $\gamma_k = E[X_t, X_{t-k}]$ denotes the autocovariance of time series values separated by k time units and $\gamma_{Xa}(k) = E[X_{t-k} e_t]$. Taking in the fact

that X_{t-k} is only dependent on shocks happening previously in time the expression is zero for negative k. It can be shown that after representing the X as a sum of infinite previous shocks and employment of recurrence the autocovariance can be written as

$$\gamma_k = \theta_1\gamma_{k-1} + \theta_2\gamma_{k-2} + \cdots + \theta_p\gamma_{k-p}. \tag{1.38}$$

And the expression for autocorrelation is then

$$\rho_k = \theta_1\rho_{k-1} + \theta_2\rho_{k-2} + \cdots + \theta_p\rho_{k-p} \tag{1.39}$$

where both expression are valid under the condition that $k \geq q + 1$.

The autocorrelation function of the ARMA(p,q) process is though similar to the autoregressive process and is dominated by exponentially decaying values. However this only applies after $q - p$ values, meaning that if $q > p$, the first $\rho_0, \rho_1, \ldots, \rho_{q-p}$ will not follow this pattern. This behaviour makes the autocorrelation graph useful for the model identification.

For the partial autocorrelation function (PACF) of the ARMA(p,q) process it is convenient to rewrite Equation 1.36 as

$$e_t = \frac{\psi(B)}{\theta(B)}X_t. \tag{1.40}$$

The characteristic polynomial representation term of the moving average part is infinite in B therefore the PACF is infinite as well, i.e. there is no cutoff lag, and behaves similarly as the PACF of a pure moving average process, i.e. the PACF is dominated by exponential decay after given lag of $p - q$ in such cases where $p > q$.

Chapter 2

Empirical likelihood

In the following chapter we present the theoretical build up for empirical likelihood, a nonparametric method of statistical inference. Unlike widely used parametric statistics, which require an assumption about the underlying distribution of the data, empirical likelihood relaxes this requirement.

The first part will be covering the classical empirical likelihood formulations build for univariate and multivariate independent and identically distributed data. However when we are dealing with time series the independence in the data is seldom satisfied, therefore the possible modifications of empirical likelihood for depended data is presented in the second part of this chapter.

2.1 Nonparametric likelihood

We will define Empirical cumulative distribution function and show, that it maximizes the nonparametric likelihood function. We will restrict this section for a discrete random variables, where the idea of empirical likelihood is the easiest to understand. The following definitions and theorems are based on the Owens publication [3] on the empirical likelihood theory.

Let X be a random variable ($X \in \mathbb{R}$), then the cumulative distribution function is $F(x) = P(X \leq x)$, where $x \in (-\infty, +\infty)$. Let $F(x-)$ denote probability $P(X < x)$, then we can write $P(X = x) = F(x) - F(x-)$. Let function $1_{A(x)}$ represents indicator of event $A(x)$, so that the function equals to 1 if proposition $A(x)$ is true, otherwise 0.

Definition 2.1. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}$ be random variables. Then the empirical cumulative distribution function (ECDF) of $X_1, X_2, \ldots, X_n$ is defined as follows*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}, \tag{2.1}$$

where $x \in (-\infty, +\infty)$.

ECDF from definition 2.1 is a step function, that increases by step $1/n$ with each one of the n data points.

Definition 2.2. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}$ be random variables, that are independent and identically distributed with common cumulative distribution function (CDF) F. The nonparametric likelihood function of F is defined as*

$$L(F) = \prod_{i=1}^{n} \left(F(X_i) - F(X_i-)\right). \tag{2.2}$$

The *nonparametric likelihood* function naturally corresponds to the classical parametric likelihood of an iid sample. I.e. it equals to the probability of observing exactly the sample $(X_1, \ldots, X_n)$.

The following theorem shows that the ECDF maximizes the likelihood from definition 2.2. Therefore ECDF is the nonparametric maximum likelihood estimator.

Theorem 2.1. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}$ be independent, identically distributed random variables with distribution function F, then*

$$L(F) < L(F_n). \tag{2.3}$$

For any distribution function F, so that $F \neq F_n$.

Proof. Let $z_1, z_2, \ldots, z_m$ be distinct values in $\{X_1, \ldots, X_n\}$. Let n_j be the count of X_i that are equal to z_j, $j \in m$. Let $p_j = F(z_j) - F(z_j-)$ and set $\hat{p}_j = n_j/n$. If for any j, $p_j = 0$, then $L(F) = 0 < L(F_n)$. Suppose now that $p_j > 0$ for all $j \in m$. We write

$$\begin{aligned}
\log\left(\frac{L(F)}{L(F_n)}\right) &= \log\left(\frac{\prod_{j=1}^{m} p_j^{n_j}}{\prod_{j=1}^{m} \hat{p}_j^{n_j}}\right) \\
&= \sum_{j=1}^{m} n_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\
&= n \sum_{j=1}^{m} \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\
&< n \sum_{j=1}^{m} \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right) = 1 - 1 = 0.
\end{aligned}$$

We used the inequality $log(x) \leq x - 1$, where the equality is met only for $x = 1$. Therefore the inequality, $L(F) < L(F_n)$, holds. □

Theorem 2.1 implies, that when there is no information about the distribution, other than observed data. ECDF is the most likely distribution generating them.

2.1.1 Empirical likelihood ratios

In parametric estimations, symbol $\hat{\theta}$ usually denotes the maximum likelihood estimation of parameter $\theta \in \Theta$. We can be build hypothesis tests and parameter confidence regions on the comparison of likelihoods. If the likelihood $L(\theta) << L(\hat{\theta})$, we reject θ from the confidence region. Let us define now, the Empirical likelihood ratio, which will be used to build parameter estimates, tests of hypothesis and set confidence intervals in the nonparametric framework.

Definition 2.3. *For a given distribution F and the nonparametric likelihood L, we define the empirical likelihood ratio*

$$R(F) = \frac{L(F)}{L(F_n)}, \tag{2.4}$$

where F_n is the ECDF from previous definition 2.1.

For further applications, it is useful to break down the Empirical likelihood ratio at this point. Consider a random sample $\{X_1, \ldots, X_n\}$ of size n identically distributed from distribution F, where all samples are distinct (i.e. $X_i \neq X_j$ for all $i, j \in 1, \ldots, n$. Denote p_i the probability that F places on X_i. Using the notation from the proof of theorem 2.1, we get

$$R(F) = \frac{L(F)}{L(F_n)} = \frac{\prod_{j=1}^{n} p_j}{\prod_{j=1}^{n} \frac{1}{n}} = \prod_{j=1}^{n} np_j. \tag{2.5}$$

Relaxing the condition, that the random sample is not necessarily distinct, in equation 2.5, we have to consider that given distinct value $z_j > 1$ occurs n_j time in our data.

$$R(F) = \frac{L(F)}{L(F_n)} = \frac{\prod_{j=1}^{m} p_j^{n_j}}{\prod_{j=1}^{m} \hat{p}_j^{n_j}} = \prod_{j=1}^{m} \left(\frac{np_j}{n_j} \right)^{n_j}. \tag{2.6}$$

In applications though, it is likely rare to observe only distinct data values, therefore consider assigning weights w_i to each observation X_i, so that

$$p_j = \sum_{k} w_k, \tag{2.7}$$

where $k = 1, \ldots, n_j$. It is clear that the weights w_i reproduce the distribution F.

Therefore it is from now on, more convenient to continue with empirical likelihood ratio 2.4, written as

$$R(F) = \prod_{i=1}^{n} nw_i, \tag{2.8}$$

where the only constraints on weights are that $w_i \geq 0$, $\sum_{i=1}^{n} w_i \leq 1$ and F places probability $\sum_{j:X_i=X_j} w_j$ on observation X_i for all $i = 1, \ldots, n$.

Similarly as with the parametric likelihood ratio, we are interested in maximizing the previously defined ratio. Consider, that given distribution F is from a set of distributions $\mathcal{F}$, ($F \in \mathcal{F}$, where $\mathcal{F}$ can be a set of all distributions on $\mathbb{R}$). We are interested in an estimation of a parameter $\theta \in \Theta$, through a function T of a distribution F, so that $\theta = T(F)$. For this purposes we define the profile empirical likelihood function of a parameter θ.

Definition 2.4. *For a given distribution $F \in \mathcal{F}$, where $\mathcal{F}$ is a set of distributions on $\mathbb{R}$, and a function of distribution $T(F) = \theta$. The profile empirical likelihood ratio function is defined as,*

$$\mathcal{R}(\theta) = \sup \{R(F) | T(F) = \theta, F \in \mathcal{F}\}. \tag{2.9}$$

To be able to make a statistical inference, we need to know the distribution of profile ELR 2.9, so we are able to set confidence regions defined as

$$\{\theta | \mathcal{R}(\theta) \geq r_o\}, \tag{2.10}$$

or in a setting when we want to reject hypothesis on parameter θ_0 based upon $\mathcal{R}(\theta_0) \geq r_o$. A possible choice for a threshold r_0 yields the Empirical likelihood theorem, which shows the asymptotic behavior of profile ELR. We restrict ourselves to the univariate case for a mean.

Theorem 2.2. *Let $X_1, X_2, \ldots, X_n$ be independent random variables identically distributed with common distribution F_0. Let $\mu_0 = E(X_1)$ and $0 \leq \mathrm{Var}(X_1) \leq \infty$. Then $-2\log(\mathcal{R}(\mu_0))$ converges in distribution to $\chi^2_{(1)}$ as $n \to \infty$.*

Proof. Proof can be found in Owen's book [3], chapter 2. □

To illustrate the definition of profile ELR, consider a case when we have a data sample $X_1, X_2, \ldots, X_n$ of random variable $X \in \mathbb{R}$ from common distribution F and we are interested in the mean of the sample μ. In the case that X is a bounded random variable by values $A, B \in \mathbb{R}$ (i.e. $-\infty < A \leq X \leq B < \infty$. We can assign the sample minimum $A_n = min_{1 \leq i \leq n} X_i$ and maximum $B_n = max_{1 \leq i \leq n} X_i$ respectively.

Then the weight w_i, used in equation 2.7, can be set to satisfy the condition

$$\sum_{i=1}^{n} w_i = 1. \tag{2.11}$$

Otherwise, if we stick with $\sum_{i=1}^{n} w_i < 1$, distribution F is putting non-negative probability $1 - \sum_{i=1}^{n} w_i > 0$ over the interval (A_n, B_n) exclusive

points $\{X_1, \ldots, X_n\}$. We can assign this probability to data points, since we are interested in maximizing the profile empirical likelihood ratio function 2.9, where the distribution set $\mathcal{F}$ can be taken as all distribution with $\sum_{i=1}^{n} w_i = 1$.

Using the condition 2.11 on weights, one can rewrite the profile empirical likelihood ratio function into a more illustrative form. That is

$$\mathcal{R}(\mu) = \max \left\{ \prod_{i=1}^{n} n w_i | \sum_{i=1}^{n} w_i X_i = \mu, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.12}$$

We can also construct the confidence intervals at given confidence levels as in classical parametric approach by

$$\{\mu | \mathcal{R}(\mu) \geq r_0\} = \max \left\{ \sum_{i=1}^{n} w_i X_i | \prod_{i=1}^{n} n w_i \geq r_0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}, \tag{2.13}$$

where r_0 is a given threshold.

2.2 Estimating equations

Up to now, we have build the Empirical likelihood framework for univariate mean, since it is much more illustrative. The theory proposed to one dimensional case, can be surely extended to a multivariate case.

Observed values $X_1, \ldots, X_n$ are now a random vectors, where $X_i \in \mathbb{R}^d$ for $d > 1$. It can be shown, that the profile empirical likelihood ratio function for a multivariate mean μ, which is now $\mu \in \mathbb{R}^d$, can be written similarly to the one dimensional version as

$$\mathcal{R}(\mu) = \max \left\{ \prod_{i=1}^{n} n w_i | \sum_{i=1}^{n} w_i X_i = \mu, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.14}$$

Except now, it is a function defined on $\mathbb{R}^d$, with range still on $\mathbb{R}$. The first constraint fixing the vectors X_i, mean μ and weights w_i, which are to be maximized, is now a vector equation.

Quinn [4] and Owen [3], proposed linking the estimating equations and empirical likelihood as a flexible way to describe various statistics of a given data set. Let $X \in \mathbb{R}^d$ be a random variable and $\theta \in \Theta \subset \mathbb{R}^p$ be a parameter of interest. Consider a real valued function $m(X, \theta) : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^s$, such that it satisfies

$$E[m(X, \theta)] = 0. \tag{2.15}$$

The function $m(X, \theta)$ is called the estimating function. The common setting, is $p = s$ when the number of restrictions is the same as number of parameters

of interest. Under these conditions the true value of θ can be estimated by solving an *estimating equation*

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta) = 0. \tag{2.16}$$

Estimating equations can be expressed through probability density function $f(X, \theta)$ by

$$m(X, \theta) = \frac{\partial}{\partial\theta}\log f(X, \theta), \tag{2.17}$$

where the resulting estimating equation (2.16) is referred to as the score function. To give a basic idea, few examples of estimating functions are listed below.

- Estimation of the mean $m(X, \theta) = X - \theta$.
- Estimation of the variance $\theta = (\mu, \sigma)$, where the equations are

$$\begin{aligned} m_1(X, \theta) &= X - \mu \\ m_2(X, \theta) &= (X - \mu)^2 - \sigma^2. \end{aligned} \tag{2.18}$$

Substituting the estimating equation for the mean, into 2.12 yields the same result as the constraint fixing observed data and it's mean, since

$$\sum_{i=1}^{n} w_i X_i = \mu \Leftrightarrow \sum_{i=1}^{n} w_i (X_i - \mu) = 0 \Leftrightarrow \sum_{i=1}^{n} w_i m(X_i, \mu) = 0. \tag{2.19}$$

Where we used the fact that, $\sum_{i=1}^{n} w_i = 1$.

Using the notation of estimating equations previously introduced, we can rewrite the profile empirical likelihood ration function 2.12, into a form

$$\mathcal{R}(\theta) = \max\left\{ \prod_{i=1}^{n} n w_i | \sum_{i=1}^{n} w_i m(X_i, \theta) = 0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.20}$$

Similarly as for the profile empirical ratio 2.9, a theorem providing the asymptotic property can be formulated.

Theorem 2.3. *Let $(X_1, \ldots, X_n) \in \mathbb{R}$ be independent random variables with common distribution F_0. For a parameter $\theta \in \Theta \subset \mathbb{R}^p$ and $X \in \mathbb{R}$ and $m(X, \theta) : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^s$. Let θ_0 inΘ be such that $\mathrm{Var}(m(X_i, \theta_0)$ is finite and has rank $q > 0$. If θ_0 satisfies $E[m(X, \theta_0)] = 0$, then $-2\log(\mathcal{R}(\theta_0))$ converges in distribution to the chi-squared distribution as n tends to infinity.*

$$-2\log(\mathcal{R}(\theta_0)) \to \chi^2_{(q)}. \tag{2.21}$$

Proof. Direct consequence of the Theorem 2.2. □

2.3 Calculation of the empirical likelihood

In this work however we are mainly interested in the calculation of the EL for various parameter values. Meaning that we need to evaluate $\mathcal{R}(\theta)$ at for each parameter of interest. Suppose that, data $(X_1, X_2, \ldots, X_n)$ are from R^d, and the parameter of interest θ is of the same dimension, i.e. $\theta \in \Theta \subset \mathbb{R}^d$.

The equation for profile empirical likelihood ratio function $\mathcal{R}(\theta)$ (2.20) is basically a maximization problem of

$$\prod_{i=1}^{n} nw_i, \tag{2.22}$$

over the convex space of vectors over the set of constraints

$$\sum_{i=1}^{n} w_i m(X_i, \theta) = 0, \quad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 1. \tag{2.23}$$

It is practical to use log transformation, since we are maximizing a monotonous function over a convex set $\{(w_1, \ldots, w_n) | \sum_{i=1}^{n} w_i = 1, w_i \geq 0\}$, so we know that a unique global maximum exists.

To solve the maximization problem, we use Lagrange multipliers (λ, γ)

$$G = \sum_{i=1}^{n} \log(nw_i) - n\lambda' \left(\sum_{i=1}^{n} w_i m(X_i, \theta) \right) + \gamma \left(\sum_{i=1}^{n} w_i - 1 \right), \tag{2.24}$$

where the multiplier $\lambda \in \mathbb{R}^d$ is a vector, $\lambda = (\lambda_1, \ldots, \lambda_d$. I.e. consider that the estimating equation is a function from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d$, which means that we need exactly $d + 1$ multipliers. One multiplier for the sum $\sum_{i=1}^{n} w_i = 1$ and d multipliers for constraints given by the estimating functions $m(X_i, \theta)$.

Solving for the first order conditions, we get

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - n\lambda' m(X_i, \theta) + \gamma = 0. \tag{2.25}$$

Multiplying the above equation by w_i and summing it over the index i gives

$$\sum_{i=1}^{n} w_i \frac{\partial G}{\partial w_i} = n + \gamma = 0, \tag{2.26}$$

which solves for one Lagrange multiplier $\gamma = n$.

Substituting this into equation 2.25, yields an equation for w_i as a function of multiplier λ

$$w_i = \frac{1}{n} \frac{1}{\lambda' m(X_i, \theta) + 1}. \tag{2.27}$$

Using the constraint that fixes weights w_i and estimating equation $m(X_i, \theta)$ from 2.23. We can write d equations for unknown vector λ

$$0 = \sum_{i=1}^{n} w_i m(X_i, \theta) = \left(\frac{1}{n} \right) \sum_{i=1}^{n} \frac{1}{\lambda' m(X_i, \theta) + 1} m(X_i, \theta). \tag{2.28}$$

This justifies, that we can think of the vector multiplier λ as a function of θ, $\lambda = \lambda(\theta)$. Therefore, to calculate the individual weights w_i associated with each data vector X_i, one must involve a numerical algorithm that searches for $\lambda(\theta)$ that satisfies

$$\sum_{i=1}^{n} \frac{1}{\lambda' m(X_i, \theta) + 1} m(X_i, \theta) = 0. \tag{2.29}$$

The search for such a vector λ, must take into an account the condition proposed on each weight $w_i \geq 0$, by enforcing this on equation 2.11, we obtain n inequality constraints for λ

$$1 + \lambda' m(X_i, \theta) > 0, \tag{2.30}$$

for all $i = 1, \ldots, n$. The case when $w_i = 0$ can be excluded, since our goal is to maximize objective function $R(\theta)$. We could conclude, that the computation of EL ratio 2.20 is straight forward. However, the numerical search for λ does not necessarily lead to the desired precision when the denominator in 2.28 approaches zero.

As a remedy Owen [3] suggested a possible workaround that takes the advantage of the complex duality. If we substitute the weights w_i (2.27) to the logarithm of EL ratio $\log R(F)$, and define L as a function of the multiplier λ

$$\log R(F) = -\sum_{i=1}^{n} \log(1 + \lambda' m(X_i, \theta)) \equiv L(\lambda). \tag{2.31}$$

We can transform the original problem of maximization over n variables w_i subject to d constraints, originally $d + 1$, but we have been able to eliminate γ in 2.26, to a minimization of function $L(\lambda)$ over d variables $\lambda = (\lambda_1, \ldots, \lambda_d)$ subject to n inequality constraints 2.30.

Even though this alternation of the original problem, can rapidly speed up the computation if $d << n$, the problem of vanishing denominator in 2.28 still holds. Owen suggests another trick, which can lead to an unconstrained minimization. By redefining $L(\lambda)$ so that it is a convex over $\mathbb{R}^d$, but it's value close enough the solution does not change.

Let a pseudo logarithm function be defined as

$$\log_\star(z) = \begin{cases} \log(z) & \text{if} \quad z \geq 1/n \\ \log(1/n) - 1.5 + 2nz - (nz)^2/2 & \text{if} \quad z < 1/n. \end{cases} \tag{2.32}$$

If we place $1 + \lambda' m(X_i, \theta)$ into function $\log_\star$ and recall the condition on the weights $w_i \leq 1$. Then arguments that are greater then $1/n$, correspond to weights, that are greater then 1, and therefore are not a part of the valid solution. The function indeed does not change its value around the solution and is quadratic otherwise.

After we rewrite $L(\lambda)$, through a $\log_\star$ function, instead of minimizing problem with n inequality constraints (2.30). We are left with minimizing

$$L_\star(\lambda) = -\sum_{i=1}^{n} \log_\star(1 + \lambda' m(X_i, \theta)), \tag{2.33}$$

without any constraints. This reformulation of the original maximization problem is a subject to convex duality and can significantly reduce the computation burden of empirical likelihood, and will be used for computation of EL in this work.

2.4 Empirical likelihood for dependent data

So far we have restricted the formulation of the empirical likelihood for independent and identically distributed data. When making a statistical inference on dependent data the assumption of statistical independence required in the Theorem 2.2 is not met. As a result the asymptotic properties of EL no longer hold and the classical formulation of EL results in an incorrect construction of confidence intervals and coverage properties.

There are two general approaches, based on the assumption made about the dependence in the data. The first case assumes that the observed data form a sample from a known time series model driven by an unobservable iid shocks. This approach takes an advantage of the time series models described in Chapter 1, where the observed variables are expressed as a function of the independent random shocks. We will refer to this approach as a *model-based* EL.

The other approach does not require any assumptions about the model generating the observed data but assumes that something is known about the range of the dependence. It assumes that observations occurring relatively close to each other carry the strongest dependence among them. On the other hand observations relatively far apart are assumed to be almost independent. This leads to the formation of independent blocks of data and therefore this approach is referred to as the *block-based* empirical likelihood.

2.4.1 Model-based empirical likelihood

With the so called *model-based* empirical likelihood approach to time series it is assumed that the observed times series $\{X_t\} \in \mathbb{R}^d$ come from a structural model, where the observed data are generated by unobserved independent random variables. This concept is well known and widely used in parametric statistics, where the serially correlated data $\{X_t\}$ are represented as unobserved random shocks, which are assumed to be independent.

To motivate this approach, in EL setting one usually needs to construct the estimating equations (Equation 2.16) which tie the parameter $\theta \in \Theta \subset \mathbb{R}^p$

of interest to the observed data $\{X_t\}$. By making an assumption about the structural model of the observed sample we can use it to construct the estimating equation based on the model to tie the parameter to the observed data as $m : \mathbb{R}^{d(p+1)} \times \Theta \to \mathbb{R}^p$.

This presentation is based on the results of Mykland [12] who introduced the generalization of the empirical likelihood for iid data to models with martingale structure, by using the concept of dual likelihood. Consider an autoregressive model that has been introduced in chapter 1, where the currently observed datum X_t is a regression of its past p values with added unobservable random disturbance e_t

$$X_t = \sum_{i=1}^{p} \psi_i X_{t-i} + e_t. \tag{2.34}$$

where $\psi = (\psi_1, \psi_2, \dots \psi_p)$ are the parameters of the model. We assume that the disturbances form a martingale difference sequence to satisfy the assumptions of the dual likelihood. Generally the class of models with martingale difference sequence is broader, but an example of such models is an iid process with zero mean and a common variance. This assumption is naturally satisfied if we take the disturbances e_t from (2.34) to be a Gaussian white noise.

A common practice in the estimation of the parameters in the autoregressive models is to use the conditional least squares which leads to the OLS estimate $\hat{\psi}$ via the minimization of

$$\hat{\psi} = \min_{\psi \in \Theta} \frac{1}{2} \sum_{t=p+1}^{n} (X_t - \psi'(X_{t-1}, \dots, X_{t-p+1}))^2. \tag{2.35}$$

By using the notation of $\mathbf{X}_t = (X_t, X_{t-1}, \dots, X_{t-p+1})$ we can write the estimate $\hat{\psi}$ as

$$\hat{\psi} = \left(\sum_{t=p+1}^{n} \mathbf{X}_{t-1} \mathbf{X}'_{t-1} \right)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_{t-1} X_t \tag{2.36}$$

where $\mathbf{X}'$ denotes the transposition of $\mathbf{X}$.

The differentiation of (2.35) with respect to the parameter ψ gives us the *score* function

$$\sum_{t=p+1}^{n} (X_t - \psi' \mathbf{X}_{t-1}) \mathbf{X}_{t-1} = \sum_{t=p+1}^{n} g_t. \tag{2.37}$$

Under the true value of parameter ψ_0, the term $X_t - \psi' \mathbf{X}_{t-1}$ can be viewed as the reformulated disturbance e_t and therefore we can write the *score* function as

$$\sum_{t=p+1}^{n} e_t \mathbf{X}_{t-1} = \sum_{t=p+1}^{n} g_t. \tag{2.38}$$

Chan el al. [11] has shown, that under $\psi = \psi_0$ the term $e_t \mathbf{X}_{t-1}$ forms the desired martingale difference sequence. Mykland defines the dual empirical likelihood ratio by writing

$$l(\psi) = -2 \sum_{i=p+1}^{n} \log(1 + \lambda' g_t), \tag{2.39}$$

where the parameter λ is defined by the relation

$$\sum_{t=p+1}^{n} \frac{g_t}{1 + \lambda' m_t} = 0. \tag{2.40}$$

Recall for now, the construction of the classical empirical likelihood for iid data. In section dedicated to the calculation of the EL, we have reformulated calculation of the EL profile to minimizing the Equation 2.39. Indeed the terms in Equation 2.39 and the problem from (2.39) are identical and one could have possibly arrived at the dual likelihood by treating the score function g_t as being iid and replacing it for the estimating functions m_t.

Chan et al. [11] studied the asymptotic properties of both stable and unstable time series. In his article he proved the following theorem which shows the asymptotic properties of the dual/empirical likelihood for stationary autoregressive processes.

Theorem 2.4. *Let the autoregressive process AR(p) be so that all the roots of its characteristic polynomial (from Equation 1.22) lie outside of the unit circle, i.e. the AR(p) process $X_1, \dots X_n$ is stable. Then the following propositions hold*

- *Term*
$$\left(\sum_{i=p+1}^{n} \mathbf{X}_{t-1} \mathbf{X}'_{t-1} \right)^{1/2} (\hat{\psi} - \psi) \tag{2.41}$$
converges in distribution to the normal distribution with zero mean and variance $\sigma^2 \mathbb{I}_p$, where $\mathbb{I}_p$ is the identity matrix of rank p.

- *The term $-l(\psi)$ from Equation 2.39 converges in distribution to $\chi^2_{(p)}$.*

All propositions as $n \to \infty$.

Proof. Can be found in the appendix of [11]. □

This results provide a possibility to build an inference for a various autoregressive models. From a computational point of view, it can be challenging that by the construction of the estimating equation from Eq. 2.37 provides only $k = n - p$ estimating equations, in contrary to classical EL, where we have the same number of estimating functions as the number of

data. A possible workaround is to set the m_t for $t = 1, \ldots, p$ to zero since for such m_t the weights assigned in EL calculation are zero and therefore do not change the empirical likelihood profile.

The discussed model-based empirical likelihood approach suffers mostly on the ability to reformulate the observed sample as a process of iid variables, or more generally as a martingale difference sequence as shown by Mykland [12]. Other applications of the model-based approach contains the generalizations of the described method to unstable AR model studied by (Chan [11]). AR models with infinite variance, i.e. $E[e_t^2] = \infty$ explored by [13] and generalized autoregressive conditional heteroskedasticity models [14].

2.4.2 Block empirical likelihood

The assumption made about the underlying model model in the previous section cannot always be made. And it can be rather restrictive, especially when little is known about the origin of the observed sample. As a possible remedy the work of Kitamura [5] presents a modified empirical likelihood based method designed to treat the dependence in the data. Kitamura does not assume any stochastic model structure but proposes to have some prior information about the character of the dependence among the data. The dependence is usually required to be weak, i.e. data far apart in the observed time perspective do not carry as much dependence among them or is assumed to be negligible. To be able to describe this dependence property, some terms about the mixing conditions need to be introduced.

Mixing in time series

To properly define the required modifications of empirical likelihood to treat the dependence in the data we must firstly define the level of dependence in time series also referred to as the mixing condition.

Consider a case that observations X_t for $t = 1, \ldots T$ are consequent observations of an infinite time series $(\ldots, X_{-1}, X_0, X_1, \ldots)$. The mixing condition describes the measure of dependence between a set of observations given at time t, i.e. $(\ldots, X_{t-1}, X_t)$, and a set of observations followed by a shift of $k \in \mathbb{N}$ time steps, that is $(X_{t+k}, X_{t+k+1}, \ldots)$.

Let a random variable A depend on the outcome of the time series before time unit t and alternatively let a random variable B depend on the result of the series X_{t+k+i}, for $i \geq 0$. The interpretation is that if the future of the time series is independent of the past, or if the time series is fully independent we get $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. For a proper definition of random variables A and B see Chapter 16 in [9]. Now define the $\alpha - mixing$ coefficient to measure the dependence in time series

$$\alpha(k) = \sup_t \sup_{A,B} |\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)|. \tag{2.42}$$

The value of $\alpha(k)$ takes values between 1 and 0. It becomes zero for fully independent data observed before time step t and after time $t+k$. When we consider that the investigated time series X_i is stationary, then the term of maximization over time step t in Eq. 2.42 can be relaxed.

Definition 2.5. *We say that a time series is α-mixing is the following condition on $\alpha(k)$ from Equation 2.42 is satisfied:*

$$\alpha(k) \to 0 \qquad \text{as} \qquad k \to \infty. \tag{2.43}$$

The term α-mixing of time series is also referred to as a strong mixing property.

Definition 2.6. *Let for a time series the assumption from Definition 2.5 hold, i.e. the series is strong mixing. Then if for some constant $\delta > 0$ the following expression holds*

$$\sum_{k=1}^{\infty} \alpha(k)^{1-\frac{1}{\delta}} < \infty, \tag{2.44}$$

we will say that the time series is weakly dependent.

Construction of the block-wise empirical likelihood

Once the concept of mixing in time series is introduced, we can start building up the block-wise empirical likelihood (BEL) as a fully non-parametric method for statistical inference on the time series. Until the end of this section it will be assumed that all the time series $X = (X_1, X_2, \ldots, X_T)$ fulfill the condition of stationarity from Chapter 1 and are assumed to satisfy the conditions of weak dependence from Equation 2.44.

The block-wise empirical likelihood is based on the idea of constructing blocks of observations rather than treating every single observation separately. This approach forms blocks of observations which are assumed to be either fully independent or its dependence is negligible from a statistical point of view. We start by constructing a block of length l of consecutive observations $B_k = \Big(X_{(k-1)m+1}, \ldots, X_{(k-1)l+l}\Big)$, where m is the number separating the observations of the process and can range from $1, \ldots, l$. It can be easily shown that the block index k is ranging from $1, \ldots, n$

$$n = \left\lceil \frac{T-l}{m} \right\rceil, \tag{2.45}$$

where $\lceil x \rceil$ denotes the closest upper integer to x.

Setting the separation length m to 1 would yield maximally overlapping blocks and results in better precision but on the other hand this can lead

to an increased computational requirements. The other extreme case would be setting the separation to $m > l$ which would leave some of the observed values X_k unused and imply into a significant loss of information.

Similarly as with the construction of the empirical likelihood for independent data we continue by an introduction of the estimating function which are modified to accommodate the blocked structure. Let $m : \mathbb{R}^d \times \Theta \to \mathbb{R}^p$, where $\Theta \subset \mathbb{R}^s$, be an estimating function as defined in Equation 2.16 binding the parameter to the data sample. Then by profiling it through the blocked data of length l we get

$$b(B_k, \theta) = \frac{1}{l} \sum_{i=1}^{l} m(X_{(k-1)m+i}, \theta), \tag{2.46}$$

where $\theta \in \Theta$ represents the parameter of interest. The function $b(B_k, \theta)$ is referred to as the k-th *smoothed moment function* to account for the fact that information in one data block is condensed into one function.

Recall that the estimating function for iid case was required to satisfy the zero mean property expressed by $E[m(x_k, \theta)] = 0$. It is of a direct consequence that if the iid estimating functions have a zero mean, its smoothed counterpart inherits this property, writing

$$E[b(B_k, \theta)] = 0. \tag{2.47}$$

Using the definition of BEL estimating function as a natural consequence we can rewrite the empirical likelihood ratio as it was introduced in the Equation 2.20 into its block-wise counterpart in the form of

$$\mathcal{R}_b(\theta) = \max \left\{ \prod_{i=1}^{n} n w_i | \sum_{i=1}^{n} w_i b(B_i, \theta) = 0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}, \tag{2.48}$$

where n accounts for the total number of blocks.

So far the construction of the block empirical likelihood ratio consisted on the idea of condensing the dependence among the neighboring observations into one data block. However the crucial property that justifies the proposed steps was proven by Kitamura [5] when he showed that the block-wise empirical likelihood ratio converges in distribution to the chi-square distribution. This property is similar to the one showed by Theorem 2.3 for the iid case of empirical likelihood. The following theorem states that when the weak time dependence condition among the data is present the adjusted term $\mathcal{R}_b(\theta)$ also converges in distribution to χ^2.

Theorem 2.5. *Let the time series X_t be weakly dependent. Let further the assumptions form [5] hold. Let additionally $l \to \infty$ so that $lT^{-1/2} \to 0$. Let $\theta_0 \in \Theta \subset \mathbb{R}^s$ satisfy the estimating equation mean from Equation 2.47, then*

$$-2 \left(\frac{T}{nl} \right) \log \mathcal{R}_b(\theta_0) \to \chi^2_s. \tag{2.49}$$

Proof. Can be found in the appendix in [5]. □

This theorem can be viewed as a justification of the block-wise EL structure and allows us to build confidence intervals and create statistical hypothesis tests on a similar basis as with the iid version of the EL.

The confidence level can be build upon profiling parameter values $\theta \in \Theta$ so that

$$\{\theta|\mathcal{R}_b(\theta) \geq r_0\} = \max\left\{\sum_{i=1}^{n} w_i b(X_i, \theta)\Big| \prod_{i=1}^{n} w_i > r_0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1\right\}, \tag{2.50}$$

for a given threshold value r_0.

The term $\frac{T}{nl}$ from Equation 2.49 is an addition in the EL theorem when compared to its iid counterpart from Eq. 2.3 and is necessary, because it accounts for the overlapping of the blocks. To see the need for it, note that there are n blocks of observations each consisting of $(X_i, X_{i+1}, \ldots, X_{i+l})$ separate observations. Altogether there are nl observations entering the block empirical likelihood function (2.48).

Consider now a special case of the block-wise EL where we decide to set the overlap factor m to 1, i.e. the fully overlapping blocks of observations. The number of blocks from Equation 2.45 is $n = \lceil T - l \rceil$ and if we further assume a blocks of length $l = 1$ the correction factor $\frac{T}{nl}$ is equal to 1 and the classical formulation of empirical likelihood becomes a special formulation of the block-wise EL with all its corresponding asymptotic properties.

To correctly choose the factors such as the overlap m and the block size l requires some tuning. Since the block-wise empirical likelihood methods are relatively new and have not been extensively studied in theory, there is not a universal approach to setting up these parameters, that would guarantee the best coverage of the BEL. However as suggested by Owen or Kitamura or Nordman, some basic guidelines can be summarized here. A common requirement for the asymptotic properties to hold is that the

$$l^{-1} + \frac{l}{T} \to 0 \quad \text{as} \quad T \to \infty. \tag{2.51}$$

Which can be interpreted that the block size is relatively small compared to the sample size but it increases significantly with the sample size. Consider a case where the blocks are not overlapping at all, i.e. setting $m = l$, and set $c^{-1} = l/T$. If we take a large sample size T and apply the iid case of empirical likelihood and assume that the coverage properties are plausible, the suggestion is that the block size l can be taken as T/c. However one should always be cautious when doing so, because for instance if the observations were iid and we would use a block of length l, instead of T observations we only end up with T/l samples and the efficient lost can become large.

On the other hand, the selection of the block overlap parameter m does not seem so crucial with large sample sizes and should be chosen between 1 and the block length l so that no observations are left out. However one should note that a small number of overlap can lead to a significant increase in computational time.

Block-wise empirical likelihood modifications

The previously introduced concept of blocking in the empirical likelihood framework proved to be a starting point for further generalizations of this approach. Recently, various modified approaches have been developed and studied to treat special cases of the EL inference in time series from modifications for smooth function models to frameworks that do not require the selection of the block length. From these variations the notable approaches with references are summarized here.

1. *Expansive block empirical likelihood (EBEL).* A method introduced by Norman et al. [8] proposes a variation of the standard BEL to use a nonstandard blocking scheme without a predetermined block length l but rather uses all possible variations of its length. However this approach does not follow the standard asymptotic properties of limiting chi-square distribution but the limit low is distribution-free and has to be obtained by simulations. The provided simulation examples show an inference on a process mean of various processes such as AR, MA, ARMA models and yield a improved coverage properties when compared to the standard BEL.

2. *Regenerative block empirical likelihood (ReBEL).* Was introduced by Harari-Kermadec [10] and designed for a nonparametric inference on Markov chains. The name is based from the regenerative structure of Markov chains and partitions them into almost independent blocks of data that can have a random length. The methodology assumes that we are able to construct a regenerative sequence withing the Markov chain with a stationary distribution. For these sequences the classical iid version of EL is applied and its corresponding chi-square limiting properties can be proven and contradictory to BEL id does not require any correction for the overlap in blocks. However even though there is no need for a tunning parameter of the block length it can be hard to estimate the regeneration times, Harari-Kermadec [10] provide some basic techniques to estimate this.

Chapter 3

Bayesian inference in time series via empirical likelihood

This chapter proposes the ideas of Bayesian estimation in time series when the framework of the empirical likelihood is employed. In the opening of this chapter we briefly recall the basic concepts of Bayesian inference in general and build upon it. In latter sections we come up with sampling algorithms (Section 3.3) from posterior distribution of time series by combining the newly proposed general samplers for independent data (Section 3.3) and the empirical likelihood designed for time series inference from Chapter 2.

3.1 Introduction to Bayesian inference

Generally in statistical inference we are interested in drawing conclusions about the observed data. The data are viewed as realizations of a random variable X defined on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. The realizations of X are called the (data) *sample*, and the set Ω is represented as the population. In this work we are mainly interested in the non-parametric statistical models, however to introduce the concept of the Bayesian inference let us consider for now the well known parametric statistical approach.

In parametric statistical models we make an assumption that the sample has been generated from a parametrized probability distribution $f(x|\theta)$, where the parameter θ is not known. Generally speaking, we want to invert the original phenomenon of observing sample governed by a fixed parameter θ, to deduce the value of parameter based on the fixed observed sample. I.e. like in the notion of a likelihood function $l(\theta|x)$ which is obtained as a rewritten probability distribution $l(\theta|x) = f(x|\theta)$, what happens is that we are basically inverting the probability to a case of a conditional distribution of parameter θ, conditioned by the observed sample x.

The inversion of probabilities is of a direct consequence of the Bayes's theorem, which loosely states the following. Let A and B be events from

the probability space, with respective conditional probabilities $P(A|B)$ and $P(B|A)$. Then their conditional probabilities are related by an equation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{P(B|A)P(A)}{P(B)}. \tag{3.1}$$

Where A^c is the complement event to the event A, and naturally the event B must satisfy $P(B) > 0$, so that equation 3.1 is well defined.

For further purposes it is more important that the Bayes' theorem can be also stated for a continuous random variables, i.e. in the sense of probability density functions. Let X and Y denote continuous random variables with a respective conditional probability density function $f(x|y)$ and let $g(y)$ denote the marginal probability density function of the random variable Y. Then the conditional density of random variable Y, conditioned by a fixed value of $X = x$ can be written as

$$g(y|x) = \frac{f(x|y)g(y)}{\int_{\mathcal{Y}} f(x|y)g(y)\mathrm{d}y}. \tag{3.2}$$

Similarly as in the previous case with the context of probability events from Equation 3.1, the denominator in Eq. 3.2 is actually a marginal distribution of X. The theorems (3.1) and (3.2), can also be viewed as an actualization of our knowledge about Y after observing X, in other words, we are updating the information on Y provided by $g(y)$ to $g(y|x)$, once that we have observed the variable X.

The concept of the probability inversion discussed previously is pivotal in the Bayesian inference. Where we are primarily interested in the inversion of a probability of observing a data sample conditioned by a parameter $\theta \in \Theta$. Consider that we place a some sort of uncertainty on the parameter θ, through a probability distribution (density function) π defined on Θ, so the parameter θ is now considered to be a random variable. We still hold, that the observed sample comes from a distribution denoted by density function $f(x|\theta)$.

The distribution of $\theta \sim \pi(\theta)$, therefore can be viewed as our prior knowledge about the parameter and is called the *prior distribution*. If we use the $\pi(\theta)$ instead of the $g(y)$ in Equation 3.2, the Bayes's theorem can be rewritten into a form

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)\mathrm{d}\theta}. \tag{3.3}$$

Where the distribution $\pi(\theta|x)$ is referred to as the *posterior distribution* of the parameter θ, i.e. the probability of observing θ given the data x. In statistical terms, we update the prior information $\pi(\theta)$ after observing a data sample to a posterior distribution $\pi(\theta|x)$.

The posterior distribution plays a key role in the Bayesian inference about parameter θ. Robert in [17], shows that the entire information about parameter, that we can obtain from the observed data, is included in the posterior distribution $\pi(\theta|x)$. Meaning that once we know the posterior distribution, the data sample cannot provide us with any additional knowledge. The summarizing statistics, such as mean, variance, median etc. can be directly evaluated from $\pi(\theta|x)$, and are noted as the posterior mean $E^{\pi}[\theta|x]$, or the posterior variance and median respectively.

That said, the whole Bayesian inference can be reduced to the problem of evaluating the posterior distribution (3.3). However this task can prove to be difficult or if not possible in many settings. The first difficulty comes in place when evaluating the denominator of (3.3), because the integration throughout the whole parameter space Θ may not be feasible. Yet this problem has been overcome recently with the introduction of the Markov chain Monte Carlo methods (which will be briefly introduced later in this chapter) since these methods can sample directly from the posterior distribution

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta). \tag{3.4}$$

when it is known only up to the normalizing constant.

The more difficult task of evaluating the posterior distribution comes in place when facing the so called *fully intractable* problems. The concept of fully intractable setting can arise in situations when the whole likelihood function $f(x|\theta)$ is unknown or it is computationally not efficient to evaluate. In these situations one is basically left with two options to obtain the posterior distribution.

- *Approximate Bayesian Computation methods.* These methods are designed for inference when the likelihood function is intractable but but it is possible to simulate data from the model. Such algorithms rely on the comparison between the simulated data for a given parameter value and then based on the *closeness* between the simulated and observed data a decision is made about the acceptance of the proposed parameter. These methods have enjoyed a relative success recently and have been extensively studied and various modifications were developed. However this work is not dedicated to the ABC samplers which have already been well reviewed e.g. [18].

- *Likelihood-free methods.* Another approach that offers itself when the parametric likelihood is not available is to use a non-parametric method. The use of non-parametric methods as a remedy for intractable setting in Bayesian inference has not been widely studied so far and from theoretical point of view it is still an open topic. An example of such a non-parametric method can be the Empirical likelihood framework

and its use in the Bayesian setting, which will be proposed and studied throughout the rest of this work.

To introduce what follows and to put it into perspective, the initial motivation to explore the use of the Empirical likelihood in the Bayesian inference actually came from the original ABC algorithm. Where it is usually considered that the likelihood function of the model is fully intractable but at the same time we are able to simulate from the model for a given parameter value. The original idea of ABC algorithm is formulated in an Algorithm 1.

input : data sample $\mathbf{x}$, prior distribution $\pi(\theta)$, likelihood $f(\mathbf{x}, \theta)$
for $i = 1$ *to* N **do**
  Generate $\theta' \sim \pi(\theta)$;
  Simulate $\mathbf{z}$ from likelihood $f(\cdot, \theta')$;
  if $(\mathbf{z} = \mathbf{x})$ **then**
    Save $\theta_i = \theta'$
  end
end

Algorithm 1: Likelihood-free rejection sampler

The issue with the this algorithm lies in the acceptance condition, where we check the simulated sample $\mathbf{z}$ against the observed sample $\mathbf{x}$. If we would be able to meet this acceptance criterion, it can be shown that the resulting accepted samples $(\theta_1, \theta_2, \dots)$ would form an exact sample from the posterior distribution $\pi(\theta|\mathbf{x})$. However one can see that such a strict acceptance criterion must lead to a great number of rejected samples, especially in a continuous setting, where theoretically cannot be any accepted samples. Various modifications of the acceptance criteria are used to compensate this issue, such that the proposed sample is accepted when

$$\rho_S(S(\mathbf{z}), S(\mathbf{x})) < \epsilon \tag{3.5}$$

where S is any chosen summary statistic and ρ_S is a metric on the range of the mapping S. This treatment increases the number of accepted sample however it comes at a cost, since the summary statistic is seldom sufficient and the parameter ϵ requires some fine tuning.

Originally as an alternative to ABC sampler, Mengersen et al. [15] formulated an idea to employ the empirical likelihood as a weight associated with each proposed sample from prior in Algorithm 1. This leads to omitting the strict rejection behaviour of the original sampler. The proposed idea of [15] would no longer be rejecting samples, but instead would keep all of the proposed samples with a respective weight equal to the empirical likelihood, i.e. a similar concept as with parametric likelihood which would assigned the proposed sample a probability with which the data set $\mathbf{x}$ would be observed.

3.2 Empirical likelihood in Bayesian inference

To proceed further and employ the concept of empirical likelihood in the Bayesian setting it is of a right place to theoretically justify to proposed ideas. The EL as it has been presented in Chapter 2 forms an alternative to the parametric likelihood function with well behaving asymptotic properties for the independent and identically distributed data.

So far the theory behind the use of the EL in the Bayesian set ting has not been studied extensively. The studies so far have been mainly based on simulation results and without much theoretically proven results. The question of the use of EL as a replacement of the parametric likelihood function in Bayesian setting has been partly discussed by Owen [3] (Section 9.4) by proposing the following idea. Let $\theta \in \Theta$ be a parameter which is tied to the observed data set $(X_1, X_2, \ldots, X_n)$ by an estimating function $m(X, \theta)$, and consider that we place a prior distribution $\pi(\theta)$ on the randomized parameter θ. In the non-parametric case no assumptions whatsoever are made about the parametric family of distributions. A natural next step would be to use the empirical likelihood $\mathcal{R}(\theta)$ from Equation 2.20 for iid data and or its modifications designed to treat time series respectively as an alternative to the parametric likelihood function $f(\mathbf{x}, \theta)$ in the Bayes's theorem for the posterior distribution

$$\pi_{el}(\theta|x) = \frac{\mathcal{R}(\theta)\pi(\theta)}{\int_\Theta \mathcal{R}(\theta)\pi(\theta)\mathrm{d}\theta} \propto \mathcal{R}(\theta)\pi(\theta). \tag{3.6}$$

Note that the EL noted by $\mathcal{R}$ cannot be viewed as a proper probability density function since its integration throughout the parametric space Θ would not equal to one and. However we can proceed further by claiming that the newly defined *empirical likelihood posterior distribution* $\pi_{el}(\theta|x)$ is proportional to $\mathcal{R}(\theta)\pi(\theta)$.

From simply computational point of view, replacing parametric with empirical likelihood can be seen as reasonable. Lazar ([23]) studied, whether the combination of empirical likelihood and prior distribution gives desired properties of posterior distribution 3.6. As a justification Lazar proposes following quantitative test to test whether the use of EL yields correct posterior distribution results.

Consider a one dimensional data sample $\mathbf{x}$ and simulate various values of

$$H(a) = \int_{-\infty}^{a} \pi_{el}(\theta|\mathbf{x})\mathrm{d}\theta. \tag{3.7}$$

This corresponds to the posterior probability of parameter θ being in the set $(-\infty, a]$. If the posterior is valid, then $H(a)$ should be distributed uniformly over the interval $(0, 1)$. Lazar tested this, by using an uniform prior distribution on the mean and EL for mean given by 2.12. The values $H(a)$ were tested against uniform distribution for samples $\mathbf{x} = (X_1, ...X_n)$

consisting of $n = 5, 10, 20, 50$ data. For increasing values of n, distribution of $H(a)$ seemed to be better fitting the expected uniform distribution. Using the Kolmogorov-Smirnov criterion for $n = 50$, gave a valid result. The qq-plots also showed a reasonable distribution fit for data samples from size $n = 20$. Author however concludes that the results should to be interpreted with some care, since the validity of the posterior inference needs to be established for each case individually.

3.3 Bayesian samplers with empirical likelihood

In this section we propose the new methods of Bayesian inference on time series with the use of the empirical likelihood function by building a sampling schemes that sample from the posterior distribution $\pi_{el}(\theta|\mathbf{x})$ defined by the Equation 3.6. In the first part we construct the possible samplers for a general iid case of statistical inference using the EL as the importance sampling weights and then propose a method of sampling directly from the posterior distribution via the Markov chain Monte Carlo methods. In the latter part we propose the combination with the empirical likelihood designed for dependent data from Chapter 2.

3.3.1 Weighted sampler

As already mentioned the original ABC algorithm (1) served as a motivation to design a algorithm that would sample from the posterior distribution when encountering a model with intractable likelihood function. The drawbacks of ABC methods were already mentioned, such that the algorithm can be too restrictive or that it requires some fine tuning of its parameters. The original idea of combining the Bayesian samplers and the framework of the empirical likelihood was initially proposed in an article [18]. The idea was to design an algorithm that would provide a sample from the posterior distribution but would still remain likelihood-free (in our case, it is better to say a parametric likelihood-free). The idea can be reformulated as follows.

At each step of the likelihood free rejection sampler 1, calculate the profile empirical likelihood ratio function given by 2.20. So for a proposed parameter θ and an observed data sample $\mathbf{x} = (X_1, X_2, \ldots, X_n)$, we want to compute the product

$$\mathcal{R}(\theta) = \max \prod_{i=1}^{n} n w_i, \tag{3.8}$$

over the set of constraints on weights w_i,

$$\left\{ w_i \geq 0, \sum_{i=1}^{n} w_i m(X_i, \theta), \sum_{i=1}^{n} w_i = 1 \right\}, \tag{3.9}$$

for $i = 1, 2, \ldots, n$.

A valuable advice is to use the simplex duality of the empirical likelihood discussed in Chapter 2 and evaluate the EL function by the minimization problem of the Lagrange multipliers to save the computation time. The weighted sampling scheme is summarized in Algorithm 2.

Data: data sample $\mathbf{x}$, prior distribution $\pi(\theta)$
for $i = 1$ *to* N **do**
  Generate $\theta' \sim \pi(\theta)$;
  Calculate the empirical likelihood ratio

$$\mathcal{R}(\theta') = \left\{ \max \prod_{i=1}^{n} n w_i \Big| w_i \geq 0, \sum_{i=1}^{n} w_i m(X_i, \theta'), \sum_{i=1}^{n} w_i = 1 \right\} \quad (3.10)$$

  Save $\theta_i = \theta'$ and it's corresponding weight $\xi_i = \mathcal{R}(\theta')$;
end

Algorithm 2: Empirical likelihood weighted sampler

The proposed algorithm scheme produces a series of pairs. Each pair consists of a sample θ_i form the prior distribution $\pi(\theta|\mathbf{x})$ and is assigned its respective weight ξ_i. These pairs of samples and their respective weights (θ_i, ξ_i) for $i = 1, \ldots N$ can be viewed in a similar manner as the output of a general importance sampling Monte Carlo algorithm.

For the empirical likelihood ratios $(\xi, \ldots, \xi_N)$ to form a proper importance sampling weights as in Monte Carlo importance sampling they need to be normalized. This can be easily achieved by a self normalization

$$\xi_i^* = \frac{\xi_i}{\sum_{i=1}^{N} \xi_i}. \quad (3.11)$$

Although this renormalization leads to a biased estimate, with large sample sizes this bias is negligible, since it is asymptotically unbiased and consistent. Therefore leading us a a weighted parameter estimates that can be obtained by

$$E_{\hat{\pi}}[\theta] = \sum_{i=1}^{N} \xi_i^* \theta_i, \quad (3.12)$$

for the mean value, and subsequently for the variance by

$$Var_{\hat{\pi}}[\theta] = \sum_{i=1}^{N} \xi_i^* \theta_i^2 - \left(\sum_{i=1}^{N} \xi_i^* \theta_i^2 \right)^2. \quad (3.13)$$

Importance resampling

As an possible improvement of the original sampler from Algorithm 2 we propose a modification by the employment of the importance resampling known from the sequential Monte Carlo methods. The disadvantage of the previous algorithm is that it only weights samples coming from the prior distribution and this can lead to situation when all the information about the posterior distribution is located in a small number of samples. Generally the preferred output of an importance sampling algorithm should be a set of samples with weights as evenly distributed as possible.

The measure of a uniformity can be expressed in the terms of the mathematical Shannon entropy

$$H(x) = -\sum_{i=1}^{N} p(x)\log p(x) \tag{3.14}$$

where the term $p(x)$ denotes the probability of observing a value x. In terms of the previous algorithm we would say that the observed values x form the sampled values θ and the normalized importance weights ξ^* are the probabilities $p(x)$. Therefore the output of the Algorithm 2 produces a sample with a low values of entropy, i.e. the measure of uncertainty is small due to the fact that all the information is densely placed on a small number of values.

As a possible remedy to this issue we propose to employ the concept of importance resampling, which can be summarized as follows.

1. Obtain pairs $((\theta_i, \xi_i))_{i=1}^N$ from Algorithm 2.
2. From the previously obtained samples $(\theta_1, \ldots, \theta_N)$ resample M values based on the multinomial distribution with probabilities equal to the normalized weights $(\xi_1^*, \ldots, \xi_N^*)$.
3. Add a white noise to the resampled values from previous step to eliminate multiply chosen samples and run the second and third step until a desired quality of the weights is reached.

The resampling technique leads to an improved quality of the resulting weights. The criterion of reaching the desired result can be for example entropy from Equation 3.14 or the measure of *effective sample size* ESS of weights

$$ESS = 1/\sum_{i=1}^{M}\{\xi_i^*/\sum_{j=1}^{M}\xi_j^*\}^2. \tag{3.15}$$

The ESS takes values between 1 and M. A fully degenerated result placing a weight 1 to one sample and zero to all other has ESS equal to 1. The maximum of M is reached for a set of uniform weights.

3.3.2 Markov chain Monte Carlo sampling

In contrast to the previous sampling scheme where the output was a pairs of samples with corresponding weights, the subsequently proposed algorithm produces a sample directly from the empirical likelihood posterior distribution $\pi_{el}(\theta|\mathbf{x})$ which was defined in Equation 3.6.

Recall that in the section 3.2 we have stressed that the term $\int_\Theta \mathcal{R}(\theta)\pi(\theta)\mathrm{d}\theta$ from the denominator of Equation 3.6 can be very difficult to evaluate if not intractable at all. However this issue of evaluating the denominator of the Bayes' theorem can cause problems even when the parametric likelihood $f(\mathbf{x}|\theta)$ is considered.

For these settings it is appropriate to employ the so called Markov chain Monte Carlo (MCMC) methods, which can sample directly from the desired probability distribution even when it is known only up to a normalizing constant. This work is not intended to cover the theory and various applications of the MCMC methods, which have been already reviewed and described extensively. A good reference that covers this topic can be found in the article by Brooks [21] and for subsequent tunning of the algorithms the review of Lewis et al. [22] is recommended. For our purposes we will make use of the Metropolis-Hastings algorithm and further modify it to utilize the empirical likelihood instead of the classical parametric likelihood.

We propose the modified Metropolis-Hastings MCMC sampler via the empirical likelihood, by using the EL version of Bayes' theorem (3.6). The proposed algorithm can be summarized in the following steps shown in (3).

Data: data sample $\mathbf{x}$, prior distribution $\pi(\theta)$, Markov kernel $q(\theta \to \theta')$
Start at θ ;
for $i = 1$ *to* N **do**
  If at θ, propose a move to θ' according to $q(\theta \to \theta')$;
  Calculate

$$h = \min\left(1, \frac{\mathcal{R}(\theta')\pi(\theta')q(\theta' \to \theta)}{\mathcal{R}(\theta)\pi(\theta)q(\theta \to \theta')}\right), \tag{3.16}$$

  Sample u from $\mathcal{U}_{[0,1]}$;
  if $u < h$ **then**
    Save $\theta_i = \theta'$ and weight $\omega_i = \mathcal{R}(\theta')$;
    Set $\theta = \theta'$;
  else
    Remain at θ;
  end
end

Algorithm 3: MCMC via Empirical likelihood

3.4 Inference on time series

Up to this point the proposed sampling algorithms assumed that the observed data sample $\mathbf{x} = (X_1, X_2, \ldots, X_n)$ was independent and identically distributed. It was already stressed in previous chapters that when treating a dependent data set as if it was iid, in the empirical likelihood framework,can lead to defected confidence regions and invalid coverage properties.

Depending on the character of the dependence among the observed data and an assumption made about the underlying model generating them, the proposed approaches will either make use of the mode-based empirical likelihood introduced in Subsection 2.4.1 or employ the block-wise empirical likelihood framework from Subsection 2.4.2.

3.4.1 Model based sampler

With the model-based empirical likelihood the approach differs from the previously presented samplers in the construction of the estimating equations tying the observed time series sample $\mathbf{x} = (X_1, X_2, \ldots, X_n)$ to the parameter of interest θ.

As an example consider the autoregressive model of order p

$$X_t = \sum_{i=1}^{p} \psi_i X_{t-i} + e_t \tag{3.17}$$

where we are interested in the inference on the regression parameter $\psi = (\psi_1, \ldots, \psi_p)$. To employ the sampling algorithms designed for iid data, take for instance the weighted sampler from Algorithm 2, the estimating equations need to be reformulated in terms of the independent and uncorrelated shocks $(e_1, \ldots, e_n)$. More generally the estimating equations need to form a martingale difference array as shown in [11], which is satisfied naturally by assuming the shocks e_t are independent and have a common variance.

Therefore the estimating equations need to be formulated in a way as was shown by Equation 2.37, placing

$$\sum_{t=p+1}^{n} (X_t - \psi' \mathbf{X}_{t-1}) \mathbf{X}_{t-1} = \sum_{t=p+1}^{n} m_t(X_t, \mathbf{X}_{t-1}, \psi) \tag{3.18}$$

where $\mathbf{X}_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})$. Note that the number of estimating functions is smaller than in the independent case, where the there is as many estimating functions as the number of observations, i.e. every observation accounts for one estimating function. With autoregressive model AR(p) which includes regressors of p time steps we are only left with $n - p$ estimating functions. This factor however does not have a significant impact with a larger number of samples.

Naturally we can employ the construction of model-based EL concept in the MCMC sampler via empirical likelihood as presented in Algorithm 3.

The idea of using the iid version of empirical likelihood in the sampler and let the estimating functions suppress the dependence is useful. However this case can only be applied to models where we are able to reformulate the estimating functions so that they form a martingale difference array, as it is possible with the AR model. This can be limiting in applications, since for some time series we are not able to apply the model-based approach.

3.4.2 Block-wise empirical likelihood sampler

In situations when the model-based approach cannot be applied or is not feasible for the given inference problem, another possibility is to limit our selves to time series classes by making an assumption about the weak dependence among observations in terms of the mixing conditions from Section 2.4.2.

For weakly dependent time series the block-wise empirical (BEL) likelihood provides a valid asymptotic behaviour and we propose to use it as a replacement of the classical EL.

As discussed previously the BEL approach requires some tuning of the parameters such as the block length l and the block separation m. Little is theoretically known about the optimal values even though some theoretical approaches have been suggested, for instance Kitamura [5],[6] or Owen [3], but most of the time the estimation of the optimal values is left to analysis of the coverage properties for a given problems character.

For the weighted sampler (2), we propose to use the block-wise empirical likelihood ratio instead of the EL weight calculated in Equation 3.10

$$\mathcal{R}_b(\theta) = \max\left\{ \prod_{i=1}^{n} n w_i \,\Big|\, \sum_{i=1}^{n} w_i b(B_i, \theta) = 0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}, \qquad (3.19)$$

where the block-wise estimating function $b(B_i, \theta)$ is constructed by joining every m-th l neighboring observations from the time series $\mathbf{x}$ as

$$b(B_k, \theta) = \frac{1}{l} \sum_{i=1}^{l} m(X_{(k-1)m+i}, \theta). \qquad (3.20)$$

For the Bayesian inference the optimal block length is still a crucial parameter to tune, however the correction factor accounting for the overlap of blocks in the Theorem 2.5 is no longer required in the proposed Bayesian samplers. To show the point, the normalizing constant ensuring the block-wise EL ratio to have the proper chi-square limit distribution is not longer required since the weights are later self normalized to provide a Monte Carlo importance sampling - like weights.

The EL Markov chain Monte Carlo sampler introduced in Algorithm 3 can be modified to use the block-wise EL function as well by changing the EL $\mathcal{R}(\theta)$ to $\mathcal{R}_b(\theta)$. The probability of accepting the proposed sample θ' from Algorithm 3 then can be written as

$$h = \min\left(1, \frac{\mathcal{R}_b(\theta')\pi(\theta')q(\theta' \to \theta)}{\mathcal{R}_b(\theta)\pi(\theta)q(\theta \to \theta')}\right). \tag{3.21}$$

It is important to keep in mind however, that when using the block-wise empirical likelihood ratio in the Algorithm 3 the resulting sample of parameters $(\theta_1, \theta_2, \ldots, \theta_N)$ does not come from the EL posterior distribution as defined by the Equation 3.6 bur rather from its block-wise version which we denote by $\pi_{bel}(\theta|\mathbf{x})$ and can be expressed via the Bayes' theorem as

$$\pi_{bel}(\theta|\mathbf{x}) = \frac{\mathcal{R}_b(\theta)\pi(\theta)}{\int_\Theta \mathcal{R}_b(\theta)\pi(\theta)\mathrm{d}\theta} \propto \mathcal{R}_b(\theta)\pi(\theta). \tag{3.22}$$

Similarly as with the weighted sampler, the correction accounting for the block overlap from Theorem 2.5 ensuring valid chi-square distribution is not necessary. Since in each step of the MCMC algorithm, only the relative change between the newly proposed sample θ' and the current sample θ is relevant for the calculation of the acceptance probability in Equation 3.21.

Chapter 4

Data examples

This chapter is dedicated to present examples of the various methods of inference on time series proposed in the previous chapters. In the first part we provide examples of the block-wise empirical likelihood as a method of statistical inference for weakly dependent time series and mainly its applications in the Bayesian setting. The second part is devoted to the examples of the model-based empirical likelihood inference on the autoregressive model.

4.1 Block-wise Bayesian empirical likelihood

At first we demonstrate the use of the empirical likelihood for time series employing the block-wise EL for weakly dependent data. For the weakly dependent class of time series it is assumed that the correlation between observations far apart is negligible.

As the illustrative example we will use the autoregressive moving average model presented in Chapter 1. The following examples will demonstrate an inference on the process mean, denoted by μ. The considered model is the centered ARMA(2,2) process, where X_t forms a regression of its two previous values and a regression of 2 previous shocks plus a current shock.

$$X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \tag{4.1}$$

where the disturbances e_t are simulated as white noise from $\mathcal{N}(0, \sigma^2)$. Time steps $t = 1, 2, \dots T$ are considered. The initial configuration of parameters is summarized in Table 4.1.

Note that the parameter values from Table 4.1 of the ARMA(2,2) process are chosen specifically so that the roots of the autoregressive characteristic polynomial $\psi(B) = (1 - 0.3B - 0.5B^2)$ lie outside of the unit circle. This implies that the conditions of stationarity from Chapter 1 are met. This is crucial for the further use of the block-wise EL to satisfy the assumptions stated in the Theorem 2.5. The graphical representation of the time series is

Parameter	ψ_1	ψ_2	θ_1	θ_2	T	σ^2	$\bar{X}$	$\hat{\sigma}$
Value	0.3	0.5	-0.5	0.1	100	1.0	0.632	1.272

Table 4.1: Parameters of ARMA(2,2) used to demonstrate the inference on process mean μ by the Bayesian samplers using the block-wise empirical likelihood.

depicted in Figure 4.1. In the following presentation all of the examples in this section will be performed on this time series.

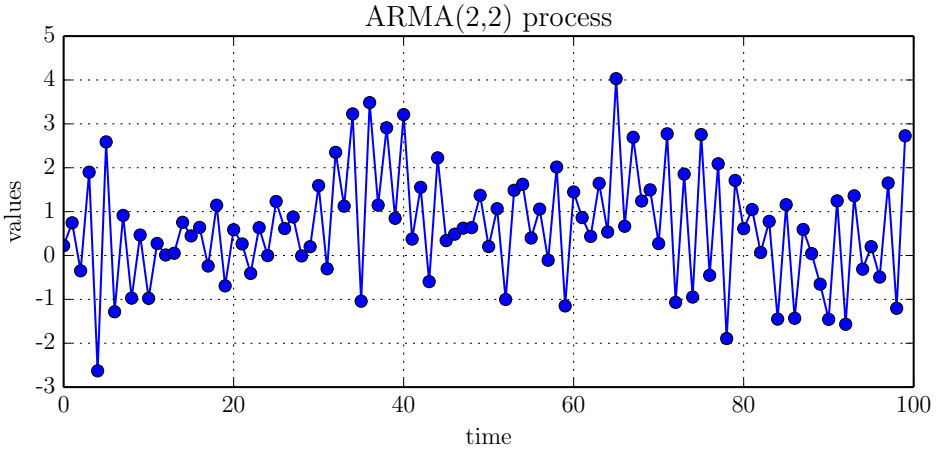

Figure 4.1: Graphical representation of the ARMA(2,2) process which will be used to provide examples of the inference using the block-wise empirical likelihood.

4.1.1 Block size calibration

The construction of the block-wise empirical likelihood in contrast with the standard empirical likelihood requires the setup of two additional parameters. To formulate the blocked estimating equations (3.20) tying the parameter of interest to the observations

$$b(B_k, \mu) = \frac{1}{l} \sum_{i=1}^{l} m(X_{(k-1)m+i}, \mu), \tag{4.2}$$

the block length l and the separation of the blocks m need to be set. In Chapter 2 we discussed that there is no generally accepted theoretical guidance to choose the block length. Usually the choice is made by performing a coverage analysis for a specific type of problem.

A common practice [3],[8] is to construct and perform analysis of the coverage rates for a various block lengths and choose the one with the best

coverage properties. For the coverage analysis, we perform $n = 2000$ separate simulations for blocks of lengths $l = 2, 3, \dots 40$. The coverage ratios are calculated for a 95% confidence interval, based on the block-wise empirical likelihood Theorem 2.5. The relation between coverage rate and block size is shown in Figure 4.2.

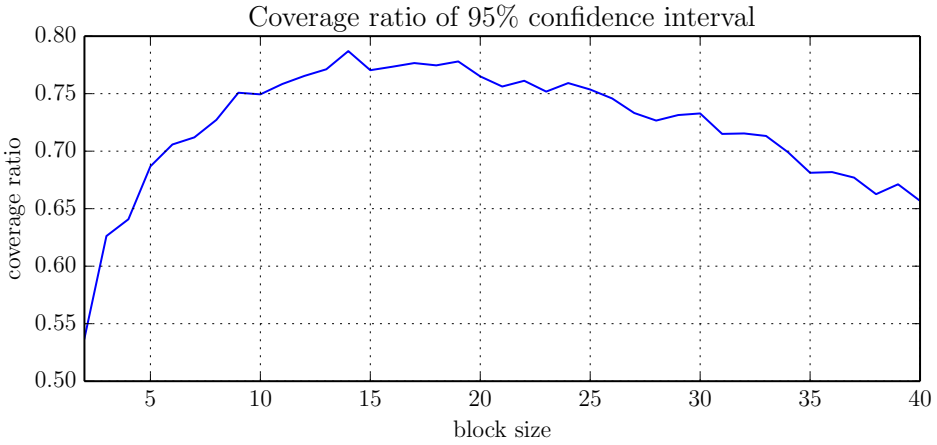

Figure 4.2: Figure showing the 95% coverage ratio of the true mean of ARMA(2,2) process as a function of block length in block-wise EL ranging from 2 to 40 blocks.

Based on the Figure 4.2 we can deduct that the best coverage is accomplished with block lengths set between 10 and 20 samples blocked into one estimating function $b(B_k, \mu)$. We proceed to further analysis with a block size set to 15 observations. In contrast to the block length the choice of block separation does not appear to be so crucial for the coverage properties and we will use the maximally overlapping blocks, i.e. setting the block separation to $m = 1$.

4.1.2 Block-wise empirical likelihood

Having chosen the block length and the separation among blocks, we can proceed with the inference on the process mean μ using the block-wise empirical likelihood function (BEL). To construct the profile of BEL ratio function

$$\mathcal{R}_b(\mu) = \max \left\{ \prod_{i=1}^{n} n w_i \,\Big|\, \sum_{i=1}^{n} w_i b(B_i, \mu) = 0, w_i \geq 1, \sum_{i=1}^{n} w_i = 1 \right\}, \qquad (4.3)$$

the value $\mathcal{R}_b(\mu)$ needs to be evaluated for a span of theoretical values of the process mean. The BEL ratio curve is plotted in Figure 4.3. The dashed

blue horizontal line depicts the 95% confidence interval and the means values falling into the confidence interval are highlighted in red.

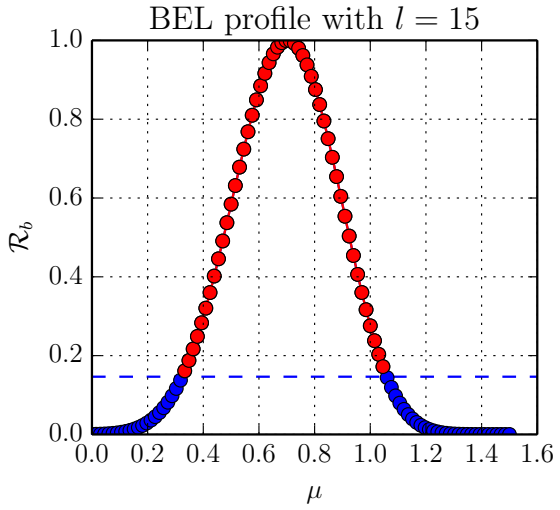

Figure 4.3: Profile of the block-wise empirical likelihood function as a value of the process mean. Blue dashed line separates the values above and below the 95% confidence level.

The overview of the estimates based on the BEL inference is summarized in Table 4.2. The maximum BEL estimate is obtained in a similar fashion as is done in parametric statistics. The mean and standard error as calculated by using the BEL as probabilities assigned to each parameter.

	$\max\mathcal{R}_b(\mu)$	mean	standard error	95% Confidence Interval
μ	0.701	0.695	0.185	$[0.324, 1.055+]$

Table 4.2: Inference on the process mean μ via the block-wise empirical likelihood with block length $l = 15$. With confidence regions constructed from Theorem 2.5

To indicate the importance of the correct block length selection, the profiles of BEL functions for two extreme cases of the block lengths are shown in Figure 4.4. Comparing the BEL profiles for block lengths 2 and 40, which proved to have a poor coverage properties (see Figure 4.2), we can deduce that the block size can rapidly degenerate the confidence regions. It can be also seen that the maximum BEL estimate is significantly shifted in Figure 4.4 compared to Figure 4.3.

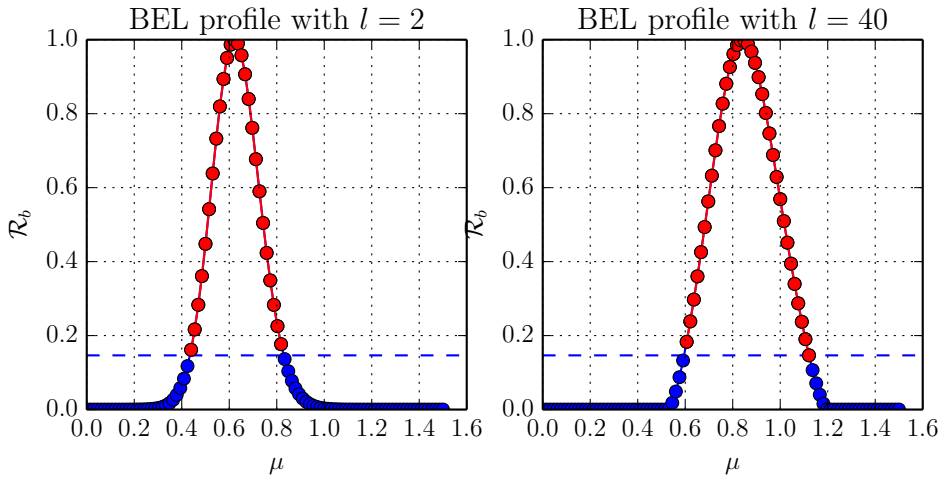

Figure 4.4: Block-wise empirical likelihood profile function depicted for the two extreme cases of block length selection. Left figure shows block length $l = 2$ and the right figure shows $l = 40$. The values falling into the 95% confidence interval are depicted in red colour.

4.1.3 Weighted Bayesian sampler

In this section we move forward and employ the Bayesian framework. We start by presenting the results of the Bayesian sampler that uses the block-wise EL function as the importance weights. This method was proposed in Section 3.4.2.

First we need to place a prior distribution $\pi(\mu)$ on the process mean. A reasonable candidate for the prior is the Normal distribution with center at the sample mean and variance equal to the sample variance.

$$\pi(\mu) = \mathcal{N}(\overline{\mathbf{x}}, s^2) = \mathcal{N}(0.632, 1.272^2). \tag{4.4}$$

To employ the Bayesian sampler with BEL function used as the importance weights, we draw μ_k for $k = 1, 2, \ldots N$, where $N = 2000$, samples from the prior (4.4) and assign each sample with weight equal to $\xi_k = \mathcal{R}_b(\mu_k)$. Figure 4.5 depicts the drawn samples $(\mu_1, \ldots, \mu_N)$ with their respective weights.

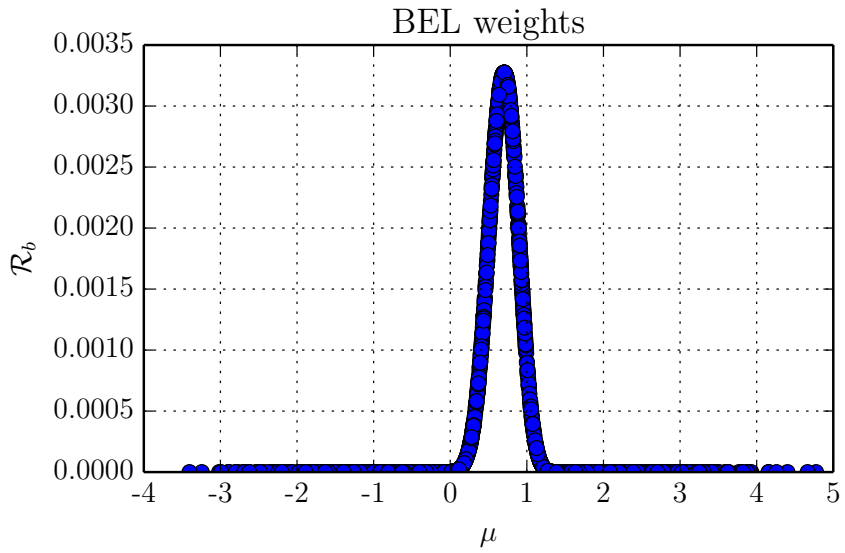

Figure 4.5: Bayesian sampler using the bock-wise empirical likelihood as weights. The figure shows 2000 samples after one run of the algorithm.

From the Figure 4.5 it can be seen that a large number of samples is placed in areas with relatively low BEL weight. To improve this, we employ a single run of the multinomial importance resampling introduced in Section 3.3.1. We proceed by resampling $(\mu_1^*, \ldots, \mu_M^*)$ where $M = 2000$ samples from the previous samples $(\mu_1, \ldots, \mu_N)$ based on the multinomial distribution with probabilities equal to ξ_k^*, where

$$\xi_k^* = \frac{\xi_k}{\sum_{i=1}^N \xi_i} \tag{4.5}$$

are self normalized BEL weights from the first run of the sampler.

After resampling the ξ_i^* samples we recalculate the BEL weights again. The samples after one run of the importance resampling with recalculated respective weights are depicted in Figure 4.6.

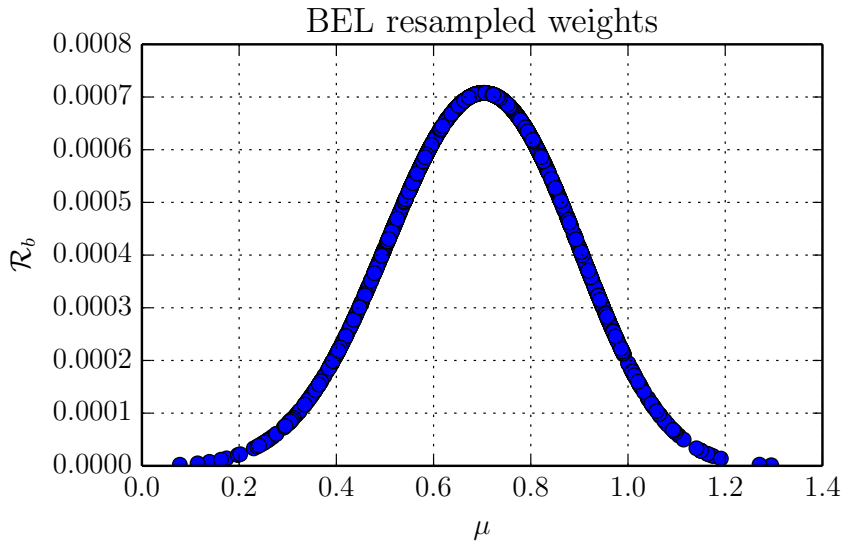

Figure 4.6: Samples after one run of multinomial resampling with weights equal to the self normalized BEL weights from the first run of the algorithm with recalculated BEL weights.

The estimates summary of the Bayesian block-wise EL sampler are presented in Table 4.3. The mean and standard error estimates are calculated by the Equations 3.12 and 3.13 respectively. To demonstrate the effect of the multinomial importance resampling the value of the effective sample size (ESS from Equation 3.15) is included. We can see that the ESS significantly increased after a single run of the importance resampling procedure.

	run	mean	std. error	ESS
μ	first	0.696	0.034	426
μ	second	0.704	0.019	1743

Table 4.3: Parameter estimates using the block-wise EL as weights in the Bayesian sampler. The first row shows the results after initial weighting of samples drawn from the prior distribution. The second row provides results after employing one run of multinomial resampling. The effective sample size (ESS) measure suggests a strong improvement after the application of the resampling.

4.1.4 Markov chain Monte Carlo Bayesian sampler

As a next example of the Bayesian inference on the process mean via block-wise empirical likelihood we provide the Markov chain Monte Carlo sampler.

In contrast to the previous sampler using the block-wise EL as importance weights, the MCMC algorithm samples directly from the BEL posterior distribution

$$\pi_{bel}(\mu|\mathbf{x}) = \frac{\mathcal{R}_b(\mu)\pi(\mu)}{\int_\Theta \mathcal{R}_b(\mu)\pi(\mu)\mathrm{d}\mu} \propto \mathcal{R}_b(\mu)\pi(\mu). \tag{4.6}$$

For the prior $\pi(\mu)$ placed on the process mean we will use the same distribution (4.4) as in the previous example.

We run the MCMC algorithm to draw $N = 10000$ samples from the posterior distribution $\pi_{bel}(\mu|\mathbf{x})$. As for any MCMC sampler it is critical that the acceptance criterion does not get stuck by rejecting the proposed samples of μ for a longer time periods. The Figure 4.7 shows the evolution of the drawn samples.

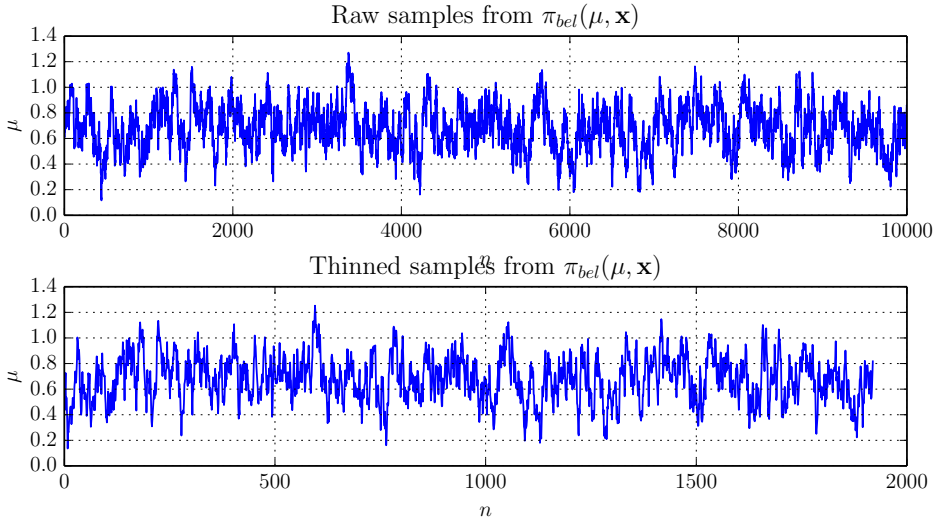

Figure 4.7: Evolution of accepted samples drawn from the posterior distribution $\pi_{bel}(\mu|\mathbf{x})$ using the MCMC sampler with block-wise empirical likelihood used as the likelihood function.

When using the MCMC algorithm to sample directly from the posterior distribution, the resulting samples are valid after the Markov chain has reached its stationary distribution. The period until stationarity is reached, is called the burn-in period. This is the cost of any MCMC algorithm and usually the first samples are not used, in our case we omitted the first 500 samples. Another cost of MCMC algorithms is the autocorrelation among the samples. This can be seen by plotting the autocorrelation function and the partial autocorrelation function shown in Figure 4.8. A common approach to treat the autocorrelation is to thin down the resulted samples and take only every n-th sample. We have used thinning by 5 samples.

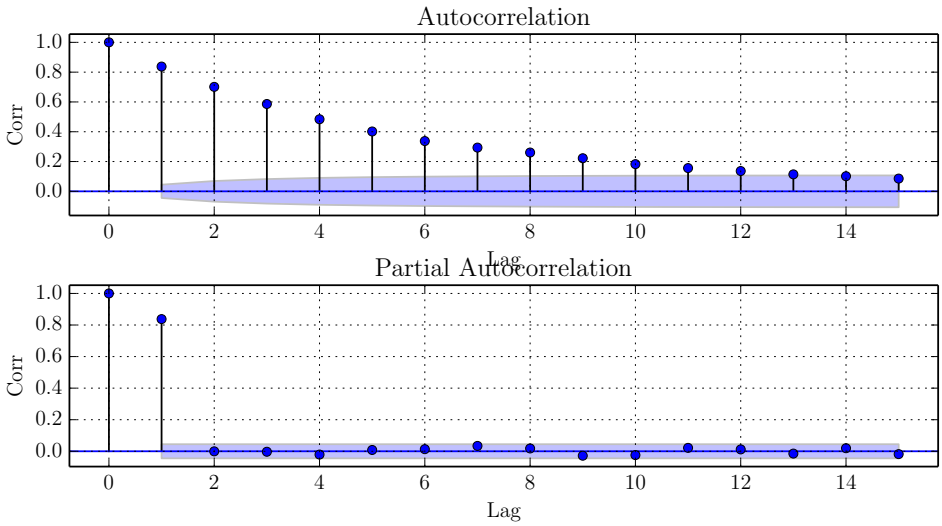

Figure 4.8: Autocorrelation function and partial autocorrelation function of samples generated by MCMC algorithm.

Since the MCMC algorithm samples directly from the posterior distribution, we can construct the histogram of the samples as shown in Figure 4.1.4. To get a sense of how the observed data from the ARMA process influence the prior information about process mean $\pi(\mu)$, the prior distribution is added to the twined y axis on the left side of the histogram.

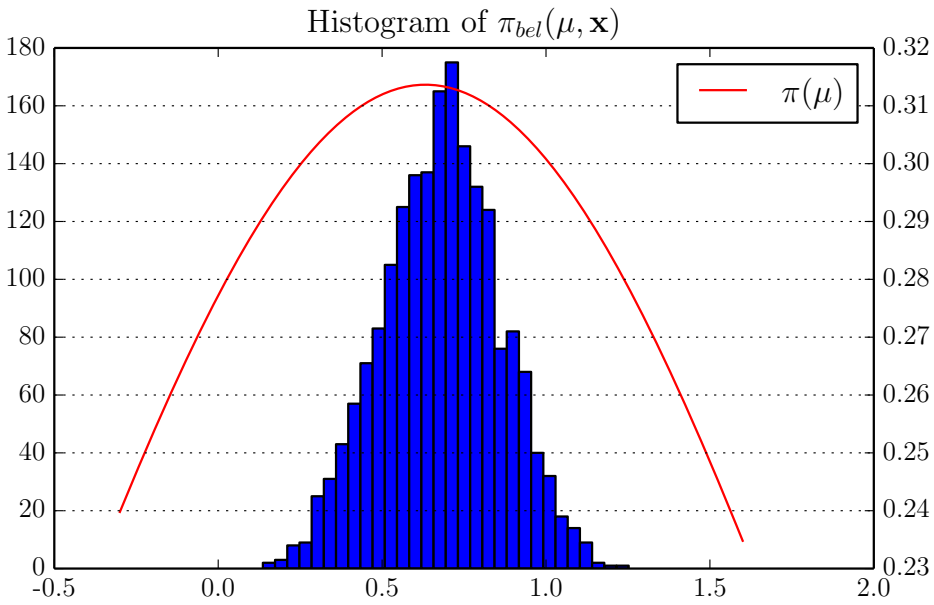

Figure 4.9: Histogram of thinned samples from the posterior distribution obtained by the BEL MCMC sampler. The red line shows the prior information placed on the process mean.

The parameter estimates using the Markov chain Monte Carlo sampler via the block-wise EL are summarized in Table 4.4.

	mean	std. error
μ	0.680	0.179

Table 4.4: Statistics calculated based on the samples from the posterior distribution $\pi_{bel}(\mu|\mathbf{x})$, provided by the MCMC sampler.

The mean estimates obtained via the methods demonstrated in this section of the ARMA(2,2) from Equation 4.1 and their respective standard error as summarized in the overview Table 4.5.

Method	μ *estimate*	*Std. error*
Sample estimates	0.632	1.272
Block-wise EL	0.695	0.185
max of Block-wise EL	0.701	-
Bayesian BEL weighted sampler	0.696	0.034
Bayesian BEL re-weighted sampler	0.704	0.019
Posterior $\pi_{bel}(\mu\|\mathbf{x})$ via MCMC	0.680	0.179

Table 4.5: Estimates overview of the mean of the ARMA(2,2) process by the methods used in this section. The maximum block-wise empirical likelihood does not provide the estimate of the standard error.

4.2 Model-based empirical likelihood for time series

This section is intended to demonstrate the inference of the model-based empirical likelihood for time series. As the example, let us consider an autoregressive model of the second order introduced in Chapter 1.

$$X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + e_t, \tag{4.7}$$

for $t = 1, \ldots, T$ where e_t denotes a white noise random shock at time t with distribution $N(0, \sigma^2)$. The model parameters used in the examples are summarized in Table 4.6.

Parameter	ψ_1	ψ_2	T	σ^2
Value	0.4	-0.5	500	1.0

Table 4.6: Configuration of the AR(2) model parameters. This is used in the following examples of inference on the parameters ψ_1 and ψ_2.

This considered model configuration is analyzed in all the examples provided later on in this section. The graphical visualization of the time series is depicted in Figure 4.10.

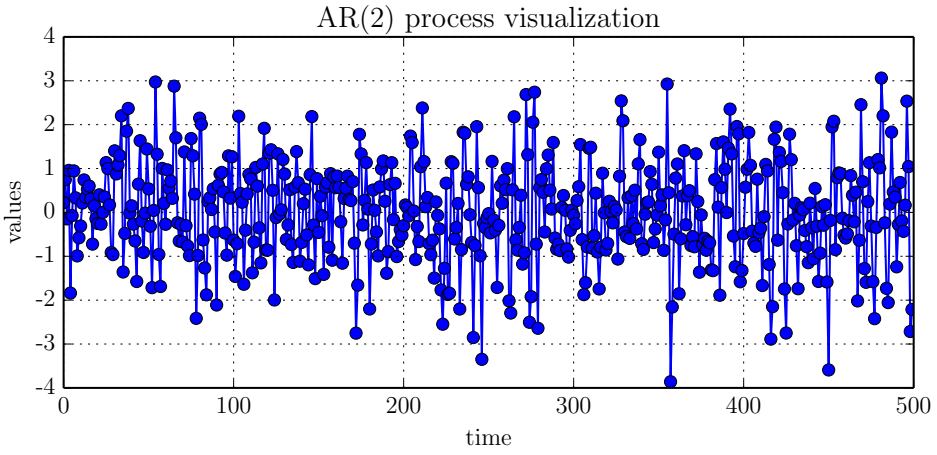

Figure 4.10: Visualization of the autoregressive model of second order with the configuration of parameters provided in Table 4.6.

For the latter comparison of results obtained via various methods of statistical inference. We provide a common statistical estimation of parameters obtained by the maximum unconditional likelihood method. The results of this model fit are summarized in Table 4.7.

parameter	*estimate*	*standard error*	95% Confidence Interval	p-value
ψ_1	0.390	0.039	$[0.313, 0.467]$	0.000
ψ_2	-0.489	0.039	$[-0.566, -0.412]$	0.000

Table 4.7: Results of the parametric inference on the regression parameters of the autoregressive model by fitting the unconditional maximum likelihood.

4.2.1 Model-based empirical likelihood inference

To proceed further with the examples, we provide an inference done by the model-based empirical likelihood introduced in Section 2.4.1. A key step to apply the model-based EL is to correctly formulate the estimating functions

$$m(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-p}; \psi) = e_t \tag{4.8}$$

in terms of the independent shocks e_t for $t = p+1, \ldots T$. As previously shown, for the example of an AR(2) process, the estimating functions should be formulated in the form

$$m(X_t, X_{t-1}, X_{t-2}; \psi_1, \psi_2) = \{X_t - \sum_{i=1}^{2} \psi_i X_{t-i}\}(X_{t-1}, X_{t-2})' \in \mathbb{R}^2. \tag{4.9}$$

For times $t = 3, \ldots 500$. The estimating functions for $t = 1, 2$ are set to zero.

The model-based EL as a function of parameters ψ_1 and ψ_2 is depicted in Figure 4.11. The parameter values falling into the 95% confidence interval, based on the EL Theorem 2.2, are distinguished in red color and the confidence level are shown as dashed blue lines.

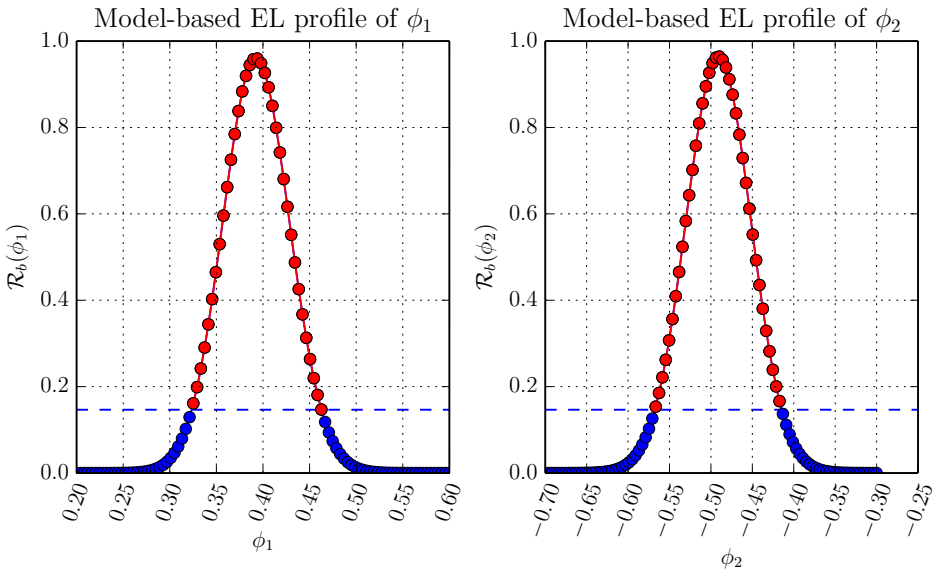

Figure 4.11: Model-based empirical likelihood profile for 2 parameters of AR(2) process. Red colour highlights the values in the 95% confidence interval.

Solely based on the values of the model-based EL function we can construct the maximum EL estimate. The mean and the standard error estimate is obtained from the Equations 3.12 and 3.13 respectively. The overview of the estimates is provided in Table 4.8 with the addition of the confidence intervals for both parameters.

parameter	$\max\mathcal{R}(\psi)$	*mean*	*standard error*	95% Confidence Interval
ψ_1	0.394	0.393	0.036	$[0.325, 0.463]$
ψ_2	-0.490	-0.491	0.040	$[-0.567, -0.417]$

Table 4.8: Estimates and statistics after inference on AR(2) process via the block-based empirical likelihood.

4.2.2 Bayesian weighted sampler

The following examples of this section we will demonstrate two algorithms, both applying the Bayesian approach. Firstly we have to place a prior distribution of both unknown parameters. We apply an independent expectation on ψ_1 and ψ_2 in terms of the Normal distribution

$$\begin{aligned} \pi(\psi_1) &= \mathcal{N}(0.394, 0.5^2), \\ \pi(\psi_2) &= \mathcal{N}(-0.491, 0.5^2). \end{aligned} \tag{4.10}$$

For both parameters we apply the same standard deviation of 0.5 and place the center of the expectation at the estimate obtained by the model-based EL inference from Table 4.8.

Using the Bayesian weighted sampler we simultaneously sample $N = 2000$ samples from both prior distributions and for each pair of samples (ψ_1^i, ψ_2^i) calculate the corresponding EL weights $\mathcal{R}(\psi_1^i, \psi_2^i)$ (the upper index i denotes the order in the sampling procedure). The samples of both parameters with their respective weights are shown in Figure 4.12.

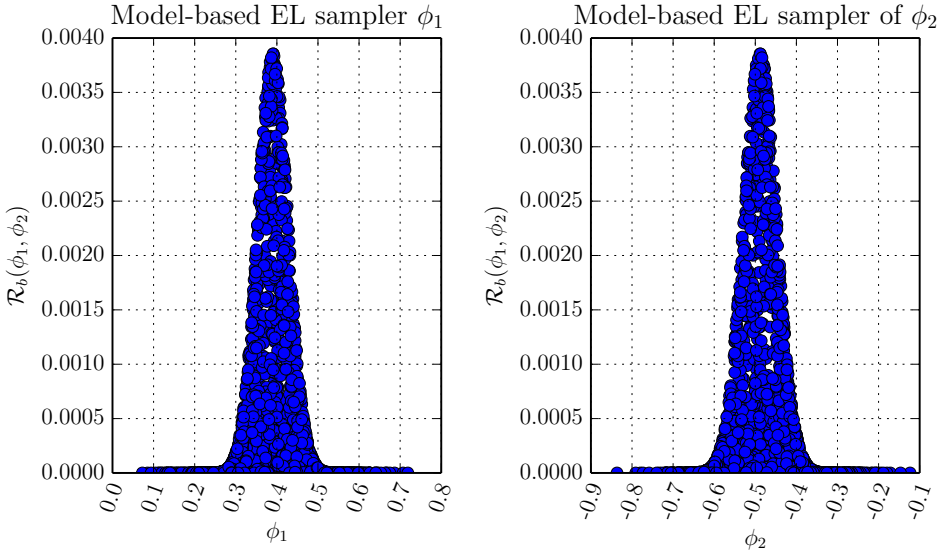

Figure 4.12: Weights calculated by the model-based empirical likelihood Bayesian sampler for 2000 samples from the prior distributions of ψ_1 and ψ_2.

The parameter estimates obtained by the Bayesian sampler using the model-based EL as the importance weights, are summarized in Table 4.9.

parameter	*mean*	*standard error*	*ESS*
ψ_1	0.389	0.002	479
ψ_2	-0.485	0.002	479

Table 4.9: Parameter estimates based on the Bayesian weighted sampler with the model-based empirical likelihood.

Similarly as has been shown with the block-wise empirical likelihood weighted Bayesian sampler in Section 4.1.3 we can significantly improve the results of the sampler by applying the multinomial importance resampling. However, now we will use the same self normalized weights $\mathcal{R}(\psi_1^i, \psi_2^i)$ for one given sample pair (ψ_1^i, ψ_2^i). The improved samples with corresponding EL weights are shown in Figure 4.13. The parameter estimates obtained after employment of the importance resampling are provided in Table 4.10.

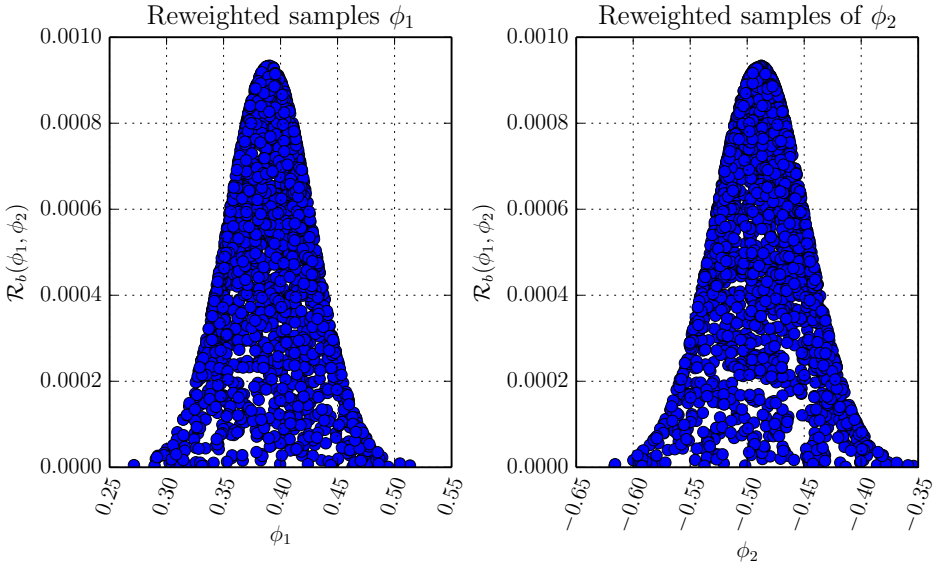

Figure 4.13: Model-based weights after employing one run of the importance resampling of data shown in Figure 4.12 using the multinomial distribution with weights as self normalized EL ratios.

parameter	*mean*	*standard error*	*ESS*
ψ_1	0.391	0.001	1552
ψ_2	-0.487	0.001	1552

Table 4.10: Parameter estimates after a single run of multinomial resampling.

By comparing the effective sample size from Tables (4.9) and (4.10) we can see that the effectiveness of the output has significantly improved.

4.2.3 MCMC Bayesian sampler

As a conclusion on the examples showing the performance of the model-based EL samplers, we provide the results of the Markov chain Monte Carlo sampler. As the prior information placed on the parameters ψ_1 and ψ_2 we take the one suggested in previous section by Equation 4.10.

We sample $N = 10000$ samples from the posterior distribution

$$\pi_{el}(\psi_1, \psi_2|\mathbf{x}) = \frac{\mathcal{R}(\psi_1, \psi_2)\pi(\psi_1)\pi(\psi_2)}{\int_{\psi_1}\int_{\psi_2}\mathcal{R}(\psi_1, \psi_2)\pi(\psi_1)\pi(\psi_2)\mathrm{d}\psi_1\mathrm{d}\psi_2}. \tag{4.11}$$

The raw and thinned samples from the posterior distribution are depicted in Figure 4.14. With the burnin period set to 400 samples. The thinning of the Markovian chain is done by taking every 5th sample which sufficiently suppresses the autocorrelation among the resulting samples.

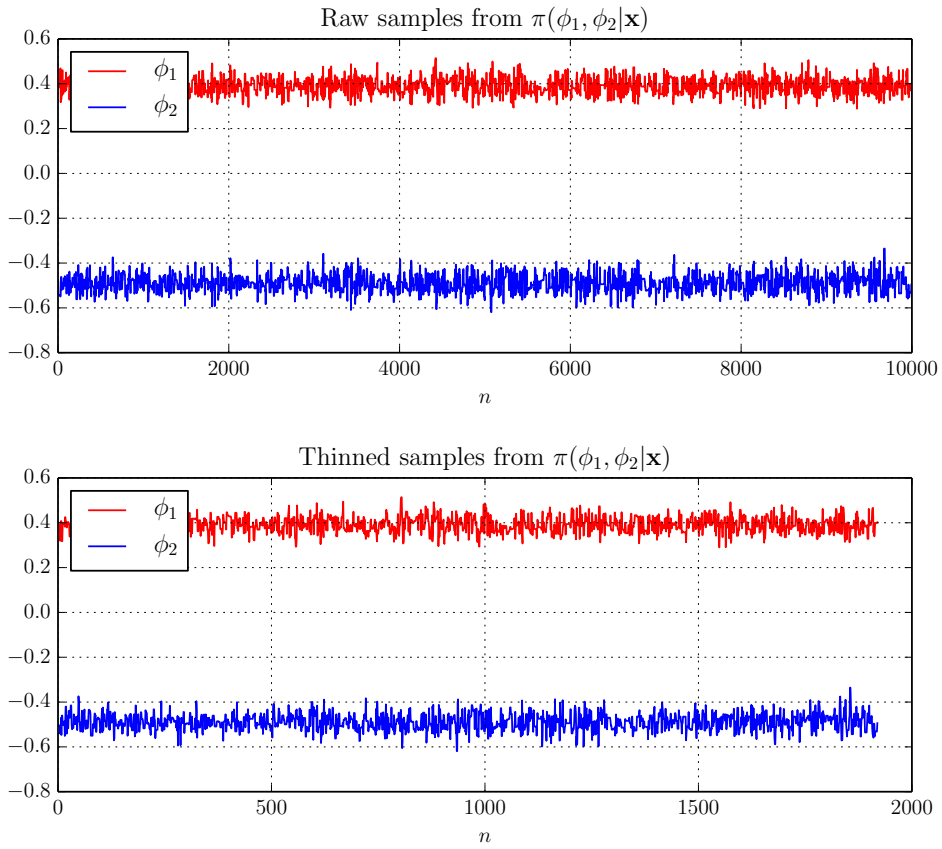

Figure 4.14: Evolution of samples drawn from the posterior distribution from Equation 4.11 by the MCMC sampler. The top row shows raw samples and the bottom row shows clean samples after cutting out the burnin period and thinning of the chain.

To provide a visual interpretation of the samples from the posterior distribution the histograms for each parameter is shown in Figure 4.15.

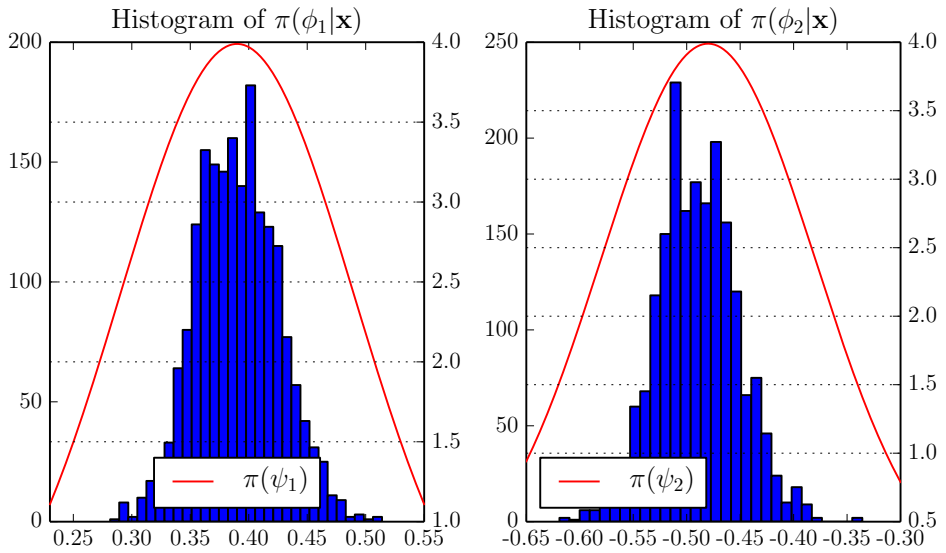

Figure 4.15: Histogram of thinned samples from the posterior distribution obtained by the MCMC algorithm using the model-based EL as the likelihood function. The red line in the figures shows the prior distribution of $\pi(\psi_1)$ and $\pi(\psi_2)$ respectively as a comparison to the posterior distribution.

The parameter estimates based on the samples from the posterior distribution (4.11) are summarized in Table 4.11.

	mean	*Standard error*
ψ_1	0.391	0.035
ψ_2	-0.490	0.036

Table 4.11: Estimates of the parameters from the posterior distribution (4.11) from samples obtained by the MCMC algorithm.

In the additional Table 4.12 we provide the summary of the estimates of ψ_1 and ψ_2 of the autoregressive model (4.7) obtained via the various methods used in this section.

method	*estimate*		*standard error*	
	ψ_1	ψ_2	ψ_1	ψ_2
Parametric MLE	0.390	-0.489	0.039	0.039
Model based EL	0.393	-0.491	0.036	0.40
Max of model based EL	0.394	-0.490	-	-
Bayesian model based BEL sampler	0.389	-0.485	0.002	0.002
Bayesian model based EL sampler	0.391	-0.487	0.001	0.001
Posterior $\pi(\psi 1_{,}\psi_2\|\mathbf{x})$ via MCMC	0.391	-0.490	0.035	0.036

Table 4.12: Parameter estimates overview of the AR(2) process by the methods demonstrated in this section. The maximum model-based EL estimate does not provide standard error estimate.

Conclusion

We have studied the Bayesian estimation of time series with the empirical likelihood function. The EL framework is a nonparametric method of statistical inference that has been recently modified as a method of inference on a time series. This work provided a summary of the concept of the EL and its modifications to accommodate dependence among data. Based on the character of the analyzed problem, the EL provides two types of approaches.

The first method assumes that the observed time series is weakly dependent and by building a blocks of observation we are able to suppress the dependence and acquire a valid EL inference. The latter method proposes that the time series come from a specific statistical model. This assumption allows us to reformulate the observations in terms of independent random variables and by a specific choice of estimating equations the EL reaches valid asymptotic properties.

For both of these EL likelihood methods for time series we proposed two types of Bayesian sampler. The weighted sampler utilizes the EL as importance weights and yields valid estimations of the posterior distribution. By an additional employment of the importance resampling the weighted sampler yields much improved results. Another introduced approach was to sample directly from the EL version of the posterior distribution via the Markov chain Monte Carlo algorithms. This method proved to provide a valid samples with comparable results to standard parametric methods.

Generally the Bayesian inference on dependent data via the EL function yields satisfactory results. However the class of time series where this approach can be applied is limited. The block-wise EL approach can only be applied to weakly dependent time series, which omits models with more complex dependence structure. On the other hand the model-based approach is limited to models, where the observed samples can be expressed as independent variables. Therefore applying the model-based EL to structures with parameters affecting an unobserved variables, such as the moving average parameters in MA models, rather difficult. Yet we can conclude that once we are able to formulate the EL function for a given problem, employing the empirical likelihood in Bayesian inference yields valid results.

Bibliography

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, *Time series analysis forecasting and control*. Prentice-Hall, Inc., Third Edition, ISBN: 0-13-060774-6, 1994.

[2] W. W. S. Wei, *Time series analysis: univariate and multivariate methods*. Temple University, Seccond edition, 2005.

[3] A. B. Owen, *Empirical Likelihood*. Chapman and Hall, ISBN: 1-58488-071-6, 2001.

[4] J. Quin, J. Lawless, *Empirical Likelihood and General Estimating Equations*. The Annals of Statistics, Vol.22, No.1, 300-325, 1994.

[5] Y. Kitamura, *Empirical likelihood methods with weakly depended processes*. The Annals of Statistics, Vol. 25, No. 5, 2084-2102, 1997.

[6] Y. Kitamura, *Empirical likelihood methods in econometrics: Theory and practice*. Yale University, 14 June 2006.

[7] D. J. Nordman, S. N. Lahiri, *A review of empirical likelihood methods for time series*. Journal of Statistical Planning and Inference, No. 155, 1-18, 2014.

[8] D. J. Nordman, H. Bunzel, S. N. Lahiri, *A nonstandard empirical likelihood for time series*. The Annals of Statistics, Vol. 41, No. 6, 3050-3073, 2013.

[9] K. B. Athreya, S. N. Lahiri, *Measure Theory and Probability Theory*. Springer, 2006.

[10] H. Harari-Kermadec, *Regenerative block empirical likelihood for Markov chains*. Journal of Nonparametric Statistics, GNST-2010-05-08.R2, 2011.

[11] N. H. Chan, C. S. Chuang, *Empirical likelihood for autoregressive models, with applications to unstable time series*. Statistica Sinicia 12, 387-407, 2002.

[12] P. A. Mykland, *Dual Likelihood*. The Annals of Statistics, Vol.23, No.2, 396-421, 1995.

[13] J. Li, W. Liang, S. He, X. Wu, *Empirical likelihood for the smoothed LAD estimator in infinite variance autoregressive models*. Statistics & Probability Letters, vol. 80, no. 17-18, 1420-1430, 2010.

[14] N. H. Chan, S. Ling, *Empirical likelihood for GARCH models*. Econometric Theory 22, 403-428, 2006.

[15] K. Mengersen, P. Pudlo, C. P. Robert, *Approximate Bayesian Computation via empirical likelihood*. arXiv:1205.5658v3, 5 Dec 2012.

[16] K. Mengersen, P. Pudlo, and C. Robert, *Bayesian computation via empirical likelihood*. Proc. Natl. Acad. Sci. U.S.A., vol. 110, no. 4, pp. 1321-1326, Jan. 2013.

[17] C. P. Robert, *The Bayesian Choice*. Springer, Second Edition, ISBN: 978-0-387-71598-8, 2007.

[18] J. Marin, P. Pudlo, C. P. Robert, R.J. Ryder, *Approximate Bayesian Computational methods*. arXiv:1101.0955v2, 27 May 2011.

[19] B. Walsh, *Markov Chain Monte Carlo and Gibbs Sampling*. Lecture Notes for EBB 596z, 2002.

[20] P. Marjoram, J. Molitor, V. Plagnol, S. Tavere *Markov chain Monte Carlo without likelihoods*. Proceedings of the National Academy of Sciences, 2003, vol. 100, no.26

[21] S. P. Brooks, *Markov chain Monte Carlo method and its applications*. The Statistician, Vol. 47, No. 1, 1998.

[22] A. E. Raftery, S. Lewis, *How many iterations in the Gibbs sampler?*. In, Bayesian Statistics 4, pp. 736-773,Oxford University Press, 1992.

[23] N. A. Lazar, *Bayesian Empirical Likelihood*. Biometrika Trust, Vol. 90, No. 2, pp. 319-326, Jun 2003.

[24] C. P. Robert, G. Casella, *Introducing Monte Carlo Methods with R*. Springer, ISBN: 978-4419-1575-7 2010.

[25] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed., Jul. 2003.