Technische Universität Wien

Vienna University of Technology

DIPLOMARBEIT

Long Memory versus Structural Breaks

A Time – Varying Memory Approach

Institut für Wirtschaftsmathematik
der Technischen Universität Wien

unter Anleitung von
O. Univ. Prof. Dipl.-Ing. Dr. techn. Manfred Deistler
durch

Georg M. Görg
georg.goerg@gmail.com

Wien, 15. Oktober 2007

.....
Georg M. Görg

Contents

Acknowledgments

While studying at the City University of New York, I had the opportunity to work on a research project with Professor Dana Draghicescu at Hunter College. During this Fall 2006 semester we wrote the paper *Nonparametric modeling of the second order structure of processes with time–varying memory.*

I submitted the paper to the US wide 2007 student paper competition of the Social Statistics, Government Statistics, and Survey Research Methods Sections held by the American Statistical Society (ASA). My paper was selected as one of the winners and I was invited to give a presentation at the 2007 Joint Statistical Meetings in Salt Lake City.

Based on ideas developed in New York City, my thesis is an extension of the paper, both theoretically and on empirical results.

I highly appreciate Professor Deistler's motivation and enthusiasm for his work as well as his mathematically deep and challenging understanding of Econometrics, which introduced me to the fascinating and wide field of Time Series Analysis. Furthermore, I want to thank Professor Deistler for continuously proposing new views on certain topics discussed in my thesis, especially the part on long memory versus structural breaks. By doing this he prepared the ground for my – a little more thorough – analysis of this question.

I also want to thank Dana Draghicescu for her ongoing support and for offering me the possibility to work on the preceding paper .

Finally I want to thank my family for their support that made it possible for me to pursue my studies the way I did.

Vienna, October 17, 2007 Georg M. Görg

Abstract

There is a broad range of literature on long memory or long range dependent processes, especially on fractionally integrated processes. Throughout the thesis different models generating long memory processes, particularly the extensively discussed fractional ARIMA model, are studied.

One of the main parts analyzes structural breaks versus long memory. To shed new light on this problem I make heavy use of an error duration model, which gives a nice view on stochastic processes in general and on short versus long memory in specific.

After presenting various estimators and tests for long range dependence, I compare short and long memory models for financial data (Dow Jones Index, Alcoa Inc., and EUR / USD exchange rate).

My contribution is a model for time-varying (long) memory and herewith I try to unify the concurring views of long memory and structural breaks.

Chapter 1

Introduction

The first steps in long memory modeling date back to 1950 when Hurst [39] studied Nile river flow data and found empirical evidence that yearly water levels of the Nile river exhibit extreme persistence that can not be captured by classic ARMA models.

Granger and Joyeux [30] established long memory in econometrics and finance in 1980 as a link between stationary $I(0)$ and integrated $I(1)$ processes. At the same time Hosking [36] developed the concept of fractional integration in biological sciences.

Parke [58] gives an interesting interpretation and very intuitive approach to long range dependence by generating long memory with an error duration model, which plays a major role in this thesis.

Some authors [3, 29, 43, 51] stress that structural changes in the mean, variance, or the whole process itself are mistaken for long memory in the data.

Both model approaches have been applied to empirical data, and one can find evidence for and against long memory (versus structural breaks). A lot of financial and macro-economic time series are subject to research of long range dependence, such as (absolute) returns, volatilities or inflation.

More recently, attention has turned to *long memory stochastic volatility* models (see Section 7.1), which exhibit some typical features of real financial data, such as stock or exchange rate returns.

The structure of the thesis is as follows: Chapter 1 introduces notation and basic definitions of time series analysis and presents essential theoretical results for a proper understanding and handling of time series. Chapter 2 deals with definitions of long memory, proper models, and resulting properties of a long memory process. Chapter 3 introduces the opposite viewpoint of structural breaks as an explanation for the observed persistence in time series.

Chapters 5 and 7 present the most common estimators for long range dependence and their application to real world data, respectively. My contribution of this thesis can be found in Chapter 6, where I demonstrate the idea and

general view of time-varying memory as an attempt to unify the long memory and structural breaks point of view.

Appendix A collects necessary theorems and lemmas, and defines (commonly used) notation for the thesis. For the sake of completeness I present essential procedures and algorithms in Appendix B.

1.1 Basic definitions of stochastic processes and time series

A time series x_t is a stochastic process observed over time. To indicate the time dependence I use the sub-script t to denote the individual observation, and T to denote the number of observations. Due to the sequential nature of the process we presume dependence between x_t and x_{t-1}. Therefore, results from classical i.i.d. statistics are not valid anymore.

Definition 1.1.1 (Stochastic process)**.** *A stochastic process is a family of random variables $(x_t|t \in \mathbb{T})$ defined on the probability space $(\Omega, \mathcal{A}, P)$.*

As the general class of stochastic processes is far too wide, to be practicable, the class of *covariance stationary* processes is introduced.

Definition 1.1.2 (Stationarity)**.** *A stochastic process x_t is called covariance (weakly) stationary if,*

i) $\mathbb{E}\, x_t = \mu < \infty \quad \forall t.$

ii) $\mathbb{V}\, x_t = \mathbb{E}(x_t - \mu)^2 = \sigma_x^2 < \infty \quad \forall t.$

iii) $cov(x_t, x_{t+k}) = \gamma(t, t+k) =: \gamma(k)$ *is independent of t for all k.*

The *autocovariance function* $\gamma(j) : \mathbb{Z} \to \mathbb{C}$ is symmetric, i.e. $\gamma(j) = \gamma(-j)$.

Classic approaches to model time series are autoregressive (AR), moving average (MA) and ARMA models. Here I give a brief overview about these processes and refer to Brockwell and Davis [8] or Hamilton [34] for a detailed discussion.

White Noise

A stochastic process ε_t satisfying

$$\begin{aligned} \mathbb{E}\, \varepsilon_t &= 0, \\ \mathbb{E}\, \varepsilon_t^2 &= \sigma_\varepsilon^2 < \infty, \\ \mathbb{E}\, \varepsilon_t \varepsilon_s &= 0 \text{ for } s \neq t, \end{aligned} \tag{1.1.1}$$

is called *white noise.*

Remark 1.1.3. *Unless stated otherwise within the text, $\{\varepsilon_t\}$ always denotes white noise.*

Autoregressive and Moving Average Processes

A process x_t satisfying the linear difference equation

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p} = \varepsilon_t, \quad \forall t \in \mathbb{Z}, \quad \phi_j \in \mathbb{R},$$

is called an *autoregressive* (AR) process of order p (provided $\phi_p \neq 0$).

A process x_t given by

$$x_t = \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} \ldots + \theta_q \varepsilon_{t-q}, \quad \forall t \in \mathbb{Z}, \quad \theta_j \in \mathbb{R},$$

is called a *moving average* (MA) process of order q (provided $\theta_0 \neq 0$ and $\theta_q \neq 0$). Without loss of generality θ_0 can be set to 1. A $MA(\infty)$ process is defined as the limit in mean square of $\sum_{j=0}^{q} \theta_j \varepsilon_{t-j}$ for $q \to \infty$ (see Example 1.2.4).

If we combine an AR and a MA system we get

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} \ldots + \theta_q \varepsilon_{t-q} \tag{1.1.2}$$

where $\theta_j, \phi_j \in \mathbb{R}$. Then (1.1.2) is an ARMA system and a solution x_t is an ARMA process.

For easier and more understandable notation we introduce the *lag* operator L, which is a useful algebraic tool in time series analysis.

Definition 1.1.4. *The lag operator L satisfies $Lx_t = x_{t-1}$.*

Defining $L^2 = LL$ we get: $L^2 x_t = x_{t-2}$. In general we have $L^k x_t = x_{t-k}$.
Thus, differencing x_t can be written as

$$x_t - x_{t-1} = (1 - L)x_t =: \Delta x_t$$

Since the $(1 - L)$ operator is used quite frequently it is abbreviated with the symbol Δ.

Integrated processes

Under certain conditions ARMA processes are stationary, but in practice one often finds non-stationary processes. Usually non-stationary processes can be transformed to stationary processes and then results from theory of stationary time series can be applied.

Differencing the process is a very common transformation and there is a vast literature on so called *integrated* processes.

Definition 1.1.5 (Integration)**.** *A process is* integrated *of order $d \in \mathbb{Z}$ if $y_t := (1 - L)^d x_t$ is stationary.*

If y_t is an ARMA process, then x_t is called an autoregressive integrated moving average process, denoted by $x_t \sim ARIMA(p, d, q)$, where p and q are the orders of the AR and MA systems, respectively.

Definition 1.1.6 (Causality)**.** *A process x_t is called causal with respect to a stationary process u_t, if there is a*

$$\Phi(L) = \phi_0 + \phi_1 L + \phi_2 L^2 + \dots$$

with $\sum_{k=0}^{\infty} |\phi_k| < \infty$ and $x_t = \Phi(L)u_t$.

Definition 1.1.7 (Invertibility)**.** *A process x_t is called invertible with respect to a stationary process u_t, if there is a*

$$\Pi(L) = \pi_0 + \pi_1 L + \pi_2 L^2 + \dots$$

with $\sum_{k=0}^{\infty} |\pi_k| < \infty$ and $u_t = \Pi(L)x_t$.

Lemma 1.1.8. *A MA(∞) process $x_t = \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k}$, $\psi_k \in \mathbb{R}$ is covariance-stationary iff*[1] *$\sum_{k=-\infty}^{\infty} \psi_k^2 < \infty$.*

Proof. The mean of x_t equals zero, independent of t.

$$\mathbb{E}x_t = \mathbb{E} \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k} = \sum_{k=-\infty}^{\infty} \psi_k \underbrace{\mathbb{E}\varepsilon_{t-k}}_{=0} = 0 \quad \forall t. \tag{1.1.3}$$

$$\begin{aligned} \mathbb{V}x_t &= \mathbb{E} \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j} \mathbb{E} \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k} = \mathbb{E} \sum_{j=-\infty}^{\infty} \psi_j^2 \varepsilon_{t-j}^2 \\ &= \sum_{j=-\infty}^{\infty} \psi_j^2 \mathbb{E}\varepsilon_{t-j}^2 = \sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} \psi_j^2. \end{aligned} \tag{1.1.4}$$

Thus, the variance is finite iff $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$.

$$\begin{aligned} cov(x_t, x_{t+k}) &= \mathbb{E} \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j} \sum_{i=-\infty}^{\infty} \psi_i \varepsilon_{t+k-i} \\ &= \mathbb{E} \sum_{i,j} \psi_j \varepsilon_{t-j} \psi_i \varepsilon_{t+k-i} = \sum_{i,j} \psi_j \underbrace{\mathbb{E}\varepsilon_{t-j}\varepsilon_{t+k-i}}_{\neq 0 \text{ for } i=k+j} \psi_i \\ &= \sigma_\varepsilon^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{k+j} =: \gamma(k) \end{aligned} \tag{1.1.5}$$

Since the autocovariance only depends on the time lag k a MA(∞) process is covariance stationary, provided that $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$. □

[1] Here and in the rest of the thesis I use *iff* as an abbreviation for *if and only if*.

The autocovariance is a non-normalized measure for the time dependence of a process. For comparable analysis it is necessary to introduce a normalized measure.

Definition 1.1.9 (Autocorrelation Function)**.** $\rho(k) := \frac{\gamma(k)}{\gamma(0)}$ *is called the autocorrelation function of* x_t*. Especially* $\rho(0) = 1$*.*

White Noise

White noise has no memory by definition, i.e. $\rho(k) = 0 \quad \forall k \geq 1$.

Autoregressive and Moving Average Processes

For a stationary and causal $AR(1)$ process

$$x_t = \rho_1 x_{t-1} + \varepsilon_t, \; |\rho_1| < 1,$$

we have $\rho(k) = \rho_1^k \quad \forall k$, i.e. the autocorrelation function decays exponentially.

For a $MA(1)$ process $x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$ we get

$$\begin{aligned}
&\gamma(0) = \mathbb{E} x_t x_t = \mathbb{E}(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_t + \theta_1 \varepsilon_{t-1}) = \sigma_\varepsilon^2 (1 + \theta_1^2) \\
&\gamma(1) = \mathbb{E} x_t x_{t-1} = \mathbb{E}(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-1} + \theta_1 \varepsilon_{t-2}) = \sigma_\varepsilon^2 \theta_1^2 \\
&\gamma(k) = 0 \text{ for } k \geq 2
\end{aligned}$$

So

$$\rho(k) = \begin{cases} 1 & \text{if } k = 0, \\ \frac{\theta_1^2}{1+\theta_1^2} & \text{if } k = 1, \\ 0 & \text{if } k \geq 2. \end{cases}$$

In general the autocorrelation function of a $MA(q)$ process displays a cutoff to 0 at lag q+1.

Integrated Processes

For a simple random walk $x_t = x_{t-1} + \varepsilon_t$ we have $\rho(k) = 1 \quad \forall k$. Nevertheless the estimated autocorrelation function of a random walk realization decays very slowly since the estimates for higher lags must use less data, but still average over the full sample size T.

Sample autocorrelations of these four processes are displayed in Figure 1.1.

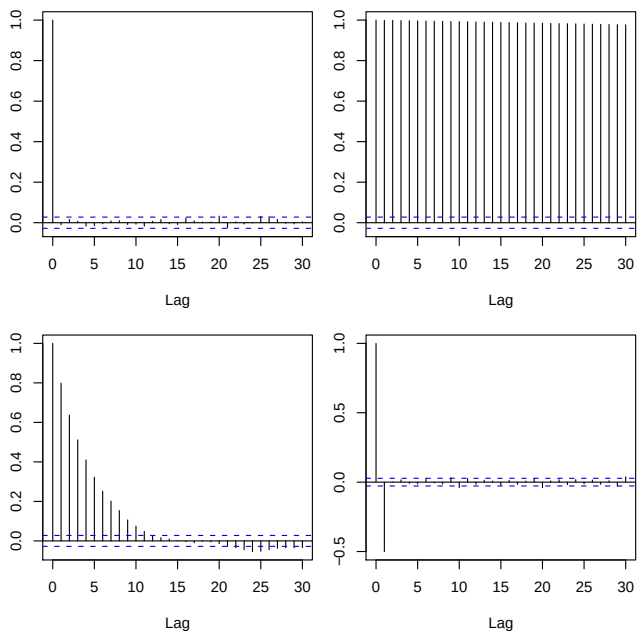


Figure 1.1: Sample autocorrelation functions for realizations (T = 5000) of: (top left) white noise ε_t; (top right) simple random walk; (bottom left) AR(1) with $\rho_1 = 0.8$; (bottom right) simple MA(1) with $\theta_1 = -0.8$

1.2 Hilbert space

Hilbert spaces are important for understanding the theory of stationary processes and provide powerful tools for interpreting them in a geometric way.[2]

Subsequently I present two special Hilbert spaces, which simplify further analysis of linear filters and spectral densities.

1. Consider the probability space $(\Omega, \mathcal{A}, P)$ underlying a stochastic process x_t and define
$$\mathcal{L}^2 := \left\{ x \in (\Omega, \mathcal{A}, P) \,\middle|\, \mathbb{E}|x|^2 < \infty \right\}, \tag{1.2.1}$$
as the space of complex-valued random variables with finite variance.

   The mapping $< x, y >= \mathbb{E}x\overline{y}$ is not an inner product on $\mathcal{L}^2$, since $\mathbb{E}|x|^2 = 0$ does not imply $x \equiv 0$. Thus, we define an equivalence relation $x \equiv y$ on $\mathcal{L}_2$ iff $x = y$ almost surely. The set of these equivalence classes is denoted by $\mathbb{L}^2(\Omega, \mathcal{A}, P)$. It is straightforward to show that $\mathcal{L}^2$ is a linear space and $< x, y >= \mathbb{E}x\overline{y}$ defined on $\mathbb{L}^2(\Omega, \mathcal{A}, P)$ is an inner product. Thus, $\mathbb{L}^2(\Omega, \mathcal{A}, P)$ is a Hilbert space.

[2]For basic definitions and Hilbert space notations see A.2. For a detailed analysis see any text on *Functional Analysis*, e.g. Yoshida [81].

From $< x, y >:= \mathbb{E}x\overline{y}$ we see that $< x, y >$ is the non-central covariance and $\|x\|^2$ is the non-central variance of the random variable x.

Thus, the stationarity conditions can be *translated* into Hilbert space notation:

$$\begin{aligned}
\mathbb{E}x_t &= < x_t, 1 >= const \\
\mathbb{E}x_t^2 &= < x_t, x_t >= \|x_t\|^2 = const \\
\mathbb{E}x_t x_{t-1} &= < x_t, x_{t-1} >= const \\
&\vdots
\end{aligned}$$

This means that all x_t are vectors in the Hilbert space $\mathbb{L}^2(\Omega, \mathcal{A}, P)$ of the same length and the angle between x_t and 1 is constant, as is the angle between x_t and x_{t-1}, and so forth. Thus, these angles do not depend on t.

2. Especially for our purposes consider the probability space $\big([-\pi, \pi], \mathcal{B}, \mu\big)$, where $\mathcal{B}$ is the Borel σ-algebra of $[-\pi, \pi]$ and μ is the normalized Lebesgue measure on $[-\pi, \pi]$, defined by

$$\mu(x) = \frac{1}{2\pi} \int_{[-\pi,\pi]} x \, \mathrm{d}\mu = \frac{1}{2\pi} \int_{-\pi}^{\pi} |x(\lambda)| \, \mathrm{d}\lambda. \tag{1.2.2}$$

Consequently

$$\mathbb{L}^2\big([-\pi, \pi], \mathcal{B}, \mu\big) = \left\{ x \in \big([-\pi, \pi], \mathcal{B}, \mu\big) \,\middle|\, \mu(x^2) < \infty \right\}, \tag{1.2.3}$$

where x is a complex-valued random variable, is a Hilbert space. Notice the equivalent interpretation of $\mathbb{L}^2\big([-\pi, \pi], \mathcal{B}, \mu\big)$ as the space of square integrable functions on $[-\pi, \pi]$

$$\mathbb{L}^2[-\pi, \pi] := \left\{ f \,\middle|\, \|f\|_2 < \infty \right\} \text{ with } \|f\|_2 := \left(\int_{-\pi}^{\pi} |f(t)|^2 \, \mathrm{d}t \right)^{\frac{1}{2}},$$

where the norm comes from the inner product $< f, g >= \int_{-\pi}^{\pi} |f(t)\overline{g(t)}| \, \mathrm{d}t$.

Likewise define the space of absolute integrable functions

$$\mathbb{L}^1[-\pi, \pi] := \left\{ f \,\middle|\, \|f\|_1 < \infty \right\} \text{ with } \|f\|_1 := \int_{-\pi}^{\pi} |f(t)| \, \mathrm{d}t.$$

Note that $\mathbb{L}^1[-\pi, \pi]$ is not a Hilbert space, as there is no mapping $< f, g >: \mathbb{L}^1[-\pi, \pi] \to \mathbb{C}$ that satisfies $\|f\|_1 := \sqrt{< f, f >}$.

To state the main result of this section we need several standard theorems. Proofs can be found in introductory books about measure theory or Fourier series (e.g. Zygmund [83]).

Theorem 1.2.1 (Riesz-Fisher). *Let $\{\mathbf{e}_t\}_{t=0}^{\infty}$ be an orthonormal sequence in a Hilbert space X. The infinite series $\sum_{j=0}^{\infty} \alpha_j \mathbf{e}_j$, $\alpha_j \in \mathbb{C}$ converges to an element $x \in X$ if and only if $\sum_{j=0}^{\infty} |\alpha_j|^2$. In that case $\alpha_j =< x, \mathbf{e}_j >$.*

Hence, for a given square-summable sequence $\left\{\alpha_j\right\}_{j=0}^{\infty} \in \mathbb{C}$ and an orthonormal sequence $\{\mathbf{e}_t\}_{t=0}^{\infty}$ in a Hilbert space X, the series $\sum_{j=0}^{\infty} \alpha_j \mathbf{e}_j$ is well defined and converges to some element $x \in X$, and it holds $\alpha_j =< x, \mathbf{e}_j >$.

On the other hand, given an element x of a Hilbert space X and an orthonormal sequence $\{\mathbf{e}_t\}_{t=0}^{\infty}$, we can consider the family of functions

$$S_n(x) = \sum_{j=0}^{n} < x, \mathbf{e}_j > \mathbf{e}_j, \; n = 0, 1, 2, \ldots. \tag{1.2.4}$$

In general the sequence $S_n(x)$ does *not* necessarily converge to $x \in X$.

Lemma 1.2.2. *Let x be an element of the Hilbert space X, and $\{\mathbf{e}_t\}_{t=0}^{\infty}$ be an orthonormal sequence in X. Then the Fourier polynomials*

$$S_n(x) = \sum_{j=0}^{n} < x, \mathbf{e}_j > \mathbf{e}_j$$

converge to $\widehat{x} \in M$, where M is the subspace of X, generated by $\{\mathbf{e}_t\}_{t=0}^{\infty}$. The difference vector $x - \widehat{x}$ is orthogonal to M.

If the orthonormal sequence is a basis of X, i.e. the generated subspace $M \equiv X$, then the space orthogonal to M is the null space; therefore, the Fourier series converges to $x \in X$ in the norm of the space X.

Lemma 1.2.3. *The sequence $\left\{e^{i\lambda t}\right\}_{t=0}^{\infty}$ is an orthonormal basis of $\mathbb{L}^2[-\pi, \pi]$. So, for every $f \in \mathbb{L}^2[-\pi, \pi]$ it holds*

$$\lim_{n\to\infty} \left\| f - S_n(f) \right\|_2 = 0,$$

where

$$S_n(f) = \sum_{s=-n}^{n} a_s e^{i\lambda s} \text{ with } a_s := \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda s} \, \mathrm{d}\lambda,$$

and norm

$$\|f\|_2 = \int_{-\pi}^{\pi} |f(\lambda)|^2 \, \mathrm{d}\lambda.$$

Finally we can state the main result of this section

Corollary 1.2.4 ($MA(\infty)$ in $\mathbb{L}^2$)**.** *For a given sequence* $\{b_j\}_{j=-\infty}^{\infty}$*, consider the process*

$$y_t := \sum_{j=-\infty}^{\infty} b_j \eta_{t-j},\ \mathbb{V}\eta_t = \sigma_\eta^2 > 0, \tag{1.2.5}$$

with a white noise input η_t*. Then* y_t *is covariance stationary if and only if* $\sum_{j=-\infty}^{\infty} |b_j|^2 < \infty$*.*

Proof. The input sequence $\{\eta_t\}_{t=-\infty}^{\infty}$ is white noise. Thus, by definition

$$\mathbb{E}\eta_t = 0,\ \mathbb{E}\eta_t^2 = \sigma_\eta^2,\ \text{and } \mathbb{E}\eta_t\eta_s = 0 \text{ for } t \neq s.$$

As η_t are elements of the Hilbert space $\mathbb{L}^2(\Omega, \mathcal{A}, P)$ with inner product

$$< x, y >:= \mathbb{E}xy,$$

η_t is an orthogonal sequence in $\mathbb{L}^2(\Omega, \mathcal{A}, P)$ with $\|\eta_t\| = \sigma_\eta$. To get an orthonormal sequence we define the normalized white noise $\varepsilon_t := \frac{\eta_t}{\sigma_\eta}$ and consequently $\alpha_j = b_j\sigma_\eta$. From theorem 1.2.1 it follows that $\sum_{j=-\infty}^{\infty} b_j\eta_{t-j} = \sum_{j=-\infty}^{\infty} \alpha_j\varepsilon_{t-j}$ converges to an element $x \in \mathbb{L}^2$ if and only if $\sum_{j=-\infty}^{\infty} |\alpha_j|^2 = \sigma_\eta^2 \sum_{j=-\infty}^{\infty} |b_j|^2 < \infty$. □

We will use this result later on to prove that the stochastic process

$$y_t := \sum_{j=-\infty}^{\infty} \binom{d}{j} \varepsilon_{t-j}$$

is well defined in $\mathbb{L}^2$.

1.3 Frequency domain

So far the analysis of a time series was motivated by the interrelation of x_t to past values x_{t-j}. But a time series x_t can also be seen as an infinite sum of sinusoidal oscillations with stochastic amplitude and frequency. For detailed discussion of spectral analysis see [8]. The main result is summarized in the *Spectral Representation Theorem.*

Theorem 1.3.1 (Spectral Representation Theorem)**.** *For every stationary process* x_t *there exists a process* $(z(\lambda)|\lambda \in [-\pi, \pi])$ *with orthogonal increments*[3] *such that*

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t}\, \mathrm{d}z(\lambda) \tag{1.3.1}$$

holds. The process $z(\lambda)$ *is almost surely uniquely determined from* x_t*.*

[3]See A.1.2 for the definition of orthogonal increments and integration with respect to such a process.

Proof. See Brockwell and Davis [8]. □

The spectral representation is just a generalization of the Fourier transform to stochastic processes, as it decomposes every stationary process into a sum of sinusoidal components with random coefficients.

If $z(\lambda)$ is the orthogonal increment process corresponding to x_t, then $F(\lambda) := \mathbb{E}z(\lambda)z(\lambda)^*$ is the *spectral distribution function* of x_t. If there exists a function $f : [-\pi, \pi] \to \mathbb{C}$ such that

$$F(\lambda) = \int_{-\pi}^{\lambda} f(\alpha)\,\mathrm{d}\alpha,$$

where α denotes the Lebesgue measure, then f is called the *spectral density* of x_t.

The autocovariance function $\gamma(s)$ can be written as ($i = \sqrt{-1}$ and $\overline{z}$ is the complex conjugate of z)

$$\begin{aligned}\gamma(s) &= \mathbb{E}x_s\overline{x_0} = \mathbb{E}\int_{-\pi}^{\pi} e^{i\lambda s}\,\mathrm{d}z(\lambda)\int_{-\pi}^{\pi}\overline{e^{i\lambda 0}\,\mathrm{d}z(\lambda)} \\ &= \int_{-\pi}^{\pi} e^{i\lambda s}\underbrace{\mathrm{d}\mathbb{E}z(\lambda)\mathrm{d}\overline{z(\lambda)}}_{\mathrm{d}F(\lambda)} = \int_{-\pi}^{\pi} e^{i\lambda s}\,\mathrm{d}F(\lambda). \end{aligned} \tag{1.3.2}$$

If the spectral density $f(\lambda)$ exists, we get

$$\gamma(s) = \int_{-\pi}^{\pi} e^{i\lambda s} f(\lambda)\,\mathrm{d}\lambda. \tag{1.3.3}$$

Thus, $\gamma(s)$ are the Fourier coefficients of $f_x(\lambda)$ and the Fourier series of $f(\lambda)$ is given by

$$f(\lambda) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty} e^{-i\lambda s}\gamma(s). \tag{1.3.4}$$

Since $\gamma(s) = \gamma(-s)$, $f(\lambda)$ is indeed a real-valued function

$$f(\lambda) = \frac{1}{2\pi}\left(\sigma_x^2 + 2\sum_{s=1}^{\infty}\cos(\lambda s)\gamma(s)\right) = \frac{\sigma_x^2}{2\pi}\left(1 + 2\sum_{s=1}^{\infty}\rho(s)\cos\lambda s\right).$$

Note that the spectral density does not necessarily exist. One sufficient (but strong) condition for the existence of the spectral density is

$$\sum_{s=-\infty}^{\infty}|\gamma(s)| < \infty. \tag{1.3.5}$$

Absolutely summable autocovariances guarantee a pointwise convergence of the Fourier series to the spectral density.

Remark 1.3.2. *The equality* $\sigma_x^2 = \gamma(0) = \int_{-\pi}^{\pi} f(\lambda)\,\mathrm{d}\lambda$ *gives a straightforward*

relation between the variance of a process and the contribution of all cycles for frequencies $\lambda \in [-\pi, \pi]$.

In general, for $-\pi \leq a < \lambda < b \leq \pi$

$$\int_a^b f(\lambda)\,\mathrm{d}\lambda$$

is the contribution of oscillations in the frequency band $\lambda \in [a, b]$ to the total variance σ_x^2 of x_t.

Example 1.3.3 (White noise)**.** *For white noise $\{\varepsilon_t\}$ we have*

$$\gamma_\varepsilon(j) = \begin{cases} \sigma_\varepsilon^2 & \text{if } j = 0, \\ 0 & \text{if } j \geq 1. \end{cases}$$

Trivially $\gamma_\varepsilon(j)$ satisfies (1.3.5) and thus

$$f_\varepsilon(\lambda) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \gamma_\varepsilon(s) e^{-i\lambda s} = \frac{1}{2\pi}\gamma_\varepsilon(0) = \frac{\sigma_\varepsilon^2}{2\pi}. \tag{1.3.6}$$

As the spectrum of ε_t is constant for all λ, every frequency has the same contribution to the variance of the process.

To conclude this part on the frequency domain it is worth noting that the spectral distribution and the spectral density, respectively, contain the same information as the autocovariance function, but the information is displayed differently. Given the purpose of analysis one approach might be easier to interpret, compute, and work with than the other one.

1.4 Linear transformations

If $\{x_t\}_{t=-\infty}^{\infty}$ is a stationary process, then

$$y_t = a(L)x_t \text{ with } a(L) = \sum_{j=-\infty}^{\infty} a_j L^j, \quad a_j \in \mathbb{R} \tag{1.4.1}$$

is called a *linear transformation* of $\{x_t\}_{t=-\infty}^{\infty}$. To ensure the existence of the sum for all stationary x_t we assume

$$\sum_{j=-\infty}^{\infty} |a_j| < \infty. \tag{1.4.2}$$

$\{a_j\}$ is called the *weighting sequence* of the linear transformation.

If $x_t = \varepsilon_t$ is white noise, then the condition on the weighting sequence can be relaxed to

$$\sum_{j=-\infty}^{\infty} |a_j|^2 < \infty.$$

1.4.1 Frequency domain

Using the spectral representation theorem, we get

$$\begin{aligned} x_t &= \int_{-\pi}^{\pi} e^{i\lambda t} \, \mathrm{d}z_x(\lambda) \\ y_t &= \sum_{j=-\infty}^{\infty} a_j x_{t-j} = \sum_{j=-\infty}^{\infty} a_j \int_{-\pi}^{\pi} e^{i\lambda(t-j)} \, \mathrm{d}z_x(\lambda) \\ &= \int_{-\pi}^{\pi} e^{i\lambda t} \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j} \, \mathrm{d}z_x(\lambda). \end{aligned}$$

Corresponding to the linear transformation (1.4.1) we define the transfer function

$$k : [-\pi, \pi] \to \mathbb{C}, \; \lambda \longmapsto \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}. \tag{1.4.3}$$

As before, $k(\lambda)$ and the weighting sequence a_j are in a one-to-one relation since $a_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\lambda j} k(\lambda) \, \mathrm{d}\lambda$.

Theorem 1.4.1. *Let x_t be stationary with spectral density f_x. If $a(L)$ is absolutely summable (or square summable for x_t white noise), then the spectral density f_y of $y_t = a(L)x_t$ and the cross spectral density f_{xy} between x_t and y_t exist, and are given by*

$$\begin{aligned} f_y &= k(\lambda) f_x(\lambda) k(\lambda)^* \\ f_{xy} &= k(\lambda) f_x(\lambda) \end{aligned} \tag{1.4.4}$$

where $k(\lambda) = a(e^{-i\lambda j}) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}$ is the transfer function and z^* is the transpose and conjugate of z.

Proof. See Brockwell and Davis [8]. □

Example 1.4.2 (ARMA)**.** *For a stationary, invertible and causal $ARMA(p, q)$ process*

$$u_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t = \sum_{j=0}^{\infty} \varphi_j L^j \varepsilon_t, \quad \Theta(z) \neq 0 \text{ and } \Phi(z) \neq 0 \text{ for } |z| \leq 1, \tag{1.4.5}$$

with no common zeros of $\Theta(z)$ *and* $\Phi(z)$*, the transfer function is given by*

$$k(\lambda) = \frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})} = \sum_{j=0}^{\infty} \varphi_j e^{-i\lambda j} \tag{1.4.6}$$

and thus

$$f_u(\lambda) = k(\lambda) f_\varepsilon(\lambda) k(\lambda)^* = \frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})} \frac{\sigma_\varepsilon^2}{2\pi} \frac{\Theta(e^{i\lambda})}{\Phi(e^{i\lambda})} = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2} \tag{1.4.7}$$

Analyzing $\lim_{\lambda \to 0} f_u(\lambda)$ we see that this limit must be equal to a constant $0 < c < \infty$. Assume $c = 0$, this implies $\lim_{\lambda \to 0} |\Theta(e^{-i\lambda})|^2 = |\Theta(e^{-i0})|^2 = \Theta(1)^2 = 0$. But this is a contradiction to our invertibility assumption and roots outside the unit circle. The same arguments for $\Phi(e^{-i\lambda})$ show that $c \neq \infty$.

Subsequently, every $ARMA(p, q)$ process satisfies the conditions of Example 1.4.2, i.e. it is stationary, causal, and invertible.

Chapter 2

Long Memory Processes

Different definitions for long range dependence are used. I work with two common and very general definitions in the time and frequency domain, respectively.[1]

Definition 2.0.3. *A stochastic process* $\{x_t\}_{t=-\infty}^{\infty}$ *with autocovariance function* $\gamma_x(j)$ *exhibits long memory iff*

$$\sum_{j=-\infty}^{\infty} \gamma_x(j) = \gamma(0) \cdot \sum_{j=-\infty}^{\infty} \rho_x(j) = \begin{cases} 0 & \textit{anti-persistent long memory,} \\ \infty & \textit{persistent long memory.} \end{cases} \tag{2.0.1}$$

Definition 2.0.4. *A stochastic process* $\{x_t\}_{t=-\infty}^{\infty}$ *with spectrum* $f_x(\lambda)$ *exhibits long memory iff*

$$\lim_{\lambda \to 0} f_x(\lambda) = \begin{cases} 0 & \textit{anti-persistent long memory,} \\ \infty & \textit{persistent long memory.} \end{cases} \tag{2.0.2}$$

Consider the spectrum (1.4.7) of a stationary, causal and invertible $ARMA(p, q)$ process u_t. As shown above

$$\lim_{\lambda \to 0} f_u(\lambda) = c \notin \{0, \infty\}.$$

Thus, stationary, causal, and invertible ARMA models do *not* exhibit long memory. In consequence they are called *short memory* models.

In the following I present parametric conditions on $\gamma(j)$ [$\rho(j)$] and $f(\lambda)$, respectively, such that processes with this autocovariance [autocorrelation] structure and spectrum exhibit long range dependence as in definition 2.0.3 or 2.0.4.

[1] In this thesis I will not consider Gegenbauer processes x_t satisfying

$$(1 - 2\nu L + L^2)^d (x_t - \mu) = \varepsilon_t,$$

where the spectral density has a pole at some frequency $\lambda \neq 0$, but only stochastic processes x_t with poles at the origin $\lambda = 0$. The interested reader is referred to [23, 31, 32].

Condition 2.0.5. *A stationary process* $\{x_t\}_{t=-\infty}^{\infty}$ *exhibits long memory if*

$$\lim_{k\to\infty} \frac{\rho_x(k)}{c_p k^{2d-1}} = 1 \tag{2.0.3}$$

Here c_p *is a constant and* $d \in \mathbb{R}$ *is the memory parameter.*[2]

Under certain assumptions, see Beran [4], there is an equivalent condition in the frequency domain

Condition 2.0.6. *A stochastic process* $\{x_t\}_{t=-\infty}^{\infty}$ *exhibits long memory if*

$$\lim_{\lambda\to 0} \frac{f_x(\lambda)}{c_f|\lambda|^{-2d}} = 1 \tag{2.0.4}$$

Here c_f *is a positive constant and* $d \in \mathbb{R}$.

Note that this is only a specification of the spectrum $f_x(\lambda)$ close to zero. Outside a neighborhood of zero we assume that $f_x(\lambda)$ is a well-behaved function. In specific $f_x(\lambda)$ is continuous and bounded for frequencies away from the origin.

Immediately we have

$$\lim_{\lambda\to 0} f_x(\lambda) = \begin{cases} 0 & \text{if } d < 0, \\ \infty & \text{if } d > 0. \end{cases} \tag{2.0.5}$$

Finally I introduce a different, but equivalent approach to long memory, which is based on the asymptotic behavior of the variance of partial sums.

Condition 2.0.7. *A stochastic process* $\{x_t\}_{t=-\infty}^{\infty}$ *exhibits long memory if*

$$\mathbb{V}\, S_T = \mathcal{O}(T^{\alpha}), \quad \alpha \neq 1,$$

where $S_T = \sum_{t=1}^{T} x_t$.[3]

As the spectral density is the limit of $\frac{1}{T}S_T$ at zero, it is clear that the partial sums condition is equivalent to the spectral condition for long memory ($\alpha = 2d + 1$).

Consider the autocorrelation function plotted in Figure 2.1(a). There is only one significant negative autocorrelation in the first 100 lags, $\rho_x(k)$ seems to decay hyperbolically, and on top of that lags up to 80 ($\approx$ 1.5 years) are siginificant.

[2]Subsequently I refer to d lying in a certain interval of $\mathbb{R}$. As the case $d = 0$ has been studied extensively, I tacitly assume $d \neq 0$.

[3]See A.3.1 for the definition of $\mathcal{O}(g(t))$ and $o(g(t))$.

2.1 Motivation for introducing the concept of long memory

Before modeling processes following a new scheme and introducing a parametric classification of long memory processes, one should motivate the occurrence of processes indicating such a pattern. There are reasonable explanations why real life processes could follow an ARMA or ARIMA model. Here I present intuitive explanations for the existence of long memory and the necessity to develop models and theory for this family of stochastic processes.

2.1.1 Aggregation

One way to address the existence of long memory starts with short memory models. Granger [28] commences with n independent stationary $AR(1)$ processes $x_{j,t}$ with the root close to the unit disk,

$$x_{j,t} = \alpha_j x_{j,t-1} + \varepsilon_{j,t}, \quad j = 1, \ldots, n,$$

where $\varepsilon_{j,t}$ are zero-mean, independent white noise and α_j are drawn from a Beta distribution on $(0, 1)$, with

$$dF(\alpha) = \frac{2}{B(p,q)} \alpha^{2p-1}(1-\alpha^2)^{q-1} d\alpha, \quad 0 \leq \alpha \leq 1, \text{ and } p > 0, q > 0.$$

If we set

$$\overline{x}_t := \sum_{j=1}^{n} x_{j,t},$$

then – for n large – $\overline{x}_t \sim I(1 - \frac{q}{2})$. For a proof see [18].

For example, consider a set of stocks and an index representing this set. As the index is a (weighted) sum of these stocks, the aggregation model could be a reasonable explanation for long range dependence of stock indices, if the corresponding stocks are stationary $AR(1)$ processes on the threshold to random walks.

2.1.2 Spectral examination

Above we have already inspected an autocorrelation function with a rather hyperbolical than exponential decay. A similar – empirically motivated – approach for the appearance of long memory gives a spectral domain examination.

Consider a process x_t, when differenced d times, results in a stationary process $u_t := (1-L)^{-d} x_t$ with spectrum $f_u(\lambda)$. x_t is then called an integrated series of order d, denoted by $x_t \sim I(d)$. The process x_t does not necessarily have a

spectrum, but from filtering theory, the spectrum – if it exists – must satisfy

$$\begin{aligned} f_x(\lambda) &= (1-e^{-i\lambda})^{-d}(1-e^{i\lambda})^{-d} f_u(\lambda) = \big(2(1-\cos\lambda)\big)^{-d} f_u(\lambda) \\ &= \left(2\sin\frac{\lambda}{2}\right)^{-2d} f_u(\lambda). \end{aligned} \tag{2.1.1}$$

Example 2.1.1 (ARMA)**.** *Previously u_t is just a stationary process, no further assumptions are needed. Suppose u_t has an $ARMA(p,q)$ representation, then*

$$\lim_{\lambda\to 0} f_u(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\lim_{\lambda\to 0}\frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2} = \frac{\sigma_\varepsilon^2}{2\pi}\frac{|\Theta(1)|^2}{|\Phi(1)|^2} = c_u, \tag{2.1.2}$$

where c_u is a constant $0 < c_u < \infty$. Ergo the behavior of the spectrum $f_x(\lambda)$ (2.1.1) for $\lambda \to 0$ does not depend on the ARMA specification in the sense that

$$\lim_{\lambda\to 0} f_x(\lambda) = c_u \lim_{\lambda\to 0}\left(2\sin\frac{\lambda}{2}\right)^{-2d}.$$

This property is the basis for the GPH estimator (see Section 5.3.2).

Now let $d \in \mathbb{R}$, $\frac{1}{2} \le d < 1$. Below I show that this implies infinite variance for the process x_t. The classical Box-Jenkins method in consequence suggests differencing x_t to get a well-behaved, covariance stationary and invertible, series. But differencing results in a spectrum

$$\begin{aligned} f_{\Delta x}(\lambda) &= |1-e^{-i\lambda}|^2 f_x(\lambda) = |1-e^{-i\lambda}|^2 |1-e^{-i\lambda}|^{-2d} f_u(\lambda) \\ &= |1-e^{-i\lambda}|^{2(1-d)} f_u(\lambda) = \big[2(1-\cos\lambda)\big]^{2(1-d)} f_u(\lambda), \end{aligned}$$

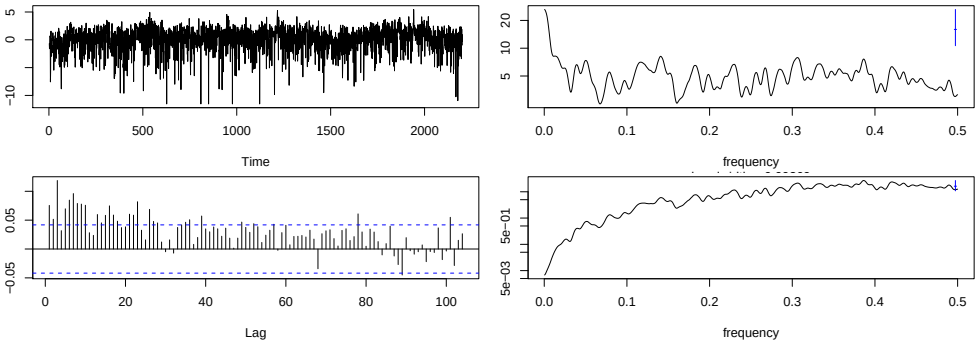
so that $\lim_{\lambda\to 0} f_{\Delta x}(\lambda) = 0$. For the same reasons as above the difference operator $(1-L)^1$ is not appropriate for time series x_t having a spectrum (2.1.1) with $\frac{1}{2} \le d < 1$, as the resulting process $u_t := (1-L)x_t$ is stationary but *not* invertible.

Therefore, neither differencing (no invertible MA representation), nor not differencing (infinite variance) seems appropriate for data having a spectrum with characteristics as in (2.1.1). These properties are illustrated in Figure 2.1(b).

Letting λ go to 0 in (2.1.1) we differ between 2 cases:

$$\lim_{\lambda\to 0} f_x(\lambda) = \begin{cases} 0 & \text{if } d < 0, \textit{ anti-persistent} \\ \infty & \text{if } d > 0, \textit{ persistent} \end{cases} \tag{2.1.3}$$

Algebraically $f_x(\lambda) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma(j)e^{-i\lambda j}$. Since $f_x(0) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma(j)$, we have that for $d > 0$ the sum of autocorrelations diverge; the *persistent* case. For $d < 0$ the autocorrelations sum up to 0; we will refer to this as the *anti-*



(a) (top) $x_t = \log(\mu_0 + r_t^2)$: logarithm of the squared weekly DJI log-returns; (bottom) autocorrelation function $\gamma_x(k)$

(b) (top) log-spectrum $f_x(\lambda)$ of the x_t; (bottom) log-spectrum $f_{\Delta x}(\lambda)$ of the differenced series $x_t - x_{t-1}$

Figure 2.1: Time and frequency analysis for transformed DJI log-returns from 1964 – 2006

persistent case.

Remark 2.1.2. *In practice one will find almost exclusively processes with persistent behavior. However, if in practice*

$$\widehat{f}_u(\omega_j) \to 0 \text{ as } \omega_j \to 0,$$

then this is an indication that the data have been over-differenced (see Figure 2.1(b)). Besides a process with negative d is an unstable status, as already a small zero-mean independent perturbation ξ_t with variance $\sigma_\xi^2 > 0$ added to an anti-persistent process

$$y_t = x_t + \vartheta_t$$

destroys *the property* $\sum_{j=-\infty}^{\infty} \gamma(j) = 0$ *since*

$$\sum_{j=-\infty}^{\infty} \gamma_y(j) = \sigma_\xi^2 + \underbrace{\sum_{j=-\infty}^{\infty} \gamma_x(j)}_{=0} = \sigma_\xi^2 > 0.$$

Of course, a persistent process remains persistent.[4]

2.2 Error duration model

Parke [58] proposes an error duration model as the driving force for long memory processes and herewith gets an elegant and very insightful representation and justification for the – non spurious – occurrence of long memory processes:

[4]See Section 7.1 for a more detailed analysis.

The basic mechanism for an error duration model is a sequence of shocks of stochastic magnitude and stochastic duration. The variable observed in a given period is the sum of those shocks that survive to that point. The distribution of the durations of the shocks determines whether or not the process is fractionally integrated. Fractional integration requires that a small percentage of the shocks have long durations.

Let $\{\varepsilon_t | t = 1, 2, \ldots\}$ be i.i.d. shocks with zero mean and constant variance σ_ε^2. Let ε_s have a stochastic duration of $n_s \geq 0$ with distribution function $F(k) = P(n_s \leq k)$. Define $g_{s,t}$ as the indicator function for the event ε_s *survives until period* t

$$g_{s,t} := \begin{cases} 1 & t \leq s + n_s, \\ 0 & t > s + n_s. \end{cases} \tag{2.2.1}$$

and p_k as the probability of ε_s surviving until period $s + k$,

$$p_k := P(g_{s,s+k} = 1), \quad k = 0, 1, 2, \ldots.$$

For easier handling note that

$$\begin{aligned} p_{k+1} &= P(g_{s,s+k+1} = 1) = \mathbb{E}\, I_{\{\text{shock duration} \geq k+1\}} \\ &= P(n_s \geq k + 1) = 1 - P(n_s \leq k) = 1 - F(k). \end{aligned}$$

Thus, p_k is the probability that any shock lasts k *or more* periods ($n_s \geq k$). In some sense the survival probabilities are similar to *covariance stationarity*, as they do not depend on s but only on the lag k.

By definition it holds $1 = p_0 \geq p_1 \geq p_2 \geq \ldots$, and for further analysis we assume $p_k \overset{k\to\infty}{\to} 0$.[5] Thus, the probabilities of surviving exactly k periods

$$\begin{aligned} \pi_k &:= P(n_s = k) = P(n_s \geq k) - P(n_s \geq k + 1) \\ &= p_k - p_{k+1} = F(k) - F(k - 1), \quad k = 1, 2, 3, \ldots, \end{aligned} \tag{2.2.2}$$

are well defined.

Set

$$y_t := \sum_{s=-\infty}^{t} g_{s,t} \varepsilon_s, \tag{2.2.3}$$

then y_t is the realization of all shocks, occurred since the infinite past, that survived until period t.

[5]Otherwise the probability of a shock having infinite duration is greater than zero and trivially the variance of y_t goes to infinity. For $p_k \equiv 1\, \forall k$ we have the simple random walk.

Claim 2.2.1. *The autocovariances $\gamma(k)$ of y_t are given by*

$$\gamma(k) = \sigma_\varepsilon^2 \sum_{i=k}^{\infty} p_i \geq 0. \tag{2.2.4}$$

Note that $\{\gamma(k)\}$ is a nonincreasing, nonnegative sequence. Therefore, only persistent memory can be achieved within this model.

Besides $p_k = \frac{\gamma(k)-\gamma(k+1)}{\sigma_\varepsilon^2} \geq 0$.

Proof. Since $g_{s,t}$ and ε_s are independent it follows that

$$\mathbb{E}y_t = \mathbb{E}\sum_{s=-\infty}^{t} g_{s,t}\varepsilon_s = \sum_{s=-\infty}^{t} \mathbb{E}g_{s,t}\mathbb{E}\varepsilon_s = \sum_{s=-\infty}^{t} \mathbb{E}g_{s,t}0 = 0.$$

$$\gamma(k) = \mathbb{E}y_t y_{t-k} = \mathbb{E}\left[\sum_{i=0}^{\infty} g_{t-i,t}\varepsilon_{t-i}\right]\left[\sum_{j=k}^{\infty} g_{t-j,t-k}\varepsilon_{t-j}\right]. \tag{2.2.5}$$

As ε_{t-i} and ε_{t-j} are independent for $i \neq j$, equation (2.2.5) reduces to

$$\gamma(k) = \sum_{i=k}^{\infty} \mathbb{E}g_{t-i,t}^2 \mathbb{E}\varepsilon_{t-i}^2. \tag{2.2.6}$$

Recall that $g_{t-i,t}$ is an indicator function only taking values 0 or 1, so trivially $g_{t-i,t}^2 = g_{t-i,t}$; thus, $\mathbb{E}g_{t-i,t}^2 = \mathbb{E}g_{t-i,t}$. $g_{t-i,t}$ represents the event that shock ε_{t-i} survives until period t, i.e. that the shock duration n_s is greater or equal to i, so we have

$$\mathbb{E}\, g_{t-i,t}^2 = \mathbb{E}\, g_{t-i,t} = P(n_s \geq i) = p_i.$$

Substituting into (2.2.6) gives

$$\gamma(k) = \sum_{i=k}^{\infty} p_i \mathbb{E}\varepsilon_{t-i}^2 = \sigma_\varepsilon^2 \sum_{i=k}^{\infty} p_i \text{ independent of t},$$

which completes the proof. □

Remark 2.2.2. *The variance of y_t equals*

$$\sigma_y^2 = \gamma(0) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} p_i = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} P(n_s \geq i).$$

As n_s is a nonnegative random variable the expectation – if it exists – is given by (see Lemma A.1.5)

$$\mathbb{E}\, n_s = \sum_{i=1}^{\infty} P(n_s \geq i) =: \nu.$$

Thus, the variance of an ED process is directly related to the expected error duration, as $\sigma_y^2 = \sigma_\varepsilon^2 (1+\nu)$. Note that $\sigma_y^2 > \sigma_\varepsilon^2$ unless $\nu = 0$, which implies the degenerate case of $n_s \equiv 0$.

Given the explicit formula for $\gamma_y(k)$ we can set restrictions on p_k such that y_t is covariance stationary.

Condition 2.2.3 (Stationarity – Non-stationarity)**.** *If ν is finite, $\mathbb{V}y_t = \sigma_\varepsilon^2(1+\nu) < \infty \quad \forall t$ and $\gamma(k) = \sigma_\varepsilon^2 \sum_{i=k}^{\infty} p_i$ independent of t. Therefore, the process y_t in (2.2.3) is covariance stationary.*

If ν is not finite, the expected error duration is infinite, as is the variance y_t; thus, y_t is non-stationary.

Analogously we get conditions on p_k for long memory features.

Condition 2.2.4 (Short Memory – Long Memory)**.** *The process defined in (2.2.3) exhibits long memory iff $\sum_{k=0}^{\infty}(k+1)\, p_k$ diverges.*

Proof. This follows directly by substitution of $\gamma(k) = \sum_{i=k}^{\infty} p_i$ into the definition of long memory ($\sum_{k=0}^{\infty} |\gamma(k)|$ diverges). □

Obviously $\sum_{i=1}^{\infty} p_i$ does not converge for $p_i = 1$, $\forall i = 1, 2, \ldots$. This degenerate distribution corresponds to the unit root case. However, the condition for nonstationarity – in the error duration sense – is much weaker than the condition for a unit root, which states that all shocks last forever. For $p_k = \frac{1}{k+1}$ the probability of having an impact for k or more periods goes to zero as k tends to infinity. But as the probability does not vanish fast enough the variance of the process tends to infinity, resulting in a non-stationary process.

2.2.1 Simulate duration driven processes

The procedure to simulate processes from the ED model is quite obvious:[6]

1. generate an i.i.d. random sample $\{\varepsilon_t\}_{t=1-K}^{T}$,
2. choose a feasible, discrete probability distribution for the survival probabilities $p_k, k = 0, 1, 2, \ldots$,
3. draw a random sample of shock durations n_s and calculate the indicator function $g_{s,t}(k)$,
4. compute $x_t = \sum_{s=1-K}^{T} g_{s,t}(k)\varepsilon_s$.

See Example 2.2.5 for a short memory ($AR(1)$), and Example 2.3.11 for a long memory (ARFIMA) realization.

[6]The described truncated method is only an approximation, as it suffers from a presampling bias [38]. Choosing a large presample size K, simulate an error duration process for $T + K$ shocks, and omitting the first K simulations, should be sufficient for our purposes.

2.2.2 Short memory and error duration

The error duration framework is of course not limited to long memory processes. Consider the $AR(1)$ process

$$x_t = \phi x_{t-1} + \eta_t,\ 0 < |\phi| < 1,\ \eta_t \sim N(0, \sigma_\eta^2), \tag{2.2.7}$$

with autocovariances

$$\gamma_x(k) = \mathbb{E} x_t x_{t-k} = \sigma_\eta^2 \frac{\phi^k}{1-\phi^2}. \tag{2.2.8}$$

The autocovariances for a realization y_t of an ED model with survival probabilities p_i and error variance σ_ε^2 satisfy

$$\gamma_y(k) = \sigma_\varepsilon^2 \sum_{i=k}^{\infty} p_i \Leftrightarrow p_k = \frac{\gamma_y(k) - \gamma_y(k+1)}{\sigma_\varepsilon^2}.$$

Therefore, $\gamma_x(k) \equiv \gamma_y(k)$ iff

$$\begin{aligned}
p_k &= \frac{\gamma_x(k) - \gamma_x(k+1)}{\sigma_\varepsilon^2} = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2} \frac{\phi^k - \phi^{k+1}}{1-\phi^2} = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2} \frac{1}{1+\phi} \phi^k \\
& p_0 \equiv 1 \Rightarrow \sigma_\varepsilon^2 = \frac{\sigma_\eta^2}{1+\phi} > 0 \Rightarrow p_k = \phi^k \text{ and } \phi > -1 \\
& p_k \geq 0 \quad \forall k \Rightarrow \phi \geq 0, \\
& p_k \geq p_{k+1} \quad \forall k \Rightarrow \phi \geq 0.
\end{aligned}$$

Again, only a certain class of $AR(1)$ processes can be reproduced, but for $0 < \phi < 1$ a realization of

$$y_t = \sum_{s=-\infty}^{t} g_{s,t} \varepsilon_s,\ \varepsilon_s \sim N\left(0, \sigma_\varepsilon^2 = \frac{\sigma_\eta^2}{1+\phi}\right) \tag{2.2.9}$$

$$p_k = P(g_{s,s+k} = 1) = \phi^k, \tag{2.2.10}$$

is not distinguishable from the realization of the $AR(1)$ model in equation (2.2.7). The survival probabilities are just the autocorrelations of x_t, which is not surprising considering the $MA(\infty)$ representation of an $AR(1)$ process. Yet, the error structure is different as $\sigma_\varepsilon^2 < \sigma_\eta^2 = \sigma_\varepsilon^2(1+\phi)$ for every $0 < \phi < 1$.

Example 2.2.5 ($AR(1)$ versus ED)**.** *Specifically consider an $AR(1)$ process with $\phi = 0.9$ and $\sigma_\eta^2 = 1$. Thus, the corresponding error duration representation is given by*

$$\begin{aligned}
y_t &= \sum_{s=-\infty}^{t} g_{s,t} \varepsilon_s, \quad \varepsilon_s \sim N(0, 0.526) \\
p_k &= P(g_{s,s+k} = 1) = 0.9^k.
\end{aligned}$$

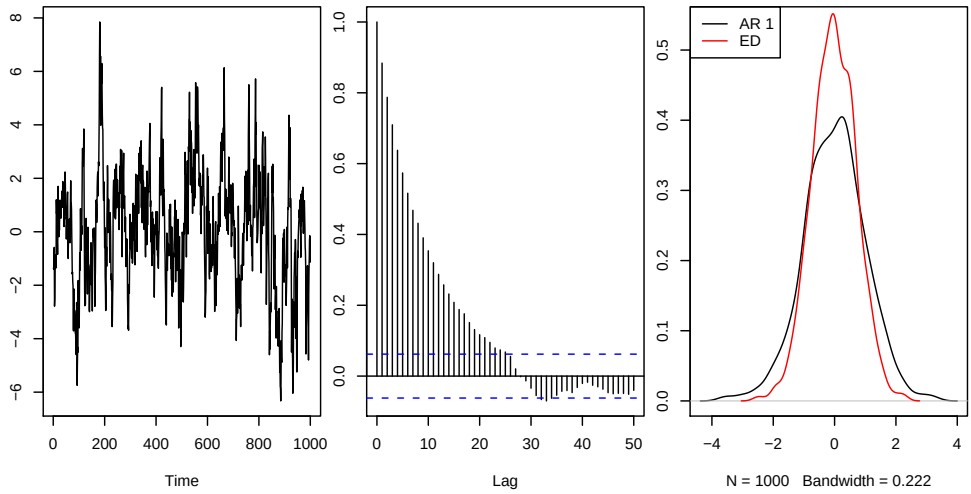


Figure 2.2: (left) y_t: simulated $AR(1)$ process from ED model; (middle) autocorrelations of y_t; (right) Densities of corresponding sample innovations.

Figure 2.2 shows one realization of this error duration model. In fact, the autocorrelations show an almost perfect exponential decay to zero. Estimating an $AR(1)$ *model for the data gives*

$$y_t = 0.113(0.273) + 0.884(0.015)y_{t-1} + \eta_t, \quad \widehat{\sigma}_\eta^2 = 1.026. \tag{2.2.11}$$

So, all the characteristics of the $AR(1)$ *model – zero mean,* $\phi = 0.9$, $\sigma_\eta^2 = 1$ *– have been reproduced.*

As I have simulated the data, I know *the innovations for the error duration model. For the* $AR(1)$ *process I did not observe the innovations, but the residuals of (2.2.11) should be a fairly good approximation. The right plot in Figure 2.2 shows the innovation densities for the two models. As already predicted above, the* $AR(1)$ *innovations have fatter tails than the ED innovations.*

2.2.3 Different view of the world

Although some AR and ED models are equivalent in the way that realizations are not distinguishable from a second order point of view, the driving forces behind a realization are quite different.

Consider a stationary, causal and invertible $MA(\infty)$ process

$$x_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}, \quad \eta_t \sim WN(0, \sigma_\eta^2). \tag{2.2.12}$$

Condition 2.2.6. *In order that* $p_k := \frac{\gamma_x(k) - \gamma_x(k+1)}{\sigma_\varepsilon^2}$ *is a valid ED probability sequence, the autocovariances of the* $MA(\infty)$ *process must satisfy*

a) $p_0 = 1 \Rightarrow 0 \overset{!}{<} \sigma_\varepsilon^2 = \gamma_x(0) - \gamma_x(1)$,

b) $p_k \geq 0 \Rightarrow \gamma_x(k) \geq \gamma_x(k+1)$,

c) $p_k \geq p_{k+1} \Rightarrow \gamma_x(k-1) - \gamma_x(k) \geq \gamma_x(k) - \gamma_x(k+1)$.

By the Cauchy-Schwarz inequality, condition a) is always satisfied for covariance stationary processes. Thus, this condition must be seen as a necessary equality for the variance of the ED innovations.

Given the autocorrelations $\gamma_x(k)$ of a $MA(\infty)$ satisfy condition 2.2.6,

$$y_t = \sum_{s=0}^{\infty} \varepsilon_s d_{s,t}, \quad \varepsilon_s \sim WN(0, \sigma_\varepsilon^2), \quad \sigma_\varepsilon^2 = \gamma_x(0) - \gamma_x(1), \tag{2.2.13}$$

is an equivalent error duration representation for an $MA(\infty)$ process x_t.

For a specific realization $\{x_t\}_{t=1}^T$ the left hand side of (2.2.12) and (2.2.13) are equal of course, but the way this realization is formed is different. For clarification, write down the first couple of observations $x_1, x_2, x_3, x_4, \ldots$ (assume that innovations prior to the first observation are equal to zero):

$$\begin{aligned}
x_1 &= \eta_1 \\
x_2 &= \eta_2 + \psi_1 \eta_1 \\
x_3 &= \eta_3 + \psi_1 \eta_2 + \psi_2 \eta_1 \\
x_4 &= \eta_4 + \psi_1 \eta_3 + \psi_2 \eta_2 + \psi_3 \eta_1 \\
&\vdots
\end{aligned}$$

Thus, in the MA model x_t is a realization of shocks η_t that have a decaying impact on future observation. So, η_3 has an impact of 1 on x_3. The next period η_3 still influences the outcome x_4, but with an impact of $|\psi_1| < 1$, and so on. By the nature of a $MA(\infty)$ model, the influence of innovation η_t on x_{t+h} equals $\psi_h > 0$ for every $h \geq 0$ independent of t.[7]

In the ED model, assume the first four shocks $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ have a duration of $(3, 1, 2, 2)$, i.e.

$$\begin{aligned}
x_1 &= \varepsilon_1 \\
x_2 &= \varepsilon_1 + \varepsilon_2 \\
x_3 &= \varepsilon_1 + 0 + \varepsilon_3 \\
x_4 &= 0 + 0 + \varepsilon_3 + \varepsilon_4 \\
&\vdots
\end{aligned}$$

[7]For a $MA(q)$ process $\eta_h = 0$ for $h > q$, thus the *influence horizon* is equal to q.

In an economical/social environment one can interpret the shocks as innovations and the random variable n_s as a stochastic *influence horizon* of an innovation. At time $t = 1$ a shock ε_1 occurs, and people *remember* the shock for three periods. At $t = 2$ a new shock ε_2 occurs, but also ε_1 is still present. As the second shock is not very important, it has no impact on $t = 3$ anymore and vanishes.

Of course, the realized values of x_t are equal for both models, but the innovations η_t in the ARFIMA case, are different to the innovations ε_s in the ED model. As we only observe x_t and neither know the reason for nor the realized values of the shocks, which influenced the process, this distinction should not bother us.

I will resume to study this idea and refine it theoretically in Section 3.3 about structural breaks versus long memory.

2.2.4 Conditional survival probabilities

An interesting insight can be gained by studying the conditional survival probability, which is the probability that a shock ε_s survives $k+1$ periods *given* it already survived k periods. By Bayes' rule

$$P(n_s \geq k+1 | n_s \geq k) = \frac{P(n_s \geq k+1 \wedge n_s \geq k)}{P(n_s \geq k)} = \frac{P(n_s \geq k+1)}{P(n_s \geq k)} = \frac{p_{k+1}}{p_k},$$

where the second to last equality follows, as the event $n_s \geq k+1$ implies $n_s \geq k$.

For an $AR(1)$ process $p_{k+1} = \phi p_k$ and thus

$$P(n_s \geq k+1 | n_s \geq k) = \phi < 1 \quad \forall k.$$

The probability of surviving the current period k equals $\phi > 0$ independent of k. This means that the probability of surviving the next period is independent of the already survived time of the shock.

Claim 2.2.7. *The conditional survival probabilities $\frac{p_{k+1}}{p_k}$ of a long memory process converge to one.*

Proof. Assume $\frac{p_{k+1}}{p_k}$ was bounded from above by $\phi < 1$, then p_k would go to zero faster than ϕ^k. But as

$$\begin{aligned}
\sum_{k=0}^{\infty}(k+1)\phi^k &= \frac{\mathrm{d}}{\mathrm{d}\phi}\int\sum_{k=0}^{\infty}(k+1)\phi^k\,\mathrm{d}\phi = \frac{\mathrm{d}}{\mathrm{d}\phi}\sum_{k=0}^{\infty}\int(k+1)\phi^k\,\mathrm{d}\phi \\
&= \frac{\mathrm{d}}{\mathrm{d}\phi}\sum_{k=0}^{\infty}\phi^{k+1} = \frac{\mathrm{d}}{\mathrm{d}\phi}\frac{\phi}{1-\phi^2} = \frac{1+\phi^2}{(1-\phi^2)^2} < \infty,
\end{aligned}$$

in contradiction to the long memory property of the underlying model. □

Therefore, a necessary condition for a long range dependence is that conditional survival probabilities tend to 1. Loosely speaking, although far more shocks die in the first couple of periods for a long memory process than for a short memory process, once an innovation reaches a certain *importance* level the probability of dying out goes to zero for k to infinity.

2.2.5 Spectral density

Under certain conditions on p_k the process y_t is a stationary, causal process. The spectral density – if it exists – is equal to

$$\begin{aligned} f_{ED}(\lambda) &= \frac{1}{2\pi}\sum_{s=-\infty}^{\infty} e^{-i\lambda s}\gamma(s) = \frac{1}{2\pi}\left[\gamma(0) + \sum_{s=1}^{\infty}\gamma(s)\cos\lambda s\right] \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\left[1 + \sum_{s=1}^{\infty}\sum_{i=s}^{\infty} p_i \cos\lambda s\right]. \end{aligned} \tag{2.2.14}$$

If there is a maximum duration S, then $\gamma(s)$ will vanish for $s > S$ and the existence is trivially given. Note that this corresponds to the $MA(q)$ case.

Although this model allows a deep understanding of the underlying structure of a process, it is of limited use in practice, as the population second moments must satisfy condition 2.2.6. Besides, even if the population autocovariances meet the conditions, the estimated sample autocovariances $\widehat{\gamma}_x(k)$ do not satisfy condition 2.2.6 in general.

2.3 ARFIMA

Until now no assumptions have been made on the process x_t itself, but only on the second moments. Here I present the widely discussed ARFIMA model, which is a natural expansion of the well known ARIMA model.[8]

Definition 2.3.1 (ARFIMA)**.** *A process x_t, with $\mathbb{E}\, x_t = \mu$, is called an autoregressive fractionally integrated moving average process with parameters p, d and q – denoted by $ARFIMA(p, d, q)$ – if x_t solves*

$$\Phi(L)(x_t - \mu) = \Theta(L)(1 - L)^{-d}\varepsilon_t. \tag{2.3.1}$$

with AR and MA lag polynomials ($\Phi(L)$ and $\Theta(L)$, respectively) of order p and q.

If d is a non-positive integer, then we get the well known $ARIMA(p, d, q)$ process. Granger and Joyeux [30], and Hosking [37] suggested that noninteger

[8] In main points and derivations of this section I follow [57] and [34].

values of d might be useful. If $d \in \mathbb{R}$, the operator $(1-L)^{-d}$ does not necessarily exist in $\mathbb{L}^2$. Thus, define

$$n(z) := (1-z)^{-d}. \tag{2.3.2}$$

Theorem 2.3.2 (Newton's generalized binomial theorem)**.** *The binomial series is the series*

$$(1+w)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} w^k = F(-\alpha, 1; 1; -w), \tag{2.3.3}$$

where

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!} = \frac{1}{k!}\frac{\Gamma(-\alpha+k)}{\Gamma(-\alpha)} = \frac{\Gamma(-\alpha+k)}{\Gamma(-\alpha)\Gamma(k)},$$

and

$$F(a,b;c;w) := \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{w^n}{n!} \text{ with } (x)_n = x(x+1)(x+2)\cdots(x+n-1)$$

is the hypergeometric function.

a) If $|w| < 1$, the series converges for any complex number α.

b) If $|w| = 1$, the series converges absolutely if and only if either $Re(\alpha) > 0$ or $\alpha = 0$.

c) If $|w| = 1$ and $w \neq -1$, the series converges if and only if $Re(\alpha) > -1$, where $Re(a+ib) = a$.

Therefore, $n(z) = (1-z)^{-d}$ is analytic in the open disk $\{z \in \mathbb{C} | |z| < 1\}$ for every $d \in \mathbb{C}$ and converges at $z = 1 (\Leftrightarrow w = -1)$ only for $d < 0 (\Leftrightarrow Re(\alpha) > 0)$. Thus, the filter $n(L) = (1-L)^{-d}$ can be expanded to the infinite series and we get the $MA(\infty)$ representation

$$x_t = \sum_{k=0}^{\infty} \binom{d}{k} \varepsilon_{t-k} = \sum_{k=0}^{\infty} n_k \varepsilon_{t-k}. \tag{2.3.4}$$

From lemma 1.1.8 we know that x_t is covariance stationary *if and only if* $\sum_{k=0}^{\infty} |n_k|^2 < \infty$. As $n_k \in \mathbb{R}$ we can omit the modulus in the convergence analysis.

Claim 2.3.3. *The sequence $\left\{\binom{d}{k}\right\}$ is square summable for $d < \frac{1}{2}$.*

Proof. Consider the Taylor expansion of $h(x) = (1+x)^{d-1}$

$$\begin{aligned}(1+x)^{d-1} &= h(0) + \frac{\partial h}{\partial x}\Big|_{x=0} \cdot x + \frac{1}{2}\frac{\partial^2 h}{\partial x^2}\Big|_{x=\delta} \cdot x^2 \\ &= 1 + (d-1)x + \frac{1}{2}(d-1)(d-2)(1+\delta)^{d-3}x^2\end{aligned}$$

for some $\delta \in (0, x)$. For $x > -1$ and $d < 1$, this implies $(1+x)^{d-1} \geq 1+(d-1)x$.

Setting $x \equiv \frac{1}{k}$, we have

$$1 + \frac{d-1}{k} \leq \left(1 + \frac{1}{k}\right)^{d-1} = \left(\frac{k+1}{k}\right)^{d-1} \tag{2.3.5}$$

$$\begin{aligned}\binom{d}{k} &= \frac{d!}{k!(d-k)!} = \left(\frac{d+k-1}{k}\right)\left(\frac{d+k-2}{k-1}\right)\cdots\left(\frac{d}{1}\right) \\ &= \left(\frac{k+d-1}{k}\right)\left(\frac{k-1+d-1}{k-1}\right)\cdots\left(\frac{k-(k-1)+d-1}{k-(k-1)}\right) \\ &= \left(1+\frac{d-1}{k}\right)\left(1+\frac{d-1}{k-1}\right)\cdots\left(1+\frac{d-1}{k-(k-1)}\right) \\ &\leq \left(\frac{k+1}{k}\right)^{d-1}\left(\frac{k}{k-1}\right)^{d-1}\cdots\left(\frac{2}{1}\right)^{d-1} = (k+1)^{d-1}\end{aligned} \tag{2.3.6}$$

Finally we get

$$\sum_{k=0}^{\infty}\binom{d}{k}^2 \leq \sum_{k=0}^{\infty}(k+1)^{2d-2} = \sum_{n=1}^{\infty}\frac{1}{n^s} = \zeta(s) \text{ with } s = 2-2d. \tag{2.3.7}$$

The Riemann zeta function $\zeta(s)$ converges for $Re(s) > 1 \Leftrightarrow d < \frac{1}{2}$, which completes the proof. □

Thus, x_t has a covariance stationary $MA(\infty)$ representation.

Theorem 2.3.4 (Stationarity, Causality and Invertibility)**.**

a) *If the roots of $\Phi(z)$ lie outside the unit circle $\left\{z \in \mathbb{C} \,\middle|\, |z| = 1\right\}$, then there is a stationary solution to (2.3.1) given by*

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}. \tag{2.3.8}$$

This solution is unique in the $\mathbb{L}^2$ sense.

b) *If the roots of $\Phi(z)$ lie outside the closed unit disk $\left\{z \in \mathbb{C} \,\middle|\, |z| \leq 1\right\}$, then the solution is causal.*

c) *If the roots of $\Theta(z)$ lie outside the closed unit disk $\left\{z \in \mathbb{C} \,\middle|\, |z| \leq 1\right\}$, then the solution is invertible.*

Proof. Define fractional noise $\nu_t := (1-L)^{-d}\varepsilon_t$ with coefficients n_j as in (2.3.3). Then $\sum_{j=0}^{\infty} n_j^2 < \infty$ and we have convergence in $\mathbb{L}^2$. Especially $\sum_{j=0}^{\infty} n_j e^{-i\lambda}$ converges to $n(e^{-i\lambda}) = (1-e^{-i\lambda})^{-d}$. Thus, ν_t is a well defined stationary process.

Since $\Phi(z) \neq 0$ for $|z| = 1$, an absolutely convergent Laurent series exists

$$\varphi(z) = \sum_{j=-\infty}^{\infty} \varphi_j z^j = \frac{\Theta(z)}{\Phi(z)},\ \delta^{-1} < |z| < \delta, \text{ for some } \delta > 1$$

and $x_t := \varphi(L)\,\nu_t$ is a stationary process. As $\nu_t = \sum_{j=-\infty}^{\infty} n_j \varepsilon_{t-j}$ is a stationary process and $\sum_{j=-\infty}^{\infty} |\varphi_j| < \infty$ we can write

$$x_t = \Psi(L)\varepsilon_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \quad \Psi(z) = \varphi(z)\,n(z). \tag{2.3.9}$$

By applying the filter $\Phi(L)$ we get

$$\Phi(L)x_t = \Phi(L)\varphi(L)n(L)\varepsilon_t = \Theta(L)n(L)\varepsilon_t.$$

Thus, $\{x_t\}$ is a stationary process that satisfies (2.3.1).

To show uniqueness, assume there is a process $\{y_t\}$ satisfying

$$\Phi(L)y_t = \Theta(L)n(L)\varepsilon_t.$$

Since $\Phi(z) \neq 0$ on the unit circle, there is an absolutely convergent Laurent series

$$\frac{1}{\Phi(z)} = \sum_{j=-\infty}^{\infty} g_j z^j = g(z)$$

and we can multiply equation (2.3) by $g(L)$ and get

$$y_t = g(L)\Theta(L)n(L)\varepsilon_t = \varphi(L)n(L)\varepsilon_t = \Psi(L)\varepsilon_t.$$

But the difference between x_t and y_t is 0 in $\mathbb{L}^2$, which proves uniqueness.

Points b) and c) can be easily derived following the comment of Beran [4], who noted that an $ARFIMA(p,d,q)$ process can be interpreted as a fractional noise process $\nu_t := (1-L)^{-d}\varepsilon_t$ passed through an $ARMA(p,q)$ filter. The definition for stationarity and invertibility is not constrained to white noise, but only to stationary processes. Since ν_t is stationary, the conditions for stationarity and invertibility are satisfied, as we already know that such a $ARMA(p,q)$ model is causal and invertible. □

Sometimes an ARFIMA process is defined as

$$\Phi(L)(1-L)^d x_t = \Theta(L)\varepsilon_t.$$

For the persistent case ($d > 0$) the solution to this equation is not unique. Let y_t be a solution, and assume W is a random variable with finite variance. Then the process $x_t := y_t + W$ is stationary and also satisfies the equation since for $d > 0$, the coefficients of $(1-z)^d = \sum_{j=0}^{\infty} \pi_j = \Pi(z)$ are absolutely summable, and $\Pi(1) = 0$.

Corollary 2.3.5. *An autoregressive fractional integrated moving average process x_t is the unique solution to (2.3.1). If the roots of the AR and MA polynomials lie outside the unit disk*

$$\Theta(z) \neq 0 \text{ and } \Phi(z) \neq 0, \quad |z| \leq 1$$

and $d \in \left(-1, \frac{1}{2}\right)$, then x_t has a covariance stationary, causal and invertible $MA(\infty)$ representation

$$\begin{aligned} x_t &= (1-L)^{-d}\frac{\Theta(L)}{\Phi(L)}\varepsilon_t = \sum_{k=0}^{\infty}\binom{d}{k}L^k\frac{\Theta(L)}{\Phi(L)}\varepsilon_t = \Psi(L)\varepsilon_t \\ &= \sum_{n=0}^{\infty}\psi_n L^n \varepsilon_t, \end{aligned} \tag{2.3.10}$$

where the weighting sequence $\{\psi_n\}$ can be calculated by matching terms.

The constraint to the interval $\left(-1, \frac{1}{2}\right)$ is not a rigid restriction as we can easily define $ARFIMA(p, d, q)$ processes with $d \notin \left(-1, \frac{1}{2}\right)$.

Definition 2.3.6. *A process y_t is called integrated of order $D \notin \left(-1, \frac{1}{2}\right)$, iff $(1-L)^{[D+\frac{1}{2}]}y_t \sim I(d)$ with $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, where $[x]$ is the smallest integer less or equal to x.*

For example, y_t is said to be integrated of order 2.3 if $(1-L)^2 y_t \sim I(0.3)$.

Interested in stationary and invertible processes and due to the manipulations above we can restrict further analysis to $d \in [-\frac{1}{2}, \frac{1}{2})$.[9] So far the solution x_t to (2.3.1) is covariance-stationary, causal and invertible for $d \in \left(-1, \frac{1}{2}\right)$. Below we see that such processes actually display long memory, as defined in (2.0.3).

[9] I will analyze the special boundary case of $d \equiv \frac{1}{2}$ within the ED model in Section 2.3.2.

2.3.1 Frequency domain analysis

From theorem 1.4.1 we obtain that $\sum_{k=0}^{\infty} n_k e^{-i\lambda}$ converges to $n(e^{-i\lambda})$ and the spectral density of x_t exists (in the $\mathbb{L}^2$ sense) and is given by

$$\begin{aligned} f_x(\lambda) &= n(e^{-i\lambda}) f_\varepsilon(\lambda) n(e^{i\lambda}) = f_\varepsilon(\lambda)(1-e^{-i\lambda})^{-d}(1-e^{i\lambda})^{-d} \\ &= \frac{\sigma_\varepsilon^2}{2\pi}\left(2\sin\frac{\lambda}{2}\right)^{-2d}. \end{aligned} \tag{2.3.11}$$

On the other hand

$$f_x(\lambda) := \frac{1}{2\pi}\sum_{s=-\infty}^{\infty} e^{-i\lambda s}\gamma_x(s), \tag{2.3.12}$$

with $\gamma_x(s) = \mathbb{E}(x_t-\mu)(x_{t-s}-\mu)$. Already knowing the explicit form of the spectral density we can calculate the autocovariances $\gamma_x(s)$ by the inverse Fourier transform.

Notice that the usually used point-wise convergence condition

$$\sum_{s=-\infty}^{\infty} |\gamma_x(s)| < \infty,$$

does not hold by definition of long memory. Although we already have a well defined spectrum in the $\mathbb{L}^2$ sense ($\sum_{s=-\infty}^{\infty} |\gamma_x(s)|^2$ is finite for $d < 0.5$), I also prove the existence of $f(\lambda)$ in $\mathbb{L}^1[-\pi,\pi]$ for $\lambda \in [\delta,\pi)$, $\delta > 0$.

Why $\mathbb{L}^2$ is not sufficient?

Knowing that the spectrum exists in $\mathbb{L}^2$, $\mathbb{L}^1$ convergence seems to be unnecesserily complex. Yet, an overseen condition – as it is somehow trivial, when speaking of $\mathbb{L}^2$ convergence – is that the spectral density $f_x(\lambda)$ must be an element of $\mathbb{L}^2$. But

$$f_x(\lambda) \in \mathbb{L}^2 \Leftrightarrow \left(\int_{-\pi}^{\pi} \big(f_x(\lambda)\big)^2 \,\mathrm{d}\lambda\right)^{\frac{1}{2}} < \infty \Leftrightarrow \int_{-\pi}^{\pi} \big(f_x(\lambda)\big)^2 \,\mathrm{d}\lambda < \infty.$$

Lemma 2.3.7. *Let $\phi(x) > 0$, $f(x)$ nonnegative, unbounded and $\lim_{x\to b} f(x) = \infty$, but integrable over every interval $[a,c]$ for $a < c < b$. Calculate*

$$\lim_{x\to b} \frac{\phi(x)}{f(x)} = c \in [0,\infty].$$

a) If $c > 0$ and $\int_a^b \phi(x)\,\mathrm{d}x$ converges, so does the integral of $f(x)$.

b) If $c < \infty$ and $\int_a^b \phi(x)\,\mathrm{d}x$ diverges, so does the integral of $f(x)$.

Proof. See Prudnikov, Marichev, and Brychkov [66], p. 739. □

Take $\phi(\lambda) = \left(\frac{\lambda}{2}\right)^{-2d}$. Then

$$\lim_{\lambda\to 0} \frac{f_x(\lambda)}{\phi(\lambda)} = \lim_{\lambda\to 0} \frac{\left(2\sin\frac{\lambda}{2}\right)^{-2d}}{\left(\frac{\lambda}{2}\right)^{-2d}} = 2^{-2d} \lim_{\lambda\to 0} \left(\frac{\sin\frac{\lambda}{2}}{\frac{\lambda}{2}}\right)^{-2d} = 2^{-2d} \notin \{0, \infty\}.$$

Therefore, $f_x(\lambda)$ converges if and only if $\phi(\lambda)$ converges.

Ergo we can set up conditions on d such that $f_x(\lambda) \in \mathbb{L}^1$ and $\mathbb{L}^2$,respectively. As

$$\|f_x\|_1 = \int_{-\pi}^{\pi} f_x(\lambda)\,\mathrm{d}\lambda < \infty \Leftrightarrow \int_{-\pi}^{\pi} \left(\frac{\lambda}{2}\right)^{-2d} \mathrm{d}\lambda = \frac{2}{2^{-2d}} \int_0^{\pi} \lambda^{-2d}\,\mathrm{d}\lambda < \infty,$$

we obtain $f_x(\lambda) \in \mathbb{L}^1$ iff $-2d > -1 \Leftrightarrow d < \frac{1}{2}$.

Yet,

$$\sqrt{\|f_x\|_2} = \int_{-\pi}^{\pi} f_x(\lambda)^2\,\mathrm{d}\lambda < \infty \Leftrightarrow \int_{-\pi}^{\pi} \left(\left(\frac{\lambda}{2}\right)^{-2d}\right)^2 \mathrm{d}\lambda = \frac{2}{2^{-4d}} \int_0^{\pi} \lambda^{-4d}\,\mathrm{d}\lambda < \infty,$$

and consequently $f_x(\lambda) \in \mathbb{L}^2$ iff $-4d > -1 \Leftrightarrow d < \frac{1}{4}$.

Thus, only settling for the trivially given $\mathbb{L}^2[-\pi, \pi]$ convergence has the immense flaw that although the process is covariance stationary for $d < \frac{1}{2}$, the corresponding spectrum is only defined for $d < \frac{1}{4}$.[10]

So, for $d \in [0, \frac{1}{4})$ we define $f_x(\lambda)$ in the $\mathbb{L}^2$ sense, and for $d \in [\frac{1}{4}, \frac{1}{2})$ in the $\mathbb{L}^1$ sense.

Compute the autocovariances of x_t

As already noted above, we will compute the autocovariances $\gamma_x(j)$ by the inverse Fourier transform.

$$\begin{aligned}
\gamma_x(j) &= \int_{-\pi}^{\pi} e^{i\lambda j} f_x(\lambda)\,\mathrm{d}\lambda \\
&= \frac{\sigma_\varepsilon^2}{2\pi} \left\{ \int_{-\pi}^{\pi} \cos(\lambda j) \left(2\sin\frac{\lambda}{2}\right)^{-2d} \mathrm{d}\lambda + i \int_{-\pi}^{\pi} \sin(\lambda j) \left(2\sin\frac{\lambda}{2}\right)^{-2d} \mathrm{d}\lambda \right\} \\
&= \frac{\sigma_\varepsilon^2}{2\pi} \int_{-\pi}^{\pi} \cos(\lambda j) \left(2\sin\frac{\lambda}{2}\right)^{-2d} \mathrm{d}\lambda = \frac{\sigma_\varepsilon^2}{\pi} \int_0^{\pi} \cos(\lambda j) \left(2\sin\frac{\lambda}{2}\right)^{-2d} \mathrm{d}\lambda \\
&= \sigma_\varepsilon^2 \frac{(-1)^j \Gamma(1-2d)}{\Gamma(1-d+j)\Gamma(1-d-j)}. \qquad (2.3.13)
\end{aligned}$$

The second line holds, since $h(\lambda) := \sin(\lambda j) \left(2\sin\frac{\lambda}{2}\right)^{-2d}$ is an odd function

[10] Additionally, in practice most long memory processes exhibit long memory with $d > \frac{1}{4}$.

on $[-\pi, \pi]$; thus, $\int_{-\pi}^{\pi} h(\lambda) = 0$. The identity

$$\int_0^{\pi} \cos(jx) \sin^{r-1}(x)\, \mathrm{d}x = \frac{\pi \cos(\frac{j\pi}{2}) \Gamma(r+1) 2^{1-r}}{r\Gamma(\frac{r+h+1}{2})\Gamma(\frac{r-h+1}{2})},$$

gives the final result.[11]

Especially

$$\sigma_x^2 = \gamma_x(0) = \sigma_\varepsilon^2 \frac{\Gamma(1-2d)}{\big(\Gamma(1-d)\big)^2}. \tag{2.3.14}$$

Therefore, the autocorrelations are

$$\rho_x(j) = \frac{\gamma_x(j)}{\gamma_x(0)} = \frac{\Gamma(j+d)\Gamma(1-d)}{\Gamma(j-d+1)\Gamma(d)}. \tag{2.3.15}$$

Applying Stirling's asymptotic approximation $\Gamma(x) \sim \sqrt{2\pi} e^{-x+1}(x-1)^{x-\frac{1}{2}}$, we get

$$\rho_x(j) \sim \frac{\Gamma(1-d)}{\Gamma(d)} j^{2d-1} \text{ for large j.}$$

Herewith the autocorrelations of an ARFIMA process display a hyperbolical decay, i.e. a slower decay as the exponential rate of an ARMA process. This hyperbolical decay and $0 < d < \frac{1}{2}$ results in non-summable autocorrelations. For a quantitative comparison see Table 2.1 and Figure 2.3. Both the autocorrelations are the same for lag 1, but the long memory autocorrelations decay very slowly ($\rho_x(100) = 0.0338$), whereas the $AR(1)$ autocorrelation at lag 100 is in fact zero.

From properties of the Gamma function $\Gamma(\cdot)$ it follows that the autocorrelations of an $ARFIMA(0, d, 0)$ process have the same sign as the memory parameter d, for all lags k. The autocorrelations for a persistent process are positive for all k, and are all negative for anti-persistent memory. On the opposite the autocorrelations of an $AR(1)$ process with $\rho_1 < 0$ alternate.

The next lemma gives sufficient and necessary conditions for the $\mathbb{L}^1$-convergence of even functions, save possibly at $\lambda = 0$.

Lemma 2.3.8 (Convergence of the Fourier series)**.** *Let $f \in \mathbb{L}^1[-\pi, \pi]$ be an even function and $S_n(x) := \sum_{j=0}^{n} a_j \cos jx$ be the partial sums of the Fourier series of f, with $a_j =< f, e^{i\lambda j} >= \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda j}\, \mathrm{d}\lambda$. If the sequence $\{a_j\}$ is nonnegative, decreasing, and $\lim_{j\to\infty} a_j = 0$, then*

$$\lim_{n\to\infty} \big\| f - S(n) \big\|_1 = 0 \quad (\textit{save possibly at } x = 0), \tag{2.3.16}$$

if and only if

$$\lim_{j\to\infty} a_j \log j = 0. \tag{2.3.17}$$

[11] See Gradshteyn and Ryzhik [25].

The norm in $\mathbb{L}^1[-\pi,\pi]$ is defined as

$$\|f\|_1 = \int_{-\pi}^{\pi} |f(x)|\,\mathrm{d}x.$$

Proof. See Zygmund [83], p. 183/184. □

If $\gamma_x(s)$ satisfies (2.3.17), then the spectrum is well defined in $\mathbb{L}^1$ for $0 < \delta \leq \lambda \leq \pi$, and for $0 \leq \lambda < \delta$ we use $\mathbb{L}^2$ convergence.

Claim 2.3.9. *The autocovariances $\gamma(s)$ of fractional noise $ARFIMA(0,d,0), d < \frac{1}{2}$ are positive and nonincreasing, and satisfy*

$$\lim_{s\to\infty} \gamma(s)\log s = 0.$$

Proof.

$$\begin{aligned}
\lim_{s\to\infty} \gamma(s)\log s &= \gamma(0)\lim_{s\to\infty}\rho(s)\log s = \gamma(0)\lim_{s\to\infty}\frac{s^{2d-1}}{s^{2d-1}}\rho(s)\log s \\
&= \gamma(0)\lim_{s\to\infty}\frac{\rho(s)}{s^{2d-1}}s^{2d-1}\log s = \gamma(0)\underbrace{\lim_{s\to\infty}\frac{\rho(s)}{s^{2d-1}}}_{=1 \text{ by Stirling}}\lim_{s\to\infty} s^{2d-1}\log s \\
&\overset{\alpha=1-2d}{=} \gamma(0)\lim_{s\to\infty} s^{-\alpha}\log s = \gamma(0)\lim_{s\to\infty}\frac{\log s}{s^{\alpha}} \\
&\overset{\text{de l'Hospital}}{=} \gamma(0)\lim_{s\to\infty}\frac{\frac{1}{s}}{\alpha s^{\alpha-1}} = \gamma(0)\lim_{s\to\infty}\frac{1}{\alpha s^{\alpha}} \\
&= 0 \text{ given } \alpha > 0 \Leftrightarrow d < \frac{1}{2}.
\end{aligned}$$

□

Thus, the spectral density of fractional white noise is well defined and we can proceed with the spectral density of a general ARFIMA process. For this rearrange

$$x_t = (1-L)^{-d}\frac{\Theta(L)}{\Phi(L)}\varepsilon_t \text{ to } x_t = \frac{\Theta(L)}{\Phi(L)}\underbrace{(1-L)^{-d}\varepsilon_t}_{=:\nu_t}.$$

Since ν_t is covariance stationary and has a well defined density f_ν we get

$$f_x(\lambda) = \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}f_\nu(\lambda) = \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}\frac{\sigma_\varepsilon^2}{2\pi}\left(2\sin\frac{\lambda}{2}\right)^{-2d}. \tag{2.3.18}$$

Using again the asymptotic approximation

$$\lim_{\lambda\to 0}\frac{\sin\frac{\lambda}{2}}{\frac{\lambda}{2}} = 1,$$

Lag j	$\rho_{I(d)}(j)$ $d=\frac{1}{4}$	$\rho_{AR(1)}(j)$ $\rho=\frac{1}{3}$	$\rho_{I(d)}(j)$ $d=-\frac{1}{4}$	$\rho_{AR(1)}(j)$ $\rho=-\frac{1}{5}$
1	0.3333	0.3333	-0.2000	-0.2000
2	0.2381	0.1111	-0.0667	0.0400
3	0.1948	0.0370	-0.0359	-0.0080
4	0.1688	0.0123	-0.0232	1.6×10^{-3}
5	0.1511	4.11×10^{-3}	-0.0166	-3.2×10^{-4}
10	0.1069	1.69×10^{-5}	-5.85×10^{-3}	1.03×10^{-7}
20	0.0756	2.86×10^{-10}	-2.07×10^{-3}	1.05×10^{-14}
50	0.0478	1.39×10^{-24}	-5.23×10^{-4}	1.125×10^{-35}
100	0.0338	1.94×10^{-48}	-1.85×10^{-4}	1.27×10^{-70}
200	0.0239	3.76×10^{-96}	-6.54×10^{-5}	1.61×10^{-140}
500	0.0151	2.75×10^{-239}	-1.65×10^{-5}	3.27×10^{-280}

Table 2.1: Exact autocorrelations for $I(d)$ and $AR(1)$ processes

we can easily prove that x_t displays long memory since

$$\begin{aligned}
\lim_{\lambda\to 0} f_x(\lambda) &= \frac{\sigma_\varepsilon^2}{2\pi}\lim_{\lambda\to 0}\left(2\sin\frac{\lambda}{2}\right)^{-2d}\frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2} \\
&= \frac{\sigma_\varepsilon^2}{2\pi}\frac{|\Theta(1)|^2}{|\Phi(1)|^2}\lim_{\lambda\to 0}\lambda^{-2d} \qquad (2.3.19)\\
&= \begin{cases} 0 & d<0, \\ \infty & d>0. \end{cases}
\end{aligned}$$

Conclusion

A process x_t satisfying definition 2.3.1 is a stationary, causal and invertible process with a well defined spectral density $f_x(\lambda)$ and exhibits long memory. As ARMA parts also play a role in this model, an ARFIMA process is a suitable and – at least equally important – a parsimonious model for short *and* long run behavior of both stationary and non-stationary time series.

2.3.2 ARFIMA versus error duration

Since the autocovariances $\gamma(k)$ for a persistent $ARFIMA(0,d,0)$ process x_t are positive and nonincreasing, we can calculate survival probabilities p_k, such that the second moments of y_t from the corresponding error duration model are equivalent to the second moments of x_t.

Claim 2.3.10. *For* $0<d\leq 1$ *the survival probabilities for an* $ARFIMA(0,d,0)$

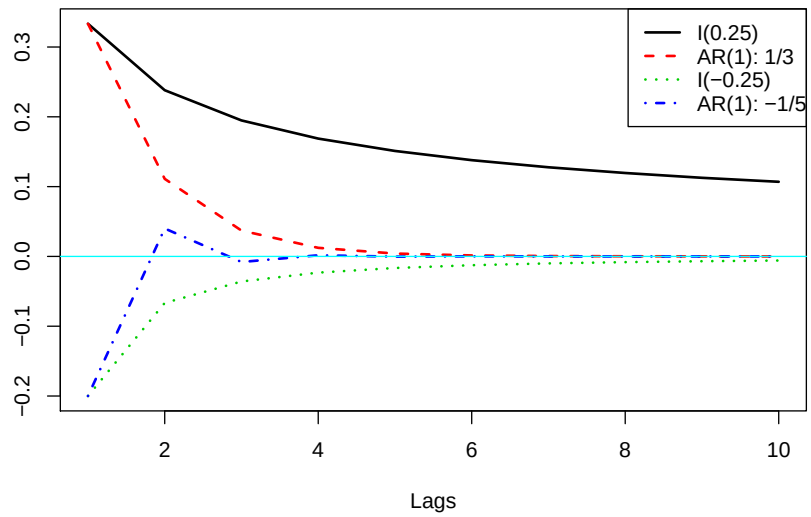


Figure 2.3: $\rho(j)$: Theoretical autocorrelation functions for $ARFIMA(0, d, 0)$ and $AR(1)$ processes.

process are

$$p_k = \frac{\Gamma(k+d)\Gamma(2-d)}{\Gamma(k+2-d)\Gamma(d)}. \tag{2.3.20}$$

Proof. See Parke [58]. □

Yet, the resulting ED process is just a special case of an $ARFIMA(p, d, q)$ process, i.e. only persistent memory can be modeled, since $p_k \geq 0$. Moreover the ED model can mimic some features of an ARFIMA model, but certain properties (e.g. asymptotic results) are not the same.

To cut a long story short, the ED model does not equal an ARFIMA model, but for our purposes it is a useful, different approach to the second moments structure of fractionally integrated processes.

Example 2.3.11 ($ARFIMA(0, 0.4, 0)$)**.** *For $T = 3000$, i.i.d. Gaussian innovations, and probabilities*

$$p_k = \frac{\Gamma(k+d)\Gamma(2-d)}{\Gamma(k+2-d)\Gamma(d)} \quad k = 1, \ldots 3000, d = 0.4,$$

we get one realization displayed in Figure 2.4. Not only the figures, but also the estimates $(\widehat{d}_{ELW}, \widehat{d}_{GPH}) = (0.36, 0.38)$ (for $m = T^{0.6} \approx 122$) demonstrate the capability of the error duration model to simulate fractionally integrated processes. A maximum likelihood estimator gives an $ARFIMA(0, d, 1)$ model for x_t,

$$(1-L)^{0.32} x_t = u_t, \quad u_t = \epsilon_t + 0.6\epsilon_{t-1}, \quad \widehat{\sigma}_\epsilon = 1.5.$$

But only this special form of p_k results in an $I(d)$ realization. Thus, the ED model is on the one hand limited as it can not mimic anti-persistent behavior,

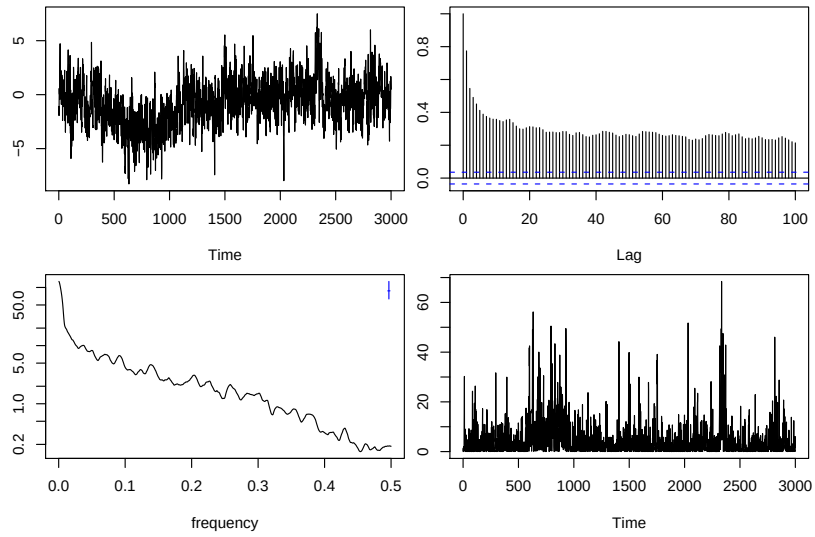


Figure 2.4: Error Duration simulation for $d = 0.4$ and $T = 3000$.

but on the other hand can simulate processes with long range dependence that are *not* fractionally integrated.

Nonetheless it must be mentioned that any long memory process y_t with a $MA(\infty)$ representation can be approximated arbitrarily well by an ARFIMA process x_t in the sense that

$$\left| \frac{f_y(\lambda)}{f_x(\lambda)} - 1 \right| < \varepsilon,$$

uniformly for $\lambda \in [-\pi, \pi]$ [see 57, p. 55].

Conditional survival probability

The survival probabilities of an $I(d)$ process satisfy the recursion

$$p_{k+1} = \frac{k+d}{k+2-d} p_k \Leftrightarrow \frac{k+d}{k+2-d} = \frac{p_{k+1}}{p_k} = P(n_s \geq k+1 | n_s \geq k).$$

As already proved above, the conditional survival probabilities for an $I(d)$ process tend to 1, in contrast to short memory models.

On the threshold

Now consider the special case of $d \equiv \frac{1}{2}$, where $I(d)$ represents a process on the threshold between stationarity and non-stationarity. The coefficients ψ_k of the MA(∞) representation of an $I(\frac{1}{2})$ process are:

$$\{\psi_k\}_{k=0}^{\infty} = \left\{ 1, \frac{1}{2}, \frac{3}{8}, \frac{15}{48}, \frac{105}{384}, \frac{945}{3840}, \ldots \right\}.$$

Trying to intuitively expand these coefficients for higher k seems impossible. On the contrary the survival probabilities for an $I(\frac{1}{2})$ process follow a very regular pattern

$$\{p_k\}_{k=0}^{\infty} = \left\{1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}, \ldots\right\},$$

and an explicit formula can be found very easily: $p_k^{I(\frac{1}{2})} = \frac{1}{2k+1}$ for $k = 0, 1, 2, \ldots$.

Both expressions are valid, and both indicate a non-stationary process because the variance of y_t is infinite. Nevertheless the ED representation gives a more natural, intuitive approach to this special border case than the MA representation.

Chapter 3

Long Memory versus Structural Breaks: Is it spurious?

Although the models in the previous chapter generate covariance stationary long memory processes, there are also other processes that display the same properties as long range dependence. Thereupon many authors [3, 51, 74, 75] point out that long memory is a spurious phenomenon, evoked by structural breaks in the series. Although we can apply estimation methods derived for stationary time series, the results might falsely indicate long range dependence, as the underlying process is in fact non-stationary.

3.1 Structural breaks

Domingo and Tonella [19] give a very good description of structural change:

> *Structural changes appear when some part or properties are lost or added to the object, some relations appear, disappear or change their form. [...] this may happen in such a small degree that the change is unnoticeable, or in such a degree that the system becomes practically a new one.*

Application of the theory of stationary time series require a constant mean μ, constant variance σ^2 and time-invariant autocovariances $\gamma(k)$. These are simplifications to develop a profound theory of time series. But real world time series often violate at least one of these conditions, e.g.

- Trending time series (e.g. world population) violate the first condition of a constant μ. By detrending the series first one can apply stationary theory and results to the new series.

- Integrated series do not have a constant variance. Consider the *simple*

random walk:

$$y_t = y_{t-1} + \varepsilon_t \Leftrightarrow y_t = \sum_{j=1}^{t} \varepsilon_j$$

$$\Rightarrow \sigma_y^2 = \mathbb{V} y_t = \mathbb{E} \sum_{j=1}^{t} \varepsilon_j \sum_{l=1}^{t} \varepsilon_l = \sum_{j=1}^{t} \sigma_\varepsilon^2 = t\sigma_\varepsilon^2$$

This implies $\lim_{t\to\infty} \sigma_y^2 = \infty$. Differencing the process y_t gives simple white noise. In practice, one often can observe that differencing integrated time series will nevertheless give a – bounded, but still – time-varying variance (e.g. stock returns).

- The autocovariance structure of a process can change over time.

The first two points, are the basis for a vast literature on tests between stochastic and deterministic trends, i.e. tests to differ between *I(1)* versus *I(0) + trend*. Likewise common tests for structural breaks versus long memory, are based on the distinction $I(d), d \in \mathbb{R}$ versus *I(0) + structural change*.

Remark 3.1.1. *Recently, the third assumption of a constant second moment structure gained high interest and a lot of current research deals with models allowing time-varying coefficients, and developing estimation techniques for e.g. time-varying spectra. In Chapter 6, I present a basic, empirically motivated approach for time-varying memory. For deeper analysis see [9, 60].*

A standard test for structural change is the $CUSUM$ test, either based on recursive OLS or standard OLS residuals.

$$TS = \sup_{\lambda \in (0,1)} |C_T(\lambda)|, \text{ where } C_T(\lambda) = \frac{1}{\widehat{\sigma}_\varepsilon T^{1/2}} \sum_{t=1}^{[\lambda T]} e_t,$$

and e_t are the (recursive) residuals of the expanding model $y_i = x_i\beta + \varepsilon_i$, $i = 1, \ldots t$. Large values of the TS lead to a rejection of a constant parameter vector β.
Krämer and Sibbertsen [43] show that TS tends to infinity (both for recursive and standard residuals) in the presence of long memory disturbances, meaning that the probability of rejecting a constant parameter vector β tends to 1.

3.1.1 Spurious long memory

Diebold and Inoue [18] give a detailed overview about three models with changes in the mean that exhibit long range dependence. They perform an extensive Monte Carlo simulation and I refer to their work for an overview of a finite sample analysis.

Mixture model

Consider the simple model

$$v_t = \begin{cases} 0 & \text{with probability } 1-p, \\ w_t \overset{i.i.d.}{\sim} N(0, \sigma_w^2) & \text{with probability } p. \end{cases}$$

Note that $\mathbb{V}\left(\sum_{t=1}^T v_t\right) = pT\sigma_w^2 = \mathcal{O}(T)$. By definition 2.0.7 v_t is $I(0)$. It is straightforward to show that for $p = \mathcal{O}(T^{2d-2})$, $\mathbb{V}\left(\sum_{t=1}^T v_t\right) = \mathcal{O}(T^{2(d-1)+1})$ and thus $v_t \sim I(d-1)$.

Now consider the more advanced mean-plus-noise model

$$\begin{aligned} x_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2) \\ \mu_t &= \mu_{t-1} + v_t \\ v_t &= \begin{cases} 0 & \text{with probability } p, \\ w_t \overset{i.i.d.}{\sim} N(0, \sigma_w^2) & \text{with probability } 1-p. \end{cases} \end{aligned}$$

A solution for μ_t is given by $\mu_t = \sum_{j=1}^t v_j$, and as v_t is $I(d-1)$ we have $\mu_t \sim I(d)$. As μ_t is uncorrelated with ε_t

$$\begin{aligned} \mathbb{V}x_t &= \mathbb{V}\mu_t + \mathbb{V}\varepsilon_t \\ \text{and } f_x(\lambda) &= f_\mu(\lambda) + f_\varepsilon(\lambda). \end{aligned}$$

Hence, x_t exhibits long memory with memory parameter d.

STOPBREAK model

Engle and Smith [20] propose the *stochastic permanent break* model,

$$\begin{aligned} x_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2) \\ \mu_t &= \mu_{t-1} + q_{t-1}\varepsilon_{t-1} \end{aligned}$$

where $q_t = q(|\varepsilon_t|)$ can be any nondecreasing function in $|\varepsilon_t|$ and bounded by zero and one, which means that bigger innovations have more permanent effects. In their study [20] use $q_t = \frac{\varepsilon_t^2}{\gamma + \varepsilon_t^2}$ with $\gamma > 0$.

Calculating $\mathbb{V}\left(\sum_{t=1}^T \Delta x_t\right)$, Diebold and Inoue [18] again show that for $\gamma = \mathcal{O}(T^\delta)$, and certain regularity assumptions on ε_t, the stochastic process x_t is $I(1-\delta)$.

Note that the standard STOPBREAK model is $I(1)$, as $\gamma_t = \gamma = \mathcal{O}(T^0)$.

Markov switching model

Let $\{s_t\}_{t=1}^T$ be a Markov process switching between state 1 and state 2 with transition matrix

$$M := \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix}, \text{ with } M_{i,j} = P(s_t = j | s_{t-1} = i) \quad \forall t.$$

For both states, s_t is a first order autoregressive processes with corresponding means μ_1 and μ_2. Now consider the sample path of an observed time series $\{y_t\}_{t=1}^T$ with conditional density depending on μ_{s_t},

$$f(y_t|s_t;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y_t-\mu_{s_t})^2}{2\sigma^2}}.$$

Hence, y_t is Gaussian white noise with a Markov switching mean (between μ_1 and μ_2 respectively) and can be written as

$$y_t = \mu_{s_t} + \varepsilon_t,$$

where $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$, and s_τ and ε_t are independent for all t and τ.

Claim 3.1.2. *For $\mu_1 \neq \mu_2$ and $p_{11} = 1 - c_1 T^{-\delta_1}$ and $p_{22} = 1 - c_2 T^{-\delta_2}$, with $\delta_1, \delta_2 > 0$, and $0 < c_1, c_2 < 1$, the process y_t is $I(\min(\delta_1, \delta_2))$.*

Proof. See [18]. □

As the transition probabilities in the standard Markov switching model do not depend on T it is $I(0)$ in contrast to the standard STOPBREAK model, which is $I(1)$.[1]

3.2 Testing long memory versus short memory

In the previous chapter I have presented various models generating *spurious* long memory. Krämer and Sibbertsen [43] and Mayoral [51, 52] propose test procedures to distinguish between long memory processes and short memory processes with random shifts in the mean.

3.2.1 Subsampling the process

Shimotsu [73] introduces a testing procedure for long memory versus short memory plus breaks by splitting up the whole sample in b subsamples and analyzing the underlying divided model. For a long memory process the subsample models are also long range dependent, and thus the local estimates should be close to

[1] Granger and Hyung [29] also use a Markov switching process to generate stochastic processes that exhibit long memory behavior.

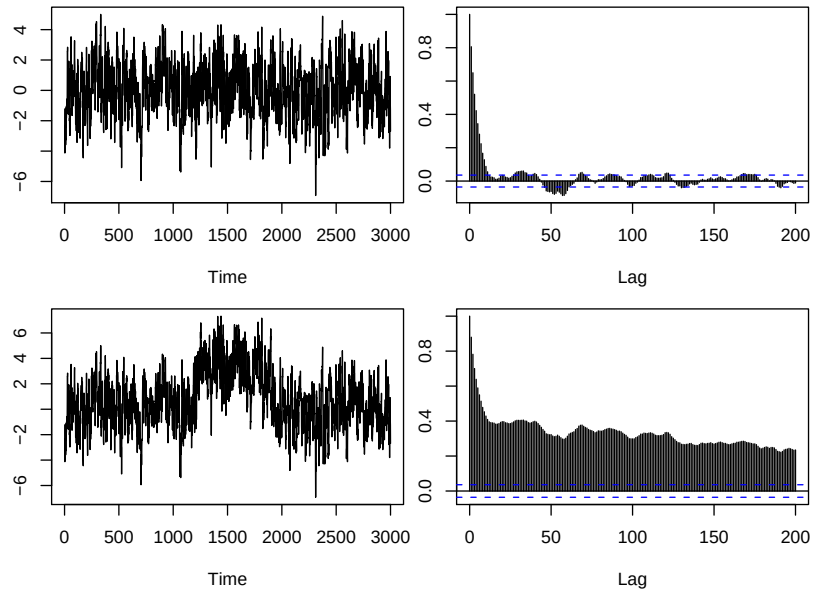


Figure 3.1: Long memory versus Structural Breaks: (top left) stationary $AR(1)$ u_t; (top right) autocorrelation function of u_t; (bottom left) Short memory plus structural breaks: $x_t = u_t + 3\, I_{1200 \le t \le 1900}$; (bottom right) autocorrelation function of x_t.

the global $\widehat{d}$. This is not the case for models with spurious long memory. [73] considers the three presented models for structural breaks and provides a Monte Carlo simulation on the behavior of the test for structural break alternatives.

Illustration of the idea

As a motivating example consider the specific model

$$x_t = (1-L)^{d_0} u_t + \delta\, I_{1200 \le t \le 1900}, \quad t = 1, 2, \ldots, 3000, \text{ with } u_t = 0.9 u_{t-1} + \varepsilon_t. \tag{3.2.1}$$

For $d_0 = 0$ and $\delta \neq 0$ the observed process is of the form $AR(1)$ *plus breaks*, whereas for $d_0 \neq 0$ and $\delta = 0$ it is an *ARFIMA(1,d,0)* process.

Assume we observe $\{x_t\}_{t=1}^{T=3000}$, generated with $d_0 = 0$ and $\delta = 3$. Therefore, the realization is not an $ARFIMA(1, d, 0)$ process, but a short memory $AR(1)$ process with an upward mean shift of magnitude 3 for $t \in [1200, 1900]$. Nevertheless the autocorrelations suggest long range dependence (see Figure 3.1), and a maximum likelihood estimate, for an underlying $ARFIMA(1, d, 0)$ model, gives $(\widehat{d}, \widehat{\phi}_1) = (0.40, 0.45)$ with standard error $(0.015, 0.023)$.

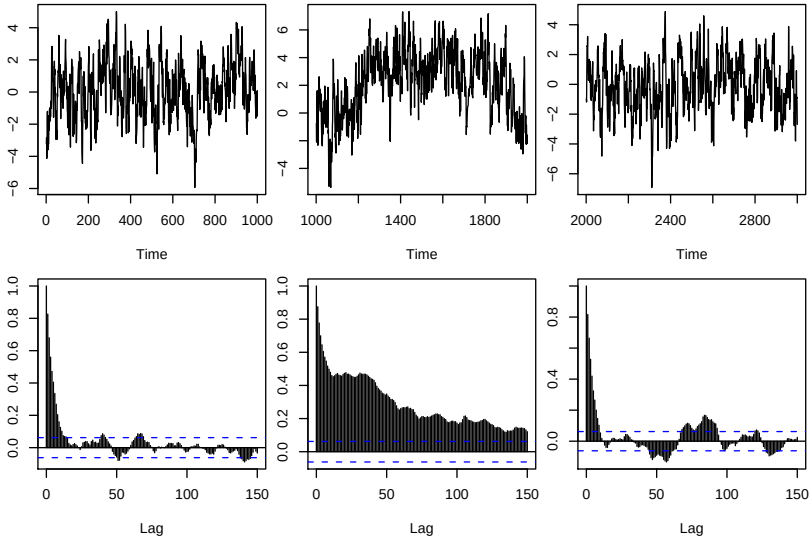


Figure 3.2: For subperiods 1 and 3, x_t is a zero mean $AR(1)$ process; for subperiod 2, it is an $AR(1)$ with shift in the mean $(0 \to 3)$.

Splitting up the model in 3 subsamples of equal size gives

$$x_t = (1-L)^{d_0} u_t, \quad t = 1, \ldots, 1000, \tag{3.2.2}$$

$$x_t = (1-L)^{d_0} u_t + \delta\, I_{1200 \le t \le 1900}, \quad t = 1001, \ldots, 2000, \tag{3.2.3}$$

$$x_t = (1-L)^{d_0} u_t, \quad t = 2001, \ldots, 3000. \tag{3.2.4}$$

The only possible spurious long memory can occur in the second subsample, as the other ones are purely $I(0)$. Hence, memory parameter estimates for subsample 1 and 3 should give $\widehat{d}^{(i)} \approx 0$, $i \in \{1, 3\}$, whereas for subsample 2, spurious long memory could occur again and $\widehat{d}^{(2)} = \tilde{d}$, in general not close to zero, and presumably $\tilde{d}$ is even greater than – the also spurious – $\widehat{d}$ for the whole sample.

The proposed test compares the global estimate $\widehat{d}$ for the whole sample, with the local estimates $\widehat{d}^{(i)}, i = 1, 2, 3$. A first approach would probably compare the average of $\widehat{d}^{(i)}, i = 1, 2, 3$ with $\widehat{d}$. It is correct that if the process is truly long range dependent $(d_0 \neq 0, \delta = 0)$, then the average of the local estimates is approximately equal to the global estimator. But note that in general the reverse is not true (see Table 3.2), as for an *I(0) plus breaks process* the lower estimates for the $I(0)$ subsample could cancel out with a substantially higher $\tilde{d}$ in the break subsample.

General study

Practically we want to construct a test between H_0 : *long memory* versus H_1 : *short memory plus breaks*. To parametrize H_0 we use the derivations above and divide the sample in b subsamples. Under H_0 (long memory) the local long memory parameters $d_0^{(i)}, i = 1, \ldots, b$ are all the same, and furthermore are equal

	$\{x_t\}_{t=1}^{3000}$	$\{x_t\}_{t=1}^{1000}$	$\{x_t\}_{t=1001}^{2000}$	$\{x_t\}_{t=2001}^{3000}$	average	d_{true}
$\widehat{d}$	0.40	0.01	0.44	0.01	0.15	0
$\widehat{\phi}_1$	0.45	0.82	0.37	0.81	0.66	0.9

Table 3.1: Local variation in the *spurious* memory parameter for x_t defined in (3.2.1)

to the global d_0. Thus, a parametric null hypothesis is given by

$$H_0 : d_0 = d_0^{(1)} = d_0^{(2)} = \cdots = d_0^{(b)}.$$

Now assume that for a sample process estimates are $\widehat{d}$ for the global parameter and $\widehat{d}^{(i)}$ for subsample i. Loosely speaking, the null hypothesis becomes less probable the larger the distance between the parameter estimates and the true parameters – *under* H_0 – gets. So, with an appropriate distance measure we get a quantitative test statistic and are able to derive certain properties of the test (e.g. critical values, asymptotic behavior). Recall that in the example above, we know the parametric structure of the short memory process u_t ($AR(1)$) and a maximum likelihood estimator can be applied. In general, we just assume short memory for u_t and make no assumptions on the type of short memory. Thus, a semi-parametric estimator for the memory parameter should be used (ELW in [73]; see 5.3.1 for details).

Remark 3.2.1. *For $a, b \in \mathbb{R}^n$, an inherent distance measure $d(a, b)$ is given by*

$$d(a, b) := \|a - b\|,$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$. Frequently used in statistic is

$$\|x\| = x^{\mathrm{T}} W x,$$

where W is a positive definite weighting matrix. A good *choice of W is crucial to get an estimator with desirable properties. For $W = I_n$ the norm reduces to the sum of squares.*

As under H_0 the local parameters are equal to the global parameter, this should also hold for the estimates. Thus, the $b \times 1$ vector

$$\theta_{d,b} = \begin{pmatrix} \widehat{d} - \widehat{d}^{(1)} \\ \widehat{d} - \widehat{d}^{(2)} \\ \vdots \\ \widehat{d} - \widehat{d}^{(b)} \end{pmatrix}$$

in connection with a weighting matrix W would be an appropriate measure. But in this setting the true parameter d_0 is not part of the test statistic, which is not

desirable as we want to derive test properties depending on the *true* memory parameter of the underlying process. It can be easily seen that $\theta_{d,b} = A_{d_0}\widehat{d}_{b,d_0}$ with

$$\widehat{d}_{b,d_0} = \begin{pmatrix} \widehat{d} - d_0 \\ \widehat{d}_1 - d_0^{(1)} \\ \vdots \\ \widehat{d}_b - d_0^{(b)} \end{pmatrix} \overset{\text{under } H_0}{=} \begin{pmatrix} \widehat{d} - d_0 \\ \widehat{d}_1 - d_0 \\ \vdots \\ \widehat{d}_b - d_0 \end{pmatrix} \tag{3.2.5}$$

$$\text{and } A_{d_0} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix} = \begin{pmatrix} \iota_b & I_b \end{pmatrix} \in \mathbb{R}^{b \times b+1}, \tag{3.2.6}$$

where ι_b is a $b \times 1$ vector of ones. [73] shows that under H_0

$$\sqrt{m}\,\widehat{d}_{b,d_0} = Z_n + bias(m), \quad Z_n \xrightarrow{d} N\left(0, \frac{1}{4}\Omega\right), \text{ where } \Omega = \begin{pmatrix} 1 & \iota_b^{\mathrm{T}} \\ \iota_b & bI_b \end{pmatrix}.$$

Recall that the initial goal is to test the distance of $\theta_{d,b} = A_{d_0}\widehat{d}_{b,d_0}$ to zero. As A_{d_0} is constant, it holds

$$\sqrt{m}\,\theta_{d,b} = A_{d_0}Z_n + A_{d_0}bias(m), \quad A_{d_0}Z_n \xrightarrow{d} N\left(0, \frac{1}{4}A_{d_0}\Omega A_{d_0}^{\mathrm{T}}\right).$$

Simple algebra shows that $A_{d_0}\Omega A_{d_0}^{\mathrm{T}} = bI_b - \iota_b\iota_b^{\mathrm{T}}$, which has rank $b-1$. Therefore, computing a generalized inverse $(A_{d_0}\Omega A_{d_0}^{\mathrm{T}})^+$ is necessary and we can consequently define a Wald statistic

$$W = 4m(A_{d_0}\widehat{d}_{b,d_0})(A_{d_0}\Omega A_{d_0}^{\mathrm{T}})^+(A_{d_0}\widehat{d}_{b,d_0})^{\mathrm{T}}, \tag{3.2.7}$$

and W has a chi-squared limiting distribution with $b-1$ degrees of freedom.

However, the variance of the ELW estimator for finite samples tends to be larger than the asymptotic bound $\frac{1}{4m}$, leading to an overrejection of the null by the Wald statistic W. Hurvich and Chen [40] propose to replace m by

$$c_m := \sum_{j=1}^{m} \nu_j^2, \quad \nu_j = \log\lambda_j - \frac{1}{m}\sum_{j=1}^{m}\log\lambda_j = \log j - \frac{1}{m}\sum_{j=1}^{m}\log j.$$

As c_m/m tends to 1 for m to infinity, asymptotic results still hold, but for finite sample it provides better estimates. After all, the corrected Wald statistic for

m	global $\widehat{d}$	average of $\widehat{d}^{(i)}$				p value for χ^2_{b-1}			
$T = 3000$		b=2	b=4	b=8	b=16	b=2	b=4	b=8	b=16
$T^{0.45} \approx 37$	0.51	0.44	0.19	0.13	0.07	0.91	0.00**	0.00**	0.00**
$T^{0.5} \approx 55$	0.51	0.46	0.32	0.20	0.07	0.50	0.02*	0.00**	0.00**
$T^{0.55} \approx 82$	0.41	0.33	0.15	0.06	0.01	0.62	0.16	0.04*	0.00**
$T^{0.6} \approx 122$	0.40	0.33	0.20	0.16	0.14	0.82	0.12	0.19	0.00**
$T^{0.65} \approx 183$	0.38	0.35	0.27	0.26	0.26	0.55	0.36	0.14	0.10

Table 3.2: Summary of the corrected Wald statistic W_c for structural break versus long memory

the null of long memory is defined as

$$W_c = 4m \cdot \frac{c_{m/b}}{m/b} A\widehat{d}_b (A\Omega A')^{+} (A\widehat{d}_b)'. \tag{3.2.8}$$

The computation of W_c is straightforward and the chi square distribution is a standard distribution, so critical values are tabulated. Under the null hypothesis of long memory $\theta_{d,b} = A\widehat{d}_b = 0$ and so W_c equals zero. Thus, small values support long memory in the data, whereas large values of W_c lead to the rejection of long memory in favor of structural breaks.

Example 3.2.2 (Continued)**.** *Recall that the series has breaks at $t = 1200$ and $t = 1900$. So, it is not surprising that the Wald statistic does not posses a lot of power for $b = 2$ (see Table 3.2), as both subsamples have one break in the mean. But by splitting up the data in more subsamples, we can reject long memory in the data. Although in this case the widespread rule of thumb $m = T^{0.5}$ seems to be a good choice for the Wald statistic, the value of m is in general crucial for useful results.*
Additionally, this example shows that jumping to conclusions by just considering the average of the subsample estimates is not always proper ($m = T^{0.6}$ and $b = 16$).

See [73] for detailed assumptions, proofs and a comprehensive Monte Carlo simulation.

3.2.2 Differencing the process

Another simple way to test for fractional integration is based on the d-th differences of a long memory process. If x_t is $I(d)$, then $u_t := \Delta^d(x_t - \mu)$ is $I(0)$ and the cumulative sum $y_t := \sum_{i=1}^{t} u_i$ is $I(1)$. If x_t is of the form $I(0)$ *plus breaks*, then in general neither u_t has short memory, nor y_t is $I(1)$. This can be tested with modified PP and KPSS tests [see 73].

The underlying H_0 model is

$$x_t - \mu_0 = \Delta^{-d_0} u_t I_{t\geq 1},\ u_t \sim I(0),$$

where $\mu_0 = \mathbb{E}\, x_t$. Note that for $d \geq 0.5$ the process is nonstationary. So, instead of seeing μ_0 as a mean value it should be interpreted as an initial state of x_t. As in [73] I use a linear combination of $\overline{x}$ (sample mean) and x_1 (initial state) as an estimator for μ_0,

$$\widehat{\mu}_d = w(d)\overline{x} + (1 - w(d))x_1,$$

where $w(d)$ is a twice differentiable weighting function with $w(d) = 1$ for $d \leq 0.5$, and $w(d) = 0$ for $d \geq 0.75$. In applications I use $w(d) := \frac{1}{2}(1 + \cos 4\pi d)$.

Thus, after estimating a consistent $\widehat{d}$ I calculate $\widehat{\mu}_0 := w(\widehat{d})$ and compute the dth differences of $x_t - \widehat{\mu}_0$

Under H_0, the differenced series u_t is $I(0)$ and the cumulative sum y_t is $I(1)$. Testing y_t for $I(1)$ we can use the test statistic Z_t by Phillips and Perron [63]. For the $I(0)$ test we have to modify the KPSS [45] test slightly, as the variance estimator is not consistent. The modified KPSS statistic is defined as

$$\begin{aligned}
\widehat{\eta}_\mu := \frac{1}{T^2}\sum_{t=1}^{T}\frac{S_t^2}{s^2(q)},\ \text{with } S_t &= \sum_{k=1}^{t} e_k, \\
\text{and } s^2(q) &= \frac{1}{T}\sum_{t=1}^{T} e_t^2 + \frac{2}{T}\sum_{s=1}^{q}\left(1 - \frac{s}{q+1}\right)\sum_{t=s+1}^{T} e_t e_{t-s} \\
&= \widehat{\gamma}(0) + 2\sum_{s=1}^{q}\left(1 - \frac{s}{q+1}\right)\widehat{\gamma}(s),
\end{aligned}$$

where e_t are the residuals from regressing $\widehat{u}_t$ on an intercept. So, e_t is the mean corrected $\widehat{u}_t$ and the variance of e_t equals the variance of the sample mean estimator. For short memory processes we have that $\mathbb{V}\,\overline{x}_T = \mathcal{O}(T)$, but if x_t is long range dependent, $\mathbb{V}\,\overline{x}_T = \mathcal{O}(T^{-\alpha})$ with $\alpha \neq 1$ (see Section 5.1.2 for details). Therefore, the classic variance estimator of e_t has to be corrected to get good asymptotic properties for the test statistic. Here I use a weighted sum (Bartlett window) of the first q autocovariances, where the choice of the truncation lag q is crucial to get an accurate test statistic. Critical values, depending on d, for Z_t and $\widehat{\eta}_t$ are presented in Table 3.3 [obtained from 73].

Suppose x_t is a realization of one of the structural break models presented above. Without going into details, it is plausible that the power of Z_t and η_μ is different for each model, as the possible structural break process x_t is either $I(0)$ (Markov switching) or $I(1)$ (STOPBREAK).[2] Overall combining both statistics,

[2]Of course these models are just two examples of a wide range of possible structural break models. But it meets the requirements to show that the power of Z_t and η_μ is different.

	Z_t			η_μ		
d	10%	5%	1%	10%	5%	1%
0.0	-2.750	-3.025	-3.556	0.347	0.460	0.736
0.1	-2.710	-2.989	-3.532	0.344	0.460	0.737
0.2	-2.678	-2.960	-3.500	0.342	0.453	0.731
0.3	-2.640	-2.932	-3.469	0.337	0.446	0.715
0.4	-2.600	-2.893	-3.432	0.335	0.440	0.702
0.5	-2.558	-2.850	-3.398	0.334	0.435	0.699
0.6	-2.475	-2.767	-3.336	0.321	0.419	0.661
0.7	-2.550	-2.838	-3.430	0.340	0.451	0.721
0.8	-2.568	-2.855	-3.430	0.348	0.463	0.743
0.9	-2.563	-2.849	-3.428	0.347	0.462	0.736
1.0	-2.563	-2.849	-3.424	0.347	0.460	0.737
1.1	-2.564	-2.850	-3.425	0.347	0.460	0.735
1.2	-2.565	-2.851	-3.426	0.347	0.460	0.735
1.3	-2.564	-2.852	-3.427	0.346	0.460	0.736
1.4	-2.564	-2.852	-3.425	0.346	0.460	0.736

Table 3.3: Critical Values for Z_t and η_μ

offers a robust test against two types of spurious long memory (originating from $I(0)$ and $I(1)$ models).

3.2.3 Relating the number of frequencies

Besides a very detailed analysis of mean-shift models and an application to a data set of almost 20,000 daily S&P 500 observations, Perron and Qu [61] also give another simple test procedure for long memory versus mean shifts.

Let $\widehat{d}_\alpha$ be the GPH estimator for $m_\alpha = T^\alpha$. Although a popular rule is $\alpha = 0.5$, the behavior of d as a function of m gives a wide range of distinctive features of $I(0)$ plus breaks and $I(d)$. For an extensive analysis of this special relation see [61], full of reflections on short memory plus mean-shift models, and an enormous amount of practicable test procedures.

Under the null of a fractionally integrated $I(d_0)$ process and some (asymptotic) conditions on d and m

$$\sqrt{m_\alpha}\left(\widehat{d}_\alpha - d_0\right) \xrightarrow{d} N\left(0, \frac{\pi^2}{24}\right).$$

For $0 < a < \frac{4}{5}$ and some $b \in (a, 1)$ consider

$$t_d(a,b) = \sqrt{\frac{T^a 24}{\pi^2}}\left(\widehat{d}_a - \widehat{d}_b\right), \tag{3.2.9}$$

then $t_d(a,b) \xrightarrow{d} N(0,1)$. This can be easily seen as

$$\begin{aligned} t_d(a,b) &= \sqrt{\frac{T^a 24}{\pi^2}}\left(\widehat{d}_a - \widehat{d}_b\right) = \sqrt{\frac{T^a 24}{\pi^2}}\left(\widehat{d}_a - \widehat{d}_0\right) - \sqrt{\frac{T^a 24}{\pi^2}}\left(\widehat{d}_b - \widehat{d}_0\right) \\ &= \sqrt{\frac{T^a 24}{\pi^2}}\left(\widehat{d}_a - \widehat{d}_0\right) - \sqrt{\frac{T^b 24}{\pi^2}}\left(\widehat{d}_b - \widehat{d}_0\right)\sqrt{\frac{T^a}{T^b}} \\ &\xrightarrow{d} N(0,1) - N(0,1)\cdot 0 \sim N(0,1), \end{aligned}$$

where the last line holds as $b > a$. Under the alternative hypothesis of short memory process with shifts in the mean, $t_d(a,b)$ tends to infinity since the limit of $\widehat{d}_a$ is strictly smaller than the limit $\widehat{d}_b$ [for a proof see 61]. Due to certain features of $d(m)$, observed in Monte Carlo studies, [61] suggest $a = \frac{1}{3}$ and $b = \frac{4}{5}$ as *good* parameters for a powerful test.

3.3 Error duration model - revisited

The ED model presented in Section 2.2 not only motivates long memory and persistent ARFIMA processes, but also sheds new light on the issue of structural breaks.

Recall that (under certain conditions on p_k) an ED model and an ARFIMA model exhibit the same features from a second moments point of view. Thus, observing a persistent process one can *choose* between an ARFIMA or an ED model. Of course, in research the ARFIMA model is much easier to estimate and implement for further analysis, such as forecasting. But the ED model is a good way to see the particular problem of structural breaks versus long memory from a different perspective.

3.3.1 Spurious structural breaks

Assume a process z_t is the sum of two independent processes x_t and y_t,

$$z_t = x_t + y_t, \tag{3.3.1}$$

where y_t is a stationary $ARMA(p,q)$ process and x_t has persistent long memory. As the processes are independent, the spectrum $f_z(\lambda)$ decomposes into

$$f_z(\lambda) = f_x(\lambda) + f_y(\lambda). \tag{3.3.2}$$

In practice, we only observe $\{z_t\}_{t=1}^T$ and have to decide whether long memory is real or spurious.

If we represent x_t in the ED framework we get the equivalent representation

$$z_t = \sum_{s=1}^{T} \varepsilon_s d_{s,t} + y_t, \tag{3.3.3}$$

where $d_{s,t}$ equals one for the interval $(s, s+n_s)$ and zero otherwise (ε_s as above). In the error duration sense $d_{s,t}$ is a stochastic indicator function for the *active* time of ε_s. Looking at it the opposite way, ε_s can be thought as a random coefficient on the stochastic mean shifting process $d_{s,t}$. Accordingly, each term $\varepsilon_s\, d_{s,t}$ can be regarded as a structural change to an otherwise stationary *short* memory $ARMA(p,q)$ process y_t.

Thus, large errors ε_s with long durations n_s might be mistaken as significant structural changes in the mean of a process, although the true process features persistent long memory.

In the following we examine a realization of a time series. Thus, let

$$M_K = \big\{ s | n_s(\omega) > K \big\}$$

be the set of all points in time, where the *realized* shock duration n_s of ε_s is greater than K. Of course it holds $M_K \cup M_K^C = \{1, \ldots, T\}$. Now rewrite (3.3.3) to

$$z_t = \sum_{s \in M_K} \varepsilon_s d_{s,t} + \sum_{s \notin M_K} \varepsilon_s d_{s,t} + y_t.$$

Both, the second and the third process are stationary, short memory processes. In the first term – as noted above – $d_{s,t}$ can be seen as an indicator function for active times of a mean process ε_s.

Finally we can write

$$z_t = \upsilon_t + \mu_t \tag{3.3.4}$$

$$\upsilon_t = \sum_{s \notin M_K} \varepsilon_s d_{s,t} + y_t \tag{3.3.5}$$

$$\mu_t = \begin{cases} \mu_1 & \text{for } 0 \le t \le \tau_1, \\ \mu_2 & 0 \le t \le \tau_2, \\ \vdots & \vdots \\ \mu_w & \text{for } 0 \le t \le \tau_w, \end{cases} \tag{3.3.6}$$

where υ_t is a short memory process (the maximum duration length equals $K-1$) and μ_t is i.i.d. with outcomes $(\mu_1, \ldots, \mu_w)$. Example 3.3.1 and Figure 3.3 give an experimental and graphical explanation of this point.

Example 3.3.1 ($ARFIMA(0, 0.4, 0)$ – Continued)**.** *Recall Example 2.3.11 of a simulated $ARFIMA(0, d, 0)$ process z_t within the ED model. As I know the*

realized stochastic durations I can compute the set $M_K = \{s|n_s(\omega) > K\}$. *I set* $K = 31$, *which is the 99.5 % quantile of* n_s. *Figure 3.3 displays the three components of* z_t:

z_t – **original series:** *the top left panel shows the original series (black line),*

μ_t – **mean shifts:** *the red line shows the spurious structural breaks in the series,*

υ_t – **short memory:** *the top right panel shows the mean corrected process* $\upsilon_t := z_t - \mu_t$.

Suppose we observe z_t, *look at the autocorrelation structure and the spectrum (see Figure 2.4), and have to decide whether the observed long range dependence arises from structural breaks, which can also be spotted in the series, or true long memory.*

Both estimators (ELW and GPH) give an estimated value of approximately 0.37, *almost constant for different values of* m. *Although the* Z_t *statistic does not reject the null of long memory in the data, the modified KPSS test rigorously rejects long memory in favor of structural breaks (5% critical value for the modified KPSS for* $d = 0.4$, *is* 0.44; *see Table 3.5).*

Consequently we compute a mean shift corrected version υ_t. *After plotting the autocorrelations, we would presumably try to fit a short memory* $ARMA(p, q)$ *model, and not consider a long memory ARFIMA model. And indeed,* $(\widehat{d}_{ELW}, \widehat{d}_{GPH}) = (0.10, 0.11)$ *(for* $m = T^{0.6} \approx 122$*) does not suggest a long memory approach. Also* η_μ *and* Z_t *do not reject* $I(0)$ *for various frequency cutoffs* m *(see Table 3.6).*

Via a Box Jenkins method we get an $ARMA(3, 1)$ *model for* υ_t

$$\upsilon_t - 0.36(0.02)\upsilon_{t-1} - 0.12(0.02)\upsilon_{t-3} = \epsilon_t + 0.51(0.02)\epsilon_{t-1}, \quad \widehat{\sigma}^2_\epsilon = 1.44. \quad (3.3.7)$$

Even the Ljung Box statistic does not reject white noise for the data. Ergo we would falsely conclude that z_t *is an* $ARMA(3, 1)$ *process with stochastic breaks in the mean, although the generating model is purely long range dependent and stationary.*

3.4 Spurious discussion?

From the results above, both theoretically and empirically, different ways of dealing with long range dependence are possible:

Lags	1	2	5	10	20	50
$\widehat{\varepsilon}$	0.98	0.97	0.98	1.00	0.73	0.68

Table 3.4: p-values of Ljung-Box statistic for residuals of (3.3.7)

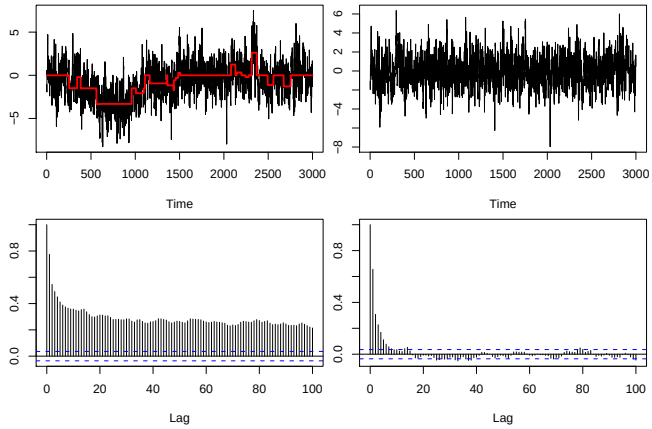


Figure 3.3: Error Duration simulation with $d = 0.4$ and $T = 3000$

m T=3000	$\widehat{d}_{GPH}$	$\widehat{d}_{ELW}$	η_μ q=5	η_μ $q_{opt} = 40$	Z_t
$T^{0.45} \approx 37$	0.45	0.43	0.43	0.60	-1.66
$T^{0.5} \approx 55$	0.34	0.41	0.52	0.68	-1.56
$T^{0.55} \approx 82$	0.38	0.39	0.66	0.79	-1.46
$T^{0.6} \approx 122$	0.37	0.36	0.93	0.98	-1.33
$T^{0.65} \approx 183$	0.36	0.37	0.91	0.96	-1.34

Table 3.5: Structural break tests for original series

m T=3000	$\widehat{d}_{GPH}$	$\widehat{d}_{ELW}$	η_μ q=5	η_μ $q_{opt} = 28$	Z_t
$T^{0.45} \approx 37$	-0.09	-0.16	0.80	0.36	-1.47
$T^{0.5} \approx 55$	-0.13	-0.13	0.60	0.29	-1.59
$T^{0.55} \approx 82$	-0.06	-0.08	0.36	0.19	-1.86
$T^{0.6} \approx 122$	0.00	-0.01	0.19	0.12	-2.28
$T^{0.65} \approx 183$	0.07	0.07	0.09	0.07	-2.94

Table 3.6: Structural break tests for spuriously demeaned series

- Long memory processes can result from a stochastic error duration model where no breaks are present theoretically.
- ARFIMA models are a natural expansion of ARIMA models, and ARFIMA processes are covariance stationary, i.e. there are neither breaks in the mean nor in the variance (given it is finite).
- On the other hand, there is a wide range of literature and Monte Carlo simulations about processes representing spurious long memory solely based on structural breaks, as seen in the three models above (see Section 3.1.1).
- The simple error duration analysis at least invites debate on the question

of spurious structural breaks in long range dependent time series.

Both, long memory and structural breaks, are theoretically appealing and demonstrate that at least two explanations for long range dependence in the data exist. And presumably one model by itself may not capture all of the persistence in a time series, i.e. the residuals under a breakpoint null model may still exhibit persistence and on the other hand, residuals from an $I(d)$ estimation may still display sudden changes in the mean or variance, respectively.

Nevertheless an ARFIMA model for a stochastic process is a parsimonious model that can explain long *and* short range dependences in time series. Even if the long memory is spurious due to actual structural breaks, ARFIMA models might provide better forecasts.[3]

However, after studying various literature about structural breaks versus long memory and describing the error duration point of view of this question, I close this part with two – despite, or actually because of their inconclusiveness – quintessential comments from recent literature:

> *We believe, however, that the temptation to jump to conclusions of "structural change producing spurious inferences of long memory" should be resisted, as such conclusions are potentially naive. Even if the 'truth' is structural change, long memory may be a convenient shorthand description, which may remain very useful for tasks such as prediction.*[4] *Moreover, at least in the sorts of circumstances studied in this paper, "structural change" and "long memory" are effectively different labels for the same phenomenon, in which case attempts to label one as "true" and the other as "spurious" are of dubious value.*
>
> Diebold and Inoue [18]

And almost simultaneous to this thesis, supporting my exemplified analysis of the structural break issue within the ED model:

> *The models [ED model] intrinsically possess both structural change and long memory, in an inextricably intertwined manner, and thus may help practitioners to view these two phenomena as a duality rather than a dichotomy.*
>
> Hsieh et al. [38]

[3]Bhardwaj and Swanson [6] find that ARFIMA models outperform ARIMA, ARMA, ARCH and related models in ex ante forecasting of absolute returns. As expected ARFIMA models perform better especially for greater forecast horizons.

[4]In a development that supports this conjecture, Clements and Krolzig [11] show that fixed-coefficient autoregressions often outperform Markov switching models for forecasting in finite samples, even when the true data-generating process is Markov switching.

Chapter 4

Forecasting

In forecasting we commence from a stationary process $x_t \in \mathbb{L}^2(\Omega, \mathcal{A}, P)$ and want to approximate future values x_{t+h} $(h > 0)$ by a feasible function of past values $x_s, s \leq t$. To get a well posed problem we have to specify the class of feasible functions and the approximation criterion in the Hilbert space $\mathbb{L}^2(\Omega, \mathcal{A}, P)$.

We consider affine functions and our rule for the *best* function is the least squares criterion. Ergo we have to solve the minimization problem

$$\min_{b, a_j \in \mathbb{R}} \mathbb{E} \overline{\left(x_{t+h} - \left(b + \sum_{j \in J} a_j x_{t-j} \right) \right)} \left(x_{t+h} - \left(b + \sum_{j \in J} a_j x_{t-j} \right) \right) \tag{4.0.1}$$

The index set J is either finite or infinite.

If the class of approximation functions is the class of measurable functions, then the general least squares approximation is the conditional expectation. For Gaussian processes the conditional expectation is in fact linear. Ergo the restriction to affine functions is less restrictive the more the distribution of the forecast errors is similar to Gaussian.

The next theorem prepares the ground for calculating the parameters b and a_j, using results from functional analysis of orthogonal subspaces in Hilbert spaces.

Theorem 4.0.1 (Projection Theorem)**.** *Let $\mathbb{H}$ be a Hilbert space and $\boldsymbol{M}$ be a closed subspace. Corresponding to every $x \in \mathbb{H}$, there is a unique decomposition*

$$x = \widehat{x} + u$$

such that $\widehat{x} \in \boldsymbol{M}$ and $u \perp \boldsymbol{M}$. Furthermore $\widehat{x}$ is the unique element of $\boldsymbol{M}$ satisfying

$$||x - \widehat{x}|| = \min_{y \in \boldsymbol{M}} ||x - y|| \tag{4.0.2}$$

Proof. See [81]. □

$\widehat{x} \in \mathbf{M}$ is called the *projection* of x on $\mathbf{M} \subseteq \mathbb{H}$ and is denoted by $\mathbf{P_M} x = \widehat{x}$. Thus, $\mathbf{P_M}$ is a mapping from $\mathbb{H}$ onto $\mathbf{M}$ and it holds $\mathbf{P_M P_M} = \mathbf{P_M}$.

Remark 4.0.2. *Note that the decomposition is unique; therefore, it is not necessary to find the explicit solution for the operator $\boldsymbol{P_M}$, which maps any $x_0 \in \mathbb{H}$ onto its projection $\widehat{x_0} \in \boldsymbol{M}$. If we find a decomposition of x_0 in $y_0 \in \boldsymbol{M}$ and v_0, with $x_0 = y_0 + v_0$* and *$y_0 \perp v_0$, then this is the projection and we know that $y_0 \equiv \widehat{x}$ and $v_0 \equiv u$.*

4.1 Prediction from a finite past

Let $\mathbb{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)$ be the space spanned by the $x_{t-j}, j = 0, \ldots, r$ and by the constant 1. Note that $\mathbb{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)$ is a Hilbert space and a subspace of $\mathbb{L}^2(\Omega, \mathcal{A}, P)$. Since $\{x_t\}_{t=-\infty}^{\infty} \in \mathbb{L}^2(\Omega, \mathcal{A}, P)$ the prediction of x_{t+h} from finite past can be seen as finding the element $\widehat{x}_{t,h} \in \mathbb{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)$ such that the distance between x_{t+h} and $\widehat{x}_{t,h}$ is minimal.

From the projection theorem 4.0.1 this is a projection of $x_{t+h} \in \mathbb{L}^2(\Omega, \mathcal{A}, P)$ on the subspace $\mathbb{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)$, i.e.

$$\widehat{x}_{t,h} = \mathbf{P}_{\mathbb{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)} x_{t+h}. \tag{4.1.1}$$

Without loss of generality we assume that the mean of x_t is equal to zero. Hence, it is clear from remark 4.0.2 that $\widehat{x}_{t,h} = \sum_{j=0}^{r} a_j x_{t-j}$ is the projection $\widehat{x} \in \mathbf{M}$ if and only if the errors $x_{t+h} - \widehat{x}_{t,h}$ are elements of $\mathbf{M}^{\perp}$, i.e. they are uncorrelated with x_{t-s}, $s = 0, \ldots, r$.

Hence, we get $r+1$ equations

$$\begin{aligned}
0 &\overset{!}{=} < x_{t+h} - \widehat{x}_{t,h}, x_{t-s} >= \mathbb{E}\left(x_{t+h} - \widehat{x}_{t,h}\right) x_{t-s}, \quad \forall s = 0, \ldots, r \\
\mathbb{E} x_{t+h} x_{t-j} &\overset{!}{=} \mathbb{E} \widehat{x}_{t,h} x_{t-s} = \mathbb{E}\left(\sum_{j=0}^{r} a_j x_{t-j}\right) x_{t-s}, \quad \forall s = 0, \ldots, r \\
\gamma_x(h+j) &\overset{!}{=} \sum_{j=0}^{r} a_j \gamma_x(s-j), \quad \forall s = 0, \ldots, r
\end{aligned}$$

The last equation can be written in matrix form

$$\begin{pmatrix} \gamma(h) & \cdots & \gamma(h+r) \end{pmatrix} = \begin{pmatrix} a_0 & \cdots & a_r \end{pmatrix} \begin{pmatrix} \gamma(0) & \cdots & \gamma(r) \\ \vdots & \ddots & \vdots \\ \gamma(-r) & \cdots & \gamma(0) \end{pmatrix} =: \begin{pmatrix} a_0 & \cdots & a_r \end{pmatrix} \Gamma_r$$

Thus, we get a solution for the coefficients a_j by inverting Γ_r. The inverse

exists, given Γ_r is not singular.[1]

Solution 4.1.1 (Yule - Walker equations)**.** *The best linear predictor of x_{t+h} is given by $\widehat{x}_{t,h} := \sum_{j=0}^{r} a_j x_{t-j}$ where*

$$\begin{pmatrix} a_0 & \cdots & a_r \end{pmatrix} = \begin{pmatrix} \gamma(h) & \cdots & \gamma(h+r) \end{pmatrix} \Gamma_r^{-1}. \tag{4.1.2}$$

In practice we only have data from the finite past, but since estimation of the autocorrelation function is not accurate for high lags and the inversion of a $T \times T$ matrix is time consuming for large T, I present the prediction from the infinite past below.

4.2 Prediction from the infinite past

Let $\mathbb{H}_x(t)$ be the Hilbert space spanned by all $x_s, s \leq t$, which is the set of all linear combinations of $x_s, s \leq t$ and their limiting elements in $\mathbb{L}^2(\Omega, \mathcal{A}, P)$. Analogously define $\mathbb{H}_\varepsilon(t)$.

Consider the stationary, causal and invertible $ARFIMA(p, d, q)$ process

$$\Phi(z)x_t = \Theta(z)\, n(z)\varepsilon_t, \quad n(z) = (1-z)^{-d} = \textstyle\sum_{j=0}^{\infty} \binom{d}{j} z^j = \sum_{j=0}^{\infty} n_j z^j$$
$$\text{and } \Phi(z)^{-1}\Theta(z) =: \varphi(z) = \textstyle\sum_{i=0}^{\infty} \varphi_i z^i.$$

As $\Phi(z), \Theta(z) \neq 0 \quad \forall |z| \leq 1,$ and $d \in [-\frac{1}{2}, \frac{1}{2})$, we have

$$x_t = \Phi(z)^{-1}\Theta(z)n(z)\varepsilon_t = k(z)\varepsilon_t = \sum_{j=0}^{\infty} k_j \varepsilon_{t-j}, \tag{4.2.1}$$

$$\varepsilon_t = n(z)^{-1}\Theta(z)^{-1}\Phi(z)x_t = r(z)x_t = \sum_{n=0}^{\infty} r_n x_{t-n}. \tag{4.2.2}$$

This implies

$$\mathbb{H}_x(t) \equiv \mathbb{H}_\varepsilon(t).$$

Going back again to the initial problem of forecasting x_{t+h} we can write

$$x_{t+h} = \underbrace{\sum_{j=0}^{\infty} k_j \varepsilon_{t+h-j}}_{x} = \underbrace{\sum_{j=0}^{h-1} k_j \varepsilon_{t+h-j}}_{u} + \underbrace{\sum_{j=h}^{\infty} k_j \varepsilon_{t+h-j}}_{\widehat{x}}. \tag{4.2.3}$$

[1] The covariance matrix of a vector of random variables $\boldsymbol{u}$ is singular iff $\boldsymbol{a}`\boldsymbol{u} = 0$ for some $\boldsymbol{a} \neq \boldsymbol{0}$. In our context, this means that the covariance matrix of a stochastic process is singular iff the random variables $x_t, \ldots, x_{t-r}$ are linearly dependent. As this is a state with Lebesgue measure zero, the computation of the inverse does not lead to problems in practice and we will not consider this special case any further.

Therefore, $\widehat{x}$ is contained in $\mathbb{H}_\varepsilon(t) = \mathbb{H}_x(t) \subseteq \mathbb{L}^2(\Omega, \mathcal{A}, P)$, implying that $\widehat{x}$ is a linear combination of past and present values $x_s, s \leq t$. u is orthogonal to all elements of $\mathbb{H}_\varepsilon(t)$, as ε_t is white noise. Since $\mathbb{H}_\varepsilon(t) = \mathbb{H}_x(t)$, the term u is also orthogonal to all elements of $\mathbb{H}_x(t)$, especially $\mathbb{E}u\widehat{x} = 0$. Therefore, the projection $x = \widehat{x} + u$ is implicitly given by equation (4.2.3).

Thus, by the projection theorem the h-step predictor and the corresponding h-step prediction error satisfy

$$\widehat{x}_{t,h} \equiv \widehat{x} \text{ and } \widehat{\varepsilon}_{t,h} \equiv u. \tag{4.2.4}$$

As $\widehat{x}_{t,h}$ is the best (in the $\mathbb{L}^2$ sense) linear approximation of x_{t+h} by past and present values of x_t, it is still necessary to express $\widehat{x}_{t,h} = \sum_{j=h}^{\infty} k_j \varepsilon_{t+h-j}$ in terms of $x_{t-j}, j \geq 0$. Since $\varepsilon_t = n(L)^{-1}\Theta(L)^{-1}\Phi(L)\, x_t = \sum_{n=0}^{\infty} r_n L^n x_t$

$$\begin{aligned}
\widehat{x}_{t,h} &= \sum_{j=h}^{\infty} k_j L^j \varepsilon_{t+h} = \sum_{j=h}^{\infty} k_j L^j \sum_{n=0}^{\infty} r_n L^n x_{t+h} = \sum_{j=0}^{\infty} k_{j+h} L^{j+h} \sum_{n=0}^{\infty} r_n L^n x_{t+h} \\
&= \sum_{j=0}^{\infty} k_{j+h} L^j \sum_{n=0}^{\infty} r_n L^n L^h x_{t+h} = \sum_{j=0}^{\infty} k_{j+h} L^j \sum_{n=0}^{\infty} r_n L^n x_t \qquad (4.2.5) \\
&=: \sum_{i=0}^{\infty} a_i(h) L^i x_t = A_h(z) x_t. \qquad (4.2.6)
\end{aligned}$$

The weighting sequence $\{a_i(h)\}$ is given by matching terms. In general the coefficients of $A_h(z)$ are different for every h.

Consequently the projection operator is explicitly given by

$$\widehat{x}_{t,h} = \mathbf{P}_{\mathbb{H}_x(t)}(x_{t+h}) = \sum_{n=0}^{\infty} a_n(h) L^n x_t \in \mathbb{H}_x(t). \tag{4.2.7}$$

The essence of any thought dealing with forecasts is *causality of events*. The Wold decomposition provides the basis for forecasting a covariance stationary process.

Theorem 4.2.1 (Wold's decomposition)**.** *Any zero-mean covariance stationary process x_t can be represented in the form*

$$x_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} + \kappa_t, \tag{4.2.8}$$

where $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. ϵ_t is white noise and represents the error in forecasting x_t with a linear function of lagged x_t:

$$\epsilon_t := x_t - \mathbb{E}(x_t \mid x_{t-1}, x_{t-2}, \ldots). \tag{4.2.9}$$

κ_t is uncorrelated with ϵ_{t-j}, for any j, but κ_t can be predicted arbitrarily well from a linear function of past values of x_t:

$$\kappa_t = \mathbb{E}(\kappa_t \mid x_{t-1}, x_{t-2}, \ldots).$$

Proof. See Brockwell and Davis [8]. □

In practice the Wold representation requires fitting an infinite number of parameters, which is impossible as $T < \infty$. A typical assumption is that $\Psi(L)$ can be written as the ratio of two finite-order polynomials (this leads to an $ARMA(p, q)$ model):

$$\sum_{j=0}^{\infty} \psi_j L^j = \frac{\Theta_q(L)}{\Phi_p(L)} =: \varphi(L).$$

With this simplification only $p + q + 1$ parameters have to be estimated (AR and MA coefficients and the variance of ϵ_t).

If we write an ARFIMA process x_t in the $MA(\infty)$ representation

$$x_t = \sum_{j=0}^{\infty} \frac{d(d+1)\ldots(d+j-1)}{j!} \epsilon_{t-j},$$

we see that this already has the form (4.2.8) with $\kappa_t = 0$ and $\psi_j = \binom{d}{j}$.[2]

Now it becomes clear why ARFIMA models are parsimonious models for long range dependent processes, as the Wold decomposition for these models depends on the memory parameter d exclusively. Thus, after estimating d, the computation of $\{\psi_j\}$ and consequently prediction of x_{t+h} is straightforward.

4.2.1 Prediction error

Since we are not only interested in the forecast itself, but also in the confidence region of the forecast we have to analyze the prediction error (4.2.4):

$$\begin{aligned}
\mathbb{E}\,\widehat{\varepsilon}_{t,h} &= \mathbb{E}\,x_{t+h} - \mathbb{E}(x_t \mid x_{t-1}, x_{t-2}, \ldots) = \mathbb{E}\sum_{j=0}^{h-1} k_j \varepsilon_{t+h-j} = 0 \\
\mathbb{V}\,\widehat{\varepsilon}_{t,h} &= \mathbb{E}\,\widehat{\varepsilon}_{t,h}^2 = \mathbb{E}\sum_{j=0}^{h-1} k_j \varepsilon_{t+h-j} \sum_{i=0}^{h-1} k_i \varepsilon_{t+h-i} \\
&= \sum_{j=0}^{h-1} k_j^2 \mathbb{E}\,\varepsilon_{t+h-j}^2 = \sigma_\varepsilon^2 \sum_{j=0}^{h-1} k_j^2.
\end{aligned}$$

Therefore, prediction is *unbiased* and the forecast variance is non-decreasing in h. Assuming a Gaussian error term and known mean, the symmetric $(1 - \alpha)$

[2]Processes with $\kappa_t = 0$ are called *purely linearly indeterministic.*

confidence interval for the forecast $\widehat{x}_{t,h}$ is given by

$$\widehat{x}_{t,h} \mp u_{1-\frac{\alpha}{2}} \sigma_\varepsilon \sqrt{\sum_{j=0}^{h-1} k_j^2}, \tag{4.2.10}$$

where $u_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ percentile of the standard Gaussian.

Remark 4.2.2 (Asymptotics)**.** *For a zero-mean covariance stationary process x_t, the prediction for the infinite future tends to $\mathbb{E}\, x_t = \mu = 0$:*

$$\lim_{h\to\infty} \widehat{x}_{t,h} = \lim_{h\to\infty} \sum_{j=h}^{\infty} k_j \varepsilon_{t+h-j} = 0,$$

The prediction error variance tends to $\mathbb{V} x_t = \sigma_x^2 < \infty$:

$$\lim_{h\to\infty} \mathbb{V}\widehat{\varepsilon}_{t,h} = \sigma_\varepsilon^2 \lim_{h\to\infty} \sum_{j=0}^{h-1} k_j^2 = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} k_j^2 = \sigma_x^2.$$

ARFIMA forecasts

For better understanding consider the $ARFIMA(0,d,0)$ case with $d \in (-1,1]$. If $d \in \left(-1, \frac{1}{2}\right)$, then x_t has an $MA(\infty)$ representation with coefficients

$$\pi_j = \binom{d}{j} = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} = \begin{cases} 0 & \text{if } j < 0, \\ 1 & \text{if } j = 0, \\ \frac{d(d+1)\cdots(d+j-1)}{j!} & \text{if } j > 0. \end{cases} \tag{4.2.11}$$

The coefficients satisfy the recursion $\pi_{j+1} = \pi_j \frac{d+j}{j+1}$.

We can differ between three cases

a) for $-1 < d < \frac{1}{2}$ the process is covariance stationary; thus, the forecast for $h \to \infty$ tends to the mean of the process, and the prediction error variance $\mathbb{V}\widehat{\varepsilon}_{t,h} \overset{h\to\infty}{\longrightarrow} \sigma_x^2 < \infty$.

b) for $\frac{1}{2} \le d < 1$ the process is mean-reverting *but* has infinite variance, so the forecast still tends to the mean of the process, but as the forecast horizon increases the forecast error variance $\sum_{j=0}^{h-1} \pi_j^2 \cdot \sigma_\varepsilon^2$ goes to infinity.

c) for $d = 1$ we get the simple random walk model, and the best forecast for x_{t+h} given the infinite past is

$$\mathbb{E}(x_{t+h} | x_t, x_{t-1}, \ldots) = x_t \text{ with } \mathbb{V}\widehat{\varepsilon}_{t,h} = \sigma_\varepsilon^2 \cdot h.$$

The optimal forecast for a random walk equals the last observations x_t for all h, and the forecast error variance grows linearly in h.

For the general ARFIMA(p,d,q) case

$$x_t = \Phi(z)^{-1}\Theta(z)\, n(z)\varepsilon_t = \sum_{j=0}^{\infty} r_j z^j \varepsilon_t. \tag{4.2.12}$$

By matching terms we get the weighting sequence r_j and proceed as above.

Chapter 5

Estimation of the Long Memory Parameter

With theoretical results in hand one is interested in estimating the long memory parameter d. From first – heuristic – detection tools, like an inspection of the ACF, to the Full Information Maximum Likelihood (FIML) estimator, one can choose between various estimation techniques. They mainly differ in their accuracy and computational burden.

5.1 Heuristic methods

As long memory implies hyperbolically decaying autocorrelations and a pole at 0 for the spectral density, taking a closer look at the sample autocorrelations and the periodogram is a first step in a broad analysis for possible long memory in the data.

5.1.1 Autocorrelation inspection

The autocorrelations of a long memory process decay hyperbolically with rate $\alpha = 2d - 1$; hence, they satisfy $\rho(k) = \mathcal{O}(k^\alpha), \alpha \in \mathbb{R}$. Thus, for high lags it should approximately hold

$$\log \widehat{\rho}(k) \approx \log c_\rho + \alpha \log k + \varepsilon, \quad k = k_{min}, \ldots, k_{max}. \tag{5.1.1}$$

The examination of the estimated autocorrelation function $\widehat{\rho}(k)$ is a first heuristic, simple and fast tool to detect long memory but has various drawbacks:

- The characteristic of a long memory process are hyperbolically decaying autocorrelations, but not the actual size of $\rho(k)$. Therefore, large lags might seem insignificant considering the usual $\pm\frac{2}{\sqrt{T}}$ band and one is tempted to reject the hypothesis of long memory in the data, although the autocorrelations decay hyperbolically (see remark 7.1.4 for a theoretical discussion and Figure 7.1 for a graphical illustration of this point).

- A hyperbolical decay is a necessary condition for long memory, but not sufficient as also e.g. realizations of structural break models exhibit such a decay.

- Beran [4] pointed out that it is very hard to distinguish between short memory and long memory for α close to -1 (or d close to 0, respectively)

Nonetheless estimating the slope α (with OLS) is a first indicator for the – possibly long memory – nature of a process.

To get a log-linear model, the lag range (k_{min}, k_{max}) has to be specified. To get satisfying and precise estimates the lag region has to be quite large. But setting k_{min} too low, might lead to perturbations from short time dependence in the data, and setting k_{max} too high, will lead to biased estimates because the autocorrelation function itself can not be estimated precisely for large k.

5.1.2 Variance method

Consider the sample mean estimator $\overline{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_t$. It holds

$$\mathbb{V}(\overline{x}_T) = \frac{\sigma_x^2}{T}\left(1 + \delta_T(\rho)\right), \text{ with } \delta_T(\rho) = \frac{1}{T}\sum_{i \neq j} corr(x_i, x_j).$$

If x_t is i.i.d., then $\delta_T(\rho) = 0$ and we have the classic result of i.i.d. statistics that the variance of the sample mean decays to zero with a rate of T^{-1}. Of course, we do not expect that observations in time are independent, so in general $\delta_T(\rho) \neq 0$.

In practice one often finds a slower convergence rate, i.e. $\mathbb{V}(\overline{x}_T) = o(T^{-\alpha})$ with $-1 < \alpha < 0$, which is in fact one definition for long range dependence with long memory parameter $d = \frac{1+\alpha}{2}$. So, an estimate for α is given by the slope estimate of

$$\log S_t^2 = \log c + \alpha \log t + \varepsilon, \quad 1 < t = t_{min}, \ldots, t_{max} < T. \tag{5.1.2}$$

where S_t^2 is the sample variance of all calculated rolling sample means using t observations. Typically, the points $\left(\log t, \log S_t^2\right)$ are scattered around a line with slope α. For short memory processes $\widehat{\alpha} \approx -1$, whereas for long memory in the data $\widehat{\alpha} \neq -1$. The relationship between the convergence rate α, the Hurst parameter H, and the memory parameter d is given in Table 5.1.

The variance method can be easily implemented, but it has drawbacks if the variance changes over time, since the estimator depends explicitly on a constant variance. Thus, a heteroskedastic process might falsely indicate long memory, because time variation is wrongly assigned to $\delta_T(\rho)$ and not to the variance of the process.

type of memory	α	$H = 1 + \frac{\alpha}{2}$	$d = H - \frac{1}{2}$
short memory	$= -1$	$= \frac{1}{2}$	$= 0$
anti-persistent	< -1	$< \frac{1}{2}$	< 0
persistent	> -1	$> \frac{1}{2}$	> 0

Table 5.1: Time series memory: Relation between α, H, and d

See Beran [4] for a detailed description of the variance method, and Görg and Draghicescu [24] for applications to exchange rates and riverflow data.

5.2 Time domain

5.2.1 R/S statistic

The rescaled range statistic was first introduced by Hurst [39] and extended by Mandelbrot [48]. The R/S statistic is the range of partial sums of deviations of a time series from its mean, normalized by its standard deviation.

Given a stochastic process x_t the classic rescaled range statistic

$$Q_T := \frac{1}{\widehat{\sigma}_x} \left[\max_{l=1,\dots,T} \left\{ \sum_{j=1}^{l} (x_j - \overline{x}_T) \right\} - \min_{j=1,\dots,T} \left\{ \sum_{l=1}^{l} (x_j - \overline{x}_T) \right\} \right], \tag{5.2.1}$$

where $\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$ is the sample mean, and $\widehat{\sigma}_x = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_t - \overline{x}_T)^2}$ is the (maximum likelihood) standard deviation estimator. A table with critical values is given in [48].

Furthermore it can be shown that asymptotically

$$\mathbb{E}\, R/S \sim c\, T^H.$$

The Hurst parameter H is another measure for long range dependence, and it holds $H = d + \frac{1}{2}$. Thus, a natural estimate of H (and d, respectively) can be obtained by calculating the R/S statistic for different values of T and then estimate a log-linear model with intercept $\log c$ and slope coefficient $H = d + \frac{1}{2}$.

I refer the interested reader to Hurst [39] for an intuitive motivation for the rescaled range statistic in a hydrological context.

5.2.2 Modified R/S statistic

If x_t is i.i.d., then $\frac{1}{\sqrt{T}}Q_T$ is weakly convergent to ν, where ν is the range of a Brownian bridge[1] on the unit interval [47]. Lo [47] stresses that normalizing the range with the sample standard deviation $\widehat{\sigma}_x$ can be misleading if there is short term dependence in x_t. For an $AR(1)$ process x_t with parameter $|\phi_1| < 1$ the normalized R/S statistic converges to $\xi \cdot \nu$, where $\xi = \sqrt{\frac{1+\phi_1}{1-\phi_1}}$.

This bias of ignoring the short term effects, has remarkable consequences on finite sample properties of the R/S statistics. In particular, Davies and Harte [13] analyze the behavior of the standard rescaled range statistics for $AR(1)$ processes with $\phi_1 = 0.3$. Although the memory parameter d equals 0 by definition the rejection rate for the 5% significance level of the Mandelbrot regression test equals 47%.

Consequently the R/S statistic should be normalized by the value of ξ. As in practice the specific short term dependence is unknown ξ can not be plugged in the formula as a given value. Desirably an invariant calculation method for a normalizing factor should be obtained that captures a broad band of short term dependence, but still is sensitive to long memory effects.

Thus, [47] introduces the *modified* R/S statistic, which accounts for possible short memory structure in the data:

$$Q_T^m := \frac{1}{\widehat{\sigma}_T(q)} \left[\max_{l=1,\ldots,T} \left\{ \sum_{j=1}^{l} (x_j - \overline{x}_T) \right\} - \min_{j=1,\ldots,T} \left\{ \sum_{l=1}^{l} (x_j - \overline{x}_T) \right\} \right], \quad (5.2.2)$$

where

$$\begin{aligned} \widehat{\sigma}_T^2(q) &= \frac{1}{T} \sum_{j=1}^{T} (x_j - \overline{x}_T)^2 + \frac{2}{T} \sum_{j=1}^{T} w_j(q) \left[\sum_{i=j+1}^{T} (x_i - \overline{x}_T)(x_{i-j} - \overline{x}_T) \right] & (5.2.3) \\ &= \widehat{\sigma}_x^2 + 2 \sum_{j=1}^{q} w_j(q) \widehat{\gamma}(j), \text{ with } w_j(q) = 1 - \frac{j}{q+1}, q < T. & (5.2.4) \end{aligned}$$

Q_T^m differs from the classic rescaled range Q_T only by the denominator. The normalizing factor relies on the subtle fact that the variance of a sum is, in general, *not* equal to the sum of the variances, but also autocovariances have to be taken into account. In principle any feasible weight function $w_j(q)$ can be used.

Under the null of any short range dependence $\frac{1}{\sqrt{T}}Q_T^m$ converges to ν, where

[1] See definition A.1.4 in the appendix.

ν is the Brownian bridge. As expected, the modified R/S statistic has the same asymptotic distribution independent of the specific short term dependence in the data.

Under long memory alternatives it can be shown that in presence of persistent ($d > 0$) long memory the R/S statistic converges in probability to infinity. For anti-persistent behavior ($d < 0$) it converges to zero in probability. In both cases, the probability of rejecting the null hypothesis of short memory approaches unity.

Although Teverovsky, Taqqu, and Willinger [79] conclude that the modified R/S statistic is a major improvement to the classic one, they also report drawbacks of the test statistic, like the tendency to reject true long memory. Their findings concentrate on the truncation lag q and its asymptotic behavior. In [47] only asymptotic properties of q (e.g. $q = \mathcal{O}(T^{0.25})$) are stated to guarantee optimality for the derived test statistic. In practice, these asymptotic conditions are no assistance for choosing an optimal value.

As q should correct the classic R/S statistic for short term effects in the data several problems arise:

- If the true process is short memory, but the truncation lag q is quite small, then the modification is too conservative and long memory would be indicated.

- Choosing q too large, *corrects* the statistic for effects that might already be contributable to the long memory structure of the process. Specifically, [79] show that asymptotically $Q_T^m \sim q^{-d}$. Therefore, the R/S statistic decreases, as q increases (for $d > 0$) and sooner or later it will be within the 95% confidence region $[0.809, 1.862]$ for short memory, reported in [47].

Consequently [79] present a data driven optimal value for q given by

$$q_{opt} = \left[\left(\frac{3T}{2} \right)^{\frac{1}{3}} \left(\frac{2\widehat{\rho}_1}{1 - \widehat{\rho}_1^2} \right)^{\frac{2}{3}} \right], \tag{5.2.5}$$

where $[x]$ is the largest integer less than or equal to x. As this optimal value is obtained under the null hypothesis of an $AR(1)$ process, it can only be an indicator for the right value of q.

However the (modified) R/S statistic is a quick way to check for long range dependence in an observed time series, but considering the difficulties with the *optimal* choice of q it should not be the exclusive criterion to test for long memory in the data.

5.2.3 Full Information Maximum Likelihood

Let θ be the vector of parameters in a parametric model for data $\mathbf{x} = (x_1, \ldots, x_T)^{\mathrm{T}}$ from a stationary time series x_t. Assume x_t has spectral density $f(\mathbf{x}|\theta)$, sample autocovariances $\gamma_{\mathbf{x}}(k|\theta)$, and a $T \times T$ covariance matrix $\Sigma_{T,\theta}$. For a Gaussian process the likelihood function is

$$L(\theta|\mathbf{x}) = (2\pi)^{-\frac{T}{2}} \frac{1}{\sqrt{|\Sigma_{T,\theta}|}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma_{T,\theta}^{-1}\mathbf{x}}. \tag{5.2.6}$$

The maximum likelihood estimator $\widehat{\theta}$ is the value θ that maximizes the likelihood function $L(\theta|\mathbf{x})$ *given* $\mathbf{x}$. This is equivalent to minimizing

$$-2\log L(\theta) = T\log 2\pi + \log|\Sigma_{T,\theta}| + \mathbf{x}^{\mathrm{T}}\Sigma_{T,\theta}^{-1}\mathbf{x}. \tag{5.2.7}$$

As for a stationary process the covariance matrix is a Toeplitz matrix it simplifies to

$$\Sigma_{i,j} = [\gamma(i-j)] \text{ for } i, j = 1, \ldots, T.$$

In order to maximize (minimize) the (log) likelihood function we have to rewrite the covariance matrix in terms of the parameters of the model. Ergo we have to derive $\gamma(s)$ explicitly for an ARFIMA process x_t. This will be done by the inverse Fourier transform

$$\gamma(s) = \frac{1}{2\pi}\int_0^{2\pi} f_x(\lambda)e^{i\lambda s}\,\mathrm{d}\lambda. \tag{5.2.8}$$

As a fractionally integrated process x_t can be seen as a linear transformation of $x_t = (1-L)^d u_t$, where u_t is an $ARMA(p,q)$ process the spectral density $f_x(\lambda) = |1 - e^{-i\lambda}|^{-2d} f_u(\lambda)$. Accordingly the calculation of $f_x(\lambda)$ (and $\gamma(s)$ consequently) is divided in two steps. First we calculate the spectral density $f_u(\lambda)$ of $u_t = (1-L)^d x_t$ and then we use the relation given above. See Sowell [76] for complete derivations and proofs.

ARMA spectral density

We assume that u_t is a covariance stationary $ARMA(p,q)$ process, thus u_t is the unique solution to

$$\Phi(L)u_t = \Theta(L)\varepsilon_t \text{ with equivalent MA}(\infty) \text{ representation } u_t = \frac{\Theta(L)}{\Phi(L)}\varepsilon_t.$$

Since all roots of $\Phi(L)$ lie outside the closed unit disk we can write $\Phi(z) = \prod_{j=1}^{p}(1-\rho_j z)$, where $|\rho_j| < 1$ for $j = 1, \ldots, p$. Thus ($z = e^{-i\lambda}$),

$$f_u(\lambda) = \frac{|\Theta(z)|^2}{|\Phi(z)|^2}\sigma_\varepsilon^2 = \sigma_\varepsilon^2|\Theta(z)|^2\prod_{j=1}^{p}(1-\rho_j z)^{-1}(1-\rho_j z^{-1})^{-1}. \tag{5.2.9}$$

After all, the spectral density of a stationary $ARMA(p,q)$ process can be written as

$$\begin{aligned} f_u(\lambda) &= \sigma_\varepsilon^2 \sum_{l=-q}^{q} \vartheta(l) z^l \sum_{j=1}^{p} \zeta_j z^p \left[\frac{\rho_j^{2p}}{1-\rho_j e^{i\lambda}} - \frac{1}{1-\rho_j^{-1} e^{i\lambda}} \right] \qquad (5.2.10) \\ \text{where } \vartheta(l) &= \sum_{s=max(0,l)}^{min(q,q-l)} \theta_s \theta_{s-l} \text{ and } \zeta_j = \left[\rho_j \prod_{i=1}^{l} (1-\rho_j \rho_i) \prod_{m \neq j} (\rho_j - \rho_m) \right]^{-1}. \end{aligned}$$

Transfer function for $(1-L)^d$

The spectral density of x_t can be written as

$$f_x(\lambda) = \underbrace{\sigma_\varepsilon^2 \sum_{l=-q}^{q} \vartheta(l) z^l \sum_{j=1}^{p} \zeta_j z^p \left[\frac{\rho_j^{2p}}{1-\rho_j e^{i\lambda}} - \frac{1}{1-\rho_j^{-1} e^{i\lambda}} \right]}_{f_u(\lambda)} (1-z)^{-d} (1-z^{-1})^{-d}. \qquad (5.2.11)$$

This is one representation of the spectral density of an ARFIMA process where the roots of the AR polynomial are simple.[2]

Analytic expression for $\gamma(s)$

Define

$$C(d,h,\rho) = \frac{1}{2\pi} \int_0^{2\pi} \left[\frac{\rho_j^{2\rho}}{1-\rho_j e^{-i\lambda}} - \frac{1}{1-\rho_j^{-1} e^{-i\lambda}} \right] (1-e^{-i\lambda})^{-d} (1-e^{i\lambda})^{-d} e^{-i\lambda h} \, \mathrm{d}\lambda$$

and embed (5.2.11) in (5.2.8), then the autocovariances satisfy

$$\gamma(s) = \sigma_\varepsilon^2 \sum_{l=-q}^{q} \sum_{j=1}^{p} \vartheta(l) \zeta_j C(d, p+l-s, \rho_j).$$

Evaluating the integral remains the biggest difficulty in the optimization. But $C(d,h,p)$ simplifies to

$$\begin{aligned} C(d,h,\rho) &= \frac{\Gamma(1-2d)\Gamma(d+h)}{\Gamma(1-d+h)\Gamma(1-d)\Gamma(d)} \\ &\quad \cdot \left[\rho^{2\rho} F(d+h,1;1-d+h;\rho) + F(d-h,1;1-d-h;\rho) - 1 \right] \text{ for } d < \frac{1}{2}. \end{aligned}$$

The advantage of this representation is that several software packages[3] possess

[2] In practice this is not a restriction as multiple roots in the lag polyonomials are a state with Lebesgue measure zero.

[3] I use the `fracdiff` method in the R package `fracdiff`. See appendix B for acknowledgments and credits for routines and packages used in the computations within this thesis.

fast algorithms to compute the *hypergeometric function*

$$F(a,b;c;z) := \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{i=0}^{\infty} \frac{\Gamma(a+i)\Gamma(b+i)}{\Gamma(c+i)\Gamma(i+1)} z^i. \tag{5.2.12}$$

Starting values

Good starting values are important such that the optimization de facto gives the global maximum for the parameter vector $\theta = (\sigma_\varepsilon^2, d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$. The following procedure should give good starting values:

1. estimate d using another *reasonable* estimator (e.g. ELW, GPH);
2. compute $u_t = (1-L)^{-\widehat{d}} x_t$ and apply Box-Jenkins methods to obtain optimal ARMA parameters;
3. compute the residuals $\varepsilon_t = \frac{\Phi(L)}{\Theta(L)} u_t$ and estimate the variance σ_ε^2.

Simulation studies show optimal properties of the maximum likelihood estimator but it is computationally extensive, since it requires inverting a $T \times T$ matrix and evaluating an integral, depending on three parameters. Furthermore, the exact structure of the underlying $ARFIMA(p,d,q)$ model must be specified. If correct, then the $FIML$ is the best we can get; if the model is misspecified, then the estimators are in general useless. Hence, approximations and robustifications for the FIML have been proposed.

5.3 Frequency domain

So far we have considered estimation techniques in the time domain. Consider the fractionally integrated process x_t satisfying $(1-L)^d x_t = u_t$, where u_t is a short memory process with spectrum $f_u(\lambda)$.[4]

It holds (see equation (2.1.1))

$$f_x(\lambda) = \left(2\sin\frac{\lambda}{2}\right)^{-2d} f_u(\lambda) = \left(4\sin^2\frac{\lambda}{2}\right)^{-d} f_u(\lambda).$$

Almost every estimator in the frequency domain uses this relation in one or the other way.

5.3.1 Whittle approximation to the MLE

Consider any stationary process x_t (not necessarily ARFIMA) with an existing spectral density $f_x(\lambda)$ and a model with parameter $\boldsymbol{\beta}$ that tries to capture the properties of x_t. A good estimator should minimize the distance between

[4] As above, assume that the spectrum of u_t is a well behaved function, especially it is bounded for frequencies away from the origin, and $\lim_{\lambda\to 0} f_u(\lambda) = f_u(0) = G \notin \{0,\infty\}$.

the *true* spectral density $f_x(\lambda)$ and a parametric approximation $f_x^{\boldsymbol{\beta}}(\lambda)$, as a function of $\boldsymbol{\beta}$. Using results from the Kullback-Leibler information divergence, which measures the difference between two *probability* distributions, we can develop an estimator for $\boldsymbol{\beta}$ by minimizing the difference between two *spectral* distributions.

It can be shown [see 59] that the asymptotic information divergence for a Gaussian process x_t is

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log f_x^{\boldsymbol{\beta}}(\lambda) + \frac{f_x(\lambda)}{f_x^{\boldsymbol{\beta}}(\lambda)} \, \mathrm{d}\lambda + const. \tag{5.3.1}$$

Since $f(\lambda)$ is unknown it is replaced with the periodogram $I_x^T(\lambda)$, obtained from a sample of T observations, and we get the Whittle function

$$\mathcal{L}_T(\boldsymbol{\beta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log f_x^{\beta}(\lambda) + \frac{I_x^T(\lambda)}{f_x^{\boldsymbol{\beta}}(\lambda)} \, \mathrm{d}\lambda + const. \tag{5.3.2}$$

The Whittle estimator is defined as (dropping *const* and $\frac{1}{4\pi}$)

$$\widehat{\boldsymbol{\beta}}_W := \arg\min_{\boldsymbol{\beta} \in B} \mathcal{L}_T(\beta) = \arg\min_{\boldsymbol{\beta} \in B} \int_{-\pi}^{\pi} \log f_x^{\boldsymbol{\beta}}(\lambda) + \frac{I_x^T(\lambda)}{f_x^{\boldsymbol{\beta}}(\lambda)} \, \mathrm{d}\lambda \tag{5.3.3}$$

The only assumption of the process x_t is Gaussianity; therefore, the optimization gives useful results for any Gaussian process with a well defined spectral density. For an underlying $AR(p)$ model it can be shown that $\widehat{\boldsymbol{\beta}}$ is equal to the solution of the Yule-Walker equations (see solution 4.1.1).

Local Whittle Estimator

For a special underlying $ARFIMA(p_0, d, q_0)$ model, we can compute the exact spectrum and perform the optimization. This might circumvent the high cost of matrix inversions of the FIML, but still it is not robust to misspecified models.

The *local* Whittle estimator uses the approximation

$$f_x^{\boldsymbol{\beta}}(\lambda) = f_u^{\boldsymbol{\beta}}(\lambda) \left(2 \sin \frac{\lambda}{2}\right)^{-2d} \approx G(d)\lambda^{-2d} \text{ for } \lambda \ll 1, \tag{5.3.4}$$

where $G(d) := f_u^{\boldsymbol{\beta}}(0) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\Theta(1)|^2}{|\Phi(1)|^2} \neq 0$ is a finite constant. The spectral density of u_t does not depend directly on d as it does not appear on the right hand side. But, in general, the AR and MA lag polynomials are different for any *given* d; therefore, $f_u^{\boldsymbol{\beta}}(0)$ changes for different values of d.

In discrete time the Whittle function reduces to

$$\mathcal{L}_T(\boldsymbol{\beta}) \longrightarrow \frac{1}{m}\sum_{j=1}^{m} \log f_x^{\boldsymbol{\beta}}(\omega_j) + \frac{1}{m}\sum_{j=1}^{m} \frac{I_x^T(\omega_j)}{f_x^{\boldsymbol{\beta}}(\omega_j)}. \tag{5.3.5}$$

Substituting (5.3.4) in (5.3.5) we get

$$\frac{1}{m}\sum_{j=1}^{m}\left[\log f_x^{\boldsymbol{\beta}}(\omega_j) + \frac{I_x^T(\omega_j)}{f_x^{\boldsymbol{\beta}}(\omega_j)}\right] \approx \frac{1}{m}\sum_{j=1}^{m}\left[\log\left(G(d)\omega_j^{-2d}\right) + \frac{\omega_j^{2d}}{G(d)} I_x^T(\omega_j)\right], \tag{5.3.6}$$

where m is some integer less than T. The local Whittle estimator is given by

$$(\widehat{G}, \widehat{d}) = \arg \min_{G\in(0,\infty),\, d\in[d_1,d_2]} Q_m(G,d) \tag{5.3.7}$$

$$\text{with } Q_m(G,d) = \frac{1}{m}\sum_{j=1}^{m}\left[\log\left(G(d)\omega_j^{-2d}\right) + \frac{\omega_j^{2d}}{G(d)} I_x^T(\omega_j)\right]$$

where $-\frac{1}{2} < d_1 < d_2 < \infty$ [see 44].

Necessary optimality conditions are

$$\frac{\partial}{\partial G} Q_m(G,d) \stackrel{!}{=} 0 \text{ and } \frac{\partial}{\partial d} Q_m(G,d) \stackrel{!}{=} 0.$$

Substituting

$$G_{opt}(d) \stackrel{!}{=} \frac{1}{m}\sum_{j=1}^{m} \omega_j^{2d} I_x(\omega_j)$$

in the objective function gives a one-dimensional optimization problem in d. The solution is the local Whittle estimator

$$\widehat{d}_{LW} = \arg \min_{d\in[d_1,d_2]} R(d), \tag{5.3.8}$$

$$\text{where } R(d) = \log G_{opt}(d) - 2d\frac{1}{m}\sum_{j=1}^{m} \log \omega_j. \tag{5.3.9}$$

This optimization must be performed numerically. For a detailed discussion see Künsch [44], Robinson [69].

Exact local Whittle estimator

The local Whittle estimator minimizes the Whittle function of the long memory process x_t, to get an optimal value for d. Phillips and Shimotsu [64] start the other way around, and transform the Whittle function for the stationary

process $u_t := (1-L)^d x_t$

$$\begin{aligned}\mathcal{L}_T^{ELW}(d) &= \frac{1}{m}\sum_{j=1}^{m}\log f_u(\omega_j) + \frac{1}{m}\sum_{j=1}^{m}\frac{I_u^T(\omega_j)}{f_u(\omega_j)} \\ &\approx \frac{1}{m}\sum_{j=1}^{m}\left(\log G(d) + \frac{I_u^T(\omega_j)}{G(d)}\right), \end{aligned} \tag{5.3.10}$$

to be dependent on x_t.

Assume for a moment that also $I_u^T(\omega_j) \approx \omega_j^{2d} I_x^T(\omega_j)$ holds. Making a change of variables and adding the Jacobian in (5.3.10) gives,

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}\left(\log G(d) + \frac{I_u^T(\omega_j)}{G(d)}\right) &= \frac{1}{m}\sum_{j=1}^{m}\left(\log G(d) + \frac{\omega_j^{2d} I_x^T(\omega_j)}{G(d)} + \log \omega_j^{-2d}\right) \\ &= \frac{1}{m}\sum_{j=1}^{m}\left(\log\left(G(d)\omega_j^{-2d}\right) + \frac{\omega_j^{2d} I_x^T(\omega_j)}{G(d)}\right).\end{aligned}$$

This equals the objective function of the classic local Whittle estimator. But [64] note that the approximation $I_u^T(\omega_j) \approx \omega_j^{2d} I_x^T(\omega_j)$ only holds for $|d| < 0.5$. Especially for $|d| > 1$, $\omega_j^{2d} I_x^T(\omega_j)$ is not a good approximation for $I_u^T(\omega_j)$, since x_t is nonstationary; therefore, the meaning of the periodogram as an estimated – in fact non-existent – spectral density gets lost. The data-driven scheme uses the definition of u_t as the dth differences of x_t, so $I_u^T(\omega_j) = I_{\Delta^d x_t}^T(\omega_j)$ and thus [64] get the objective function

$$Q_m^{ELW}(G,d) = \frac{1}{m}\sum_{j=1}^{m}\left[\log(G\omega_j^{-2d}) + \frac{1}{G} I_{\Delta^d x}(\omega_j)\right]. \tag{5.3.11}$$

As above, the two dimensional problem can be reduced to a one dimensional problem, and the ELW estimator is

$$\widehat{d}_{ELW} = \min_{d\in[d_1,d_2]} R(d), \tag{5.3.12}$$

$$\text{where } R(d) = \log G_{opt}(d) - 2d\frac{1}{m}\sum_{j=1}^{m}\log\omega_j \text{ with } G_{opt}(d) = \frac{1}{m}\sum_{j=1}^{m} I_{\Delta^d x}(\omega_j).$$

Phillips and Shimotsu [65] state that the exact local Whittle estimation (numerical optimization) lasts about ten times longer than the local Whittle estimator, but now also non stationary processes beyond $d = 1$ can be estimated accurately.

5.3.2 GPH

Geweke and Porter-Hudak [22] use the exact relation

$$f_x(\lambda) = \left(4\sin^2\frac{\lambda}{2}\right)^{-d} f_u(\lambda), \tag{5.3.13}$$

as the basis for a simple (log) linear regression, with slope coefficient $-d$.

Let $\{x_t\}_{t=1}^T$ be a sample of size T, $\omega_j = \frac{2\pi j}{T} \quad j = 0, \ldots, T-1$ be the harmonic ordinates, and let $I_x^T(\omega_j)$ denote the periodogram. As the above relation holds for all $\lambda \in [-\pi, \pi]$, it especially holds for $\lambda = \omega_j$. Hence,

$$f_x(\omega_j) = \left(4\sin^2\frac{\omega_j}{2}\right)^{-d} f_u(\omega_j) = \left(4\sin^2\frac{\omega_j}{2}\right)^{-d} f_u(\omega_j) \cdot \left(\frac{f_u(0)}{f_u(0)}\frac{I_x^T(\omega_j)}{I_x^T(\omega_j)}\right)$$

Rearranging terms and applying $\log(\cdot)$ gives

$$\begin{aligned}
\log I_x^T(\omega_j) &= \log f_u(0) - d\log 4\sin^2\frac{\omega_j}{2} + \log\frac{f_u(\omega_j)}{f_u(0)} + \log\frac{I_x^T(\omega_j)}{f_x(\omega_j)} \\
&= \left(\log f_u(0) + \mathbb{E}\,\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)}\right) - d\log 4\sin^2\frac{\omega_j}{2} \\
&+ \left(\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)} - \mathbb{E}\,\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)}\right) + \log\frac{f_u(\omega_j)}{f_u(0)}
\end{aligned} \tag{5.3.14}$$

Log-linear model

Equation (5.3.14) is quite similar to a simple linear model $y_j = \beta_0 \cdot 1 + \beta_1 \cdot x_j + \varepsilon$, $j = 1, \ldots, m$, with

- dependent variable $y_j = \log I_x^T(\omega_j)$,
- intercept $\beta_0 \cdot 1 = \left(\log f_u(0) + \mathbb{E}\,\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)}\right)$,
- explanatory variable $x_j = \log 4\sin^2\frac{\omega_j}{2}$ with slope coefficient $\beta_1 = -d$,
- zero-mean disturbances $\varepsilon = \left(\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)} - \mathbb{E}\,\log\frac{I_x^T(\omega_j)}{f_x(\omega_j)}\right)$ with variance σ_ε^2, and
- the term $\log\frac{f_u(\omega_j)}{f_u(0)}$ becomes negligible, as attention is drawn to ω_j close to zero and $\lim_{\omega_j \to 0}\log\frac{f_u(\omega_j)}{f_u(0)} = 0$.

Consequently the GPH estimator is the negative value of the slope estimate $\widehat{\beta}_1^{OLS}$. Since this is a simple linear regression

$$\widehat{\beta}_1^{OLS} = \widehat{corr}(x, y) = \frac{\sum_{j=1}^{m}(x_j - \widehat{\mu}_x)(y_j - \widehat{\mu}_y)}{\sum_{j=1}^{m}(x_j - \widehat{\mu}_x)^2}$$

we have that

$$\widehat{d}_{GPH} := -\widehat{\beta}_1^{OLS} = -\frac{\sum_{j=1}^{m}(\log 4\sin^2\frac{\omega_j}{2} - \widehat{\mu}_x)(\log I(\omega_j) - \widehat{\mu}_y)}{\sum_{j=1}^{m}(\log 4\sin^2\frac{\omega_j}{2} - \widehat{\mu}_x)^2}. \tag{5.3.15}$$

As the approximation only holds for ω_j close to zero, the linear model must be estimated for ω_j close to zero, where m specifies the maximum number of frequencies to include in the sample.

In practice, there is no guideline for the optimal choice of m. As can be seen in the proposition below, the higher m the smaller the variance; but it is also clear that if m becomes too large, the last term in (5.3.14) is not negligible and therefore the OLS estimate of d gets more and more biased.

Proposition 5.3.1 (Consistency of GPH)**.** *For Gaussian processes $\widehat{d}_{GPH}$ is asymptotically normal with*

$$\sqrt{m}\,(\widehat{d}_{GPH} - d) \xrightarrow{d} N\left(0, \frac{\pi^2}{24}\right), \tag{5.3.16}$$

as long as $m \to \infty$ for $T \to \infty$, and $\frac{m \log m}{T} \to 0$ for $T \to \infty$.

Proof. See Hurvich, Deo, and Brodsky [41]. □

Note that $m = c \cdot T^{\alpha}, 0 < \alpha < 1$ satisfies the conditions in proposition 5.3.1. As a rule of thumb $m = T^{0.5}$ is common practice, but as [41] point out, the choice of $m = T^{0.5}$ can perform considerably inferior to the optimal MSE minimizing choice of m, which they report as $m_{opt} = \mathcal{O}(T^{0.8})$.

5.4 Comparison of estimators – Monte Carlo

Although the presented estimators would give good estimates in theory, the main problem of every estimator is the disability to draw a line between short memory and long memory effects. In the frequency domain this is reflected by the difficult choice of m. To overcome the risk of choosing the *wrong* maximum frequency, Taqqu and Teverovsky [78] suggest to rather look at d as a function of m, and choose that $m = m^*$, where $\widehat{d}(m)$ seems to be constant in a neighborhood of m^*.

5.4.1 Feasible properties of an estimator

A desired property of an estimator for the memory parameter is invariance to differencing the series. More formally, let $T(x_t)$ be an estimator of the memory parameter. If $x_t \sim I(d)$, then $T(x_t)$ should be approximately equal to d. Additionally $T(\Delta x_t)$ should be close to $d-1$ since $\Delta x_t \sim I(d-1)$ by definition.

Unfortunately this intuitive requirement for a test statistic does not hold for every memory parameter estimator. For example if we simulate a long memory process $x_t \sim I(0.3)$ and $\widehat{d} = 0.29$, we expect that for the differenced process $y_t = (1-L)x_t \sim I(0.7)$ we get an estimate of $\widehat{d} = 0.71$. As the proposed estimators have different feasible intervals where estimates are unbiased, it is clear that they do no fulfill these basic requirements.

5.4.2 Monte Carlo simulation

Almost every proposed estimator or test statistic comes with a Monte Carlo simulation, showing superior properties compared to other estimators. I simulated $x_t \sim I(d)$ processes over a grid of $d \in \{-0.5, -0.3, \ldots 0.7, 0.9\}$ and compare four common estimators for the long memory parameter.

A first look at the properties of the estimators for a sample size of $T = 500$ with 50 replication shows (Figure 5.1) that only the GPH and ELW estimator perform accurately outside the *persistent and stationary* interval $(0, 0.5)$ (here $m = 0.5$).[5] The R/S statistic is far off the true values and maximum likelihood estimation for $d \notin [0, 0.5]$ was not doable, as it obviously suffers from numerical problems (`fracdiff` package in R 2.5.0 [67]).

Thus, we will restrict the advanced Monte Carlo simulation – with sample sizes of $T = 500$ and $T = 1000$, and 1000 replications – to the interval $(0, 0.5)$ where a comparison between all four estimators is worthwhile.

Figure 5.2 and Table 5.2 show:

- ELW performs better than FIML and GPH, but on this level the bias is in fact 0 for the three estimators. The R/S statistic does not prove to be a good estimator for the long memory parameter at all, although we just considered fractional noise with no short term perturbations.
- The GPH standard deviation is about 5 (!) times larger than for ELW or FIML.
- For $d \in \{0.1, 0.2, 0.3, 0.4\}$ the FIML has a slightly smaller standard deviation than the ELW. The very small standard errors at the boundary values of $d = 0$ and 0.5 for FIML are dubious since, as one can see from

[5] The y-axis of all figures in this section represents $\widehat{d} - d_{true}$, so we can concentrate on the horizontal line at 0.

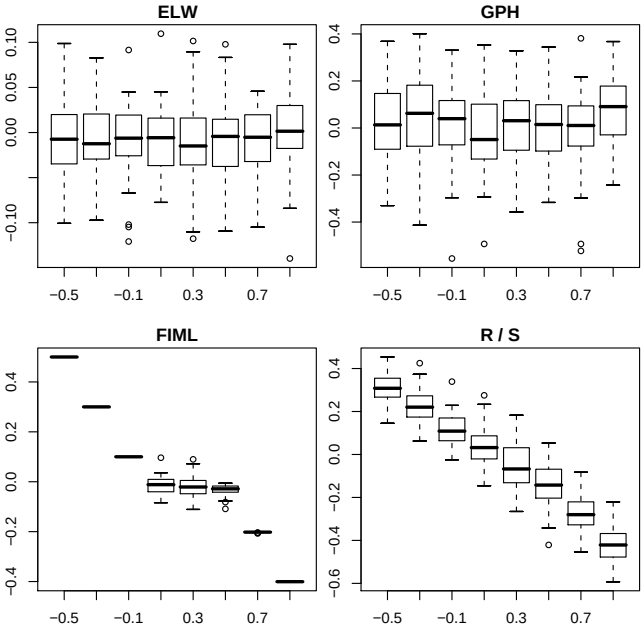


Figure 5.1: Comparison of four commonly used estimators for the long memory parameter with $x_t \sim ARIMA(0, d, 0)$ and d on the interval $(-0.5, 0.9)$

Figure 5.2, the small variance results from feasible optimal values limited to the interval $[0, 0.5]$.

To check if the variance of GPH remains higher for $d \notin \{0, 0.5\}$, I perform another simulation with $T = 500$ observations and d ranging from -0.7 to 1.3 with step-size 0.2.

Analyzing Figure 5.3, and Tables 5.3 and 5.4 we can observe the same pattern as before:

- ELW and GPH are unbiased estimators for $d \in \{-0.5, -0.3, \ldots, 0.9\}$.

true d	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
ELW bias	−0.002	−0.002	−0.001	−0.000	−0.002	−0.002
ELW variance	0.027	0.028	0.027	0.028	0.028	0.027
FIML bias	0.007	−0.005	−0.004	−0.005	−0.008	−0.023
FIML variance	0.013	0.026	0.025	0.026	0.025	0.014
GPH bias	0.000	−0.003	0.005	0.007	0.016	0.016
GPH variance	0.138	0.141	0.14	0.134	0.140	0.141
R/S bias	0.066	0.032	−0.000	−0.039	−0.082	−0.125
R/S variance	0.078	0.085	0.085	0.089	0.092	0.086

Table 5.2: Summary statistics (mean and standard deviation) for 4 memory parameter estimators: ELW, FIML, GPH, and R/S. Sample size $T = 1000$ and 1000 replications for every d.

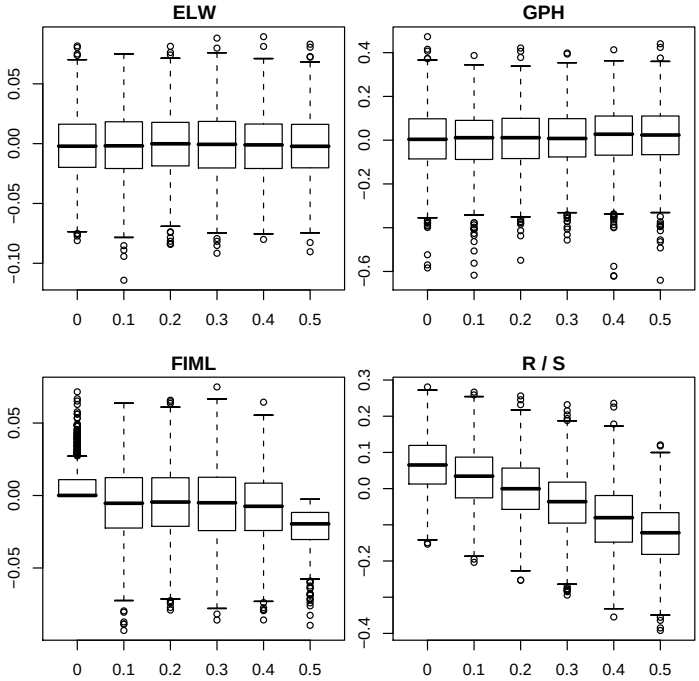


Figure 5.2: Comparison of four commonly used estimators for the long memory parameter with $x_t \sim ARIMA(0, d, 0)$ and d on the interval $[0, 0.5]$. Sample size $T = 1000$ and 1000 replications for every d.

true d	**-0.7**	**-0.5**	**-0.3**	**-0.1**	**0.1**
ELW bias	-0.006	-0.007	-0.006	-0.005	-0.004
ELW variance	0.04	0.040	0.039	0.039	0.039
GPH bias	0.120	0.045	0.012	0.003	-0.002
GPH variance	0.206	0.185	0.176	0.172	0.171

Table 5.3: a) Estimator comparison: sample size $T = 500$ and 1000 replications for every d.

For $d > 1$ both estimators have difficulties to estimate an accurate d. Whereas the ELW estimator overestimates the coefficient, the GPH statistic underestimates the memory parameter.

For $d = -0.7$ the GPH estimator has a slight bias, and presumably this bias will increase for $d < -0.7$.

- Again the GPH standard deviation is about 4 times larger than for ELW (for $d \in \{-0.5, -0.3, \ldots, 0.9\}$).

After all, we face a tradeoff between a fast estimator (fast OLS regression for GPH / slow numerical optimization for ELW) versus small standard errors (data driven ELW).

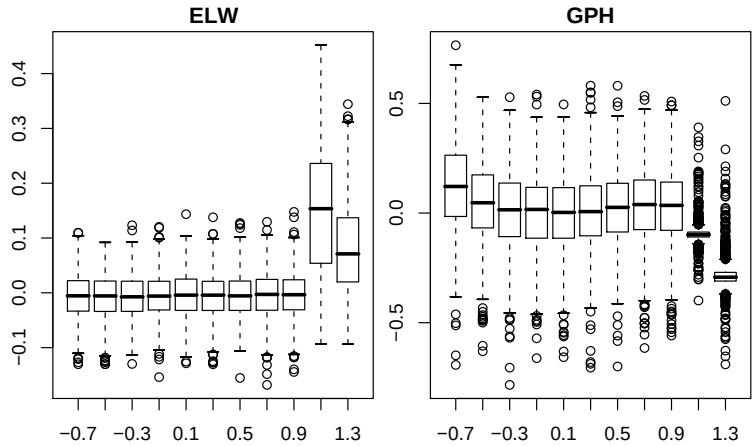


Figure 5.3: Comparison of the ELW and GPH estimator with $x_t \sim ARIMA(0, d, 0)$ and d on the interval $[-0.7, 1.3]$. Sample size $T = 500$ and 1000 replications for every d.

true d	**0.3**	**0.5**	**0.7**	**0.9**	**1.1**	**1.3**
ELW bias	−0.005	−0.004	−0.004	−0.004	0.154	0.081
ELW variance	0.040	0.039	0.041	0.04	0.115	0.08
GPH bias	0.008	0.021	0.033	0.026	−0.093	−0.278
GPH variance	0.177	0.165	0.171	0.167	0.057	0.088

Table 5.4: b) Estimator comparison: sample size $T = 500$ and 1000 replications for every d.

Chapter 6

Time – Varying Memory

In an economical/financial framework one can interpret d as a measure of memory length of investors or the influence of the past in the market. By imposing a constant memory parameter in such processes, investors would always take the same amount of past information into account when making their investment decisions. This view of the world is neither reasonable nor supported by empirical data.

Here I present an empirically motivated approach to capture time variation in the memory parameter. Although many authors [6, 29] detect different memory parameters for subperiods of analyzed data, they do not pursue this issue but deduce possible breaks in the mean, and consequently refer to spurious long memory.[1]

Based on my point of view on the spurious long memory discussion, I do not consider locally changing memory parameters as an instantaneous indication for structural breaks, but as a potential demand for time-varying memory models.

This is not supposed to be a complete theoretical analysis of time-varying long memory models, but rather a collection of ideas on time-varying memory, its implications for the defined stochastic process, and possible practical applications to get better forecasts and more accurate confidence intervals. For more theoretical considerations I refer the reader to [27, 42], which are – to my knowledge – the only theoretical considerations of time-varying long memory models. Subsequently I will present simulations of time-varying (long) memory processes, also performed in [60].

Time-varying memory is probably similar to locally stationary processes [see 12], although the length of memory is not restricted to the case where $\mathbb{V}x_t < \infty$. Thus, non stationary processes are also allowed in this concept.

[1] See Morana and Beltratti [54] for a structural breaks *and* long memory modeling of exchange rate volatility.

6.1 Bounded variation

In the standard model the memory parameter is restricted to a constant value in $\mathbb{R}$, but there are no restrictions on the value of d. If the parameter d is allowed to change, then this variation should reflect some behavior of the underlying forces. From an economical perspective, time variation in d displays a constantly changing behavior of people in the market, but this variation is presumably limited to a certain range (d_{min}, d_{max}). I distinct between (at least) two possible variations in time:

a) The process starts in a state A and evolves into a final equilibrium state B. In our context this means that starting from state $A \widehat{=} d_{-\infty} \in \mathbb{R}$, the memory parameter tends – in a *nice* way – to the final state $B \widehat{=} d_{\infty} \in \mathbb{R}$.

b) The process changes back and forth between two (or more) states in a repetitive way. Correspondingly d_t follows a periodic path in some bounded interval.

In the first case, consider a process starting in a totally chaotic fashion (random walk) and as time goes on the process tends to an equilibrium, e.g. white noise. Here d goes from 1 to 0 in some well-behaved manner.

For the periodic case, I am expecting that natural boundaries are 0, 0.5 and 1, as these values represent thresholds where the structure of the process changes rigorously. (0 ... short/long memory; 0.5 ... finite/infinite variance; 1 ... mean reversion / no mean reversion). Thus, in a market people change their behavior if d_t reaches these bounds, and this consequently results in a change of d_t.

6.2 Different memory measure

Although ARFIMA models are possibly one of the best models for estimation and prediction for long range dependent processes, one flaw for a time-varying consideration is that for the wide class of long memory processes, d is in the open set $(0, 1)$, whereas for short memory models, which represent just as well a broad class of processes, d is reduced to the point $d = 0$ (see Table 6.1). Thus, for a reflective explanation I suggest a different measure and the ED model is again of great help.

6.2.1 Time – varying stochastic duration

In the error duration model an extension to time variation can be achieved by a varying distribution function $F(k) = F_s(k)$ for n_s.

Consequently $p_k = p_{(s,k)}$ is the probability that *shock* ε_s *survives until* $s + k$,

$$p_{(s,k)} := P(g_{s,s+k} = 1),\ k = 0, 1, 2, \ldots \sim F_s.$$

Memory	Properties	ARFIMA	ED	$p_k = \left(\frac{1}{k+1}\right)^a$
No	MR, FV	d=0	$p_k = 0$ for $k \geq 1$	$a = \infty$
Short	MR, FV	d=0	$\sum_{k=0}^{\infty}(k+1)p_k < \infty$	$2 < a < \infty$
Long	MR, FV	$0 < d < 0.5$	$\sum_{k=0}^{\infty}(k+1)p_k = \infty$, $\sum_{k=0}^{\infty} p_k < \infty$	$1 < a \leq 2$
Long	MR, no FV	$0.5 \leq d < 1$	$\sum_{k=0}^{\infty}(k+1)p_k = \infty$, $\sum_{k=0}^{\infty} p_k = \infty$	$0 < a \leq 1$
Unit Root	no MR, no FV	$d = 1$	$p_k = 1 \quad \forall k$	$a = 0$

Table 6.1: Different memory and its classification. MR ... mean reversion; FV ... finite variance.

For further analysis we use a specific structure of p_k depending on the *amnesia* parameter $a \in \mathbb{R} \cup \{-\infty, \infty\}$,

$$P(n_s \geq k) = p_k := \left(\frac{1}{k+1}\right)^a \in [0,1] \quad \forall a \text{ and } k \geq 0. \tag{6.2.1}$$

Claim 6.2.1. *For $a \in [0, \infty]$ the probabilities in (6.2.1) are feasible in the error duration sense.*

A classification of memory depending on a is given by the last column in Table 6.1.

Proof. A sequence $\{p_k\}$ is a suitable probability measure in the error duration sense if

1. $p_0 = 1$
2. $p_{k+1} \leq p_k \quad k = 0, 1, 2, \ldots$

The first condition holds for every $a \in \mathbb{R} \cup \{-\infty, \infty\}$. The second condition implies

$$\left(\frac{1}{k+2}\right)^a \leq \left(\frac{1}{k+1}\right)^a \Leftrightarrow \left(1 - \frac{1}{k+2}\right)^a \leq 1.$$

As $\left(1 - \frac{1}{k+2}\right) < 1 \quad \forall k$, it follows that only $a \in [0, \infty]$ give proper ED probabilities.

No memory – White noise: For white noise $p_0 = 1$ and $p_k = 0 \quad \forall k \geq 1$. It is clear that only $a = \infty$ gives this result.

Short memory: The ED model possesses short memory iff $\sum_{k=0}^{\infty}(k+1)p_k < \infty$. Since

$$\sum_{k=0}^{\infty}(k+1)\left(\frac{1}{k+1}\right)^a = \sum_{k=0}^{\infty}(k+1)^{1-a} = \sum_{n=1}^{\infty}\frac{1}{n^{a-1}}.$$

it follows that only for $a - 1 > 1 \Leftrightarrow a > 2$ the sum converges.

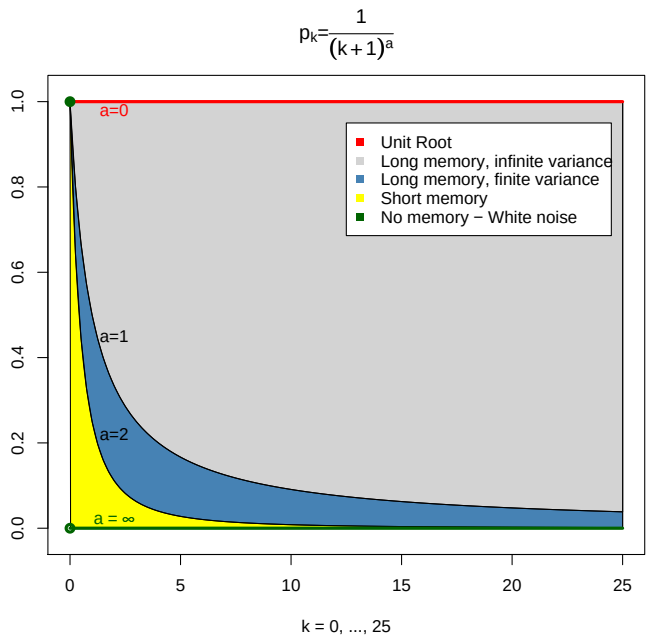


Figure 6.1: Classification of types of memory

Long memory, finite variance: For $a \leq 2$ the process exhibits long memory. The variance of an ED process equals $\sum_{k=0}^{\infty} p_k = \sum_{n=1}^{\infty} \frac{1}{n^a}$. Hence, the variance is finite for $a > 1$.

Long memory, infinite variance: For $0 < a \leq 1$ we have long memory with infinite variance.

Unit root: As a unit root means that every shock lasts forever (no mean reversion and infinite variance), i.e. $p_k = 1 \quad \forall k$, it follows that $a = 0$.

□

As this classification only depends on convergence results it can be used for every feasible probability sequence p_k that satisfies $p_k = \mathcal{O}((k+1)^{-a})$.

Within this classification, processes with short memory, long memory with finite variance, and long memory with infinite variance are all represented by (half-)open sets in $\mathbb{R}$ (see Table 6.1 and Figure 6.1). So, contrary to the ARFIMA classification short memory models do not reduce to a set with Lebesgue measure zero.

Remark 6.2.2. *Note the connection of the long memory condition and the variance of the process to Riemann's zeta function* $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$. *(Zeta distribution ?)*

Remark 6.2.3. *Also note the similarity of the ED model to compound Poisson processes and renewal theory.*

6.2.2 The model

As noted above, a time varying memory parameter d in the ARFIMA model would lead to *ugly* behavior of the path of d. With the *amnesia* parameter we have a nice smooth parameter for different memory models and we can easily formulate time-varying properties of an ED model.

The standard ED model is based on i.i.d. shocks with stochastic durations n_s and corresponding probability distribution function $F(k) = P(n_s \leq k)$.

$$y_t := \sum_{s=-\infty}^{t} g_{s,t}\varepsilon_s, \quad \varepsilon_s \sim IID(0, \sigma_\varepsilon^2) \tag{6.2.2}$$

$$\text{and } g_{s,t} := \begin{cases} 1 & t \leq s + n_s, \\ 0 & t > s + n_s. \end{cases} \tag{6.2.3}$$

Already noted in condition 2.2.3 the variance of an ED process y_t equals $1 + \nu$, where $\nu = \mathbb{E}\, n_s$.

Suppose the random variable *shock duration* is switching between various states with corresponding distribution functions $F_1, F_2, \ldots, F_Q$. If the expected values $\nu_1, \nu_2, \ldots, \nu_Q$ all equal ν, the variance of the process remains constant over time. But note that as the underlying probability distribution changes, the covariance structure also varies.

Example 6.2.4. *For simplicity I consider a special class of distribution functions, namely the negative binomial distribution function with parameters* $0 < p < 1$ *and* $r > 0$. *For a negative binomial distributed random variable* X *it holds*

$$P(X = k) = f(k; r, p) = \frac{\Gamma(r+k)}{k!\,\Gamma(r)}\, p^r\, (1-p)^k \text{ for } k = 0, 1, 2, \ldots.$$

We have

$$\mu = \mathbb{E}\, X = r\frac{1-p}{p} \text{ and } \sigma^2 = \mathbb{V}\, X = r\frac{1-p}{p^2} = \frac{1}{p}\mu.$$

Now suppose the error duration switches from $f_1(k; r^{(1)}, p^{(1)})$ *to* $f_2(k; r^{(2)}, p^{(2)})$. *Set* $(r^{(1)}, p^{(1)}) = (5, 1/2)$ *and* $(r^{(2)}, p^{(2)}) = (5/3, 1/4)$, *then*

$$\mu_1 = 5\frac{1/2}{1/2} = 5, \quad \sigma_1^2 = \frac{1}{1/2}\mu_1 = 2\mu_1 = 10 \text{ and} \tag{6.2.4}$$

$$\mu_2 = \frac{5}{3}\frac{3/4}{1/4} = 5, \quad \sigma_2^2 = \frac{1}{1/4}\mu_2 = 20 = 2\sigma_1^2. \tag{6.2.5}$$

Thus, in either state $\mathbb{V}\, y_t = \sigma_\varepsilon^2(1+5) = 6\,\sigma_\varepsilon^2$, *but the durations in state 2 have fatter tails than in the first state resulting in higher autocavariances in the second state than in the first one* ($\gamma_2(k) \geq \gamma_1(k)$ *for high lags* k).

Definition 6.2.5. *A time-varying memory ED model is given by*

$$\begin{aligned} y_t &:= \sum_{s=-\infty}^{t} g_{s,t}\varepsilon_s, \quad \varepsilon_s \sim IID(0,\sigma_\varepsilon^2), \\ g_{s,t} &:= \begin{cases} 1 & t \leq s+n_s, \\ 0 & t > s+n_s. \end{cases} \\ \text{and } n_s &\sim F_s(k) \text{ for an } F_s \text{ in a set of distribution functions } \left\{F_j\right\}_{j\in J}, \\ & \quad \text{with } J \text{ either a finite or infinite index set.} \end{aligned} \tag{6.2.6}$$

For the *amnesia* classification, the index set is infinite and $F_s(k) \sim p_{s,k} = \left(\frac{1}{k+1}\right)^{a(s)}$, where $a(s)$ is the trajectory of the parameter in time.

Example 6.2.6. *Suppose we observe a process with the following varying behavior (see Figure 6.2):*

1. nice *behavior (mean reverting (MR) and finite variance (FV); short memory) in the beginning ($2 < a \leq \infty$),*
2. *memory lengthens ($1 < a \leq 2$),*
3. *the process goes to a state of infinite variance, but still MR (no unit root) ($0 < a < 1$),*
4. *Then the variance is decreasing again to a finite level ($1 < a \leq 2$), but does not stay long and*
5. *increases again to infinite variance ($0 < a < 1$),*
6. *at this point a break in the mean occurs ($a = 0$),*
7. *after the break the variance is still high but decreasing ($0 < a < 2$),*
8. *and finally the process behaves* nicely *again ($2 < a \leq \infty$).*

6.3 Time varying ARFIMA

Replacing d by d_t in

$$(1-L)^d(x_t-\mu) = \Psi(L)\varepsilon_t \longrightarrow (1-L)^{d_t}(x_t-\mu) = \Psi(L)\varepsilon_t \Rightarrow x_t-\mu = \sum_{j=0}^{\infty} w_j(t)\varepsilon_{t-j} \tag{6.3.1}$$

provides a natural expansion from static memory to dynamic memory.

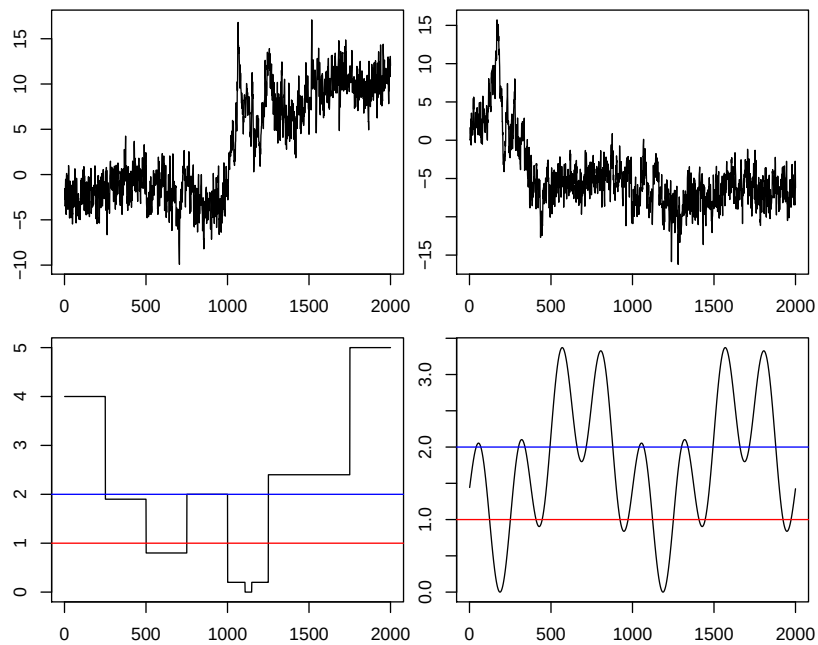


Figure 6.2: (left) Realization of the time varying memory model described in Example 6.2.6; (right) Harmonically changing memory with structural breaks

Although the simulation of time varying $d = d_t$ might seem straightforward, there are certain difficulties that arise. The biggest problem is that as the sample is always finite, the sum stops at a certain maximum lag. Thus, even if d_t equals 1 for some $t = t^*$, after a certain time frame, this obvious structural break in the series will be forgotten again. Additionally time-varying ARFIMA simulation is computationally much more extensive than time-varying ED simulation.

But for the estimation of time-varying memory parameters I will use the ARFIMA framework, since the literature on long memory estimators for ARFIMA models is excessive.

6.4 Estimating time variation

Using a rolling window approach, I integrate usual estimation methods of the long memory parameter d in a time-varying framework.

Let $L = 2b+1$ be the window size and b denote the bandwidth. This special choice simplifies notation in the text below. A varying estimator for a sample $\{x_t\}_{t=1}^T$ is given by

$$\widehat{d}_t(x_t) = \widehat{d}(\mathbf{y}) \text{ for the subsample } \mathbf{y} = \{x_{t-2b}, \ldots, x_t\}\,, \tag{6.4.1}$$

and $\widehat{d}$ is one of the classic estimators, such as ELW, FIML, or GPH.

The use of moving windows needs a huge amount of computational work, since d has to be estimated for every window. To lower the computation time, one can use more heuristic, and thus less computationally skilled, methods as the time-variance or ACF method to determine the variation in d. Besides, estimating $\widehat{d}$ not for every single t, but for a grid $\{0 < L, L+s, L+2s, \ldots, T\}$ with step size $s > 1$, will also decrease computation time (trivially by the factor $\frac{1}{s}$). As the window length is $2b+1$, we can not start at $t = 1$ but with $t = b+1$. If a time-varying memory approach seems justifiable, local estimates for d_t should be computed with the most accurate estimators as ELW, FIML or GPH.

6.5 Analyzing time variation

After estimating d_t, forecasting d_t itself is interesting to get even better predictions for x_t and more accurate confidence intervals. Before proceeding, we have to decide whether d_t should be modeled deterministically or stochastically. This will determine the mehtods to forecast d_t; either non-parametrically or with parametric models.

6.5.1 Parametric or non-parametric

In parametric analysis d_t is subject to time series analysis. Thus, fitting (seasonal) ARIMA models – with possible explanatory variables, e.g. variance of the process – not only provide a basis for further analysis, but also help to get a better understanding for the process' memory.

From empirical work it becomes clear that d_t is a combination of smooth functions and very erratic behavior. Therefore, applying parametric models for d_t does not seem reasonable. Hence, methods from non-parametric analysis like polynomial regression and splines can be used to model the time variation.

I think the decision between parametric or non-parametric modeling is strongly related to the question whether d_t is periodic or tending to an equilibrium state. For parametric modeling of the time-variation see Granger and Zhuanxin [27], who consider $AR(1)$ and Markov switching models for d_t.

6.5.2 Forecasting d

Recall that d_t is just another notation for the function

$$\begin{aligned} d: \quad & \mathbb{N} \to \mathbb{R} \\ & t \to d(t). \end{aligned}$$

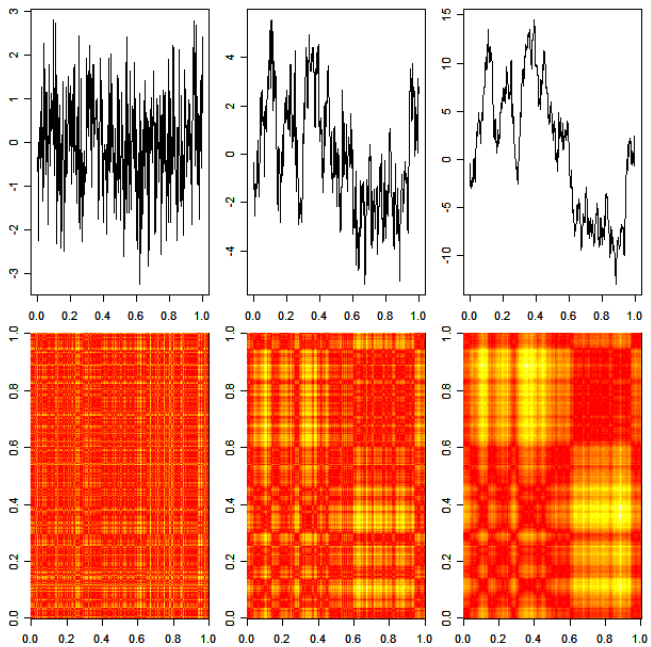


Figure 6.3: Recurrence plots with same i.i.d. innovations: (left) fractional noise with $d = 0.2$; (middle) fractional noise with $d = 0.7$; (right) random walk, $d = 1$

For a well behaved d_t the Taylor expansion around the point α is given by

$$d(t) = \sum_{n=0}^{\infty} \frac{d^{(n)}(\alpha)}{n!}(t-\alpha)^n \text{ where } d^{(n)} \text{ is the n-th derivative of d(t).}$$

Expanding $d(t+h)$ in a Taylor series around t

$$d(t+h) = \sum_{n=0}^{\infty} \frac{d^{(n)}(t)}{n!} h^n$$

we intuitevly get a forecast for d_{t+h} by truncation of the infinite power series at some maximum lag K:

$$\widehat{d}_{t+h} = \sum_{n=0}^{K} \frac{d^{(n)}(t)}{n!} h^n. \tag{6.5.1}$$

Given a sequence of forecasted values $\widehat{d}_{t+h}$ we can compute a time-varying weighting sequence $\left\{w_j(t+h)\right\}_{t=1-h}^{T}$ and consequently forecast x_{t+h} by

$$\widehat{x}_{t,h} = \sum_{j=h}^{\infty} w_j(t+h)\varepsilon_{t+h-j}. \tag{6.5.2}$$

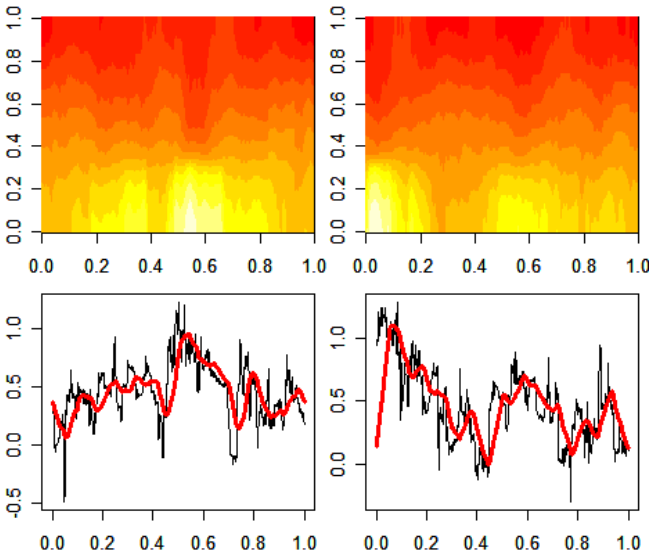


Figure 6.4: (top) Time – varying log spectral densities; (bottom) Time – varying $\widehat{d}_{GPH}$ and smoothed version (red line); window length = 150, step size = 1.

6.6 Graphical detection tools

Here I present two graphical tools that help detect time-varying structures in a process.

Recurrence plot: A recurrence plot displays the closeness of points in time. This is not only interesting for the time variation, but also for the distinction between $I(d)$ processes with $d \in [\frac{1}{2}, 1)$ and $I(1)$ processes, as the first is mean reverting, but the second is not. Thus, there should be a distinct difference between the recurrence plot of a long memory and a random walk (see Figure 6.3).

Time-varying spectrum: Computing the spectrum for different time periods and plotting a heat map (see Figure 6.4) or a 3-dimensional graph gives a first indication about possible time-variation.

Figure 6.4 displays the estimated time variation of d_t. The moment d_t hits the limit of 1 corresponds to a break in the series ($\approx 0.58 * 3000 = 1160$). Reconsidering Example 6.2.6 and Figure 6.2 shows that the moving window approach detects the break almost perfectly.

6.7 Conclusion

Of course the presented ideas do not stand on a fully developed theoretical basis, but are empirically motivated approaches for time series modeling. As noted above, a time-varying model (not restricted to varying memory) is also a more

realistic way of seeing the world. Undoubtedly, consideration of time-varying models is much more complex than just writing down a model and simulating data. Nevertheless, it is a first step towards a more profound, – and in my opinion – promising analysis of time-varying memory.

Chapter 7

Applications

In this chapter I will apply the theory described above to real data. The goal is to estimate parameters of a model for x_t and predict future values $x_{t+l}, l \geq 1$.

As an underlying model for financial data I use a stochastic volatility model presented below and used in several publications (e.g. Deo and Hurvich [15]). All computations and plots have been done with R 2.5.0 [see 67].

7.1 LMSV model

There is a considerably large literature about ARCH models, which model *conditional heteroskedasticity* by an autoregressive model. Extending this to long memory models might seem straightforward, but as there are certain restrictions on the parameters some difficulties occur for defining such a long memory heteroskedastic model. I will not go into detail here, but refer to [7, 16].

Breidt et al. [7] and Harvey [35] introduce the *long memory stochastic volatility model* (LMSV) for modeling volatility, which is an extension of the basic stochastic volatility model proposed by Melino and Turnbull [53].

Definition 7.1.1 (Stochastic Volatility Model)**.** *The* stochastic volatility model *for r_t is defined as*

$$r_t = \sigma_t \varepsilon_t, \quad \sigma_t = \sigma e^{\frac{v_t}{2}}, \quad \sigma > 0 \tag{7.1.1}$$

where $\varepsilon_t \sim IID(0,1)$ is independent of v_t.

Common choices for ε_t are Gaussian processes or processes with a normalized t-distribution and n degrees of freedom. If $v_t \sim ARIMA(p,d,q)$, we have a *long memory* stochastic volatility model.

Definition 7.1.2 (Log normal distribution)**.** *A random variable Y has a log normal distribution if $X = ln(Y)$ has a normal distribution. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the probability density function for Y is*

$$f(Y) = \frac{1}{\sqrt{2\pi}\sigma Y} e^{-\frac{ln(Y)-\mu}{2\sigma}}.$$

Corollary 7.1.3 (Moment generating function)**.** *The moment generating function* $\mathbb{E}y^n = \int_0^\infty y^n f(y)\,\mathrm{d}y$ *of a log normal distribution is given by*

$$MG_n(\mu,\sigma) := e^{n\mu + \frac{n^2\sigma^2}{2}}. \tag{7.1.2}$$

7.1.1 Properties of LMSV

If the long memory process v_t is a zero mean Gaussian process with autocovariance function $\gamma_v(j)$, then

$$\begin{aligned}
\mathbb{E}r_t &= \mathbb{E}\sigma_t\varepsilon_t = \mathbb{E}\sigma_t\mathbb{E}\varepsilon_t = \mathbb{E}\sigma_t \cdot 0 = 0 && (7.1.3)\\
\mathbb{V}r_t &= \mathbb{E}r_t^2 = \mathbb{E}\sigma_t^2\varepsilon_t^2 = \mathbb{E}\sigma_t^2\mathbb{E}\varepsilon_t^2 \\
&= \mathbb{E}\sigma_t^2 = \sigma^2\mathbb{E}e^{v_t} = \sigma^2 MG_1(0,\gamma_v(0)) = \sigma^2 e^{\frac{\gamma_v(0)}{2}} && (7.1.4)\\
cov(r_t, r_{t+k}) &= 0 \text{ for } k \neq 0, \text{ since } \varepsilon_t \text{ is i.i.d..} && (7.1.5)
\end{aligned}$$

Therefore, r_t is covariance stationary white noise, which is compatible with the efficient market hypothesis.

Another compelling feature of the proposed model, with Gaussian i.i.d. ε_t, is a positive excess kurtosis.

$$\begin{aligned}
\gamma_2 &:= \frac{\mathbb{E}r_t^4}{\left(\mathbb{E}r_t^2\right)^2} - 3 = \frac{\mathbb{E}\sigma_t^4\varepsilon_t^4}{\left(\mathbb{E}\sigma_t^2\varepsilon_t^2\right)^2} - 3 = \frac{\mathbb{E}\sigma_t^4 \overbrace{\mathbb{E}\varepsilon_t^4}^{=3}}{\left(\mathbb{E}\sigma_t^2 \underbrace{\mathbb{E}\varepsilon_t^2}_{=1}\right)^2} - 3 \\
&= 3\left(\frac{\sigma^4\mathbb{E}(e^{v_t})^2}{\left(\sigma^2\mathbb{E}e^{v_t}\right)^2} - 1\right) = 3\left(\frac{\sigma^4 MG_2(0,\gamma_v(0))}{\sigma^4 MG_1(0,\gamma_v(0))^2} - 1\right) \\
&= 3\left(\frac{e^{2\mu + \frac{4\gamma_v(0)}{2}}}{\left(e^{1\mu + \frac{\gamma_v(0)}{2}}\right)^2} - 1\right) = 3\left(e^{\gamma_v(0)} - 1\right) > 0. && (7.1.6)
\end{aligned}$$

Note that γ_2 is always positive, as $\gamma_v(0) = \sigma_v^2 > 0$. Also known as *leptokurtic*, a positive excess kurtosis represents *fat tails* in a density and is quite characteristic for financial data.

Interested in the volatility of r_t we analyze the squared returns $y_t := r_t^2$.

$$\begin{aligned}
\mathbb{E}y_t &= \mathbb{E}r_t^2 = \sigma^2 e^{\frac{\gamma_v(0)}{2}} && (7.1.7)\\
\mathbb{V}y_t &= \mathbb{E}y_t^2 - (\mathbb{E}y_t)^2 = \mathbb{E}r_t^4 - \left(\sigma^2 e^{\frac{\gamma_v(0)}{2}}\right)^2 = 3\,\sigma^4\mathbb{E}e^{2v_t} - \sigma^4 e^{\gamma_v(0)} \\
&= \sigma^4\left(3\,MG_2(0,\gamma_v(0)) - e^{\gamma_v(0)}\right) = \sigma^4\left(3\,e^{2\gamma_v(0)} - e^{\gamma_v(0)}\right) \\
&= \sigma^4 e^{\gamma_v(0)}\left(3\,e^{\gamma_v(0)} - 1\right) && (7.1.8)\\
cov(y_t, y_{t+k}) &= \sigma^4 e^{\gamma_v(0)}\left(e^{\gamma_v(k)} - 1\right) \text{ for } k \neq 0. && (7.1.9)
\end{aligned}$$

Squaring equation (7.1.1) and applying the logarithm we get a simple (log) linear model

$$x_t = \mu + \zeta_t + v_t, \tag{7.1.10}$$

where $\mu = \log(\sigma^2) + \mathbb{E}\log(\varepsilon_t^2)$ and $\zeta_t = \log(\varepsilon_t^2) - \mathbb{E}\log(\varepsilon_t^2)$ is i.i.d. with zero mean and variance $\sigma_\zeta^2 = \mathbb{E}\left(\log(\varepsilon^2) - \mathbb{E}\log(\varepsilon_t^2)\right)^2$. Disregarding the mean μ, x_t is a signal plus noise model with a long memory signal v_t uncorrelated to the (non-Gaussian) noise ζ_t.

For x_t we have

$$\begin{aligned}
\mathbb{E}x_t &= \mu \\
\mathbb{V}x_t &= \mathbb{V}\zeta_t + \mathbb{V}v_t = \sigma_\zeta^2 + \gamma_v(0) \\
cov(x_t, x_{t+k}) = \gamma_x(k) &= \begin{cases} \gamma_v(0) + \sigma_\zeta^2 & \text{if } k = 0, \\ \gamma_v(k) & \text{if } k \geq 1. \end{cases}
\end{aligned}$$

Remark 7.1.4. *Note that*

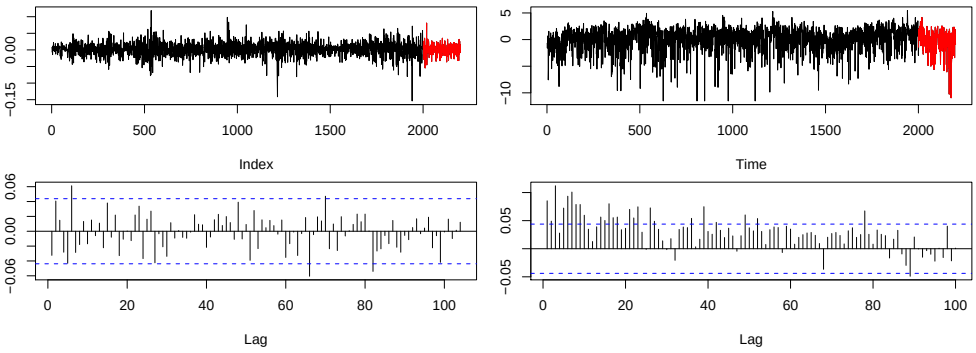
$$\rho_x(k) = \begin{cases} 1 & \textit{if } k = 0, \\ \frac{\gamma_v(k)}{\gamma_v(0)+\sigma_\zeta^2} & \textit{if } k \geq 1. \end{cases} \tag{7.1.11}$$

The autocorrelation function explicitly depends on the error variance σ_ζ^2: the higher the variance, the smaller the autocorrelations for $k \geq 1$.

Suppose σ_ζ^2 tends to ∞, then the autocorrelations go to 0 and we might be entrapped into rejecting long memory in the data in favor of white noise. See Granger and Marmol [26] for a detailed discussion.

If we set up an $ARFIMA(p, d, q)$ model for v_t

$$v_t = (1-L)^{-d}\Phi(L)\Theta(L)\eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2),$$



(a) (top) r_t: weekly DJI log-returns from 1964 - 2006; (bottom) autocorrelation function $\rho_r(k), k \neq 0$

(b) (top) x_t: transformed weekly DJI log-returns from 1964 - 2006; (bottom) autocorrelation function $\rho_x(k), k \neq 0$

Figure 7.1: DJI in the time domain

then the spectral density of the signal plus noise process x_t is given by

$$\begin{aligned} f_x(\lambda) &= f_v(\lambda) + f_\zeta(\lambda) \\ &= \frac{\sigma_\eta^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|1-e^{-i\lambda}|^{2d}|\Phi(e^{-i\lambda})|^2} + \frac{\sigma_\zeta^2}{2\pi}, \quad -\pi \leq \lambda \leq \pi. \end{aligned} \tag{7.1.12}$$

Although there is a gap between lag 0 and lag 1, the autocovariance function of x_t is the same as for v_t for $k \geq 1$. Since estimation methods, both in the time and frequency domain, draw attention to high lags and low frequencies, respectively, I estimate the long memory parameter d for x_t and then the ARMA parameters for the residual series.

7.2 Weekly DJI log returns

I study weekly log returns r_t (see Figure 7.1(a)) of the Dow Jones Index from July 6th, 1964 – September 5th, 2006[1], giving a total number of 2200 observations. For out-of-sample forecasts comparison, I split up the data in a training set $\{x_1, \ldots, x_{2000}\}$ and a test set $\{x_{2001}, \ldots, x_{2200}\}$. All tests below only use the training data.

As noted above, I manipulate the data to

$$x_t = \log(\mu_0 + r_t^2), \quad \mu_0 \ll 1,$$

and presuppose that the process is generated by a LMSV model (7.1.10). The constant μ_0 is added to r_t^2 for numerical issues, as $\log r_t^2$ gives undefined values for returns equal to 0. In this application $\mu_0 = 10^{-9}$.

[1]Source: *Yahoo Finance* – http://finance.yahoo.com.

H_0	H_1	Test	Lags / q	Statistic	p - value
$I(0)$	$I(1)$	KPSS	31	0.49	0.04
$I(1)$	$I(0)$	Phillips - Perron	25	-45.44	0.01
$I(1)$	$I(0)$	Augmented DF	12	-9.76	0.01
$I(1)$	bounded process	Bounded R/S	4	0.36	0.01

Table 7.1: DJI: Unit root tests

As we can see from the autocorrelation function of the original data, r_t seems to be white noise, although there is a slight 1 month (4-5 weeks) dependence in the data. The transformed series x_t exhibits the patterns described in remark 7.1.4. In specific the very small autocorrelations for $k \geq 1$, due to a presumably large error variance σ_ζ^2. Nevertheless we can make out a very slow decay for higher lags autocorrelations (see Figure 7.1(b)).

7.2.1 Unit root

Inspecting the data in Figure 7.1(b) indicates that

- there is no apparent global trend in the data; thus, unit root tests do not need a trend component in their null hypothesis.
- the data are skewed to the left, since a considerable amount of times the returns r_t are very close, or even equal, to zero; therefore, the Gaussian condition will not be met.
- the autocorrelations do not show a typical *exponential* decay to zero.

Nevertheless I perform three widely used unit root tests (Dickey and Fuller [17], Kwiatkowski et al. [45], Phillips and Perron [63]). Additionally Cavaliere [10] proposed a unit root test based on the range of a process, which is a generalization of the R/S statistic (Section 5.2.1).

The results of the tests (see Table 7.1) confirm the guess of fractional integration, as *both*, the $I(0)$ and the $I(1)$ hypothesis, can be rejected at a 0.5 and 0.01 level respectively. Additionally the test procedures detect significant autocorrelations for large lags, as they include lagged values from 10 to 31.

7.2.2 Long memory

Analyzing several unit root tests the *true* d seems to lie in the interval $(0, 1)$. Although the observed data exhibits fat tails a value of $d < \frac{1}{2}$ is reasonable. As the structure of the stationary process u_t is not known a priori, I apply the semi-parametric estimators ELW and GPH to get good starting values for the FIML procedure.

$m = T^{\alpha}$	0.4	0.45	0.5	0.55	0.6	0.65	0.7
GPH	0.33	0.31	0.27	0.29	0.28	0.27	0.26
ELW	0.42	0.37	0.31	0.33	0.25	0.25	0.24

Table 7.2: DJI: semiparametric estimators for different values of m.

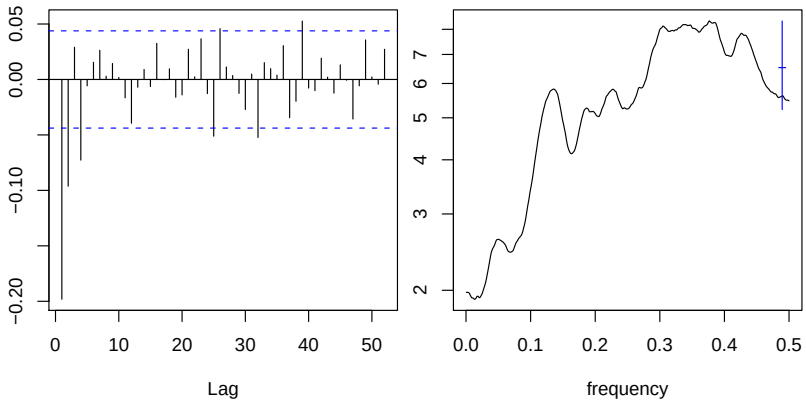


Figure 7.2: (left) Autocorrelation function of u_t (without $\rho(0) = 1$); (right) Spectrum of u_t

Semiparametric Estimation

Considering Table 7.2, $d_0 = 0.29$ seems to be a good starting value for $\widehat{d}$. Figure 7.2 shows the autocorrelation function of $u_t := (1-L)^{-0.29}x_t$, with significant lags at $k = 1, 2, 4$, and the spectral density with a peak around $\omega_j = 0.136$. This corresponds to a period of $\frac{1}{\omega_j} = 7.35$. As we deal with weekly data, this corresponds to a cycle length of almost 2 months (8 weeks), which is important for the variation of Dow Jones volatility.

FIML

I choose an $ARFIMA(4, d, 2)$ model as the underlying structure of the FIML estimation. After various modifications I get

$$(1-L)^{0.3}(1 - 0.92L + 0.55L^2 + 0.14L^4)x_t = (1 - 1.17L + 0.66L^2)\varepsilon_t, \quad \widehat{\sigma}^2_{\varepsilon} = 7.204 \tag{7.2.04}$$

with standard errors $(NA, 0.21, 0.13, 0.02, 0.21, 0.16)$ for $(d, \phi_1, \phi_2, \phi_4, \theta_1, \theta_2)$.

The Ljung-Box statistic for the residuals, squared residuals, and absolute residuals is presented in Table 7.3. The residuals seem to be white noise, and also no heteroskedasticity effects are present.

7.2.3 Time variation

Figure 7.4 displays the time – varying spectrum and Figure 7.3 time varying memory parameter $\widehat{d}_t$.

Lags	1	2	5	10	20	50
$\widehat{\varepsilon}$	0.81	0.79	0.78	0.96	0.99	0.57
$\widehat{\varepsilon}^2$	0.58	0.85	0.52	0.82	0.86	0.97
$\lvert\widehat{\varepsilon}\rvert$	0.66	0.52	0.10	0.36	0.42	0.70

Table 7.3: p-values of Ljung-Box statistic for the residuals of (7.2.1)

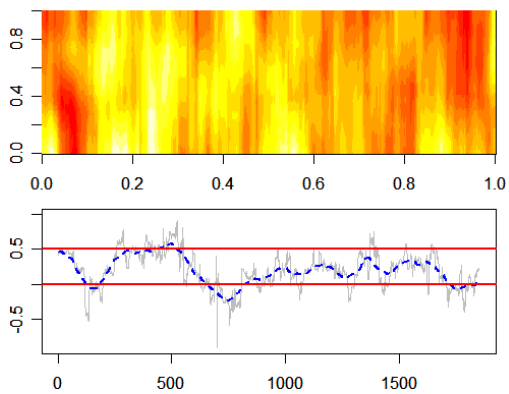


Figure 7.3: DJI: (top) Time – varying spectrum with window length = 150 (≈ 3 years) and step size 1; (bottom) Time varying memory parameter d_t and smoothed version .

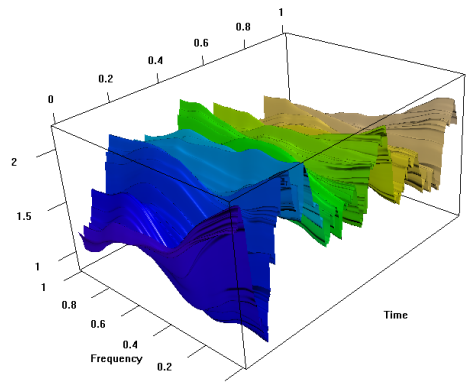


Figure 7.4: DJI: Time varying spectrum with window length = 150 (≈ 3 years) and step size 4 (1 month)

7.2.4 Model comparison

I compare different models and try to measure the goodness of fit by descriptive and adequate criteria. As our final goal are better forecasts I use the following criteria with regard to the out-of-sample errors $e_i = x_{t+i} - \widehat{x}_{t,i}, \quad i = 1, \ldots, h$ for different time horizons:

RMSE relative mean square error $= \frac{MSE}{MSE_\mu}$,

RMdAE relative median error of prediction $= \frac{MdAE}{MdAE_\mu}$.

Here I compare models to the naive forecast, i.e. $\widehat{x}_{t,h} = \mu$, where μ is the mean of the training sample.

One should expect better forecast for medium and larger forecast horizons with ARFIMA models.

I consider the following models

naive model $x_t = \mu$

ARMA(p,q) $x_t = \frac{\Theta(L)}{\Phi(L)}\varepsilon_t$

ARIMA(p,1,q) $(1 - L)x_t = \frac{\Theta(L)}{\Phi(L)}\varepsilon_t$

ARFIMA(p,d,q) $(1-L)^d x_t = \frac{\Theta(L)}{\Phi(L)}\varepsilon_t \quad d \in \mathbb{R}$

I select the *best* ARIMA models by minimizing the AIC value. Doing this, I get a short memory ARIMA(5,0,1) and an integrated ARIMA(5,1,2) model for the training data.

Table 7.4 shows the out-of-sample accuracy measures for four different models. Using the mean squared error as an accuracy measure we see that the ARIMA model performs better than ARFIMA, and better than ARMA in the beginning, but as the forecast horizon increases ARIMA and ARMA change positions and the ARFIMA model is again in the middle of the two models.

As mean square procedures (also classic OLS) are non-robust to outliers and this series actually exhibits some outliers ($r_t \approx 0$), the $RMSE$ might not be a reasonable accuracy measure. Therefore, I also consider the $RMdAE$, which is more robust to outliers than the RMSE.

Here we can see that the long memory model is superior to ARMA and ARIMA models especially for the mid term horizon.

Horizon h	1	2	4	10	25	50	100	150	200
	MSE								
mean	2.16	2.82	2.46	2.69	3.31	2.68	3.88	3.88	5.05
	RMSE								
ARMA	1.83	0.93	0.72	0.62	0.75	0.87	0.97	0.98	0.99
ARIMA	2.40	1.03	0.76	0.54	0.65	0.94	1.22	1.26	1.23
ARFIMA	2.61	1.12	0.82	0.62	0.69	0.89	1.07	1.08	1.06
	MdAE								
mean	1.47	1.67	1.67	1.68	1.47	1.27	1.46	1.36	1.34
	RMdAE								
ARMA	1.35	0.93	0.72	0.69	0.76	0.88	0.89	0.95	0.94
ARIMA	1.55	0.92	0.55	0.57	0.64	0.81	0.82	0.93	0.89
ARFIMA	1.62	0.96	0.57	0.55	0.62	0.77	0.82	0.89	0.90

Table 7.4: DJI: Forecast accuracy measured by RMdAE and RMSE

7.3 Weekly Stock returns

As a stock index can be seen as a sum of different stocks, the long memory in the data is probably a result of aggregation. Thus, I analyze weekly stock log returns of Alcoa Inc., for the same period as the Dow Jones Index, i.e. from July 15th, 1964 to September 5th, 2006 (see Figure 7.6).

As the data from Yahoo were not corrected for splits I corrected the data to current prices, i.e. for *obvious* stock splits (drop in price of $\approx 50\%$) I multiplied the data before this point by a factor of 0.5. Starting from the end, I proceeded

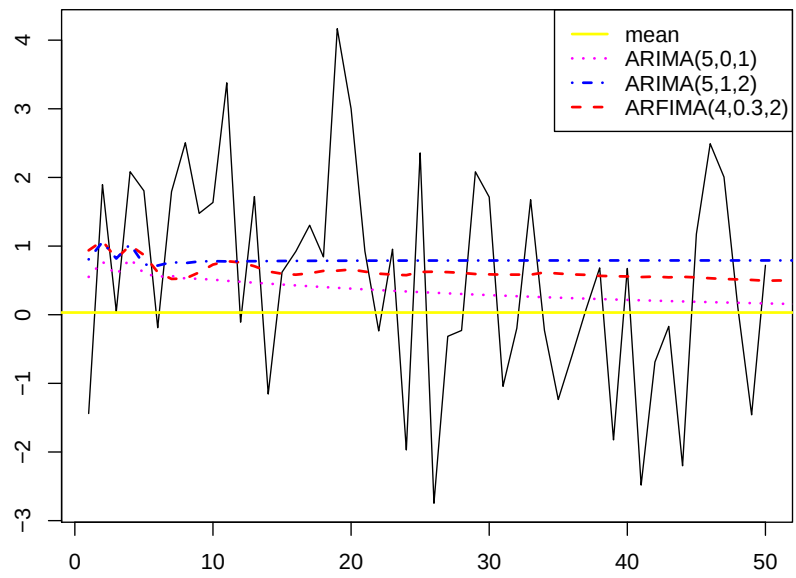
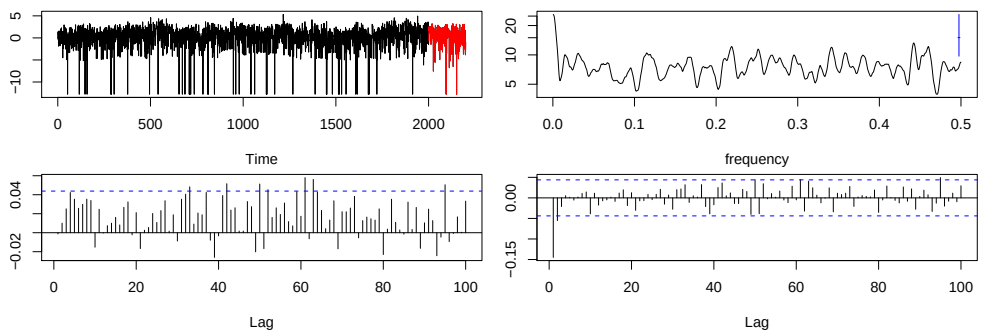


Figure 7.5: DJI: Comparison of forecasts for a time horizon of 50 periods



(a) Alcoa: (top) $x_t = \log(\mu_0 + r_t^2)$: log-arithm of the squared weekly log-returns; x_t; (bottom) autocorrelation function of (bottom) autocorrelation function of the training set (black line)

(b) Alcoa: (top) log-spectrum $f_x(\lambda)$ of the $u_t := (1-L)^{-0.15}$

Figure 7.6: Time and frequency analysis for transformed Alcoa log-returns from 1964 – 2006

until no obvious breaks were present (in total 6 stock splits in this period).

Considering Figure 7.6 we can spot three characteristics:

- There is no apparent global trend in the data; thus, unit root tests do not need a trend component in their null hypothesis.
- A serious issue for the robustness of the model are 49 returns equal to 0. Adding $\mu_0 = 10^{-9}$ to the squared returns, unfortunately does not straighten out this problem.
- A first look at the autocorrelations indicate white noise, but the spectrum

H_0	H_1	Test	Lags / q	Statistic	p - value
$I(0)$	$I(1)$	KPSS	10	0.352	0.10
$I(1)$	$I(0)$	Phillips - Perron	25	-45.51	0.01
$I(1)$	$I(0)$	Augmented DF	12	-11	0.01
$I(1)$	bounded process	Bounded R/S	4	0.40	0.01

Table 7.5: Unit Root tests performed on weekly DJI log–returns.

shows quite clearly that white noise is *not* a proper model for this time series. Also notice that about 80% of the autocorrelations are positive.

7.3.1 Unit root

The data show similar properties as the DJI time series and unit root tests can reject $I(1)$ rigorously. But the KPSS test can reject short memory $I(0)$ only on the 10% level. Thus, we will presumably get a lower $\widehat{d}$ than for the DJI data, or even an $I(0)$ process.

7.3.2 Long memory

Semiparametric Estimation

And indeed, the values in Table 7.6 indicate a lower memory parameter. I use $\widehat{d} = 0.15$ as a starting point for further analysis. The autocorrelation function of $u_t = (1-L)^{-0.15}x_t$ suggests a $MA(2)$ process for the differenced series (see Figure 7.6(b)). To include possible autoregressive behavior in the market, the FIML estimator is based on the $ARFIMA(5, d, 2)$ model.

$m = T^\alpha$	0.4	0.45	0.5	0.55	0.6	0.65	0.7
GPH	0.40	0.39	0.24	0.15	0.12	0.14	0.10
ELW	0.59	0.50	0.24	0.23	0.15	0.16	0.16

Table 7.6: Alcoa: semiparametric estimators for different values of m

FIML

After several estimations with restricted parameters and different lag lengths and comparing AIC values, the FIML estimation procedure gives the final model

$$(1-L)^{0.17}(1-0.35L)x_t = (1-0.52L)\varepsilon_t, \quad \widehat{\sigma_\varepsilon^2} = 7.47, \tag{7.3.1}$$

with standard errors $(0.02, 0.11, 0.10)$ for (d, ϕ_1, θ_1).

7.3.3 Model comparison

Again I compare the forecasting accuracy for different time horizons. An $IMA(1,1)$ seems to be the best model for stochastic volatility except for very large forecast horizons.

Horizon h	1	2	4	10	25	50	100	150	200
	MSE								
mean	0.36	3.38	3.93	3.22	3.62	4.48	5.96	5.70	6.27
	RMSE								
ARMA(2,1)	3.34	0.82	0.72	0.64	0.71	0.89	0.96	0.97	0.99
ARIMA(0,1,1)	4.67	0.75	0.59	0.48	0.55	0.86	0.97	1.02	1.05
ARFIMA(1, 0.17, 1)	3.91	0.82	0.73	0.68	0.75	0.90	0.96	0.98	0.99
	MdAE								
mean	0.60	1.56	2.07	1.64	1.66	1.60	1.69	1.62	1.65
	RMdAE								
ARMA(2,1)	1.83	1.02	0.79	0.74	0.85	0.79	0.89	0.90	0.92
ARIMA(0,1,1)	2.16	1.00	0.75	0.69	0.80	0.77	0.83	0.83	0.86
ARFIMA(1, 0.17, 1)	1.98	1.03	0.78	0.76	0.89	0.83	0.91	0.89	0.91

Table 7.7: Alcoa: Forecast accuracy measured by RMdAE and RMSE

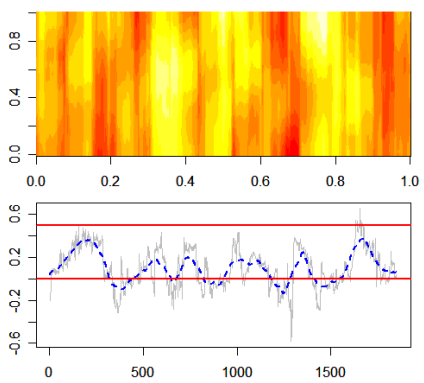


Figure 7.7: Alcoa: (top) Time – varying spectrum with window length = 150 ($\approx$ 3 years) and step size 1; (bottom) Time varying memory parameter d_t and smoothed version .

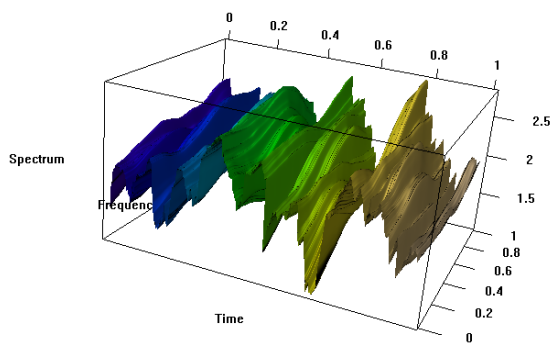


Figure 7.8: Alcoa: Time varying spectrum with window length = 150 ($\approx$ 3 years) and step size 1

H_0	H_1	Test	Lags / q	Statistic	p - value
$I(0)$	$I(1)$	KPSS	31	1.55	0.10
$I(1)$	$I(0)$	Phillips - Perron	25	-46.91	0.01
$I(1)$	$I(0)$	Augmented DF	12	-9.72	0.01
$I(1)$	bounded process	Bounded R/S	4	0.09	0.01

Table 7.8: EUR / USD : Unit Root tests performed on absolute daily exchange rates

m	η_μ				$\widehat{d}_{ELW}$	Z_t
$T = 2000$	q_{opt}=1	q=2	q=4	q=8		
$T^{0.40} \approx 21$	0.03	0.04	0.05	0.07	0.29	-1.30
$T^{0.45} \approx 31$	0.05	0.06	0.08	0.10	0.24	-0.89
$T^{0.50} \approx 45$	0.03	0.04	0.05	0.07	0.29	-1.28
$T^{0.55} \approx 66$	0.03	0.04	0.05	0.07	0.28	-1.26
$T^{0.60} \approx 95$	0.04	0.05	0.06	0.08	0.26	-1.07

Table 7.9: η_μ and Z_t for absolute daily EUR - USD returns

7.4 EUR / USD daily exchange rate

Here I consider the dailz returns of the EUR / USD exchange rate between March 25th, 1999 and September 14th, 2007 (2200 observations)[2]. Missing values have been replaced with the last available observation such that the time series possesses equidistant observations. Again I split up the sample in a training and test set with 2000 and 200 observations, respectively.

As the problem of outliers in this (daily) time series is even more severe than in the previous ones, I analyze absolute returns $x_t = |r_t|$.

7.4.1 Unit root and structural breaks

Both, $I(0)$ and $I(1)$, can be rejected on the 1% level from the four tests.

I also apply the modified KPSS and PP test, for different values of η_μ and for different values of q and m, but not one single constellation can reject constant long memory in the data in favor of structural breaks.

7.4.2 Long memory

Using a starting value of $\widehat{d} = 0.35$ we get a process with significantly autocorrelations for the first 3 lags (see Figure 7.9(a)) and lag 6; thus, I start the FIML estimation with an underlying $ARFIMA(6, d, 3)$ structure (see Table 7.11).

In this case the FIML procedure faces the problem already noted above. The numerical optimization procedure can not evaluate the Hessian matrix. Since

[2]Source http://sdw.ecb.int, [21].

	0.4	0.45	0.5	0.55	0.6	0.65
GPH	0.48	0.32	0.41	0.43	0.33	0.21
ELW	0.29	0.24	0.29	0.28	0.26	0.17

Table 7.10: EUR / USD: semiparametric estimators for different values of m

	Estimate	Std. Error	z value	Pr(> |z|)
d	0.381	0.000	Inf	0.000
ar1	0.690	0.000	2278.080	0.000
ar2	−1.015	0.017	−59.551	0.000
ar3	0.407	0.000	Inf	0.000
ar4	0.109	0.043	2.556	0.011
ar5	−0.009	0.013	−0.729	0.466
ar6	0.028	0.001	53.538	0.000
ma1	1.084	0.000	Inf	0.000
ma2	−1.187	0.038	−31.357	0.000
ma3	0.803	0.025	32.330	0.000

Table 7.11: EUR / USD: FIML estimates with an underlying $ARFIMA(6, d, 3)$ model

the `fracdiff` package does not support fixed parameters in the FIML estimation, I apply the filter $(1 - L)^{0.38}$ to x_t and analyze the – presumably short memory – residual series u_t.

Figure 7.9(b) displays the reciprocal values of the AR and MA roots. The close MA and AR roots on the right half plane make clear why the FIML procedure has problems evaluating the Hessian.

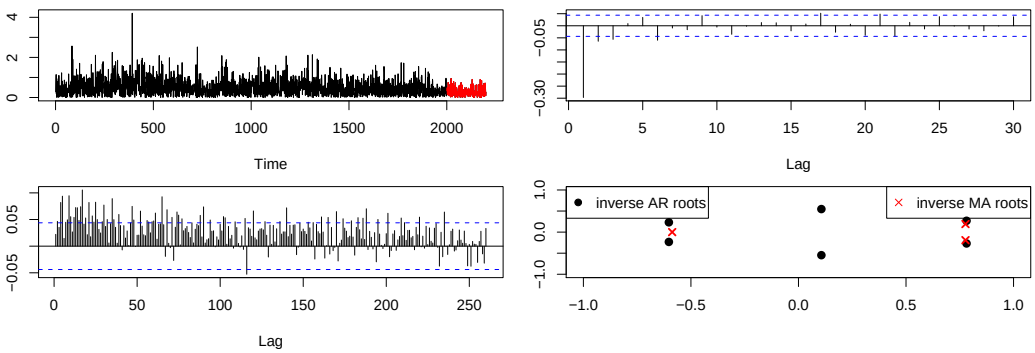
In the $ARMA(6, 3)$ model for u_t, only ϕ_6 and θ_1 are significant at the 5% level. Consequently, I estimate restricted models and finally get an $ARMA(6, 1)$ model for u_t and consequently

$$(1 - L)^{0.38}(1 - 0.30L + 0.06L^6)x_t = (1 - 0.70L)\varepsilon_t, \quad \widehat{\sigma_\varepsilon^2} = 0.16, \tag{7.4.1}$$

with standard errors $(NA, 0.04, 0.02, 0.03)$ for $(d, \phi_1, \phi_6, \theta_1)$.

7.4.3 Time - variation

For the absolute exchange rate returns the time variation approach shows its capacities, as this time series would be a textbook example for the reflections about bounded variation of d_t made in Section 6.1. Here I estimate the time-varying memory parameter d_t and a time-varying spectrum with window length of 250 days ($\approx$ 1 year) and step size 5 days. The blue dashed line in Figure 7.10 is the Nadaraya Watson kernel estimator with bandwidth 13 ($13 \cdot 5 = 13$



(a) (top) $x_t := |r_t|$: daily EUR / USD squared returns from 1999 - 2007; (bottom) autocorrelation function $\rho_x(k), k \neq 0$

(b) (top) $u_t = (1-L)^{-0.38}x_t$: autocorrelation function of the FIML estimated short memory $\rho_u(k), k \neq 0$; (bottom) inverse complex roots of the $AR(6)$ and $MA(3)$ polynomial

Figure 7.9: EUR / USD in the time domain

Horizon h	1	2	4	10	25	50	100	150	200
	MSE								
mean	0.02	0.01	0.03	0.05	0.08	0.08	0.09	0.09	0.09
	RMSE								
ARMA(7,1)	0.08	1.17	0.53	0.78	0.79	0.74	0.72	0.74	0.78
ARIMA(1,1,1)	9.86	14.3	6.94	4.49	3.45	3.59	3.71	3.80	3.74
ARFIMA(6,0.38,1)	0.07	1.06	0.55	0.79	0.81	0.77	0.74	0.75	0.78
	MdAE								
mean	0.15	0.08	0.10	0.21	0.27	0.26	0.27	0.27	0.27
	RMdAE								
ARMA(7,1)	0.28	1.28	1.13	0.90	0.75	0.79	0.76	0.81	0.83
ARIMA(1,1,1)	3.14	5.08	4.12	2.06	1.90	2.12	2.11	2.11	2.12
ARFIMA(6,0.38,1)	0.27	1.22	0.92	0.90	0.84	0.85	0.83	0.84	0.85

Table 7.12: EUR / USD: Forecast accuracy measured by RMdAE and RMSE

weeks $\widehat{=}$ 3 months), and normal density kernels. These are scaled such that their quartiles (viewed as probability densities) are at $\pm 0.25 \cdot$ bandwidth.

Figure 7.11 shows that the spectral density is substantially higher in the beginning than at the end. This can also be seen by almost white (high values) versus red (low values) areas in the heat map.

It is not unreasonable that the variance decay also has an impact on the memory parameter since the variance is a function of d (besides other parameters). So, ceteris paribus $d_0 < d_1$ implies $\mathbb{V}\left(I(d_0)\right) < \mathbb{V}\left(I(d_1)\right)$. In practice this relation must be seen the other way around; thus, a decaying variance gives decaying memory parameters d, all other equal. As already stated above, d_t

is not any random function of t but can be well approximated with a periodic function with cycle length of about 120 weeks.

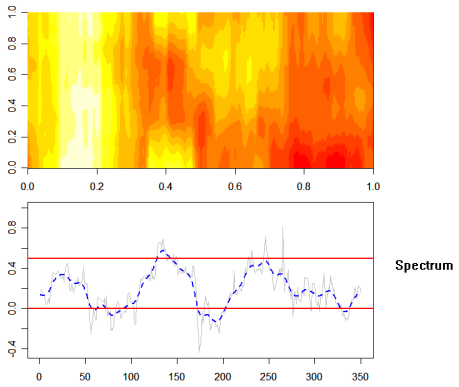


Figure 7.10: EUR / USD: (top) Time – varying spectrum with window length = 250 (≈ 1 year) and step size 5; (bottom) Time varying memory parameter d_t and smoothed estimates

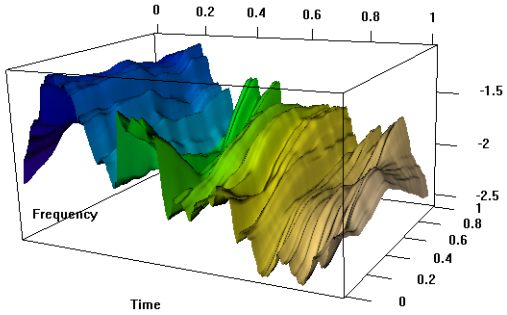


Figure 7.11: EUR / USD: Time varying spectrum with window length = 250 (≈ 1 year) and step size 5 (1 week)

7.5 Usefulness of long memory modeling

After all, a model should capture the structure of a process, its reaction to shocks, and provide a reasonable understanding of the process. In Section 2.1 I laid out different ideas why a model that allows long term dependence might be useful for analyzing real world data.

In fact the comparison of Alcoa and the Dow Jones Index is a good example for the role of aggregation with respect to long term dependence, as the DJI has a substantially higher memory parameter than Alcoa.

Also noticeable is that an $I(d)$ model is not only theoretically in between $I(0)$ in $I(1)$, but also the forecast accuracy of ARFIMA models is (with few exceptions) just between ARMA and ARIMA models. In the beginning integrated models perform better as they principally use the last observation as there forecast. Thus, if the actual series is far off the mean of the process, ARIMA models perform better as they forecast the last observed level. As the horizon increases, the process reverts to the mean and ARMA and ARFIMA models gain accuracy for large time horizons.

Although for these time series an ARFIMA approach is not the *best* model to accurately forecast the out-of-sample series, it must be noted that the partly better performing ARIMA models possess infinite variance, which – looking at

the time varying approach – does not seem to be a suitable property of the data.

Besides the time-varying approach also delivers quite interesting results to the question of finite or infinite variance in financial data.

All three time series are driven by forces that somehow do not *allow* infinite variance in the process and thus every time the (theoretical) variance approaches the threshold $d = 0.5$, the market counteracts the transmission to infinite variance and the memory parameter decreases again below 0.5.

Overall the *best* model choice to use for an analysis of a process depends on the forecast horizon and the importance of accurately describing the second moments of the process. Especially interesting are ARFIMA models with $0.5 \leq d < 1$ as they are mean reverting, but possess infinite variance. Neither ARMA nor ARIMA models can accomplish this combination. But in this analysis not one of the series exhibits this behavior.

Chapter 8

Conclusion and Outlook

The proposed models for time-varying memory certainly need further theoretical and empirical investigation. Disregarding the competing theories of structural breaks and long memory the proposed concept is a powerful tool to analyze processes that locally change between stationarity and non-stationarity.

The concept of time-varying memory has a variety of theoretical implications, mainly dealing with structural breaks in and changing variance of the process. Even though the error duration approach shows that time-varying memory does not necessarily imply a heteroskedastic variance, I am convinced that time-varying memory models might have similar fields of applications as GARCH models.

Elaborating the theory of time-varying memory, to get a well defined concept for stochastic processes, remains a task for future work.

List of Figures

List of Tables

List of Data Sources

1. Yahoo Finance
   http://finance.yahoo.com
2. European Central Bank - Statistical Data Warehouse: EUR / USD exchange rate data
   http://sdw.ecb.int

Appendix A

Theorems and Proofs

The appendix gives an overview about certain subjects and notation needed in the thesis. As most of the presented theory can be found in any basic text book on the subject, the standard theorems are stated without proof or reference.

A.1 Probability theory

Definition A.1.1 (Probability Space)**.** *A probability space* $(\Omega, \mathcal{A}, P)$ *is a measure space with a measure* P *that satisfies the probability axioms:*

i) $P(E) \geq 0 \qquad \forall E \in \Omega.$

ii) $P(\Omega) = 1$

iii) Any countable sequence of pairwise disjoint events $E1, E2, \ldots$ *satisfies* $P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i)$.

Definition A.1.2 (Orthogonal increments)**.** *A stochastic process* $(z(\lambda)|\lambda \in [-\pi, \pi])$ *with random variables* $z(\lambda) : \Omega \to \mathbb{C}^n$ *is called a* process of orthogonal increments *if*

i) $z(-\pi) = 0$ *almost everywhere and* $z(\pi) = x_0$ *a.e.*

ii) $\lim_{\epsilon \to 0} z(\lambda + \epsilon) = z(\lambda)$ *for* $\lambda \in [-\pi, \pi)$ *(right continuity).*

iii) $\mathbb{E} z(\lambda)^* z(\lambda) < \infty \quad \forall \lambda [-\pi, \pi]$

iv) $\mathbb{E}(z(\lambda_4) - z(\lambda_3))(z(\lambda_2) - z(\lambda_1))^* = 0$ *for all* $\lambda_1 < \lambda_2 \leq \lambda_3 < \lambda_4$.

Definition A.1.3 (Stochastic Integral)**.** *For a given deterministic function* $g : [-\pi, \pi] \to \mathbb{C}$ *and a partition* $-\pi = \lambda_1^n < \lambda_2^n < \ldots < \lambda_n^n = \pi$ *of the interval* $[-\pi, \pi]$*, define the finite sum*

$$I_n(g) = \sum_{i=0}^{n-1} g(\lambda_i^n)(z(\lambda_{i+1}^n) - z(\lambda_i^n)).$$

If for all sequences of partitions with $\max_i(\lambda_{i+1}^n - \lambda_i^n) \to 0$ *the limit for* $n \to \infty$ *in mean square sense of* $I_n(g)$ *exists and is the same for every partition, then*

$$I(g) = \int_{-\pi}^{\pi} g(\lambda)\,\mathrm{d}z(\lambda) := l.i.m_{n\to\infty} I_n(g)$$

is the stochastic integral of g with respect to $z(\lambda)$.

Using linearity and continuity of the expectation and properties of $z(\lambda)$, it can be shown that the stochastic integral features similar properties as the Riemann integral (for proofs see Brockwell and Davis [8]):

- interchangeability of expectation and integration: $\mathbb{E}\, I(g) = \mathbb{E} \int_{-\pi}^{\pi} g(\lambda)\,\mathrm{d}z(\lambda) = \int_{-\pi}^{\pi} g(\lambda)\,\mathrm{d}\mathbb{E}z(\lambda)$.
- $\mathbb{E}\, I(g)I(h)^* = \int_{-\pi}^{\pi} g(\lambda)h(\lambda)\,\mathrm{d}F(\lambda)$ where $F(\lambda) = \mathbb{E}z(\lambda)z(\lambda)^*$.

Definition A.1.4 (Brownian motion)**.** *Brownian motion is a Gaussian process* $B_d(\lambda)$ *with stationary increments and variance* $\mathbb{E}B_d^2(\lambda) = k\lambda^{2d+1}$*, where* k *is a positive constant.*

Brownian motion is self-similar with parameter d, i.e. $c^{-d}B_d(c\lambda)$ has the same distribution as $B_d(\lambda)$ for every c. For standard Brownian motion $d = \frac{1}{2}$. Lo [47] defines a Brownian bridge B_d^0 as

$$B_d^0(\lambda) := B_d(\lambda) - \lambda B_d(1).$$

Lemma A.1.5. *For a nonnegative random variable* X *that takes only integer values it holds* $\mathbb{E}\, X = \sum_{n=1}^{\infty} P(X \geq n)$.

Proof. We have

$$\mathbb{E}\, X = \sum_{k=1}^{\infty} kP(X = k) = \sum_{k=1}^{\infty}\sum_{n=1}^{k} P(X = k) = \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} P(X = k) = \sum_{n=1}^{\infty} P(X \geq n),$$

after interchanging order of summation. □

A.2 Hilbert spaces

Definition A.2.1 (Inner Product)**.** *A mapping* $< \cdot, \cdot >: \mathbb{H} \times \mathbb{H} \to \mathbb{C}$ *where* $\mathbb{H}$ *is a linear space is an* inner product *if*

i) $< a_1x_1 + a_2x_2, y >= a_1 < x_1, y > +a_2 < x_2, y >$ *for all* $a_1, a_2 \in \mathbb{C}$ *and* $x_1, x_2, y \in \mathbb{H}$

ii) $< x, y >= \overline{< x, y >}$

iii) $< x, x >\geq 0$ *and* $< x, x >= 0 \Leftrightarrow x \equiv 0$.

The inner product defines in a natural way a norm on $\mathbb{H}$, given by $\|x\| = \sqrt{< x, x >}$

Definition A.2.2 (Hilbert Space)**.** *A set* $\mathbb{H}$ *is a* Hilbert space *if*

i) $\mathbb{H}$ *is a linear space,*

ii) with an inner product,

iii) which is complete in the norm defined by the inner product.

Two elements f and g of $\mathbb{H}$ are *orthogonal*, i.e. $f \perp g$ iff $< f, g >= 0$.

Example A.2.3 ($\mathbb{L}^2$)**.** *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space and define*

$$\mathbb{L}^2 := \left\{ x \in (\Omega, \mathcal{A}, P) | x \text{ is a complex-valued random variable with } \mathbb{E}|x|^2 < \infty \right\}.$$

If we define an inner product by $< f, g >:= \mathbb{E} f \overline{g}$*, then* $\mathbb{L}^2$ *is a Hilbert space on* $\mathbb{R}$.[1]

The norm of $x \in \mathbb{L}^2$ *is equal to the noncentral variance* $\mathbb{E}\, x \overline{x} < \infty$*. And the distance between* x *and* y *is defined as*

$$||x - y|| = \sqrt{\mathbb{E}(x - y)\, \overline{(x - y)}}.$$

Therefore, convergence in this space is mean-square convergence.

An infinite series $\sum_{j=0}^{\infty} x_j$ converges to $x \in \mathbb{H}$ iff the partial sums $s_t = \sum_{j=0}^{t} x_j$ converge to x_t,

$$\lim_{t \to \infty} \|x - s_t\| = 0.$$

Definition A.2.4 (Convergence in r-th mean)**.** *The sequence of random variables* $(x_k | k \in \mathbb{N})$ *converges to* x_0 in the r-th mean *or* in the $\mathbb{L}_r$ sense *if*

$$\mathbb{E}|x_0|^r < \infty, \quad r \geq 1$$

and

$$\lim_{k \to \infty} \mathbb{E}|x_k - x_0|^r = 0.$$

Although r can be any integer, $r = 1$ and $r = 2$ are the most commonly used convergence concepts.

For $r = 1$ we say that x_k converges in the mean to x_0, and for $r = 2$ we say that x_k converges to x_0 in the mean square sense and denote this convergence by

$$l.i.m_{k \to \infty}\, x_k = x_0.$$

In the following I present some useful results for Hilbert spaces.

[1] As already mentioned in the text, the random variable x is a representantive element of the equivalence class of all random variables that equal x, except on a set with Lebesgue measure zero.

Lemma A.2.5 (Cauchy-Schwarz Inequality)**.** *For all $x, y \in \mathbb{H}$ it holds $< x, y > \leq \|x\|\|y\|$. Equality only if $x = a \cdot y$ for some $a \in \mathbb{C}$ or $y = 0$.*

Corollary A.2.6. *Without loss of generality assume $\mathbb{E}x_t = 0$, then for $x_t, x_{t-s} \in \mathbb{L}^2$ with inner product $< x_t, x_{t-s} > := \mathbb{E}x_t x_{t-s}$ the Cauchy Schwarz Inequality becomes $\gamma(s) = \mathbb{E}x_t x_{t-s} \leq \sqrt{\mathbb{V}x_t}\sqrt{\mathbb{V}x_{t-s}} = \sigma_x^2 = \gamma(0)$. Therefore, if $\mathbb{V}x_t < \infty$, then $\gamma(s) < \infty$, which guarantees the existence of the series $\sum_{j=-\infty}^{\infty} b_j b_{j-s}$ for a stationary process.*

Lemma A.2.7 (Continuity of the Inner product)**.** *If $x_n \to x$ and $y_n \to y$ in $\mathbb{H}$, then $< x_n, y_n > \to < x, y >$.*

A.2.1 Fourier series

Definition A.2.8. *If a sequence $\{\mathbf{e}_t\}_{t=1}^{\infty}$ in a Hilbert space satisfies $< \mathbf{e}_t, \mathbf{e}_s > = 0$ for $s \neq t$ and $\|\mathbf{e}_t\| = 1$, then $\{\mathbf{e}_t\}_{t=1}^{\infty}$ is called* orthonormal.

Lemma A.2.9. *If $f, g \in \mathbb{L}^2$, then $fg \in \mathbb{L}^1$.*

Proof.

$$\begin{aligned} 0 &\leq (f \pm g)^2 = f^2 \pm 2fg + g^2 \\ &\Rightarrow \pm 2fg \leq f^2 + g^2 \\ &\Leftrightarrow 2|fg| \leq |f|^2 + |g|^2. \end{aligned}$$

As the left and right hand side are greater or equal to zero, it must also hold

$$\|fg\|_1 = \int |fg| \, \mathrm{d}\mu \leq \int |f|^2 \, \mathrm{d}\mu + \int |g|^2 \, \mathrm{d}\mu = \|f\|_2 + \|g\|_2 < \infty.$$

□

A.3 Analysis

Definition A.3.1 (Landau symbols)**.** *A function $f(t) = \mathcal{O}(g(t))$ iff*

$$\lim_{t \to \infty} \frac{f(t)}{g(t)} = c \notin \{0, \infty\}.$$

A function $f(t) = o(g(t))$ iff

$$\lim_{t \to \infty} \frac{f(t)}{g(t)} = 0.$$

Theorem A.3.2. *For a well behaved, smooth function $f(x)$ on a compact interval K, it holds*

$$f(x) = \sum_{j=0}^{\infty} a_j (x - x_0)^j \text{ with } a_j = \frac{f^{(j)}(x_0)}{j!} \tag{A.3.1}$$

where $f^{(j)}(x_0)$ is the j-th derivative of $f(x)$, evaluated at x_0.

Appendix B

Code

B.1 R and packages

The following packages in R 2.5.0[1] have been used throughout the analysis:

- zoo
- uroot
- urca
- rgl
- fracdiff
- longmemo
- xtable
- forecast
- its

B.2 Algorithms

Here I present the essential procedures I use throughout the thesis; any improvements are welcome. For use in research please cite this thesis as the source for the obtained programs.

B.2.1 Estimation

Exact Local Whittle estimation

[1] available at http://www.r-project.org and [67].

```
################################
# ELW ... Exact Local Whittle Estimation
################################ Input
# wts ... one dimensional time series
# m ... number of frequencies included;
either natural number or as exponent for T^m
# plot ... should the result be plotted
# d.start ... starting value for optimization; default = GPH
################################ Output
# d ... ELW estimate of the memory parameter
################################
require(fracdiff)

ELW=function(wts, m=ceiling(length(wts)^0.5),
     plot=FALSE, d.start=c("automatic")){
T=length(wts)
if (m <1) m = ceiling(T^m)
glob.exp=log(m)/log(T)

if (d.start == "automatic") d.start=fdGPH(wts,bandw.exp=glob.exp)$d

R_d=function(d){
T=length(wts)
if (m <1) m = ceiling(T^m)
x=diffseries(wts, d=d)
I_diff.x=spectrum(x, plot=F)$spec[1:m]
G=mean(I_diff.x)
a=mean(log(1:m))
Q=log(G)-2*d*(log(2*pi/T)+a)
return(Q)
}

est=optim(d.start,R_d, method = c("BFGS"))
d.hat=est[1]$par
######################### Plot
if (plot) {
l=25
R.d=rep(0,l)
d=seq(d.hat-0.25*d.hat, d.hat+0.25*d.hat,length=l)
for (i in 1:l)
{
R.d[i]=R_d(d[i])
}
plot(d,R.d)
```

```
}
##########################
return(d=d.hat)
}
```

B.2.2 Coefficient conversion

I(d) to MA conversion

```
############################
# FRACtoMA ... Fractional noise to MA(\infty) representation
############################ Input
# d ... memory parameter
# q.max ... truncation lag for MA coefficients
########################### Output
# coeff.ma ... MA coefficients
############################
FRACtoMA=function(d, q.max) {
coeff.ma=NULL
coeff.ma[1]=d
for (k in 2:q.max) {
coeff.ma[k]=(k-1+d)/(k)*coeff.ma[k-1]
}
return(coeff.ma)
}
```

ARFIMA(p,d,q) to MA conversion

```
##############################
# ARFIMAtoMA ... ARFIMA(p,d,q) to MA(\infty) representation
############################## Input
# d ... memory parameter
# ar ... AR part of the model; as vector
# ma ... MA part of the model; as vector
# lag.max ... maximum lag number to match terms for z-transform
# plot ... plot the Cauchy product results
############################## Output
# c ... MA(\infty) coefficients of an ARFIMA(p,d,q) model
##############################
source("FRACtoMA.r")
source("Cauchy.product.r")

ARFIMAtoMA=function(d, ar=0, ma=0, lag.max=250, plot=TRUE) {
a=c(1, ARMAtoMA(ar=ar, ma=ma, lag.max))
```

```
eta=c(1, FRACtoMA(d, lag.max))
c=Cauchy.product(a,b=eta, plot)
return(c)
}
```

B.2.3 Prediction

Implemented *matching terms* for forecasting from the infinite past (see equation (4.2.5)).

Predict ARFIMA processes

```
##################################
# predict.ARFIMA ... predict ARFIMA processes from infinite past
################################## Input
# wts.train ... trainings data / original data
# d ... memory parameter
# ar ... ar coeffs in the form c(1,-ar) x_t = eps_t
# ma ... ma coeffs in the form x_t = c(1,ma) eps_t
# h.step ... forecast horizon
# tol ... cut-off tolerance for coefficients different to 0
# lag.max ... max number of lags used for the infinite forecast coeffs
################################## Output
# wts.pred ... predicted values of wts.train
##################################
predict.ARFIMA=function(wts.train,d=0,ar=0, ma=0,
     h.step, tol=10^(-4), lag.max=200) {

phi=ar
theta=ma

k=ARFIMAtoMA(d=d,ar=phi, ma=theta, lag.max, plot=FALSE)
r=ARFIMAtoMA(d=-d, ma=-phi, ar=-theta, lag.max, plot=FALSE)

k.trunc=k[1:(which(abs(k)<tol)[1]-1)]
r.trunc=r[1:(which(abs(r)<tol)[1]-1)]

A=matrix(0, nrow=h.step, ncol=length(k.trunc)+length(r.trunc)-1)

wts.pred=rep(0,h.step)
for (h in 1:h.step) {
A[h,]=c(Cauchy.product(k.trunc[(h+1):length(k.trunc)],
r.trunc,plot=FALSE),rep(0,h-1))
}
```

```
wts.pred=A%*%as.matrix(rev(wts.train[(length(wts.train)-ncol(A)+1):
             length(wts.train)]))
return(wts.pred)
}
```

B.2.4 Simulation

Fractional noise – I(d)

Although the `fracdiff` package provides a simulation procedure for ARFIMA(p,d,q) processes, d is limited to the interval $[-\frac{1}{2}, \frac{1}{2})$. By cumulating and differencing the series appropriately `arfima_sim` expands the feasible values to any $d \in \mathbb{R}$.

```
################################
# I_d_sim ... simulate fractional noise I(d); d arbitrary
################################ Input
# T ... sample size
# d ... memory parameter
# innov ... innovations
################################ Output
# x ... simulated I(d) process
################################

I_d_sim=function(T, d, innov=rnorm(T)) {
d.bk=d
data=fracdiff.sim(n=T, d=d-ceiling(d), innov=innov)$series
int.order=floor(d)

if (d>=0) {
while (ceiling(d)>0) {
data=cumsum(data)
d=d-1
}
d=d.bk
}

if (d<0) {
while (ceiling(d)<0) {
data=diff(data)
d=d+1
}
d=d.bk
}
```

```
x=ts(data)
return(x)
}
```

ARFIMA(p,d,q)

```
##############################
# arfima_sim ... simulate ARFIMA processes for arbitrary d in R
############################## Input
# T ... sample size
# d ... memory parameter
# ar ... optional AR parameters
# ma ... optional MA parameters
# innov ... innovations
############################## Output
# arfima_d ... simulated ARFIMA process
##############################

source("I_d_sim.r")
arfima_sim=function(T, d, ar=c(0), ma=c(0), innov=rnorm(T)) {
v=I_d_sim(T=T,d=d, innov=innov) # simulate I(d) process with arbitrary d
if (ar ==0 && ma ==0) arfima_d=v
if (ar!=0) arfima_d=arima.sim(n=T, innov=v,list(ar = ar))
if (ma!=0) arfima_d=arima.sim(n=T, innov=v,list(ma = ma))
if (ar!=0 && ma!=0) arfima_d=arima.sim(n=T, innov=v,list(ar = ar, ma=ma))
return(arfima_d)
}
```

(Time – varying) Error duration simulation

```
###########################
# ED_sim_vector ... Error duration simulation; time variation allowed
###########################
# T ... sample length
# p.exp ... amnesia parameter (exponents) OR defined probabilities
# sd ... error standard deviation
# plot ... plot different characteristic features
# seeds ... specific set.seed()
# variation ... time varying ED; default is FALSE
###########################
# A ... matrix with column 1: shocks \varepsilon_s;
        column 2: corresponding durations n_s
# series ... simulated series
```

```
###########################
require(Runuran)

ED_sim_vector=function(T=1000, p.exp=1,nar=0, nma=0, plot=TRUE,
       seeds=0, sd=1, variation=FALSE){

prob=prob.control=NULL
p.i=NULL
K.max=T

g=function(x, a=p.exp) {
return(a*1/(1+x)^(a+1))
}

dpmf <- new("unuran.discr",pmf=g,lb=1,ub=K.max)
unr <- unuran.new(dpmf, "dgt")
n_s <- unuran.sample(unr, 2*T)

########################### dynamic memory
if (variation != FALSE) {
if (T!=length(p.exp)) stop("the length of time varying exponents
   must equal the length of the simulated process.")
p.exp=c(rep(p.exp[1], T), p.exp)

n_s=NULL
for (t in 1:length(p.exp)) {

g=function(x, a=p.exp[t]) {
return(a*1/(1+x)^(a+1))
}

dpmf <- new("unuran.discr",pmf=g,lb=1,ub=K.max)
unr <- unuran.new(dpmf, "dgt")
n_s[t] <- unuran.sample(unr, 1)
if (p.exp[t] < 0.01) n_s[t]=K.max
}
}

############################## simulate ED
if (seeds!=0) {set.seeds=seeds}
eps=rnorm(length(n_s), 0, sd)
A=matrix(0, ncol=2, nrow=length(eps))
A[,1]=eps
A[,2]=n_s
```

```
x.sim=rep(0,length(eps))
for (j in 1:length(eps)) {
ind=j:(j+A[j,2])
x.sim[ind]=x.sim[ind]+eps[j]
}
x.sim=na.omit(x.sim)
x.sim.trim=x.sim[(T+1):(2*T)]

############## PLOTS

if (plot==TRUE) {
par(mfcol=c(3,2))
plot(x.sim.trim, type="l", xlab="Time", ylab="", main="x_t")
plot((x.sim.trim-mean(x.sim.trim))^2, type="l",
xlab="Time", ylab="", main="x_t")
plot(n_s[(T+1):(2*T)], main="", ylab="duration length",
xlab="t", type="h", col=c(rep(1, K.max),rep(2,K.max)))
title(paste("Shock duration at time t"))
abline(K.max,0)
acf(x.sim.trim,T^0.8, main="Series x_t")
spectrum(x.sim.trim,sqrt(T),sqrt(T))
spectrum(diff(x.sim.trim),sqrt(T),sqrt(T))

par(mfrow=c(1,1))
}
return(list(A=A, series=x.sim.trim))
}
```

Bibliography

[1] Arteche, J. (2002, April). Gaussian Semiparametric Estimation in Long Memory in Stochastic Volatility and Signal Plus Noise Models. Technical report, Universidad del Pais Vasco - Departamento de Economía Aplicada III (Econometría y Estadística). available at http://ideas.repec.org/p/ehu/biltok/200202.html.

[2] Basak, G., N. Chan, and W. Palma (2001). The approximation of long-memory processes by an ARMA model. *Journal of Forecasting 20*(6), 367 – 389.

[3] Baum, C. F., J. T. Barkoulas, and M. Caglayan (1999, November). Long memory or structural breaks: can either explain nonstationary real exchange rates under the current float? *Journal of International Financial Markets, Institutions and Money 9*(4), 359 – 376. available at http://ideas.repec.org/a/eee/intfin/v9y1999i4p359-376.html.

[4] Beran, J. (1994). *Statistics for Long-Memory Processes.* Chapman and Hall/CRC.

[5] Beran, J. (1996). Testing for change of the long – memory parameter. *Biometrika 83(3)*, 627 – 638.

[6] Bhardwaj, G. and N. Swanson (2004, September). An Empirical Investigation of the Usefulness of ARFIMA Models for Predicting Macroeconomic and Financial Time Series. Departmental Working Papers 200422, Rutgers University, Department of Economics. available at http://ideas.repec.org/p/rut/rutres/200422.html.

[7] Breidt, F. J., N. Crato, and P. de Lima (1998). The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics 83*(1-2), 325 – 348. available at http://ideas.repec.org/a/eee/econom/v83y1998i1-2p325-348.html.

[8] Brockwell, P. J. and R. A. Davis (1991). *Time series: Theory and methods* (2nd ed.). Springer Series in Statistics.

[9] Cajueiro, D. O. and B. M. Tabak (2004). Testing for Fractional Dynamics in the Brazilian Term Structure of Interest Rates. Technical report, Banco Central do Brazil.

[10] Cavaliere, G. (2001). Testing the unit root hypothesis using generalized range statistics. *The Econometrics Journal 4*(1), 70–88.

[11] Clements, M. P. and H. Krolzig (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal 1*, C47 – C75. available at http://ideas.repec.org/a/ect/emjrnl/v1y1998iconferenceissuepc47-c75.html.

[12] Dahlhaus, R. (2000). A Likelihood Approximation for Locally Stationary Processes. *The Annals of Statistics 28*(6), 1762 – 1794.

[13] Davies, R. and D. Harte (1987). Tests for Hurst Effect. *Biometrika 74*, 95 – 101.

[14] Davis, R. A., T. Lee, and G. A. Rodriguez-Yam (2006). Structural breaks estimation for non-stationary time series signals. *Journal of the American Statistical Association 101*(473), 223 – 239.

[15] Deo, R. and C. Hurvich (2000). Estimation of Long Memory in Volatility. Technical report, Stern School of Business, New York University.

[16] Deo, R., C. Hurvich, and Y. Lu (2006, January). Forecasting Realized Volatility Using a Long Memory Stochastic Volatility Model: Estimation, Prediction and Seasonal Adjustment. *Journal of Econometrics 127*(0501002), 29 – 58. available at http://ideas.repec.org/p/wpa/wuwpem/0501002.html.

[17] Dickey, D. and W. Fuller (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association 74*(366), 427 – 431.

[18] Diebold, F. X. and A. Inoue (2001, 11). Long Memory and Structural Change. *Journal of Econometrics 105*, 131 – 159.

[19] Domingo, C. and G. Tonella (2000). Towards a theory of structural change. *Structural Change and Economic Dynamics 11*(1-2), 209 – 225.

[20] Engle, R. and A. D. Smith (1999). Stochastic Permanent Breaks. *The Review of Economics and Statistics 81*(4), 553 – 574.

[21] European Central Bank (2006). Statistical Data Warehouse: EUR/CHF exchange rate data. Available at: http://sdw.ecb.int.

[22] Geweke, J. and S. Porter-Hudak (1983). The Estimation and Application of Long Memory Time Series Models. *Journal of Time Series Analysis 4*, 221 – 238.

[23] Giraitis, L. and R. Leipus (1995). A generalized fractionally differencing approach in long-memory modeling. *Lithuanian Mathematical Journal 36*(1), 53 – 65.

[24] Görg, G. M. and D. Draghicescu (2007). Nonparametric modeling of the second order structure of processes with time-varying memory. In *2007 JSM Proceedings, Government Statistics, Social Statistics, and Survey Research Methods Sections*, Alexandria, VA. American Statistical Association.

[25] Gradshteyn, I. S. and I. M. Ryzhik (2007). *Tables of Integrals, Series, and Products* (7th ed.). Academic Press.

[26] Granger, C. and F. Marmol (1997, November). The Correlogram of a Long Memory Process Plus a Simple Noise. ftp://weber.ucsd.edu/pub/econlib/dpapers/ucsd9729.pdf. Discussion paper.

[27] Granger, C. and D. Zhuanxin (1996). Varieties of long memory models. *Journal of Econometrics 73*(1), 61–77.

[28] Granger, C. W. J. (1980, October). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14*(2), 227 – 238. available at http://ideas.repec.org/a/eee/econom/v14y1980i2p227-238.html.

[29] Granger, C. W. J. and N. Hyung (1999, jun). Occasional Structural Breaks and Long Memory. http://citeseer.ist.psu.edu/granger99occasional.html.

[30] Granger, C. W. J. and R. Joyeux (2001). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis 1*, 15 – 30.

[31] Gray, H. L., N. Zhang, and W. A. Woodward (1989). On generalized fractional processes. *Journal of Time Series Analysis 10*(3), 233 – 257.

[32] Gray, H. L., N. Zhang, and W. A. Woodward (1994). On generalized fractional processes. A correction. *Journal of Time Series Analysis 15*(5), 561 – 562.

[33] Greene, W. H. (2002). *Econometric Analysis*. Prentice Hall.

[34] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

[35] Harvey, A. C. (1998). *Long memory in stochastic volatility*, Chapter 12, pp. 307 – 320. Butterworth-Heineman.

[36] Hosking, J. R. (1981). Fractional Differencing. *Biometrika 68*, 165 – 176.

[37] Hosking, J. R. (1984). Modeling Persistence in Hydrological Time Series Using Fractional Differencing. *Water Resources Research 20*(12), 1898 – 1908.

[38] Hsieh, M., C. Hurvich, and P. Soulier (2007, may). Asymptotics for Duration-Driven Long Range Dependent Processes. Technical report, New York University, Universite Parix X.

[39] Hurst, H. E. (1950). Long-term Storage Capacity of Reservoirs. *Transaction of the American Society of Civil Engineers 116*, 770 – 808.

[40] Hurvich, C. and W. Chen (2000). An Efficient Taper for Potentially Overdifferenced Long-memory Time Series. *Journal of Time Series Analysis 21*(2), 155–180.

[41] Hurvich, C., R. Deo, and J. Brodsky (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis 19*(1), 19–46.

[42] Jensen, M. J. and B. Whitcher (2000). Time-Varying Long-Memory in Volatility: Detection and Estimation with Wavelets. available at http://citeseer.ist.psu.edu/jensen00timevarying.html.

[43] Krämer, W. and P. Sibbertsen (2002). Testing for structural change in the presence of long memory. *International Journal of Business 1*(3), 235 – 242.

[44] Künsch, H. R. (1987). Statistical properties of self-similar processes.

[45] Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics 54*(1-3), 159 – 178.

[46] Ljung, G. M. and G. Box (1978). On a measure of lack of fit in time series models. *Biometrika 65*(2), 297 – 303.

[47] Lo, A. (1991). Long memory in stock market prices. *Econometrica 59*(5), 1279 – 1313.

[48] Mandelbrot, B. (1972). Statistical methodology for non-periodic cycles: from the covariance to R/S analysis. *Annals of Economic and Social Measurement 1*(3), 259 – 290.

[49] Martens, M., D. van Dijk, and M. de Pooter (2004, June). Modeling and Forecasting S&P 500 Volatility: Long Memory, Structural Breaks and Nonlinearity. Tinbergen Institute Discussion Papers 04-067/4, Tinbergen Institute. available at http://ideas.repec.org/p/dgr/uvatin/20040067.html.

[50] Mayoral, L. (2004). A new minimum distance estimation procedure of ARFIMA processes. Forthcoming in Econometrics Journal.

[51] Mayoral, L. (2005, October). Is the observed persistence spurious? A test for fractional integration versus short memory and structural breaks. Economics Working Papers 956, Department of Economics and Business, Universitat Pompeu Fabra. available at http://ideas.repec.org/p/upf/upfgen/956.html.

[52] Mayoral, L. (2006). Testing for Fractional Integration Versus Short Memory with Trends and Structural Breaks. Working paper.

[53] Melino, A. and S. M. Turnbull (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics 45*(1-2), 239 – 265. available at http://ideas.repec.org/a/eee/econom/v45y1990i1-2p239-265.html.

[54] Morana, C. and A. Beltratti (2004). Structural change and long range dependence in volatility of exchange rates: either, neither or both. *Journal of Empirical Finance 11*(5), 629–658.

[55] Newey, W. and D. McFadden (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics 4*, 2111 – 2245.

[56] Nielsen, M. (2005). Multivariate Lagrange Multiplier Tests for Fractional Integration. *Journal of Financial Econometrics 3*(3), 372 – 398. available at http://ideas.repec.org/p/aah/aarhec/2002-18.html.

[57] Palma, W. (2007). *Long-Memory Time Series: Theory and Methods*. Wiley Series in Probability and Statistics.

[58] Parke, W. R. (1999, November). What is Fractional Integration ? *The Review of Economics and Statistics 81*(4), 632 – 638. available at http://ideas.repec.org/a/tpr/restat/v81y1999i4p632-638.html.

[59] Parzen, E. (1982). Autoregressive Spectral Estimation.

[60] Percival, D. B. and W. Constantine (2002). Exact simulation of time-varying fractionally differenced processes. Technical report, Journal of Computational and Graphical Statistics.

[61] Perron, P. and Z. Qu (2004). An Analytical Evaluation of the Log-Periodogram Estimate in the Presence of Level Shifts and its Implications for Stock Market Volatility. Technical report, Working Paper, Department of Economics, Boston University.

[62] Philips, P. C. (1999, December). Discrete Fourier Transforms of Fractional Processes. Cowles Foundation Discussion Papers 1243, Cowles Foundation, Yale University. available at http://ideas.repec.org/p/cwl/cwldpp/1243.html.

[63] Phillips, P. and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika 75*(2), 335 – 346.

[64] Phillips, P. C. and K. Shimotsu (2004). Local Whittle estimation in non-stationary and unit root cases. *Annals of Statistics 32*(2), 656 – 692.

[65] Phillips, P. C. and K. Shimotsu (2005). Exact local Whittle estimation of fractional integration. *Annals of Statistics 33*(4), 1890 – 1933.

[66] Prudnikov, A., O. Marichev, and I. Brychkov (1992). *Integrals and Series.* Gordon and Breach Science Publishers.

[67] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

[68] Rinne, H. and K. Specht (2002). *Zeitreihen - Statistische Modellierung, Schätzung und Prognose.* Verlag Franz Vahlen Muenchen.

[69] Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics 23*(5), 1630 – 1661.

[70] Robinson, P. M. (1995b). Log-periodogram regression of time series with long range dependence. *The Annals of Statistics 23*(3), 1048 – 1072.

[71] Robinson, P. M. (2003). *Time series with long memory.* Oxford.

[72] Saric, B. (2000, May). Expansion of real valued meromorphic functions into Fourier trigonometric series. Provided by the Smithsonian/NASA Astrophysics Data System.

[73] Shimotsu, K. (2006, December). Simple (but effective) tests of long memory versus structural breaks. Working Papers 1101, Queen's University, Department of Economics. available at http://ideas.repec.org/p/qed/wpaper/1101.html.

[74] Sibbertsen, P. (2001). Long Memory versus Structural Breaks: An overview. *Statistical Papers 45*, 465 – 515.

[75] Sibbertsen, P. (2004, 09). Long memory in volatilities of German stock returns. *Empirical Economics 29*(3), 477 – 488. available at http://ideas.repec.org/a/spr/empeco/v29y2004i3p477-488.html.

[76] Sowell, F. B. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics 53*(1-3), 165 – 188. available at http://ideas.repec.org/a/eee/econom/v53y1992i1-3p165-188.html.

[77] Stoev, S., M. S. Taqqu, C. Park, G. Michailidis, and J. S. Marron (2006). LASS: a tool for the local analysis of self-similarity. *Computational Statistics and Data Analysis 50*, 2447 – 2471.

[78] Taqqu, M. and V. Teverovsky (1996). Semi-Parametric Graphical Estimation Techniques for Long-Memory Data. *Lecture Notes in Statistics 115*, 420 – 432.

[79] Teverovsky, V., M. Taqqu, and W. Willinger (1999). A critical look at Lo's modified R/S statistic. *Journal of Statistical Planning and Inference 80*(1), 211 – 227.

[80] Tiao, G. C. and R. S. Tsay (1994). Some Advances in Nonlinear and Adaptive Modelling in Time Series. *Journal of Forecasting 13*, 109 – 131.

[81] Yoshida, K. (1980). *Functional Analysis*. Springer.

[82] Yu, D., P. Zhou, and S. Zhou (2007). On L1 Convergence of Fourier Series Under MV BV Condition. eprint arXiv: 0704.1865.

[83] Zygmund, A. (2003). *Trigonometric Series*. Cambridge University Press.