Financial time series

Gerencsér László Vágó Zsuzsanna Gerencsér Balázs

ISBN 978-963-308-161-7

Contents

Preface

Over the last few decades financial mathematics has become an area that attracted mathematicians, economists, econometricians, physicists, psychologists and many more. The main reason for this is the emergence of new technical ideas that may help people to understand the delicate nature of risk more fully, and to to find ways to reduce it.

Understanding and reducing risk has been a major motivation for such classical studies as the Markowitz-model in portfolio theory that captures the trade-off between returns and risk, see Markowitz [41]. Another classic example is the the Capital Asset Pricing Model (CAPM) of Sharp [50], that quantifies the relationship between the return and the risk of a financial instrument under certain ideal market conditions.

A major breakthrough in finance was the emergence of so-called derivative instruments that are defined in terms of fundamentals, such as the future price of wheat, or the future exchange rate between the US Dollar and the Euros. Writing a contract for having the option to buy say 10,000 US Dollar at a fixed exchange rate one year from now is an excellent mean of reducing the risk of the buyer. Trading in derivatives today makes up a significant fraction of the overall trade on stock exchanges, with options on foreign currencies making up as much as 90% plus of all the trade. Pricing of derivative instruments, such as buy options, has become the prime area of research in financial mathematics, prompted by the seminal papers of Black and Scholes [6] and Merton [42].

Modelling risk calls for modelling the dynamics of financial data, such as returns of share prices, foreign exchange rates and stock indices etc. A variety of models have been proposed, starting with the simplest classic model of Louis Bachelier, the founding father of modern financial mathematics. He thought that price movements over any fixed equidistant subdivision of time are independent and identically distributed (i.i.d.). This hypothesis lead him to model the price process by a so-called Wiener-process (w_t) with drift:

$$S_t = bt + w_t.$$

A Wiener-process (w_t) is a mathematical model of diffusion, with w_t denoting the random position of a particle at time t. Its main characteristics are that the increments $w_{t'} - w_t,\ t < t'$ are stochastically independent of the prehistory of w prior to time t, moreover these increments have Gaussian distribution with 0 mean and variance $t' - t$.

Unfortunately, this model would imply that at some time we may have negative

prices. A better model has been proposed by Paul Samuelson in which the assumed that the returns, rather that the increments, are considered i.i.d. This lead him to modelling the price of an asset as a geometric Brownian motion with drift:

$$S_t = \exp\{bt + \sigma w_t\},$$

where b is the fixed, guaranteed log-return, while the Brownian motion w_t takes care of uncertainty.

Although widely accepted in option pricing, this model fails to capture some basic features of log-price processes obtained from data collected on financial markets. These so-called stylized facts include among others the phenomenon of so-called *v*olatility clustering, heavy-tailed distributions, skewness of distributions and sudden price movements. Therefore we are going to present a variety of alternative models that are constructed by mathematical speculations so as to have the potential to exhibit some of these stylized facts.

The complexity of financial time series is exhibited on the figures below. On the first figure we present historic data of the prices of an individual stock, namely IBM stock prices in the period of 1991-2011.

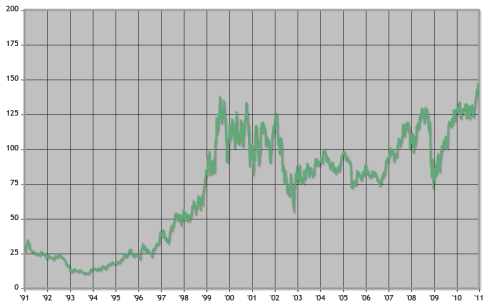

Figure 1: Historic daily closing prices of IBM stocks, 1991-2011

In the second figure we present historic data of the prices of an index, representing the overall dynamics of a collection of stock prices, namely the NASDAQ index values in the period of 1991-2011:

These models are called technical models, as opposed to fundamental models in which the minute details of the market, in particular the behavior of the agents, are described. The advantage of using technical models is that they lead to tractable mathematical problems. In addition, simulation results show, that financial data generated using an assumed micro-structure of the market, can be superbly described, in a statistical sense, by an appropriately fitted technical model.

An early powerful alternative to the random walk model is the so-called *linear model*, in which the dynamics of the market with its millions of small feedback effects is reflected in the fact that we the model has a a non-zero, (but fading) memory. Linear models are

Figure 2: Historic NASDAQ index values 1991-2011

quite acceptable for preprocessed data, having non-zero means, linear trends or periodic cycles removed by appropriate methods. Linear models also have a highly developed and sophisticated theory thanks to their multiple relevance in circuit theory, communication and systems and control theory, with major contribution by the R. Kalman.

Linear models are quite acceptable for preprocessed data, having non-zero means, linear trends or periodic cycles removed by appropriate methods. A neuralgic point in using a linear model is that it is completely specified by second order statistics, and finer properties of the return processes may not be reflected. In particular, within the class of linear models all second order stationary orthogonal processes (often called white noise precesses) are statistically equivalent, or indistinguishable.

To model the phenomenon of volatility clustering we would need a to consider a finer model structure in which today's returns influence tomorrow's conditional volatility. On the other hand we wold like to get a model which is mathematically tractable. A breakthrough in dealing with the above problem was provided by the classical paper of Engle [22], in which the so-called ARCH (autoregressive conditional heteroscedasticity) model was introduced. The proposed application area in that paper was the analysis of macroeconomic data. Standard generalizations of ARCH models are the so-called GARCH (generalized ARCH) models, introduced by Bollerslev [7]. It turned out that GARCH models are excellent candidates for modelling financial data with exhibiting stochastic volatility.

The extraordinary attention paid to these models in the academic community is due to the fact that this a technical model defined via a relatively simple dynamics, yet leading to a variety of interesting theoretical problems, and at the same time it is versatile enough when fitting to real data. Robert Engle is the the winner of the 2003 Nobel Memorial Prize in Economic Sciences, sharing the award with Clive Granger, "for methods of analyzing economic time series with time-varying volatility (ARCH)".

Another important development in modeling financial time is the construction of high frequency (continuous time) model classes allowing to model shocks or jumps in the price process. Thus we come to model classes using so-called Lévy processes. A rough idea of the latter can be given as follows: take the limit of a so-called compound Poisson process

which has a finite number of jumps in a finite interval at times the number of which follows a Poisson distribution. There are a number of models using Lévy processes to model financial time series, such as the Variance Gamma or the CGMY model. The special feature of these models is that the characteristic function rather than the distribution function of the price is known explicitly, possibly modulo a few unknown constants.

In this course we discuss selected topics of the theory of stochastic processes with special attention to areas used in modelling financial time series, as discussed above. The basic question in connection with financial time series is very simple: predict future values of financial instruments as accurately as possible to support a decision to buy or sell. While there are powerful model-free methods for prediction based on the patterns of past ups and downs, prediction theory and practice is still dominated by model-based approaches. In this setting we first try to understand the mechanism by which our data has been generated.

First we need to construct potential classes of models that may be appropriate for modeling. Finally, the properties of these models have to be understood and a theory of prediction has to be developed. Then we have to derive methods to describe real data by a single element of the proposed model class. This last step is called *estimation* in the mathematical statistics literature, while the terminology *identification* is accepted in the engineering literature.

As for the selection of the course material we should note that the development of the theory of stochastic processes was significantly inspired by telecommunication and later by systems and control theory, and the interaction between engineering and finance is not over yet.

The course is suitable for students with a solid basic training in basic probability theory and introductory functional analysis. To assist the learning process many of the smaller mathematical facts are formulated as exercises. Some of these are marked with a * to indicate that their solution may require a bit more than straightforward, one-minute application of known facts.

Moreover, course material is supplemented with a number of sophisticated interactive simulation programs available at

`http://digitus.itk.ppke.hu/~vago/Financial_Time_Series/`

It is hoped that experimenting with these programs will help the student to develop a feeling for the variety of behaviors of data generated by our models.

Chapter 1

Basic concepts

1.1 Wide sense stationary processes

A discrete time stochastic process $y = (y_n)$ is simply a sequence of random variables over a fixed probability space $(\Omega, \mathcal{F}, P)$. The subscript n indicates time, the range of which is assumed to be typically $-\infty < n < +\infty$ for the sake of mathematical convenience. When we speak about a random variable we assume that it is real valued, unless explicitly stated otherwise. Complex valued random variables will indeed play an eminent role in our discussions. Depending on the application area a stochastic process may be called a *time series* (economics) or a *random signal* (telecommunication and control).

A key property of a stochastic process is the dependence structure between the random variables y_n. Dependence is what makes prediction possible. Another key property of a stochastic process is a kind of statistical homogenity in time, which again makes prediction possible.

The simplest measure of dependence is covariance or correlation. Thus in most part of the course we will restrict ourselves to processes such as that

$$\mathrm{E}(y_n^2) < +\infty \qquad \text{for all} \quad n.$$

Equivalently, we have $y_n \in L_2(\Omega, \mathcal{F}, P)$ for all n. Here $L_2(\Omega, \mathcal{F}, P)$ stands for the Hilbert space of equivalence classes of real valued random variables ξ with $\mathrm{E}(\xi^2) < \infty$. In this course we will identify a class of a.e. identical random variables with a representative of its class.

To model statistical homogenity first we will assume that for some m

$$\mathrm{E}y_n = m \qquad \text{for all} \quad n.$$

A good measure of dependence is the covariance

$$\mathrm{Cov}(y_{n+\tau}, y_n).$$

Statistical time-homogenity then would mean that the above covariance is independent of n. Thus we arrive at the following concept:

Definition 1.1. *A real valued stochastic process $y = (y_n)$, $-\infty < n < +\infty$ is called* wide sense stationary*, w.s.st. for short, if $\mathrm{E}(y_n^2) < +\infty$ for all n, and for some m*

$$\mathrm{E}y_n = m, \qquad \text{for all} \quad n,$$

and

$$r(\tau) = r^y(\tau) = \mathrm{Cov}(y_{n+\tau}, y_n)$$

is independent of n. The function $r^y(\tau)$ is called the autocovariance function*.*

Three examples of so-called AR(1) processes (see Chapter 5) are shown below:

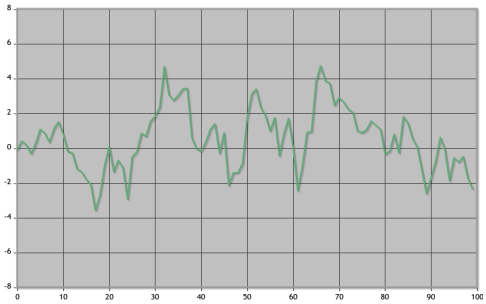

Figure 1.1: AR(1) process with an almost unstable positive pole

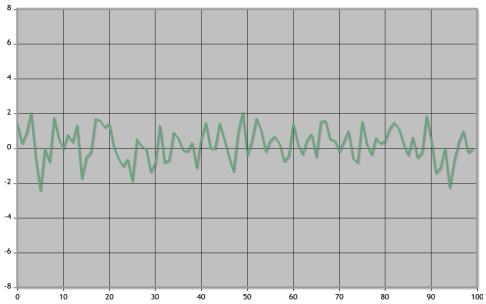

Figure 1.2: AR(1) process with a stable positive pole

An alternative terminology is that $y = (y_n)$ is a *second order stationary process* or *weakly stationary process*. A standard assumption will be in this course that the expectation of y_n is 0 for all n, i.e.

$$\mathrm{E}y_n = 0 \qquad \text{for all} \quad n.$$

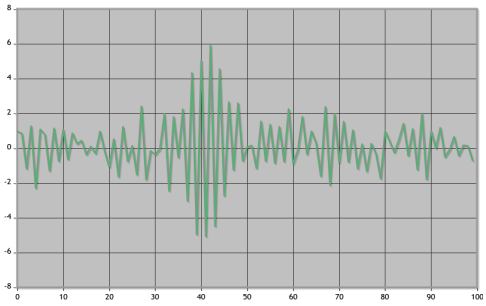

Figure 1.3: AR(1) process with an almost unstable negative pole

Then the autocovariance function is defined as

$$r(\tau) = r^y(\tau) = \mathrm{E}(y_{n+\tau} y_n).$$

Remark. The assumption $\mathrm{E}y_n = 0$ is not restrictive. If we have a general w.s.st. process $y = (y_n)$, then the process $y' = (y'_n)$ defined by $y'_n = y_n - m$ will be a zero mean w.s.st. process.

Note, that the autocovariance function is symmetric, or even:

$$r(\tau) = r(-\tau),$$

and that

$$r(0) = \mathrm{E}(y_n^2) = \sigma^2 = \text{const.}$$

If we allow y_n to be *complex* valued then the above definition, restricted to the case of zero mean processes, is modified by requiring that $\mathrm{E}|y_n|^2 < +\infty$ for all n, and, assuming $m = 0$,

$$\mathrm{E}y_n = 0 \quad \text{for all} \quad n,$$

and

$$r(\tau) = \mathrm{E}(y_{n+\tau} \overline{y}_n)$$

is independent of n, where $\overline{y}$ denotes the complex conjugate. For complex w.s.st. processes we have

$$r(\tau) = r^y(\tau) = \overline{r(-\tau)}.$$

The condition $\mathrm{E}|y_n|^2 < +\infty$ will also be expressed by saying that $y_n \in L_2^c(\Omega, \mathcal{F}, P)$, the Hilbert space of complex valued random variable ξ with $\mathrm{E}|\xi|^2 < \infty$. Here again we will consider identical r.v.-s which are identical a.e.

The autocovariance function of an AR(1)-process is fairly trivial. The closest non-trivial example is the autocovariance function of a so-called AR(2) process, two examples for which are given below:

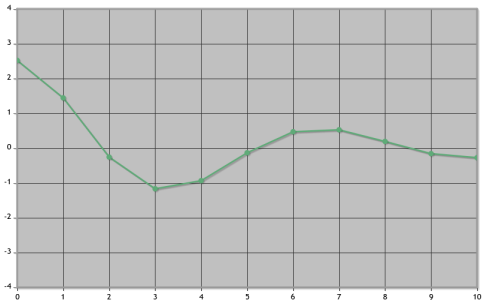

Figure 1.4: Autocovariance of an AR(2) process with a pair of complex poles of length 0.8 and argument $\pm 0.3\pi$.

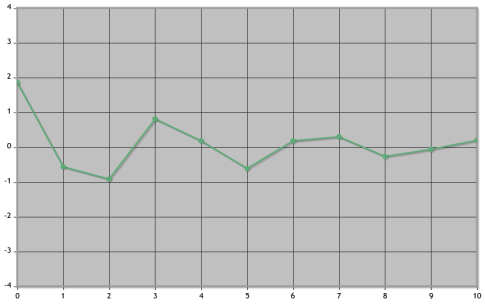

Figure 1.5: Autocovariance of an AR(2) process with a pair of complex poles of length 0.8 and argument $\pm 0.6\pi$.

Exercise 1.1. *Let (y_n) be a wide sense stationary process and let us define*

$$u_n = a_1 y_{n-1} + \cdots + a_p y_{n-p}, \qquad \text{with } a_k \text{ real}, \quad k = 1, \dots p.$$

Show that (u_n) is also a wide sense stationary process.

Let us compute $\mathrm{E}u_n^2$. We have

$$\mathrm{E}u_n^2 = \sum_{k=1}^{p} \sum_{l=1}^{p} a_k a_l r(l-k).$$

Define the $p \times p$ matrix $R = (R_{k,l})$ by

$$R_{k,l} = r(l-k), \qquad k, l = 1, \dots p, \tag{1.1}$$

and the p vector

$$a = (a_1, \dots, a_p)^T.$$

Then we can write

$$\mathrm{E}u_n^2 = a^T R a.$$

The matrix R is obviously symmetric and positive semi-definite. In addition, its elements are constant along any sub-diagonal:

$$R = \begin{pmatrix} r(0) & r(1) & \dots & \dots & r(p-1) \\ r(-1) & r(0) & r(1) & \dots & r(p-2) \\ r(-2) & r(-1) & r(0) & \dots & r(p-3) \\ \vdots & \vdots & \ddots & \ddots & \\ r(-p+1) & r(-p+2) & \dots & \dots & r(0) \end{pmatrix}$$

Such a matrix is called a *Toeplitz matrix*.

Exercise 1.2. *Set*

$$Y = (y_{n-1}, \dots, y_{n-p})^T. \tag{1.2}$$

Prove that the matrix R defined under (1.1) can also be written as $R = \mathrm{E}(YY^T)$.

Exercise 1.3. *Using the representation $R = \mathrm{E}(YY^T)$ prove that R is symmetric and positive semidefinite.*

To sum up our findings we have:

Proposition 1.2. *The matrix R defined under (1.1) is a symmetric, positive semi-definite Toeplitz matrix.*

For complex-valued processes we have a similar result:

Exercise 1.4. *Let $y = (y_n)$ be a complex-valued w.s.st. process with auto-covariance function $r^y(\cdot)$. Show that the matrix $R = (R_{k,l})$ defined by*

$$R_{k,l} = r^y(l-k), \qquad k, l = 1, \dots p$$

is a Hermitian, positive semi-definite Toeplitz matrix.

1.2 Orthogonal processes and their transformations

How do we get a wide sense stationary process? Let us start with the simplest possible w.s.st. process (e_n) called w.s.st. orthogonal process. This is a w.s.st. process characterized by

$$\mathrm{E}e_n e_m = \sigma^2 \delta_{n,m} \qquad \text{for all} \quad n,\ m. \tag{1.3}$$

Below we plot the graph of an orthogonal process:

The last condition can be expressed in the geometry of Hilbert spaces by saying that the random variables e_n and e_m are orthogonal for $n \neq m$. This explains the terminology. In terms of the autocovariance function we may say that

$$r^e(\tau) = 0 \qquad \text{for} \quad \tau \neq 0.$$

To summarize:

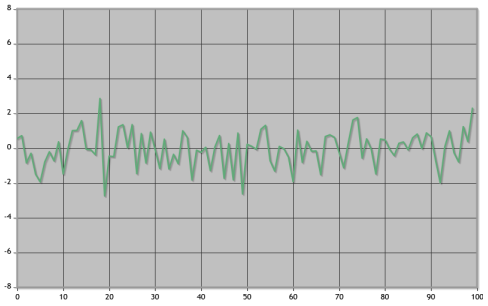

Figure 1.6: The graph of an orthogonal process

Definition 1.3. *A w.s.st. process $e = (e_n)$ satisfying (1.3) is called a* w.s.st. orthogonal process.

An alternative terminology is that $e = (e_n)$ is a white noise process.

As a practical example we mention that an often used assumption in modeling financial time series is that the return processes are identically and independently distributed, in short, they are i.i.d. They can also be assumed to have zero mean after discounting. If, in addition, they have finite second moments, then they form a w.s.st. orthogonal process. Although it should be mentioned, that the price dynamics based on the assumption of i.i.d. returns fails to reproduce some basic features of a real price process. In fact, daily return data from the past, called also historical data, exhibit a small correlation. In addition, the assumption on i.i.d. returns would lead to a price process the variance of which is unbounded, contradicting basic theoretical speculations. This contradiction is particularly obvious with agricultural products, the prices of which are tied to meteorological data, which are inherently bounded.

Now let us consider a practical example from engineering for a mechanism through which more general w.s.st. processes are generated from w.s.st. orthogonal processes. Let us think of (e_n) as the vertical displacement of a road surface properly normalized by its mean, measured at equidistant points not too close to each other. Then we may rightly assume that the e_n-s are uncorrelated. Now if a car rolls along this road then the unevenness of the road surface causes the body of the car (via a damping device) to vibrate. An exact description of this effect is given in the literature on so-called half-car models. Denoting the vertical displacement of the body of the car by (y_n) properly normalized by its mean, the overall picture is that (e_n) is transformed into (y_n). Naive physics suggests that past values of e may effect the present value of y. Assuming that this transformation is linear we arrive at the following representation of y_n:

$$y_n = \sum_{k=0}^{\infty} h_k e_{n-k}. \tag{1.4}$$

Here the h_k-s are the so-called impulse responses of the "system", represented by the body of the car, mapping e to y.

To ensure that (y_n) given in (1.4) is well defined in some sense we resort to the Hilbert space theory. Assuming that

$$\sum_{k=0}^{\infty} h_k^2 < +\infty,$$

the right hand side of (1.4) converges in $L_2(\Omega, \mathcal{F}, P)$. Thus y_n is well defined. To see this, for a fixed n consider the random variables

$$y_{n,N} = \sum_{k=0}^{N} h_k e_{n-k}, \qquad N \in \mathbb{N}.$$

Exercise 1.5. *Show that the sequence $(y_{n,N})$ is a Cauchy sequence in L_2 norm.*

Indeed, for $N < M$ we have by the orthogonality of (e_k)

$$\|y_{n,M} - y_{n,N}\|^2 = \sum_{k=N+1}^{M} h_k^2 < \varepsilon,$$

if N is large enough, since by assumption $(h_k) \in \ell_2$.

In a Hilbert space every Cauchy sequence is convergent, thus y_n is well defined.

Exercise 1.6. *Show that $y = (y_n)$ given in (1.4) is a wide sense stationary process.*

The variance of (y_n) is obtained as

$$\mathrm{E}(y_n^2) = \sum_{k=0}^{\infty} h_k^2 \sigma^2.$$

As we shall see later the class of wide sense stationary processes given by (1.4) is not only an interesting instance, but actually most w.s.st. processes of practical interest fall into this class.

A special case of the above is a wide sense stationary process, which is obtained by taking a finite moving average of an orthogonal process $e = (e_n)$. Thus let $e = (e_n)$ be a wide sense stationary orthogonal process and define

$$y_n = \sum_{k=0}^{r} c_k e_{n-k}. \tag{1.5}$$

Definition 1.4. *A wide sense stationary process defined by (1.5) is called a moving average or MA process.*

If $c_r \neq 0$, then r is called the order of the process. If we wish to emphasize the order, we say that y is an MA(r) process. In engineering terminology we would say that y_n is obtained by passing an orthogonal process through a finite impulse response (FIR) filter.

It is interesting to note that even in the simplest cases of $r = 1$ and $r = 2$ the visual variety of the graphs of MA processes is remarkable. In all the examples below we take $c_0 = 1$. For $r = 1$ and $c_1 < 0$ we take a weighted difference of the white noise process resulting in similar processes:

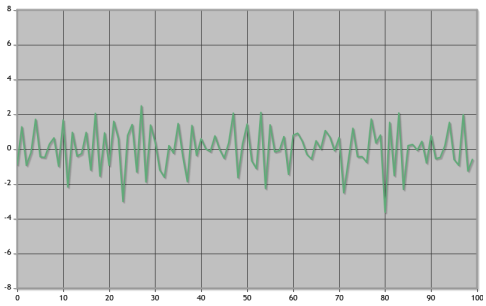

Figure 1.7: MA(1) process with a medium positive zero, $c_0 = 1, c_1 = -0.6$.

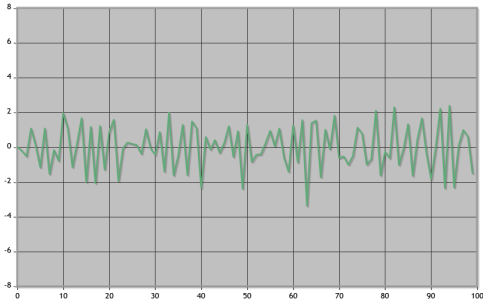

Figure 1.8: MA(1) process with a large positive zero, $c_0 = 1, c_1 = -0.9$.

For $r = 1$ and $c_1 > 0$ we take a weighted sum of the white noise process resulting in a kind of averaging and smoother trajectories:

For $r = 2$ the combination of the above models results in an enhanced effect:

1.3 Prediction

A fundamental problem of the theory of time series is the prediction of future values. In the case of a one-step ahead prediction we would like to predict y_n for some fixed n knowing past values (y_{n-i}) for $i = 1, 2, \ldots$. In other words we assume that the complete, infinite past of y up to time $n - 1$ is known. This assumption is a matter of convenience for certain theoretical arguments.

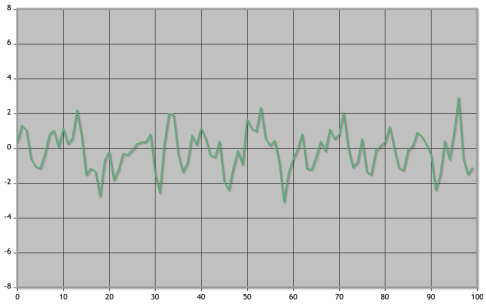

Figure 1.9: MA(1) process with a large positive zero, $c_0 = 1, c_1 = 0.6$.

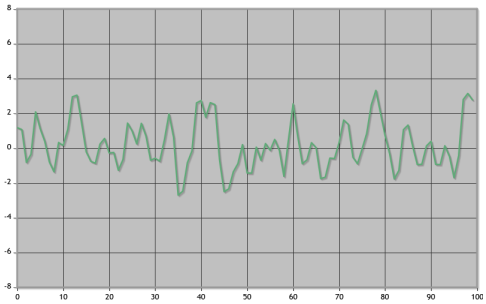

Figure 1.10: MA(1) process with a large positive zero, $c_0 = 1, c_1 = 0.9$.

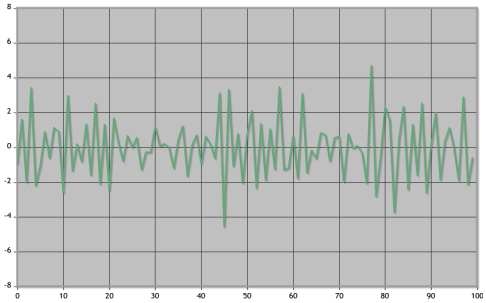

Figure 1.11: MA(2) process with two positive zeros, 0.6 and 0.9.

However in practice we have to work with finite segment of data. Therefore we consider first the following problem: predict y_n based on the finite segment of past data $y_{n-1}, \ldots, y_{n-p}$. We restrict ourselves to linear prediction of the form:

$$\widehat{y}_n = \sum_{k=1}^{p} \alpha_k y_{n-k}, \tag{1.6}$$

with the coefficients $\alpha_1, \ldots, \alpha_p$ still to be specified. The quality of our predictor is

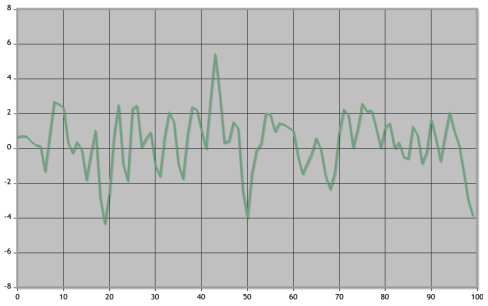

Figure 1.12: MA(2) process with two negative zeros, -0.6 and -0.9.

measured by mean square error (MSE):

$$J(\alpha) = \mathrm{E}(\widehat{y}_n - y_n)^2.$$

Minimizing $J(\alpha)$ with respect to α yields the least squares (LSQ) predictor of y_n within the class of linear predictors.

Obviously, $J(\alpha)$ is quadratic in $\alpha = (\alpha_1, \ldots, \alpha_p)$, hence its minimization is trivial by direct arithmetic. However, more instructive way of solving this minimization problem is to use a geometric approach.

A geometric approach to LSQ will simplify the solution of the prediction problem. Let us consider the linear space spanned by the random variables $(y_n, y_{n-1}, \ldots, y_{n-p})$, i.e. consider all random variables of the form:

$$u = \sum_{k=0}^{p} \alpha_k y_{n-k}.$$

Let us denote this linear space by $H_{n,n-p}$, or, if we want to stress the dependence on y, write $H^y_{n,n-p}$. Since y is a wide sense stationary process and thus $\mathrm{E}y_l^2 < +\infty$ for all l, $H^y_{n,n-p}$ is a finite dimensional subspace of $L_2(\Omega, \mathcal{F}, P)$ equipped with a scalar product:

$$\langle \xi, \eta \rangle = \mathrm{E}\xi\eta.$$

From now on, $H^y_{n,n-p}$ will denote the space above equipped with the scalar product inherited from $L_2(\Omega, \mathcal{F}, P)$. Thus $H^y_{n,n-p}$ is a Euclidean space. The problem of best linear prediction is then equivalent to the following geometric problem: find the orthogonal projection of y_n on the subspace $H^y_{n-1,n-p}$.

We will use the following notations: if H is a Hilbert space and H' is a subspace of H (i.e. a linear subspace of H which is closed), then for $y \in H$ the projection of y onto H' will be denoted by

$$\widehat{y} = (y|H').$$

The projection $\widehat{y}$ is uniquely defined by the following two properties:

$$\widehat{y} \in H' \qquad \text{and} \qquad (y - \widehat{y}) \perp u \qquad \forall u \in H'.$$

The existence and uniqueness of such a projection is a fundamental result of the Hilbert-space theory. In the case of Euclidean spaces we refer to basic linear algebra.

The best linear predictor of y_n in terms of $y_{n-1}, \dots, y_{n-p}$ is then uniquely defined by the orthogonal properties:

$$(y_n - \widehat{y}_n) \perp y_{n-j}, \qquad j = 1, \dots p.$$

This can be written equivalently as

$$\mathrm{E}\widehat{y}_n y_{n-j} = \mathrm{E} y_n y_{n-j}, \qquad j = 1, \dots p.$$

Substituting (1.6) for $\widehat{y}$ and working out the left hand side we get

$$\mathrm{E} \sum_{k=1}^{p} \alpha_k y_{n-k} y_{n-j} = \sum_{k=1}^{p} \alpha_k r(j-k).$$

Introducing the notation

$$r = (r(1), \dots r(p))^T$$

the right hand side becomes simply r. Thus the above equation becomes

$$R\alpha = r,$$

and we arrive at the following result:

Proposition 1.5. *Assume that R is nonsingular. Then the LSQ linear predictor of y_n in terms of $y_{n-1}, \dots, y_{n-p}$ is uniquely defined as*

$$\widehat{y}_n = \sum_{k=1}^{p} \alpha_k y_{n-k},$$

where $\alpha = (\alpha_1, \dots, \alpha_p)$ is the solution of

$$R\alpha = r.$$

We shall discuss prediction in more detail in Chapter 5. Here we present only two graphs of two AR(2) processes and their predictor, marked yellow:

Remark. The coefficients of the best linear predictor are uniquely determined if R is nonsingular, or equivalently, if R is positive definit. What if R is singular? Then we have a non-zero vector v such that

$$v^T R v = 0,$$

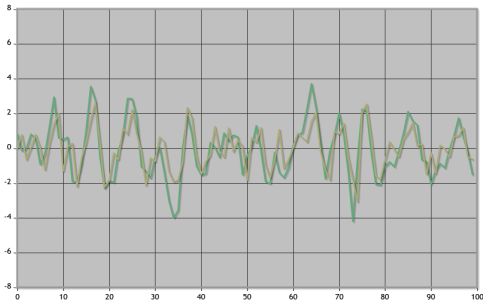

Figure 1.13: Prediction of an AR(2) process with a pair of complex poles of length 0.8 and argument $\pm 0.3\pi$

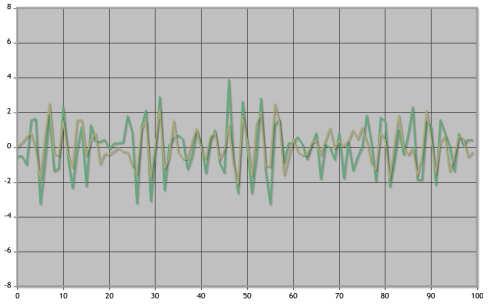

Figure 1.14: Prediction of an AR(2) process with a pair of complex poles of length 0.8 and argument $\pm 0.6\pi$

which is equivalent to writing that

$$\mathrm{E}\left|\sum_{k=1}^{p} v_k y_{n-k}\right|^2 = 0.$$

Exercise 1.7. *Show that if R is singular then y_n can be predicted with 0 error.*

We would not want to consider stochastic processes which can be predicted with 0 error as truly random. We shall therefore single out such processes with the name "singular processes". For details see below.

We conclude this section with two non-trivial exercises:

Exercise 1.8. * *Show that for the process (y_n^r) we have*

$$H_{-\infty}^{y^r} = \{0\}.$$

Exercise 1.9. * *Show that*

$$H_{-\infty}^{y} = H_{-\infty}^{y^s}.$$

Chapter 2

Prediction, innovation and the Wold decomposition

2.1 Prediction using the infinite past

Let us now consider the prediction problem with $p = \infty$. Predicting y_n using infinite past looks impractical at first. However, it provides us with a fundamental insight into the structure of the process, Moreover, as we shall see, it can often be realized by a finite recursion. To formulate the problem, we need a little more care. Define the linear space

$$\mathcal{L}_{n-1}^{y} = \left\{ \sum_{i=1}^{k} \alpha_i y_{n-i} : \text{ with } \alpha_1, \ldots, \alpha_k \text{ real, } k = 1, 2, \ldots \right\}.$$

Obviously

$$\mathcal{L}_{n-1}^{y} \subset L_2(\Omega, \mathcal{F}, P).$$

The closure of $\mathcal{L}_{n-1}^{y}$ in $L_2(\Omega, \mathcal{F}, P)$ is a subspace of $L_2(\Omega, \mathcal{F}, P)$ which will be denoted by H_{n-1}^{y}. Formally we write

$$H_{n-1}^{y} = \mathrm{cl}\mathcal{L}_{n-1}.$$

Thus H_{n-1}^{y} consists of all random variables that can be approximated arbitrarily well in $L_2(\Omega, \mathcal{F}, P)$-sense by linear combinations of the form

$$\sum_{k=1}^{p} \alpha_k y_{n-k}.$$

Then the best linear predictor of y_n in terms of infinite past is defined as

$$\widehat{y}_n = (y_n | H_{n-1}^{y}).$$

A key object in the theory of wide sense stationary processes is the prediction error

$$e_n = y_n - (y_n|H^y_{n-1}).$$

Loosely speaking the random variable e_n expresses the information content of y_n not contained in $(y_{n-1}, y_{n-2}, \dots)$ when we restrict ourselves to linear approximations.

Definition 2.1. *The process*

$$e_n = y_n - (y_n|H^y_{n-1})$$

is called the innovation process *of* (y_n).

Exercise 2.1. *Prove that*

$$(y_n|H^y_{n-1}) = \lim_{p\to\infty}(y_n|H^y_{n-1,n-p}) \qquad \text{in} \quad L_2(\Omega, \mathcal{F}, P).$$

Proposition 2.2. (e_n) *is a wide sense stationary orthogonal process.*

Proof. Obviously

$$y_n - (y_n|H^y_{n-1,n-p})$$

is a wide sense stationary process. (*Why?*) Taking limit we get, that (e_n) is wide sense stationary. The orthogonality of (e_n) is obvious. □

Processes with the property that a finite segment of past values is sufficient to compute the best linear predictor $\widehat{y}_n = (y_n|H^y_{n-1})$ are of particular interest. If y is such a process then we can write

$$y_n = \sum_{k=1}^{p} a_k y_{n-k} + e_n, \tag{2.1}$$

where (e_n) is the innovation process of (y_n).

Definition 2.3. *A wide-sense stationary process* (y_n) *satisfying (2.1) is called an* autoregressive *or* AR process.

If $a_p \neq 0$, then p is called the order of the process. If we wish to emphasize the order, we say that y is an AR(p) process.

2.2 Singular processes

A "*truly random*" process has a non-trivial innovation, i.e. $e_n \neq 0$. Thus $H_n^y \supset H_{n-1}^y$ in a strict sense. A process with zero innovation would thus be an anomaly. For such a process we would have

$$H_n^y = H_{n-1}^y \qquad \text{for all} \quad n.$$

Exercise 2.2. *Show that if $H_n^y = H_{n-1}^y$ for a single n, then $H_n^y = H_{n-1}^y$ for all n.*

Definition 2.4. *A process y is* singular, *if*

$$H_n^y = H_{n-1}^y$$

holds for all n.

It follows that y_n can be arbitrarily accurately approximated by linear combinations of the form $\sum_{k=1}^{p} \alpha_k y_{n-k}$, in the L_2 sense. The question arises: "How we can construct such a process?"

Consider the complex-valued process:

$$y_n = \xi e^{in\omega}, \qquad n = 0, \pm 1, \pm 2, \ldots,$$

where $\omega \neq 0$ is a fixed frequency, ξ is an eventually complex-valued random variable with

$$\mathrm{E}\xi = 0, \qquad \mathrm{E}|\xi|^2 = \sigma^2 < +\infty.$$

Exercise 2.3. *Show that the above process is wide sense stationary:*

$$\mathrm{E}y_n = \mathrm{E}\xi e^{in\omega} = 0,$$

and

$$\mathrm{E}y_{n+\tau}\overline{y_n} = \mathrm{E}\xi e^{i(n+\tau)\omega}\overline{\xi} e^{-in\omega} = \sigma^2 e^{i\tau\omega}$$

is independent of n.

Note that the autocovariance function does not decay in absolute value, as τ increases, indicating a strong dependence between past and present.

The above process is not "truly random" in the sense that two values of y, say y_0 and y_1, uniquely determine the (*random* value of) ξ and ω, and thus the complete future of y is known. Formally, we are tempted to assume that ω is obtained from $e^{i\omega} = y_1/y_0$. But thus ω would be a non-linear(!) function of y. In spite of this, our intuition is right.

Exercise 2.4. *Show that y is singular, i.e.*

$$H_n^y = H_{n-1}^y \qquad \text{for all} \quad n.$$

A simple proof is obtained by applying the result given in the following exercise:

Exercise 2.5. *Let y be a w.s.st. process such that $\dim(H_n^y) < \infty$ for some n. Then y is singular.*

Consider now a finite sum of complex-valued singular processes of the form

$$y_n = \sum_{k=1}^{m} \xi_k e^{in\omega_k} \tag{2.2}$$

where $\omega_k \neq \omega_{k'}$ for $k \neq k'$. Assume that

$$\mathrm{E}\xi_k = 0, \qquad \mathrm{E}|\xi_k|^2 = \sigma_k^2 < +\infty \qquad \text{for all} \quad k.$$

In addition, assume that the random coefficients are mutually orthogonal, i.e.

$$\mathrm{E}\xi_k \bar{\xi}_{k'} = 0 \qquad \text{for} \quad k \neq k'.$$

The last condition is crucial in proving the following result:

Proposition 2.5. *The process defined by (2.2) is a wide sense stationary process.*

Proof. Obviously we have $\mathrm{E}y_n = 0$ for all n. The autocovariance function for y_n

$$\mathrm{E}y_{n+\tau}\bar{y}_n = \sum_{k=1}^{m}\sum_{k'=1}^{m} \mathrm{E}\xi_k\bar{\xi}_{k'} e^{i(n+\tau)\omega_k} e^{-in\omega_{k'}} = \sum_{k=1}^{m} \mathrm{E}|\xi_k|^2 e^{i\tau\omega_k} = \sum_{k=1}^{m} \sigma_k^2 e^{i\tau\omega_k} = r(\tau)$$

is indeed independent of n. □

Exercise 2.6. *Prove that the process y defined above by (2.2) is singular, i.e.*

$$H_n^y = H_{n-1}^y \qquad \text{for all} \quad n.$$

(*Hint*: Apply previous Exercise.)

The variance of y_n is obtained by setting $\tau = 0$:

$$\mathrm{E}|y_n|^2 = \sum_{k=1}^{n} \sigma_k^2$$

We conclude, that the contribution of the frequency ω_k to the variance of y_n is σ_k^2. In telecommunication $\mathrm{E}|y_n|^2$ is the energy of the random signal. Correspondingly, the values σ_k^2 show how the energy of y_n is spread along different frequencies.

This simple and to some extent artificial construction of a wide sense stationary process can be significantly extended. In fact, we will see that by an appropriate extension of the representation given in (2.2) we can recover *any* wide sense stationary process. Needless to say that in such an extension the anomaly of singularity may disappear.

Remark. It would be wrong to believe that all singular processes are of the form given in (2.2). There are examples for singular processes, where singularity can not be established by direct inspection.

A simple example for a *r*eal-valued singular process is given by

$$y_n = \cos(\omega n + \varphi) \qquad \omega \neq 0,$$

where φ is a random phase with uniform distribution on $[0, 2\pi]$.

Exercise 2.7. *Show that (y_n) is a wide sense stationary process.*

Exercise 2.8. *Show that (y_n) is singular. (*Hint*: Apply the identity $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$).*

2.3 Wold decomposition

Let us now consider a process y which is not singular, i.e.

$$H_n^y \supset H_{n-1}^y$$

in a strict sense, or equivalently, its innovation process

$$e_n = y_n - (y_n | H_{n-1}^y)$$

is not zero. We have seen that $e = (e_n)$ is a wide sense stationary orthogonal process. Write

$$y_n = e_n + (y_n | H_{n-1}^y).$$

Now decompose the second term $(y_n | H_{n-1}^y)$ as

$$(y_n | H_{n-1}^y) = v_{n-1} + (y_n | H_{n-2}^y),$$

where $v_{n-1} \perp (y_n | H_{n-2}^y)$. Note that any random variable v such that $v_n \in H_n^y$ and $v_n \perp H_{n-1}^y$ can be written as

$$v_n = c e_n$$

with some real c. (*Why?*). Then we can write

$$v_{n-1} = c_1 e_{n-1}.$$

Continuing this argument we get

$$y_n = \sum_{k=0}^{p} c_k e_{n-k} + (y_n | H^y_{n-p-1}),$$

with $c_0 = 1$. The question now arises, how to deal with the residual term.

Let us define the Hilbert space of distant past, or prehistory of y as

$$H^y_{-\infty} = \bigcap_{m \geq 0} H^y_{-m}.$$

Lemma 2.6. *For any random variable $\xi \in L_2(\Omega, \mathcal{F}, P)$ we have*

$$\lim_{m \to \infty} (\xi | H_{-m}) = (\xi | H_{-\infty}).$$

Exercise 2.9. * *Prove Lemma 2.6*

To prove Lemma 2.6, we first formulate a dual result:

Lemma 2.7. *Let $H_n \subset L_2(\Omega, \mathcal{F}, P)$ be a monotone increasing sequences of Hilbert subspaces, i.e. $H_n \subset H_{n+1}$. Let*

$$H_\infty = \text{cl } (\cup_n H_n),$$

with cl *denoting the closure. Then for any $x \in L_2(\Omega, \mathcal{F}, P)$ we have*

$$\lim_{n \to \infty} (x | H_n) = (x | H_\infty).$$

Exercise 2.10. * *Prove Lemma 2.7.*

Define

$$y^s_n = (y_n | H^y_{-\infty}) \tag{2.3}$$

and

$$y^r_n = \sum_{k=0}^{\infty} c_k e_{n-k}$$

with the limit interpreted in $L_2(\Omega, \mathcal{F}, P)$. Then we arrive at the following decomposition of y:

$$y_n = y^r_n + y^s_n.$$

Exercise 2.11. *Show that the processes (y_n^s) and (y_n^r) are orthogonal, $y^s \perp y^r$, meaning that*

$$y_n^s \perp y_m^r \qquad \text{for all} \quad n,\ m.$$

(*H*int. Note that for any $v \in H_{-\infty}^y$ and any k we have $v \perp e_k$.)

Exercise 2.12. *Show that for the process $y^r = (y_n^r)$ we have*

$$H_{-\infty}^{y^r} = \{0\}.$$

(*Hint*: First show that $H_n^{y^r} \subset H_n^e$ for all n, then show that $H_{-\infty}^e = \{0\}$. The latter follows from the fact that $e_n \perp H_{-\infty}^e$ for all n.)

Definition 2.8. *A process (y_n) is called* completely regular *if*

$$H_{-\infty}^y = \{0\}.$$

Proposition 2.9. *The process (y_n^s) defined under (2.3) is singular and*

$$H_{-\infty}^y = H_{-\infty}^{y^s}.$$

Proof. First we show that the infinite past of the off-spring process y^s is contained in the the infinite past of y, i.e.

$$H_{-\infty}^y \supset H_{-\infty}^{y^s}.$$

To see this, note that the definition $e_n = y_n - (y_n | H_{n-1}^y)$ implies $e_n \in H_n^y$ for all n, which in turn implies $H_n^e \subset H_n^y$ for all n. But then the representation of y_n^r as a linear combination of past values of e implies

$$y_n^r \in H_n^y, \qquad \text{for all} \quad n,$$

and hence for $y_n^s = y_n - y_n^r$ we also have

$$y_n^s \in H_n^y, \qquad \text{for all} \quad n.$$

We conclude from here that

$$H_n^{y^s} \subset H_n^y \qquad \text{for all} \quad n,$$

which implies $H_{-\infty}^{y^s} \subset H_{-\infty}^y$, as stated. □

Now to prove the opposite inclusion, note that the decomposition of y as $y_n = y_n^r + y_n^s$ with $y^r \perp y^s$ implies

$$H_{-n}^y = H_{-n}^{y^r} \oplus H_{-n}^{y^s}$$

for all n, where $\oplus$ denotes orthogonal direct sum. Now formally taking intersection over n, and using the fact that $H^{y^r}_{-\infty} = \{0\}$ would give the result. This formal argument can be elaborated as follows. Let $v \in H^y_{-\infty}$. Then $v \in H^y_{-n}$ for all n. Let us write v as

$$v = v^r_{-n} + v^s_{-n}$$

with $v^r_{-n} \in H^{y^r}_{-n}$ and $v^s_{-n} \in H^{y^s}_{-n}$. Then

$$v^r_{-n} = (v|H^{y^r}_{-n}).$$

Now letting n tend to infinity we get that

$$\lim_{n\to\infty} v^r_{-n} = \lim_{n\to\infty} (v|H^{y^r}_{-n}) = (v|H^{y^r}_{-\infty})$$

by Lemma 2.6. But $H^{y^r}_{-\infty} = \{0\}$, hence $v^r_{-n} \to 0$ and we conclude that

$$v = \lim_{n\to\infty} v^s_{-n}.$$

Obviously, the right hand side belongs to $H^{y^s}_{-m}$ for all m and hence it belongs to $H^{y^s}_{-\infty}$.

Thus we arrive to the following result, which is known as the Wold decomposition of a w.s.st. process:

Proposition 2.10. *Any wide sense stationary process can be decomposed as*

$$y_n = y^r_n + y^s_n,$$

where (y^r_n) is completely regular, (y^s_n) is singular and $y^s \perp y^r$. Moreover

$$H^y_{-\infty} = H^{y^s}_{-\infty}.$$

The singular component of the process contains 'a priori' randomness, or randomness in the distant past or prehistory of y.

To complete the above result, consider now a completely regular process y. By the construction of e we have $H^e_n \subset H^y_n$. On the other hand the representation

$$y_n = \sum_{k=0}^{\infty} c_k e_{n-k}$$

implies $H^y_n \subset H^e_n$. Thus we arrive at the following result.

Proposition 2.11. *Let (e_n) be the innovation process of a completely regular process (y_n). Then*

$$y_n = \sum_{k=0}^{\infty} c_k e_{n-k} \qquad \text{with} \quad c_0 = 1,$$

with $\sum_{k=0}^{\infty} c_k^2 < \infty$ and

$$H^e_n = H^y_n \qquad \text{for all} \quad n.$$

Chapter 3

Spectral theory, I.

3.1 The need for a spectral theory

Let us revisit the problem of prediction. Let (y_n) be a completely regular stochastic process that can be written in the form

$$y_n = \sum_{k=0}^{\infty} h_k e_{n-k},$$

where $h_0 = 1$, and (e_n) is the innovation process of y. Then the LSQ predictor of y_n is given by

$$\widehat{y}_n = \sum_{k=1}^{\infty} h_k e_{n-k}.$$

By this the problem of prediction seems to be solved. But in fact this is not the case: we would like to express $\widehat{y}$ in terms of y rather than in terms of the (unobserved) e.

Let us simplify the problem and assume that the representation of (y_n) is actually an MA(r) process:

$$y_n = \sum_{k=0}^{r} h_k e_{n-k}.$$

A useful tool for future calculation is the backward shift operator acting on doubly infinite sequences as follows:

$$(q^{-1}y)_n = y_{n-1}.$$

Introducing a polynomial of q^{-1} as

$$H(q^{-1}) = \sum_{k=0}^{r} h_k q^{-k},$$

the defining equation for the MA process above can be rewritten as

$$y = H(q^{-1})e. \tag{3.1}$$

Similarly, the process $\widehat{y}$ formed of the one-step ahead predictors $\widehat{y}_n$ can be defined via

$$\widehat{y} = \big(H(q^{-1}) - 1\big)\, e.$$

To express e via y a formal procedure, often adapted in the engineering literature, is to invert (3.1) as

$$e = H^{-1}(q^{-1})y.$$

To see if a meaning can be given to this step it is best to see an example. Let

$$y_n = e_n + ce_{n-1}.$$

Then

$$e_n = -ce_{n-1} + y_n,$$

and iterating this equation we get, assuming $|c| < 1$,

$$e_n = \sum_{k=0}^{\infty} (-c)^k y_{n-k}.$$

Exercise 3.1. *Show that the right hand side above is well defined.*

However, the situation becomes much more complicated for higher order MA models, so the interpretation of the operator $H^{-1}(q^{-1})$ needs extra care.

To find an appropriate interpretation to our formal procedure let us make a brief excursion to the theory of linear time invariant (LTI) systems. A linear time invariant system is defined by

$$y_n = \sum_{k=0}^{\infty} h_k u_{n-k}, \qquad n \geq 0.$$

Here $u = (u_n),\ n \geq 0$ is the input process, $y = (y_n),\ n \geq 0$ is the output process, and the coefficients h_k are called *impulse responses* of the linear system. A standard tool for studying the linear time invariant systems is the so-called *z-transform*. Briefly speaking, consider a linear system such that

$$\sum_{k=0}^{\infty} |h_k| < \infty.$$

Consider a deterministic and bounded input signal $u = (u_n)$. Then, as it is easily seen, the output signal $y = (y_n)$ will also be bounded. Define

$$U(z^{-1}) = \sum_{k=0}^{\infty} u_k z^{-k}, \qquad Y(z^{-1}) = \sum_{k=0}^{\infty} y_k z^{-k} \qquad \text{and} \quad H(z^{-1}) = \sum_{k=0}^{\infty} h_k z^{-k}$$

with $|z| > 1$. Then we have the very simple multiplicative description of our linear time invariant system as follows:

$$Y(z^{-1}) = H(z^{-1})\, U(z^{-1}).$$

To extend this ingenious device to two sided processes we run into the problem of choosing the range for z. Neither $|z| > 1$, nor $|z| < 1$ would do. The only option, with a vague hope of success is to try $|z| = 1$. Thus we are led to the study of a formal object of the form

$$\sum_{n=-\infty}^{\infty} y_n e^{-in\omega}.$$

The ultimate objective of spectral theory of w.s.st. processes is to give a meaning to this formal object.

3.2 Fourier methods for w.s.st. processes

Obviously, the infinite sum above is unlikely to converge in any reasonable sense. To see how a meaning can be given, consider our benchmark example for a *singular* process given as

$$y_n = \sum_{k=1}^{m} \xi_k e^{in\omega_k}. \tag{3.2}$$

A natural first alternative to the infinite sum above is a finite sum appropriately weighted:

$$\xi_N(\omega) = \frac{1}{2N+1} \sum_{n=-N}^{N} y_n e^{-in\omega}. \tag{3.3}$$

Proposition 3.1. *We have*

$$\begin{aligned} \lim_{N\to\infty} \xi_N(\omega_k) &= \xi_k, \\ \lim_{N\to\infty} \xi_N(\omega) &= 0 \qquad \text{for} \quad \omega \neq \omega_k \end{aligned}$$

in the sense of $L_2(\Omega, \mathcal{F}, P)$ *and also w.p.1.*

Exercise 3.2. *Prove Proposition 3.1.*

Corollary 3.2. *The spectral weights* σ_k^2 *can be obtained as*

$$\sigma_k^2 = \mathrm{E}|\xi_k|^2 = \lim_{N\to\infty} \mathrm{E} \left| \frac{1}{2N+1} \sum_{n=-N}^{+N} y_n e^{-in\omega_k} \right|^2.$$

Exercise 3.3. *Prove the above corollary.*

This follows simply from the fact that $\xi_N(\omega_k) \to \xi_k$ in $L_2(\Omega, \mathcal{F}, P)$. Now the question arises, whether the above arguments can be extended to general processes.

Let us consider a *general* wide sense stationary process (y_n). We can ask ourselves: does the finite, normalized Fourier series

$$\frac{1}{2N+1} \sum_{n=-N}^{N} y_n e^{-in\omega} =: \xi_N(\omega)$$

have a limit in any sense? For a start we can ask a simpler question: does the sequence

$$\mathrm{E} \left| \frac{1}{2N+1} \sum_{n=-N}^{N} y_n e^{-in\omega} \right|^2$$

have a limit? Forgetting about the normalization by $1/(2N+1)$ and expanding the above expression we can express the above expectation as

$$\mathrm{E} \sum_{n=-N}^{N} \sum_{n'=-N}^{N} y_n \overline{y}_{n'} e^{-in\omega} e^{in'\omega} = \sum_{\tau=-2N}^{2N} r(\tau)\ e^{-i\omega\tau}(2N+1-|\tau|). \tag{3.4}$$

Indeed, the value of $\mathrm{E}(y_n \overline{y}_{n'})$ depends only on $\tau = n - n'$, and the number of occurrences of $r(\tau)$ is $2N+1-|\tau|$. (To double-check this note that the number of occurrences of $r(0)$ is $2N+1$, while the number of occurrences of $r(\pm 2N)$ is 1.)

Now at this point we shall make the *assumption* that

$$\sum_{\tau=-\infty}^{+\infty} r^2(\tau) < +\infty. \tag{3.5}$$

This assumption would then enable us to apply Fourier theory. (Note, however, that with this assumption our benchmark examples for singular process are excluded!) Now, the right hand side of (3.4) can be related to the partial sums of the Fourier series of $r(\tau)$ as follows. Defining

$$s_n(\omega) = \sum_{\tau=-n}^{+n} r(\tau) e^{-i\omega\tau}$$

we can write the r.h.s. of (3.4) as

$$\sum_{n=0}^{2N} s_n(\omega).$$

Now assumption (3.5) implies that

$$f(\omega) = \sum_{\tau=-\infty}^{+\infty} r(\tau) e^{-i\omega\tau}$$

is well-defined in the sense that the right hand side converges in $L_2[0, 2\pi] = L_2([0, 2\pi], d\omega)$, where $d\omega$ stands for the standard Lebesgue-measure. It follows by the celebrated *Fejér's theorem* that the Fourier series of $f(\omega)$ also converges in the Cesaro sense a.s., i.e.

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} s_n(\omega) = f(\omega) \qquad \text{a.s.}$$

Thus we come to the following conclusion.

Proposition 3.3. *Under condition (3.5)*

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{E} \left| \sum_{n=-N}^{N} e^{-in\omega} y_n \right|^2 = f(\omega)$$

exists a.s. on $[0, 2\pi]$ *w.r.t. the Lebesgue-measure, and we have*

$$f(\omega) = \sum_{\tau=-\infty}^{+\infty} r(\tau) e^{-i\omega\tau}$$

where the r.h.s. converges in $L_2[0, 2\pi]$ *and also in Cesaro sense a.s. with respect to the Lebesgue-measure.*

3.3 Herglotz's theorem

From Proposition 3.3 immediately get a special form of the celebrated Herglotz's theorem:

Proposition 3.4. *Under condition (3.5) we have*

$$r(\tau) = \frac{1}{2\pi} \int_0^{2\pi} e^{i\omega\tau} f(\omega) d\omega$$

with some $f(\omega) \geq 0$, $f(\omega) \in L_2[0, 2\pi)$. *In particular*

$$r(0) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) d\omega.$$

A remarkable fact is that the latter proposition, in a slightly modified form, is true for *any autocovariance sequence*. More exactly we have the following result:

Proposition 3.5. ***(Herglotz's theorem)*** *Let $r(\tau)$ be the autocovariance function of a wide sense stationary process. Then we have*

$$r(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\omega\tau} dF(\omega),$$

where $F(\cdot)$ is a nondecreasing, left-continuous function with finite increment on $[0, 2\pi)$, with $F(0) = 0$. We have

$$r(0) = \frac{1}{2\pi}F(2\pi)$$

in particular.

The function F is called *the spectral distribution function*, and the corresponding measure dF is the spectral measure. The integral above can be interpreted as a Riemann-Stieltjes integral. The distribution function is like a probability distribution function except that we may have $F(2\pi) \neq 1$.
Remark. At this point we need to recall that there is a dichotomy in defining a probability distribution function. If ξ is a random variable then its distribution function may be defined either as $F(x) = P(\xi < x)$ or $F(x) = P(\xi \leq x)$, depending on local traditions. In the former case F is continuous from left, in the latter case F is continuous from right. In this lecture we will assume that F is left-continuous. Thus the dF measure assigned to an interval $[a, b)$ is $F(b) - F(a)$. Let us now see the proof of Herglotz's theorem.

Proof. Let us truncate the autocovariance sequence $r(\tau)$ by setting

$$r_N(\tau) = \begin{cases} r(\tau) & \text{for} \quad |\tau| \leq N \\ \\ 0 & \text{for} \quad |\tau| \geq N+1. \end{cases}$$

Exercise 3.4. *Show that the truncated $r_N(\tau)$ sequence itself is an auto-covariance sequence.*

Obviously, $r_N(\tau)$ is a positive semi-definite sequence. Hence it is an auto-covariance sequence.

Then we have by the special form of Herglotz's theorem that

$$r_N(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{-i\tau\omega}\ f_N(\omega)\ d(\omega)$$

with $f_N(\omega) \geq 0$. Defining $F_N(\omega)$ by

$$F_N(\omega) = \int_0^{\omega} f_N(\lambda)\ d\lambda$$

we have

$$r_N(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{-i\omega\tau} dF_N(\omega), \quad \text{for} \quad |\tau| \leq N.$$

Obviously F_N is nondecreasing for all N. Now we have

$$\frac{1}{2\pi}\int_0^{2\pi} dF_N(\omega) = F_N(2\pi) = r(0) < \infty \quad \forall N.$$

Now we can refer ro Helly's theorem stating that there exists a subsequence N_k and a monotone nondecreasing function $F(x)$ such that

$$\lim_{k\to\infty} F_{N_k}(x) = F(x)$$

at all continuity point of F, and $F(2\pi) = r(0)$. This is also expressed as saying that F_{N_k} converges to F weakly, or formally writing:

$$dF_{N_k} \quad \Rightarrow \quad dF.$$

It then follows by classical results of introductory probability theory that for every bounded continuous function g we have

$$\lim_{n\to\infty} \int_{\mathbb{R}} g \; d\mu_n = \int_{\mathbb{R}} g \; d\mu.$$

For $g(\omega) = e^{in\omega}$ in particular we have for any fixed τ

$$r(\tau) = \lim_{N\to\infty} \frac{1}{2\pi}\int_0^{2\pi} e^{-i\omega\tau} dF_N(\omega) = \frac{1}{2\pi}\int_0^{2\pi} e^{-i\omega\tau} dF(\omega).$$

Q.e.d. □

3.4 Effect of linear filters

To see the power of Herglotz's theorem, we present the following simple result:

Proposition 3.6. *Let (y_n) be a wide sense stationary process. Then for any finite linear combination $\sum_{k=0}^{p} a_k y_{n-k}$ we have*

$$\mathrm{E}\left|\sum_{k=0}^{p} a_k y_{n-k}\right|^2 = \frac{1}{2\pi}\int_0^{2\pi} |A(e^{i\omega})|^2 \; dF(\omega), \tag{3.6}$$

where

$$A(e^{i\omega}) = \sum_{k=0}^{p} a_k e^{-i\omega k}.$$

Exercise 3.5. *Prove the above proposition.*

To extend the above result, the question can be raised: how we can conveniently express the auto-covariance function of the process

$$v_n = \sum_{k=0}^{m} h_k y_{n-k}, \qquad h_k \in \mathbb{R}, \tag{3.7}$$

obtained from y by applying a finite impulse response (FIR) filter. Or in other terms: how do we get the spectral distribution of v from that of y?

Note that we have:

$$r^v(\tau) = \mathrm{E} v_{n+\tau} v_n = \sum_{k=0}^{m}\sum_{l=0}^{m} h_k h_l \; y_{n+\tau-k} y_{n-l} = \sum_{k=0}^{m}\sum_{l=0}^{m} h_k h_l \; r^y(l-k+\tau).$$

Now expressing $r^y(l-k+\tau)$ via Herglotz's theorem we get, after interchanging the sum and the integrand,

$$r^v(\tau) = \frac{1}{2\pi}\int_0^{2\pi} \sum_{k=0}^{m}\sum_{l=0}^{m} h_k h_l \; e^{i\omega(l-k+\tau)} \; dF^y(\omega) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\omega\tau} \left|\sum_{k=0}^{m} h_k e^{i\omega k}\right|^2 \; dF^y(\omega).$$

Introducing

$$H(e^{-i\omega}) = \sum_{k=0}^{m} h_k e^{-i\omega k}$$

we come to the following conclusion:

Proposition 3.7. *For the spectral distribution of the process v we have*

$$dF^v(\omega) = \left|H(e^{-i\omega})\right|^2 \; dF^y(\omega).$$

If y has a spectral density $f^y(\omega)$ then v also has a spectral density and we have

$$f^v(\omega) = \left|H(e^{-i\omega})\right|^2 \; f^y(\omega).$$

The complex valued function $H(e^{-i\omega})$ is called the *transfer function* or *frequency response function* of the FIR filter. If $|H(e^{-i\omega})| > 1$ than the energy contained at frequency ω will be amplified, for $|H(e^{-i\omega})| < 1$ it will be attenuated. For appropriately chosen weights h_k the filter H may be such that $|H(e^{-i\omega})|$ is close to zero except for a small band around a specific frequency ω_0. An ideal case would be when

$$\left|H(e^{-i\omega})\right| = 1 \qquad \text{for} \quad \omega_0 - \delta < \omega < \omega_0 + \delta,$$

and 0 otherwise. Such a filter is called *band-pass filter*. It is readily seen that such a filter can not be represented as an FIR filter. (Why?)

Chapter 4

Spectral theory, II.

4.1 First construction of the spectral representation measure

Let us now return to the Fourier transform of the series of (y_n) itself:

$$\sum_{n=-N}^{N} y_n e^{-in\omega}.$$

At this point assume that

$$\sum_{\tau=0}^{\infty} r^2(\tau) < \infty.$$

Let the spectral density of (y_n) be denoted by $f(\omega)$. We can not expect that the Fourier transform of (y_n) will converge in any reasonable sense (unless $y_n \equiv 0$). On the other hand, if $y_n \equiv 1$ then, using a heuristics that can be made precise, (y_n) is the inverse Fourier transform of the Dirac delta function assigning a unit mass at the point 0. Hence the Fourier transform of (y_n) itself is the the Dirac delta function. While the Dirac delta function is a generalized function, its integral is an ordinary function (namely the unit step function). These observations motivate us to consider the integrated process

$$\zeta_N(\omega') = \int_0^{\omega'} \sum_{n=-N}^{N} y_n e^{-in\omega} \, d\omega.$$

We can write

$$\zeta_N(\omega') = \sum_{n=-N}^{N} y_n \int_0^{\omega'} e^{-in\omega} \, d\omega = \sum_{n=-N}^{N} y_n c_n,$$

Note that $\frac{1}{\sqrt{2\pi}}c_n$ can be interpreted as the Fourier coefficients of the characteristic function of the interval $[0,\omega')$, which we denote by $\chi_{[0,\omega')}$.

Let us now compute $\mathrm{E}|\zeta_N(\omega')|^2$. We have

$$\mathrm{E}|\zeta_N(\omega')|^2 = \mathrm{E}\left|\sum_{n=-N}^{N} y_n c_n\right|^2 = \frac{1}{2\pi}\int_0^{2\pi} |C_N(e^{i\omega})|^2 \; f(\omega)\; d(\omega), \tag{4.1}$$

where

$$C_N(e^{i\omega}) = \sum_{n=-N}^{N} c_n e^{in\omega}.$$

Write $C_N(e^{i\omega})$ as

$$C_N(e^{i\omega}) = 2\pi \sum_{n=-N}^{N} \frac{c_n}{(2\pi)^{1/2}} \; \frac{e^{in\omega}}{(2\pi)^{1/2}}.$$

Now the latter sum is the Fourier series of the characteristic function $\chi_{[0,\omega')}(\cdot)$. Thus

$$\lim_{N\to\infty} \frac{1}{2\pi} C_N(\omega) = \chi_{[0,\omega')}(\omega)$$

in $L_2[0,2\pi)$. Since $f(\omega)$ is an element of $L_2[0,2\pi)$, and the scalar product in $L_2[0,2\pi)$ is continuous in its variables, we conclude from (4.1) that

$$\lim_{N\to\infty} \mathrm{E}|\zeta_N(\omega')|^2 = 2\pi \int_0^{2\pi} \chi_{[0,\omega')}(\omega) f(\omega) d\omega = 2\pi F(\omega').$$

By similar arguments we can see that if we now look at the increments of $\zeta_N(\omega')$ on two non-overlapping intervals $[a,b)$ and $[c,d)$ contained in $[0,2\pi)$ then we get

$$\lim_{N\to\infty} \mathrm{E}(\zeta_N(a) - \zeta_N(b))(\overline{\zeta}_N(d) - \overline{\zeta}_N(c)) = 2\pi \int_0^{2\pi} \chi_{[a,b)}\overline{\chi}_{[a,b)}(\omega) f(\omega) d\omega = 0.$$

Using the same train of thought it is easily seen that $\zeta_N(\omega')$ is a Cauchy sequence in $L_2^c(\Omega,\mathcal{F},P)$, hence it converges to some element of $L_2^c(\Omega,\mathcal{F},P)$ denoted by $\zeta(\omega')$:

$$\lim_{N\to\infty} \zeta_N(\omega') = \zeta(\omega').$$

Furthermore, if we take two non overlapping intervals $[a,b)$ and $[c,d)$ then we have

$$\mathrm{E}(\zeta(a) - \zeta(b))(\overline{\zeta}(d) - \overline{\zeta}(c)) = d\omega = 0.$$

We will express this fact by saying that $\zeta(\omega')$ is a process of orthogonal increments. To summarize our findings:

Theorem 4.1. *Let $y = (y_n)$ be a w.s.st. process with autocovariance function such that*

$$\sum_{\tau=0}^{\infty} r^2(\tau) < \infty.$$

Then

$$\lim_{N\to\infty} \int_0^{\omega'} \sum_{n=-N}^{N} y_n e^{-in\omega}\, d\omega = \zeta(\omega'),$$

in $L_2^c(\Omega, \mathcal{F}, P)$, where $\zeta(\omega')$ is a process with orthogonal increments. Moreover, denoting the spectral distribution function of $y = (y_n)$ by F we have

$$\mathrm{E}|\zeta(\omega')|^2 = 2\pi F(\omega').$$

4.2 Random orthogonal measures. Integration

The question is now raised: how we can represent a general wide sense stationary process as an integral of weighted trigonometric functions $e^{in\omega}$ in the form

$$y_n = \int_0^{2\pi} e^{in\omega} d\zeta(\omega),$$

where $d\zeta(\omega)$ is a random weight defined as some kind of random measure. Thus $d\zeta(\omega)$ is a substitute for the random coefficients ξ_k appearing in the definition of singular processes of the form $\sum_k \xi_k e^{i\omega_k n}$. Recalling the conditions imposed on ξ_k we define "orthogonal random measures $d\zeta(\omega)$" via the stochastic processes of orthogonal increments. The definition of the latter is obvious, it is almost a tautology:

Definition 4.1. *A complex valued stochastic process $\zeta(\omega)$ in $[0, 2\pi]$ is called a* process with orthogonal increments, *if it is left continuous, $\zeta(0) = 0$, $\mathrm{E}|\zeta(a)|^2 =: F(a) < \infty$ for all $0 \le a < 2\pi$, and for any two non-overlapping intervals $[a, b)$ and $[c, d)$ contained in $[0, 2\pi)$ we have*

$$\zeta(d) - \zeta(c) \;\perp\; \zeta(b) - \zeta(a).$$

The "measure $d\zeta$" assigning the value $\zeta(b) - \zeta(a)$ to an interval $[a, b)$ is called a random orthogonal measure. *The function F is called* the structure function. *It is assumed that F is left continuous.*

From the definition it follows that $F(0) = 0$.

Exercise 4.1. *Prove, that for any $0 \le a < 2\pi$ we have*

$$F(b) - F(a) = \mathrm{E}\,|\zeta(b) - \zeta(a)|^2,$$

thus F is monotone nondecreasing.

(*Hint:* Write $[0, b)$ as the union of $[0, a)$ and $[a, b)$ and apply Pythagoras theorem.)

Let now $d\zeta(\omega)$ be a random orthogonal measure on $[0,\, 2\pi)$, and let f be a possibly complex valued step function of the form

$$f(\omega) = \sum_{k \in K} \lambda_k \chi_{[a_k, b_k)},$$

where K is a finite set, and the intervals $[a_k, b_k)$ are non-overlapping. Then define

$$I(f) = \int_0^{2\pi} f(\omega) d\zeta(\omega) = \sum_{k \in K} \lambda_k (\zeta(b_k) - \zeta(a_k)).$$

Thus $I(f)$ is a random variable which is obviously in $L_2^c(\Omega, \mathcal{F}, P)$, where ^{c} indicates that we consider the L_2 space of complex valued functions.

Exercise 4.2. *Let f, g be two left continuous step functions on $[0,\, 2\pi]$. Then*

$$\mathrm{E} I(f)\overline{I(g)} = \int_0^{2\pi} f(\omega)\overline{g(\omega)}\, dF(\omega). \tag{4.2}$$

(*Hint*: Take a common subdivision for f and g.)

Let H_s^c be the set of complex-valued left-continuous step-functions on $[0,\, 2\pi)$. Obviously H_s^c is a linear space and $H_s^c \subset L_2^c([0,\, 2\pi],\, dF)$. Thus (4.2) can be restated saying that stochastic integration as a linear operator

$$I \;:\; H_s^c \to L_2^c(\Omega, \mathcal{F}, P)$$

is an isometry.

Exercise 4.3. *Prove that*

$$y_n = \int_0^{2\pi} e^{in\omega} d\zeta(\omega)$$

is w.s.st.

4.3 Representation of a wide sense stationary process

Perhaps the most powerful tool in the theory of w.s.st. processes is the following spectral representation theorem, which will be used over and over again in this course.

Theorem 4.2. *Let (y_n) be a wide sense stationary process. Then there exists a unique random orthogonal measure $d\zeta(\omega)$, such that*

$$y_n = \int_0^{2\pi} e^{in\omega} d\zeta(\omega).$$

The process ζ is called *the spectral representation process* of (y_n).

Proof. Assuming that (y_n) can be represented as stated we have

$$\mathrm{E}(y_{n+\tau}\overline{y_n}) = \int_0^{2\pi} e^{i(n+\tau)\omega} e^{-in\omega} dF(\omega) = \int_0^{2\pi} e^{in\tau} dF(\omega).$$

Thus the structure function of $d\zeta$ is necessarily determined by the spectral distribution of y, denoted by $F^y(\omega)$, as follows:

$$dF(\omega) = dF^y(\omega) \cdot \frac{1}{2\pi}.$$

Now, integration with respect to $d\zeta$ defines an isometry I from $L_2^c(dF) = L_2^c([0, 2\pi), dF)$ into $L_2^c(\Omega, \mathcal{F}, P)$. Conversely, if such an isometry I is given, then it defines an orthogonal random measure on $[0, 2\pi)$ with structure function F simply by setting

$$\zeta(a) = I(\chi_{[0,a)}).$$

Thus finding the spectral representation process $\zeta(\omega)$ is equivalent to finding the isometry I from $L_2^c(dF)$ into $L_2^c(\Omega, \mathcal{F}, P)$.

Now, the assumed representation for (y_n) implies the following specifications for I:

$$I(e^{in\omega}) = y_n. \tag{4.3}$$

From here we could argue as follows: to get $\zeta(a) = I(\chi_{[0,a)})$ write

$$\chi_{[0,a)} = \sum_n c_n e^{in\omega}, \tag{4.4}$$

where convergence on the right hand side is assumed to take place in $L_2^c(dF)$. Then, by the continuity of I, we would get

$$\zeta(a) = \sum_n c_n y_n.$$

The difficulty with this argument is to actually find the representation of $\chi_{[0,a)}$ as given under (4.4), when convergence of the right hand side is required in a possibly strange norm defining $L_2^c(dF)$. Therefore we follow another line of thought. Consider the set of

specifications (4.3) prescribed for I. Let us now extend the definition of the yet undefined isometry I to the linear space

$$H_e^c = \{g(\omega) : \ g(\omega) = \sum_{n\in N} c_n e^{in\omega}, \quad c_n \in \mathbb{C}, \quad N \subset \mathbb{Z} \text{ finite}\}$$

Define for $g \in H_e^c$

$$I(g) = \sum_{n\in N} c_n y_n.$$

Consider H_e^c as a linear subspace of $L_2^c(dF)$. The linear extension of I is well-defined if $I(g)$ is independent of the representation of g. This is equivalent to saying that $g = 0$ in $L_2^c(dF)$ implies

$$I(g) = 0 \quad \text{in} \quad L_2^c(\Omega, \mathcal{F}, P).$$

Exercise 4.4. *The above implication.*

The last argument also shows that I is an isometry from H_e^c to $L_2^c(\Omega, \mathcal{F}, P)$. Since H_e^c is a dense linear subspace in $L_2^c(dF)$, I can be extended to a linear isometry mapping from $L_2^c(dF)$ into $L_2^c(\Omega, \mathcal{F}, P)$ in a unique manner. As said above, the orthogonal random measure itself is obtained by setting

$$\zeta(a) = I(\chi_{[0,a)}).$$

Exercise 4.5. *Show that the structure function of the random orthogonal measure ζ is F, predetermined by the spectral distribution function of y.*

Let now I' denote the isometry from $L_2^c(dF)$ to $L_2^c(\Omega, \mathcal{F}, P)$ defined by integration w.r.t. $d\zeta$:

$$I'(g) = \int_0^{2\pi} g(\omega) d\zeta(\omega).$$

Then I and I' agree on all characteristic functions $\chi_{[0,a)}$, and therefore I and I' agree on all step functions. Since the latter are dense in $L_2^c(dF)$, we conclude that $I = I'$. It follows, that

$$y_n = I(e^{in\omega}) = I'(e^{in\omega}) = \int_0^{2\pi} e^{in\omega} d\zeta(\omega)$$

as stated.

4.4 Change of measure

Let $d\zeta(\omega)$ be a random orthogonal measure on $[0, 2\pi)$ with the structure function $F(\omega)$. Let $g \in L_2^c(dF)$, and define

$$\eta(\omega) = \int_0^\omega g(\omega')d\zeta(\omega') \qquad 0 \leq \omega < 2\pi.$$

Exercise 4.6. *Show that $d\eta(\omega)$ is a random orthogonal measure, with the structure function*

$$dG(\omega) = |g(\omega|^2 dF(\omega).$$

The corresponding random orthogonal measure will be written as

$$d\eta(\omega) = g(\omega)d\zeta(\omega).$$

Let now $h(\omega)$ be a function in $L_2^c(dG)$. Note that we have taken an element of a new Hilbert-space, defined by dG. Then

$$\int_0^{2\pi} h(\omega)d\eta(\omega)$$

is well-defined. Now we have the following, intuitively obvious-looking result:

Proposition 4.2. *We have*

$$\int_0^{2\pi} h(\omega)d\eta(\omega) = \int_0^{2\pi} h(\omega)g(\omega)d\zeta(\omega).$$

Proof. The proposition is obviously true if h is a characteristic function $\chi_{[0,a)}(\omega)$. Since both sides of the stated equality are linear in h, it follows that the proposition is true whenever h is a step function. Now let h be an arbitrary function in $L_2^c(dG)$ and let h_n be a sequence of step functions converging to h in the corresponding Hilbert-space norm. Then

$$\int_0^{2\pi} h_n(\omega)d\eta(\omega) \quad \to \int_0^{2\pi} h(\omega)d\eta(\omega) \qquad \text{in } L_2^c(\Omega, \mathcal{F}, P).$$

On the other hand, the assumed convergence

$$\int_0^{2\pi} |h_n(\omega) - h(\omega)|^2 dG(\omega) = \int_0^{2\pi} |h_n(\omega) - h(\omega)|^2 \ |g(\omega)|^2 \ dF(\omega) \ \to \ 0$$

implies

$$\int_0^{2\pi} |h_n(\omega)g(\omega) - h(\omega)g(\omega)|^2 dF(\omega) \ \to \ 0.$$

But then, the isometry property of stochastic integral w.r.t. $d\zeta$ gives

$$\int_0^{2\pi} h_n(\omega)g(\omega)d\zeta(\omega) \ \to \ \int_0^{2\pi} h(\omega)g(\omega)d\zeta(\omega),$$

and the proposition follows. □

4.5 Linear filters

Let us now consider the effect of *linear filters* on the spectral representation process. Let (u_n) be a wide sense stationary process with spectral representation process $d\zeta^u(\omega)$. Define the process (y_n) via a FIR filter as

$$y_n = \sum_{k=0}^{m} h_k u_{n-k}.$$

Then (y_n) is a wide sense stationary process.

Exercise 4.7. *Show that the spectral representation process of y is given by*

$$d\zeta^y(\omega) = H(e^{-i\omega})d\zeta^u(\omega),$$

where

$$H(e^{-i\omega}) = \sum_{k=0}^{m} h_k e^{-ik\omega}.$$

Let us now consider the infinite linear combination

$$y_n = \sum_{k=0}^{\infty} h_k u_{n-k}.$$

We have seen that the r.h.s is well defined (converges in $L_2^c(\Omega, \mathcal{F}, P)$), if the infinite series

$$H(e^{-i\omega}) = \sum_{k=0}^{\infty} h_k e^{-ik\omega}$$

is well defined in $L_2^c(dF)$.

Proposition 4.3. *The spectral representation process of (y_n) is given by*

$$d\zeta^y(\omega) = H(e^{-i\omega})d\zeta^u(\omega).$$

Proof. Truncate the infinite sum defining y_n at m, i.e. define

$$y_n^m = \sum_{k=0}^{m} h_k u_{n-k}.$$

Then the spectral representation of $y^m = (y_n^m)$ is given as

$$y_n^m = \int_0^{2\pi} e^{in\omega} H^m(e^{-i\omega})d\zeta^u(\omega), \tag{4.5}$$

where

$$H^m(e^{-i\omega}) = \sum_{k=0}^{m} h_k e^{-ik\omega}.$$

Now letting m tend to infinity, the l.h.s. of (4.5) converges to y_n in $L_2^c(\Omega, \mathcal{F}, P)$.

Exercise 4.8. *Show that the integrand on the right hand side converges to* $e^{in\omega}H(e^{-i\omega})$ *in* $L_2^c(dF^u)$.

Thus the corresponding integral w.r.t. $d\zeta^u$ will converge to

$$\int_0^{2\pi} e^{in\omega}H(e^{-i\omega})d\zeta^u(\omega)$$

in $L_2^c(\Omega, \mathcal{F}, P)$. This proves the claim. □

Chapter 5

AR, MA and ARMA processes

5.1 Autoregressive processes

Consider now a process that is implicitly defined via the equation

$$y_n + a_1 y_{n-1} + \ldots + a_p y_{n-p} = e_n, \tag{5.1}$$

where (e_n) is a w.s.st. orthogonal process.

A shorthand notation is

$$A(q^{-1})y = e \tag{5.2}$$

where q^{-1} is the backward shift operator, and

$$A(q^{-1}) = \sum_{k=0}^{p} a_k q^{-k}, \qquad a_0 = 1.$$

Proposition 5.1. *Assuming that $A(e^{-i\omega}) \neq 0$ for all ω, the equation (5.2) has a unique wide sense stationary solution (y_n). The process (y_n) has a spectral density equal to*

$$f(\omega) = \frac{\sigma^2(e)}{|A(e^{-i\omega})|^2}.$$

Proof. Assume that a wide sense stationary solution (y_n) exists. Let the spectral distribution process of e and y be denoted by $d\zeta^e(\omega)$ and $d\zeta^y(\omega)$, respectively. Then

$$A(e^{-i\omega})d\zeta^y(\omega) = d\zeta^e(\omega),$$

from which we get formally

$$d\zeta^y(\omega) = \frac{1}{A(e^{-i\omega})}\, d\zeta^e(\omega). \tag{5.3}$$

Recall, that the spectral density function of (e_n) is

$$f^e(\omega) = \frac{\sigma^2(e)}{2\pi}.$$

If $A(e^{-i\omega}) \neq 0$ for all ω, then $1/A(e^{-i\omega})$ is in $L_2^c(d\omega)$, where $d\omega$ is the spectral measure of e, modulo a constant multiplier. Indeed, we have

$$\int_0^{2\pi} \frac{1}{|A(e^{-i\omega})|^2} d\omega = \oint_S h(z)\, dz, \qquad \text{with } h(z) = \frac{1}{|A(z)|^2},$$

where S is the complex unit circle and $h(z)$ is continuous on S, thus the line integral is well defined.

Hence the right hand side of (5.3), as a new random orthogonal measure, is well defined. It follows, that the wide sense stationary process

$$y_n = \int_0^{2\pi} e^{in\omega} \frac{1}{A(e^{-i\omega})} d\zeta^e(\omega).$$

is well defined. If there is a solution, then it must be equal to the process y just defined. By this uniqueness is proved. On the other hand, it is easily seen that the process y is indeed a solution of (5.1). Namely,

$$(Ay)_n = \int_0^{2\pi} A(e^{-i\omega}) e^{in\omega} \frac{1}{A(e^{-i\omega})} d\zeta^e(\omega) = \int_0^{2\pi} e^{in\omega} d\zeta^e(\omega) = e_n.$$

This proves the existence of a solution, and the proposition is proved. □

The graphs of a two newly selected AR(2) processes are shown on the figures below together with their autocovariance functions:

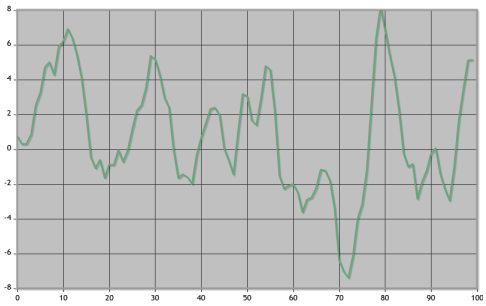

Figure 5.1: AR(2) process with two almost unstable complex poles whose real part is positive, with very small delays. The actual values are: length 0.8 and argument $\pm 0.1\pi$.

More complex behaviors can be noticed on the graphs of AR(4) processes below:

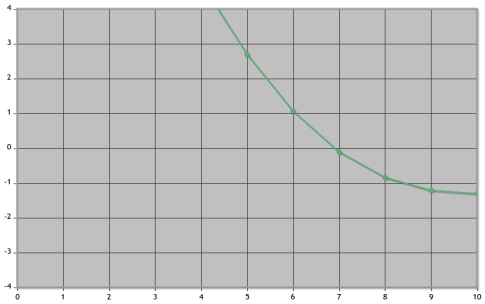

Figure 5.2: Autocovariance of AR(2) process with two almost unstable complex poles whose real part is positive, with very small delays. The actual values: a pair of complex poles with length 0.8 and argument $\pm 0.1\pi$.

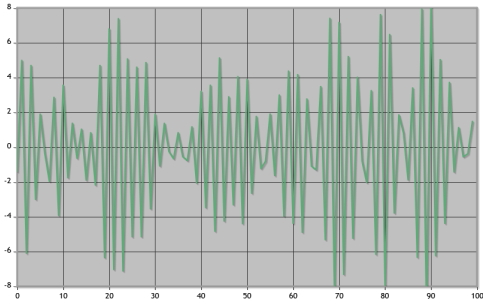

Figure 5.3: AR(2) process with two almost unstable complex poles whose real part is negative, with very large delays. The actual values: length 0.8, argument $\pm 0.9\pi$.

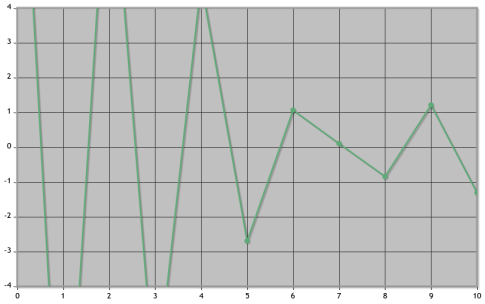

Figure 5.4: Autocovariance of AR(2) process with two almost unstable complex poles whose real part is negative, with very large delays. The actual values: length 0.8, argument $\pm 0.9\pi$.

The above derivation demonstrates the supreme power of spectral representation: the outline of the proof is easily obtained by formal arguments, which then are easily filled with rigorous technical details. It is not clear at this point if there is any other more direct method that would yield the proposition. An exception is the case $p = 1$, but even

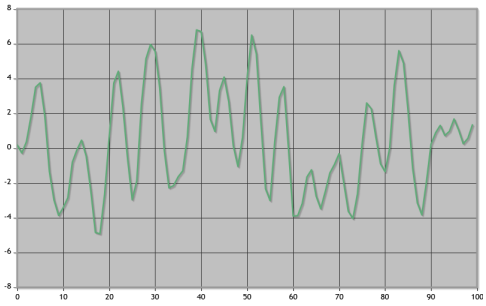

Figure 5.5: AR(4) process with two positive poles and two almost unstable complex pose whose real part is positive. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.3\pi$.

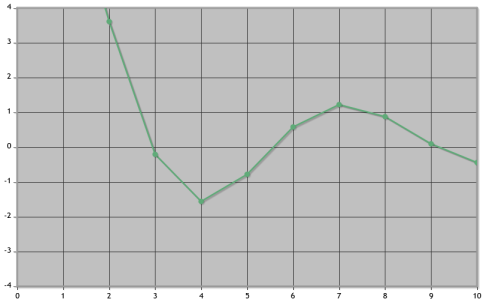

Figure 5.6: Autocovariance of AR(4) process with two positive poles and two almost unstable complex pose whose real part is positive. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.3\pi$.

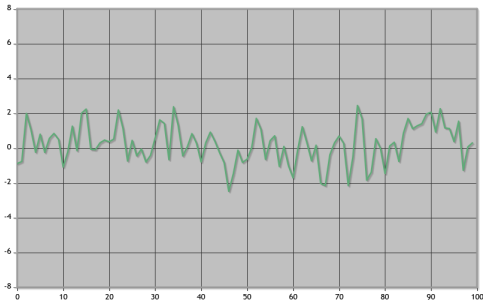

Figure 5.7: AR(4) process with two positive poles and two almost unstable complex pose whose real part is negative. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.6\pi$.

there we may run into unexpected challenges.

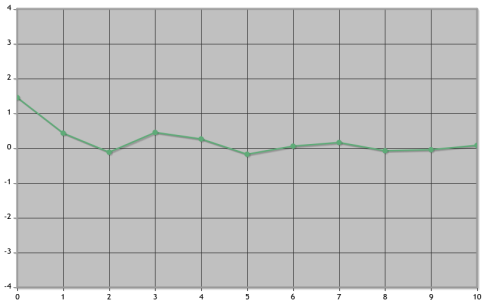

Figure 5.8: Autocovariance of AR(4) process with two positive poles and two almost unstable complex pose whose real part is negative. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.6\pi$.

5.2 A classic result: the Yule-Walker equations

AR processes are of particular interest due to the fact that a finite segment of its past values is sufficient to compute the best linear one-step ahead predictor. Let y be a stable AR(p) process, satisfying

$$y_n + \sum_{k=1}^{p} a_k y_{n-k} = e_n, \tag{5.4}$$

where (e_n) is the innovation process of (y_n). Recall that if $a_p \neq 0$ then p is called *the order of the process*. The question can be put: how the auto-covariances of (y_n), say $r(k) := \mathrm{E}(y_{n+k}y_n)$ are to be computed.

Let $0 \leq k \leq p$ and multiply equation (5.4) by y_{n-k}. Taking expectation, and using the that $r(k) = r(-k)$ we get the following system of linear equations:

$$\begin{aligned}
r(0) + a_1 r(1) \cdots + a_p r(p) &= \sigma^2(e) \\
r(1) + a_1 r(0) \cdots + a_p r(p-1) &= 0 \\
&\vdots \\
r(k) + a_1 r(k-1) \cdots + a_p r(p-k) &= 0 \\
&\vdots \\
r(p) + a_1 r(p-1) \cdots + a_p r(0) &= 0.
\end{aligned}$$

We have $p+1$ linear equation for the first $p+1$ auto-covariances. This set of equations is called *Yule - Walker equations*. The coefficient matrix of this systems of linear equations is

$$\begin{pmatrix} 1 & a_1 \ldots a_p \\ a_1 & 1 \ldots a_{p-1} \\ . & .... \\ a_p & a_{p-1} \ldots 1 \end{pmatrix}$$

This is a so-called circulant matrix the eigenvalues of which are known to be of the form $A(\gamma^k)$, where $\gamma = e^{2\pi/(p+1)}$. Now if $A(z) \neq 0$ on $|z| = 1$ the no eigenvalues are 0, hence the Yule-Walker equation has a unique solution. Recall, that in Proposition 5.1 this was the condition for the unique wide sense stationary solution of the AR process. Thus the first $p+1$ auto-covariances are uniquely determined.

For $k > p$ the auto-covariance $r(k)$ can be computed recursively in terms of previous values of $r(\cdot)$ using the same arguments as above: multiplying equation (5.4) by y_{n-k} and taking expectation we get

$$r(k) = -a_1 r(k-1) - \cdots - a_p r(k-p).$$

*S*pecial case. For $p = 1$ the Yule-Walker equations consist of two equations:

$$\begin{aligned} r(0) + ar(1) &= \sigma^2 \\ r(1) + ar(0) &= 0. \end{aligned}$$

The solution is well-known:

$$r(0) = \frac{\sigma^2}{1-a^2}, \qquad r(1) = -a\frac{\sigma^2}{1-a^2},$$

and further on $r(k) = (-a)^k \dfrac{\sigma^2}{1-a^2}$ for $k > 0$.

5.3 The AR(1) process

Thus let us now consider an AR(1) process defined by

$$y_n + ay_{n-1} = e_n. \tag{5.5}$$

Let us first assume that $|a| < 1$. Recall the examples below from Chpater 1:

Then $A(z^{-1}) = 1 + az^{-1} \neq 0$ for $|z| = 1$, and by the above theorem a unique solution exists. The existence and uniqueness of the solution can directly be seen as follows. First assume that a w.s.st. solution exists. Then iterating (5.5) we get after m steps

$$y_n = \sum_{k=0}^{m-1} (-a)^k e_{n-k} + (-a)^m y_{n-m}.$$

The residual term $(-a)^m y_{n-m}$ tends to 0 for $m \to \infty$ in $L_2(\Omega, \mathcal{F}, P)$, thus we get

$$y_n = \sum_{k=0}^{\infty} (-a)^k e_{n-k}. \tag{5.6}$$

It is easy to see that the right hand side is indeed convergent in $L_2(\Omega, \mathcal{F}, P)$ for $|a| < 1$. This proves uniqueness. It is also easy to see that the process defined by (5.6) is indeed a solution of (5.5), proving existence.

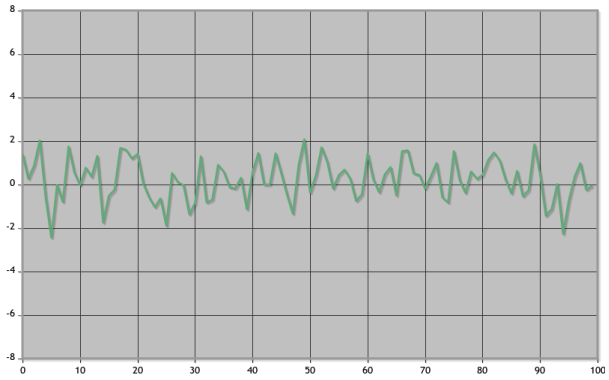

Figure 5.9: AR(1) process with a stable positive pole

Exercise 5.1. *Prove that* (y_n) *defined by (5.6) does indeed satisfy (5.5).*

Equations (5.5) and (5.6) also imply that $H_n^e \subset H_n^y$ and $H_n^y \subset H_n^e$ and thus

$$H_n^e = H_n^y.$$

It follows that (e_n) is the innovation process of (y_n).

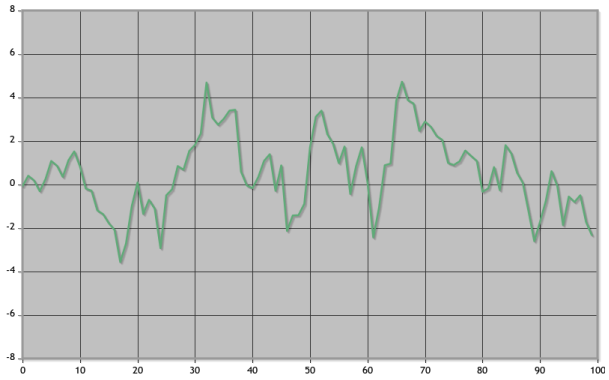

Figure 5.10: AR(1) process with almost unstable positive pole

Let us now consider an AR(1) process defined by (5.5) with $|a| > 1$. Then, again, $A(z^{-1}) = 1 + az^{-1} \neq 0$ for $|z| = 1$, thus a unique solution of (5.5) exists. However, iterating (5.5) as above does not yield a representation of the form (5.6). Nevertheless, if we rewrite (5.5) in the form

$$y_{n-1} = -\frac{1}{a}y_n + \frac{1}{a}e_n,$$

and iterate this equation *forward* in time we get that y_n must be of the form

$$y_n = \sum_{k=0}^{\infty}(-\frac{1}{a})^k\frac{1}{a}e_{n+k+1}.$$

Exercise 5.2. *Show that the r.h.s. is well-defined, and* (y_n) *does indeed satisfy (5.5).*

Thus y is expressed via the *future* of the wide sense stationary orthogonal process e. Therefore we conclude that

$$y_n \notin H_n^e,$$

and thus (e_n) is **not** the innovation process of (y_n).

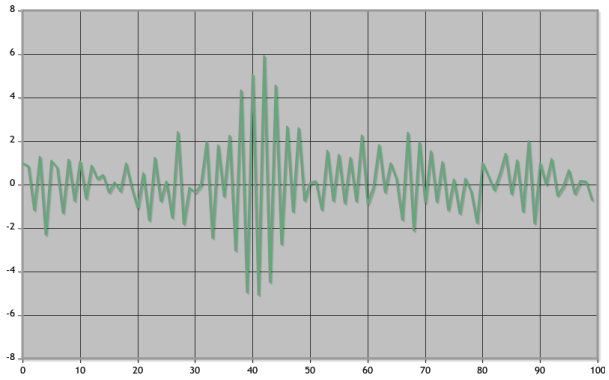

Figure 5.11: AR(1) process with almost unstable negative pole

5.4 Stable AR systems

Let us now consider a higher order AR process defined by

$$A(q^{-1})y = e \tag{5.7}$$

with $\deg A = p > 1$. We can ask ourselves: under what conditions e is the innovation process of y.

Obviously, a backward iteration of (5.7), that worked nicely for $p = 1$, is not easily manageable. Hence we will settle the issue within the framework of spectral representation. Obviously, we have $H_n^e \subset H_n^y$ as before. To prove the opposite inclusion, $H_n^y \subset H_n^e$, we need to express y_n in terms of the past of e_n. This will be certainly possible, if the rational function $1/A(e^{-i\omega})$ can be expanded into a power series of $e^{-i\omega}$. The possibility of such an expansion is clearly related to the position of the zeros of $A(z^{-1})$.

Definition 5.2. *A polynomial of* z^{-1}, $A(z^{-1}) = \sum_{k=0}^{p} a_k z^{-1}$, *with* $a_0 = 1$ *is called* stable *if*

$$A(z^{-1}) \neq 0 \quad \text{for} \quad |z| \geq 1.$$

Lemma 5.3. *If A is a stable polynomial with $a_0 = 1$ then*

$$1/A(e^{-i\omega}) = \sum_{k=0}^{\infty} h_k e^{-ik\omega}, \qquad h_0 = 1,$$

where convergence on the right hand side is uniform in ω.

It follows immediately that the r.h.s converges also in $L_2^c([0, 2\pi], d\omega)$.

Proof. Consider the function $1/A(z^{-1})$ with z taking its values in $\mathbb{C}$. The equation $A(z^{-1}) = 0$ has, counted with multiplicity, exactly p roots. Hence, the stability of $A(z^{-1})$ implies that $1/A(z^{-1})$ is analytic in $\{z : |z| > 1 - \varepsilon\}$ with some $\varepsilon > 0$, or equivalently, analytic in z^{-1} for $\{z^{-1} : |z^{-1}| < 1 + \varepsilon\}$ with some $\varepsilon > 0$. Therefore it can be expanded into a Taylor series of z^{-1} around 0

$$1/A(z^{-1}) = \sum_{k=0}^{\infty} h_k z^{-k} \tag{5.8}$$

which is convergent uniformly for $|z^{-1}| < 1 + \varepsilon$ with some $\varepsilon > 0$. It follows that (5.8) converges uniformly for $|z| = 1$. Finally, $h_0 = 1$ follows by evaluating the two sides of (5.8) for $z^{-1} = 0$. □

Let us now consider the AR(p) process (y_n) defined by

$$A(q^{-1})y = e \tag{5.9}$$

where (e_n) is a wide sense stationary orthogonal process, and $A(q^{-1})$ is a polynomial of q^{-1} with $\deg A = p$ and $a_0 = 1$.

Proposition 5.4. *If the polynomial $A(z^{-1})$ is stable then e is the innovation process of y.*

Proof. Obviously $e_n \in H_n^y$, thus we need only to prove that $y_n \in H_n^e$. Now

$$d\zeta^y(\omega) = \frac{1}{A(e^{-i\omega})} d\zeta^e(\omega) = (\sum_{k=0}^{\infty} h_k e^{-ik\omega}) d\zeta^e(\omega).$$

Thus

$$y_n = \int_0^{2\pi} e^{in\omega} (\sum_{k=0}^{\infty} h_k e^{-ik\omega}) d\zeta^e(\omega).$$

Since the infinite series on the right hand side (multiplicated by $e^{in\omega}$) converges in $L_2^c([0, 2\pi], d\omega)$, we can interchange the integration and the summation to get

$$y_n = \sum_{k=0}^{\infty} h_k \int_0^{2\pi} e^{in\omega} e^{-ik\omega} d\zeta^e(\omega) = \sum_{k=0}^{\infty} h_k e_{n-k}.$$

□

Remark. The converse result also holds: if $A(z^{-1})$ is a polynomial with $A(z^{-1}) \neq 0$ for $|z| = 1$ and the AR(p) process (y_n) defined by (5.9) has (e_n) for its innovation process, then $A(z^{-1})$ is stable.

5.5 MA processes

A simple class of wide sense stationary processes is obtained by taking a finite moving average of an orthogonal process $e = (e_n)$. Thus let $e = (e_n)$ be a wide sense stationary orthogonal process and define

$$y_n = \sum_{k=0}^{m} c_k e_{n-k} \tag{5.10}$$

Exercise 5.3. *Show that (y_n) is a wide sense stationary process.*

Definition 5.5. *A wide sense stationary process defined by (5.10) is called a* moving average *(MA)* process*, or more precisely MA(m) process.*

In addition to the examples given in Chapter 1 the graphs of a simulated MA(4) process are displayed below:

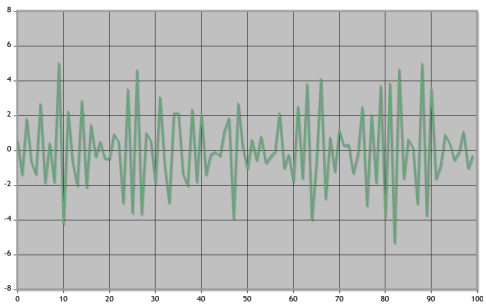

Figure 5.12: MA(4) process with small positive zeros. The actual values are $0.1, 0.3, 0.5, 0.7$.

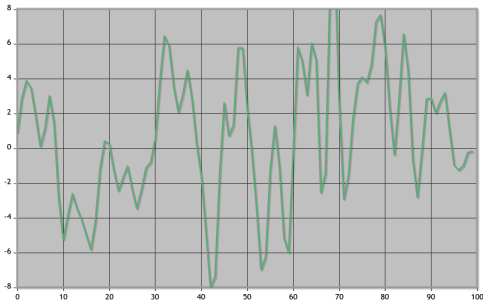

Figure 5.13: MA(4) process with large negative zeros. The actual values are $-0.6, -0.7, -0.8, -0.9$.

In engineering terminology we would say that y_n is obtained by passing an orthogonal process through a finite impulse response (FIR) filter.

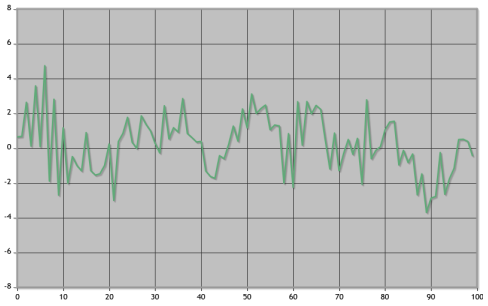

Figure 5.14: MA(4) process with large complex zeros.The actual values are: lengths equal to 1, arguments $\pm 0.3\pi$ and $\pm 0.7\pi$.

The autocovariance function of this process can be computed as follows:

$$r(k) = \begin{cases} \sum_{j=0}^{m} c_j^2 & k = 0 \\ \sum_{j=0}^{m-k} c_j c_{j+k} & 0 < k \leq m \\ 0 & k > m \\ c_{-k} & k < 0. \end{cases}$$

Thus the coefficients $(c_0, \ldots, c_m)$ uniquely determine autocovariances $(r(0), r(1), \ldots, r(m))$. In the other direction uniqueness is not true. We show a simple example.
Example. Consider the w.s.st. processes:

$$y_n = e_n + be_{n-1} \tag{5.11}$$
$$z_n = be_n + e_{n-1}. \tag{5.12}$$

Assume $b \neq 1$, then the two processes are different. Nevertheless, their autocovariances are the same, we have

$$r(k) = \begin{cases} 1 + b^2 & k = 0 \\ b & k = 1 \\ 0 & k > 1 \\ r(-k) & k < 0 \end{cases}.$$

Rearrange (5.11) to get e_n and iterate this equation. We get the following formal series

$$e_n = y_n - be_{t-1} = y_n - b(y_{n-1} - be_{n-2}) = \cdots = \sum_{k=0}^{\infty} (-1)^k b^k y_{n-k}. \tag{5.13}$$

Similarly, from equation (5.12) we get

$$e_n = \frac{1}{b}(z_n - e_{n-1}) = \frac{1}{b}z_n - \frac{1}{b}(\frac{1}{b}z_{n-1} - e_{n-2}) = \cdots = \frac{1}{b}\sum_{k=0}^{\infty}(-1)^k \frac{1}{b^k} z_{n-k}. \tag{5.14}$$

We get two formal infinite series reconstruction of (e_n). A process is called *invertible* if the reconstruction of (e_n) is possible. In this example it is equivalent with the convergence of the formal infinite series.

Exercise 5.4. *Show that (5.13) is well-defined, if* $\sum_{k=0}^{\infty} b^{2k} < \infty$, *i.e.* $|b| < 1$. *Similarly, (5.14) is well-defined if* $|b| > 1$.

Hence, although the autocovariances of both processes are the same, only one of them is invertible, depending on the value of b.

5.6 ARMA processes

Let us now consider the combination of AR and MA processes.

Definition 5.6. *A wide sense stationary process is called an* ***ARMA*** *process if it satisfies the difference equation*

$$A(q^{-1})y = C(q^{-1})e, \tag{5.15}$$

where (e_n) *is a w.s.st. orthogonal process, and* A, C *are polynomials of the backward shift operator* q^{-1}. *The degrees* p *and* r *of* $A(q^{-1})$ *and* $C(q^{-1})$, *respectively, are called the* orders *of the ARMA process.*

Writing

$$A(q^{-1}) = \sum_{k=0}^{p} a_k q^{-k}, \qquad C(q^{-1}) = \sum_{k=0}^{r} c_k q^{-k}$$

we assume that $a_0 = c_0 = 1$. If the degrees of $A(q^{-1})$ and $C(q^{-1})$ are p and q, respectively, then also $a_p \neq 0$ and $c_r \neq 0$. If we want to stress the orders we say that (y_n) is an ARMA(p, r) process.

The graphs of simulated ARMA$(2, 2)$ processes together with their auto-covariance functions are displayed on the figures below:

Straightforward extensions of Propositions 5.1 and 5.4 are the following results:

Proposition 5.7. *Assume that* $A(z^{-1}) \neq 0$ *for* $|z| = 1$. *Then there is a unique w.s.st. process satisfying (5.15).*

Exercise 5.5. *Prove the above proposition.*

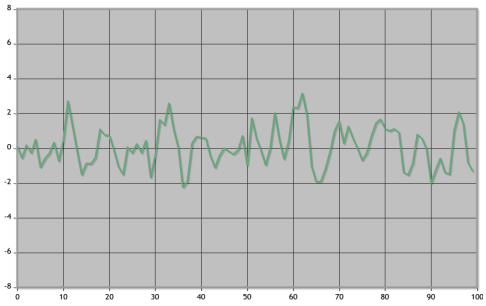

Figure 5.15: ARMA$(2,2)$ process with similar AR poles and MA zeros. The actual poles: length 0.8, arguments $\pm 0.3\pi$, the actual zeros: length 0.9 arguments $\pm 0.4\pi$.

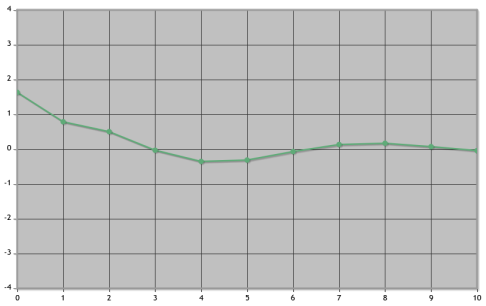

Figure 5.16: Autocovariance of ARMA$(2,2)$ process with similar AR poles and MA zeros. The actual poles: length 0.8, arguments $\pm 0.3\pi$, the actual zeros: length 0.9 arguments $\pm 0.4\pi$.

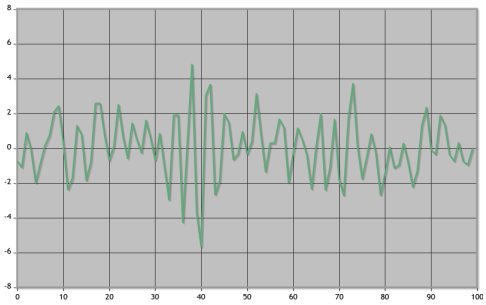

Figure 5.17: ARMA$(2,2)$ process with complex AR poles with small negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.6\pi$, the actual zeros: $-0.6, -0.9$.

Proposition 5.8. *Assume that $A(z^{-1})$ and $C(z^{-1})$ are stable polynomials. Then (e_n) is the innovation process of (y_n).*

Exercise 5.6. *Prove the above proposition.*

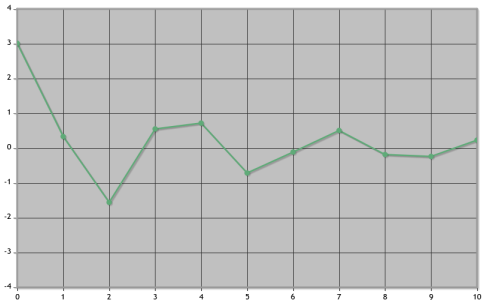

Figure 5.18: Autocovariance of ARMA$(2, 2)$ process with complex AR poles with small negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.6\pi$, the actual zeros: $-0.6, -0.9$.

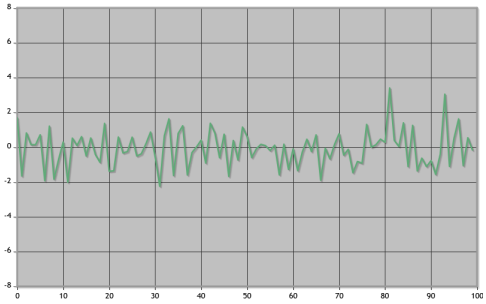

Figure 5.19: ARMA$(2, 2)$ process with complex AR poles with large negative real parts combined with two real negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.9\pi$, the actual zeros: $-0.6, -0.9$.

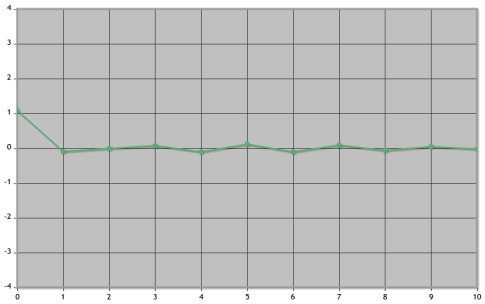

Figure 5.20: Autocovariance of ARMA$(2, 2)$ process with complex AR poles with large negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.9\pi$, the actual zeros: $-0.6, -0.9$.

5.7 Prediction

Now we have all the machinery to revisit the prediction problem in general, and for ARMA processes in particular. Let (y_n) be a *completely regular* w.s.st. process with innovation process (e_n). Then we can write

$$y_n = \sum_{k=0}^{\infty} h_k e_{n-k} \tag{5.16}$$

with $\sum_{k=0}^{\infty} h_k^2 < \infty$. The one-step ahead predictor of y_n is then given by

$$\widehat{y}_n = \sum_{k=1}^{\infty} h_k e_{n-k}. \tag{5.17}$$

Here we have taken into account that $H_{n-1}^y = H_{n-1}^e$, and hence

$$(y_n | H_{n-1}^y) = (y_n | H_{n-1}^e).$$

To complete the above argument we have to express e in term of y. This can be done best in the spectral domain. Let us write (5.16) in the form

$$d\zeta^y(\omega) = H(e^{-i\omega}) d\zeta^e(\omega).$$

Then

$$d\zeta^e(\omega) = H^{-1}(e^{-i\omega}) d\zeta^y(\omega). \tag{5.18}$$

Note that the right hand side is well defined, since $H^{-1}(e^{-i\omega}) \in L_2^c(dF^y(\omega))$, we have $dF^y(\omega) = |H(e^{-i\omega})|^2 d\omega$. The one-step ahead predictor given by (5.17) can be described by

$$d\zeta^{\widehat{y}}(\omega) = (H(e^{-i\omega}) - 1) d\zeta^e(\omega). \tag{5.19}$$

Combining (5.18) and (5.19) we get the following result:

Proposition 5.9. *Let $y = (y_n)$ be a completely regular wide sense stationary process given by (5.16). Then its one step ahead predictor $\widehat{y} = (\widehat{y}_n)$ is obtained via the spectral representation measure*

$$d\zeta^{\widehat{y}}(\omega) = (1 - H^{-1}(e^{-i\omega})) d\zeta^y(\omega).$$

The above result provides a general solution to the prediction problem for regular processes. A shortcoming of this result that it is formulated in the spectral domain.

It is important to stress that $H^{-1}(e^{-i\omega})$ may not be written as an infinite series $\sum_{k=0}^{\infty} g_k e^{-ik\omega}$ which is convergent in $L_2^c(dF^y(\omega))$. In other words, a linear filter in frequency domain (amounting to re-weighting the frequencies) does not necessarily have a time domain representation. However, the situation is simplified considerably in the case of ARMA processes.

Let (y_n) be a wide sense stationary ARMA process defined by

$$A(q^{-1})y = C(q^{-1})e \tag{5.20}$$

where $A(z^{-1})$ and $C(z^{-1})$ are *stable* polynomials. Then e is the innovation process of y, and setting

$$H(e^{-i\omega}) = C(e^{-i\omega})/A(e^{-i\omega})$$

we can write (5.20) as

$$d\zeta^y(\omega) = H(e^{-i\omega})d\zeta^e(\omega).$$

The stability of $A(z^{-1})$ implies that

$$H(e^{-i\omega}) = \sum_{k=0}^{\infty} h_k e^{-ik\omega}$$

where the right hand side converges in $L_2(d\omega)$. Applying Proposition 5.9 we get

$$d\zeta^{\widehat{y}}(\omega) = (1 - A(e^{-i\omega})/C(e^{-i\omega}))d\zeta^y(\omega).$$

Multiplying both sides by $C(e^{-i\omega})$ and converting the resulting equality to time domain we get the following result:

Proposition 5.10. *Let (y_n) be a wide sense stationary ARMA process given by (5.20), where $A(z^{-1})$ and $C(z^{-1})$ are stable. Then the one-step ahead prediction process $\widehat{y}$ is defined by the equation*

$$C(q^{-1})\widehat{y} = (C(q^{-1}) - A(q^{-1}))y$$

Note that $a_0 = c_0 = 1$ implies that the constant term of $(C(q^{-1}) - A(q^{-1})$ is 0, and hence the result of its action on y at any time n can be computed using only values of y up to time $n-1$. Thus we do get a genuine one-step ahead predictor.

The prediction of three ARMA processes, using AR(4) approximation, are displayed below. The predicted processes are displayed in yellow.

In dealing with actual data prediction is based on models based on the data, therefore the exact true dynamics is unknown. It is therefore an interesting experiment is to see the effect of parametric uncertainty onto prediction. In the figures below we display three AR(4)-processes together with their predictors based on artificially and randomly perturbed models. It is interesting to note that these misspecified predictors perform remarkably well:

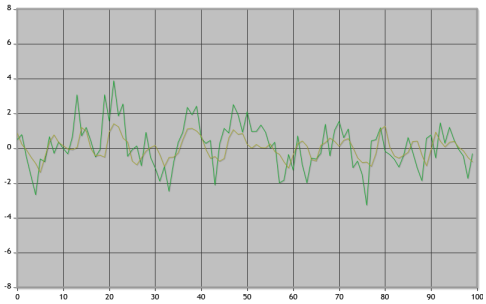

Figure 5.21: AR(4) Prediction of ARMA$(2, 2)$ process with similar AR poles and MA zeros. The actual poles: length 0.8, arguments $\pm 0.3\pi$, the actual zeros: length 0.9 arguments $\pm 0.4\pi$.

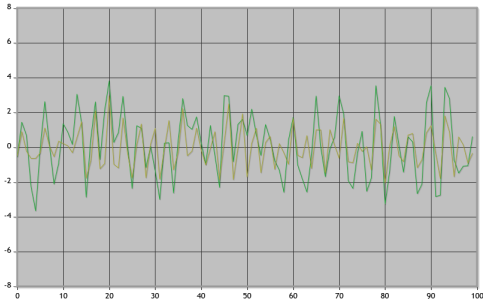

Figure 5.22: AR(4) Prediction of ARMA$(2, 2)$ process with complex AR poles with small negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.6\pi$, the actual zeros: $-0.6, -0.9$.

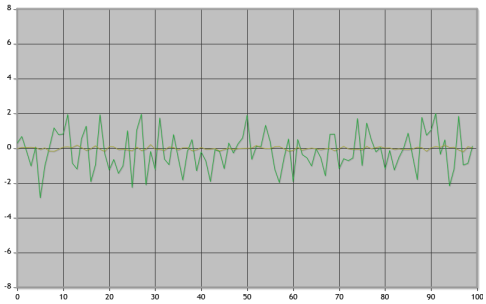

Figure 5.23: AR(4) Prediction of ARMA$(2, 2)$ process with complex AR poles with large negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.9\pi$, the actual zeros: $-0.6, -0.9$.

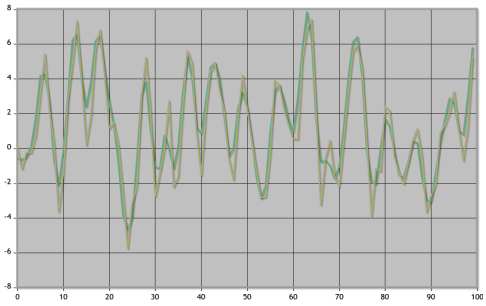

Figure 5.24: Prediction of AR(4) process with two positive poles and two almost unstable complex pose whose real part is positive with 10% perturbation of the coefficients. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.3\pi$.

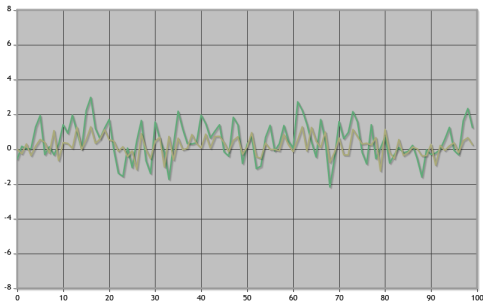

Figure 5.25: Prediction of AR(4) process with two positive poles and two almost unstable complex pose whose real part is negative with 10% perturbation of the coefficients. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.6\pi$.

5.8 ARMA processes with unstable zeros

Let us now consider the problem of predicting an ARMA process when $C(z^{-1})$ not necessarily stable.

An innocent looking example is

$$y_n = e'_n - c'e'_{n-1}$$

with $|c'| > 1$. The obvious thing to do is to find an alternative representation of (y_n) in terms of its innovation process, which we denote by (e_n). To see how this can be done we consider a general MA process (y_n) given by

$$y = C'(q^{-1})e' \tag{5.21}$$

where C' is a polynomial of q^{-1} and (e'_n) is a w.s.st. orthogonal process. Assuming

$\sigma^2(e') = 1$ the spectral density of (y_n) is then given by

$$f(\omega) = |C'(e^{-i\omega})|^2.$$

This can also be obtained by restricting the complex-function

$$g(z) = C'(z^{-1})C'(z)$$

to $|z| = 1$, since the coefficients of C' are assumed to be real. Let us now assume that $C'(z^{-1})$ has an unstable root, say γ'. Then factorizing $C'(z^{-1})$ we will have a factor of the form

$$c'(z^{-1}) = 1 - \gamma' z^{-1}.$$

The effect of this factor in $g(z)$ is

$$c'(z^{-1})c'(z) = (1 - \gamma' z^{-1})(1 - \gamma' z) = (z - \gamma')(z^{-1} - \gamma'). \tag{5.22}$$

Thus the zeros of the above rational function are γ' and $1/\gamma'$. It is now clear that we can swap the role of γ' and $1/\gamma'$ without changing the function (5.22). Indeed, setting

$$c(z^{-1}) = (1 - (\gamma')^{-1} z^{-1}) \cdot \gamma' = \gamma' - z^{-1}$$

we have

$$c'(z^{-1})c'(z) = (z - \gamma')(z^{-1} - \gamma') = c(z)c(z^{-1}) = c(z^{-1})c(z).$$

The main benefit of this transformation is that the (first order) polynomial $c(z^{-1})$ is now stable. Replacing all factors of $C'(z^{-1})$ by stable ones we come to the following conclusion.

Proposition 5.11. *Let $C'(z^{-1})$ be a polynomial such that $C'(z^{-1}) \neq 0$ for $|z| = 1$. Then $\exists$ a stable polynomial $C(z^{-1})$ with $\deg C = \deg C'$ such that*

$$C'(z^{-1})C'(z) = C(z^{-1})C(z). \tag{5.23}$$

The decomposition of the spectral density of y in the form given by the r.h.s. of (5.23) is called *spectral factorization*, and $C(e^{-i\omega})$ is called *a stable spectral factor*. Setting $z = e^{i\omega}$ we get a factorization of the spectral density of y, denoted by $f(\omega)$, as

$$f(\omega) = C(e^{-i\omega})C(e^{i\omega}) = |C(e^{-i\omega})|^2.$$

Now, to find the innovation process of y, we will not invert (5.21), but rather define a new process e by

$$d\zeta^e(\omega) = \frac{1}{C(e^{-i\omega})} d\zeta^y(\omega). \tag{5.24}$$

It is readily seen that the r.h.s. is well defined since $1/C(e^{-i\omega})$ is in $L_2^c(dF^y(\omega))$. Moreover the spectral density of e is

$$\left|\frac{C'(e^{-i\omega})}{C(e^{-i\omega})}\right|^2 = 1.$$

Thus e is a w.s.st. orthogonal process. Since

$$y = C(q^{-1})e,$$

with $C(z^{-1})$ stable, e is the innovation process of y. To summarize we obtained the following result:

Proposition 5.12. *Let $y = (y_n)$ be an MA process given in (5.21). Assume that $C'(z^{-1}) \neq 0$ for $|z| = 1$. Let $C(e^{-i\omega})$ be the stable spectral factor of $f(\omega)$, and define $e = (e_n)$ by (5.24). Then e is a w.s.st. orthogonal process,*

$$y = C(q^{-1})e,$$

and e is the innovation process of y.

The above procedure can be extended to ARMA processes in a straightforward manner. Let (y_n) be a w.s.st. ARMA process given by

$$A'(q^{-1})y = C'(q^{-1})e',$$

where the polynomials $A'(z^{-1})$ and $C'(z^{-1})$ are not necessarily stable, but $A'(z^{-1}) \neq 0$ and $C'(z^{-1}) \neq 0$ for $|z| = 1$, and the process e' is a w.s.st. orthogonal process with $\sigma^2(e') = 1$. The spectral density of y is then given by

$$f(\omega) = \left|\frac{C'(e^{-i\omega})}{A'(e^{-i\omega})}\right|^2.$$

Let $A(e^{-i\omega})$ and $C(e^{-i\omega})$ be the stable spectral factors of the denominator and the numerator, respectively. Then

$$f(\omega) = \left|\frac{C'(e^{-i\omega})}{A'(e^{-i\omega})}\right|^2 = \left|\frac{C(e^{-i\omega})}{A(e^{-i\omega})}\right|^2.$$

The rational function $C(e^{-i\omega})/A(e^{-i\omega})$ is called a *stable spectral factor* of f. Now define the w.s.st. process e by

$$d\zeta^e(\omega) = \frac{A(e^{-i\omega})}{C(e^{-i\omega})}d\zeta^y(\omega) = \frac{A(e^{-i\omega})}{C(e^{-i\omega})} \cdot \frac{C'(e^{-i\omega})}{A'(e^{-i\omega})}d\zeta^{e'}(\omega).$$

Note that the transfer function

$$G(e^{-i\omega}) = \frac{A(e^{-i\omega})}{C(e^{-i\omega})} \cdot \frac{C'(e^{-i\omega})}{A'(e^{-i\omega})}$$

is such that

$$G(e^{-i\omega})G(e^{i\omega}) = |G(e^{-i\omega})|^2 = 1 \tag{5.25}$$

for all ω. A transfer function satisfying (5.25) is called *all-pass*, indicating that all frequencies are passed through the filter corresponding to G with unchanged energy.

It is readily seen that the new process $e = (e_n)$ is a w.s.st. orthogonal process. This is formally and generally stated in the next exercise.

Exercise 5.7. *Let G be an all-pass transfer function, and let e' be a w.s.st. orthogonal process. Then the process e defined by*

$$d\zeta^e(\omega) = G(e^{-i\omega})d\zeta^{e'}(\omega)$$

is also a w.s.st. orthogonal process.

The simplest example for an all-pass function was obtained above in factoring the spectral density of a MA(1) process. This is obtained by taking a first order polynomial $C'(z^{-1}) = 1 - \gamma' z^{-1}$, and swapping its zero γ' for $(\gamma')^{-1}$, to get $C(z^{-1}) = \gamma' - z^{-1}$. Then

$$G(z^{-1}) = \frac{C'(z^{-1})}{C(z^{-1})} = \frac{1 - \gamma' z^{-1}}{\gamma' - z^{-1}}$$

is all-pass.

Remark. Our result on the spectral factorization of $C'(z^{-1})C'(z)$ does not cover the case when $C'(z^{-1}) = 0$ for some $|z| = 1$. Consider e.g. the process

$$y_n = e_n - e_{n-1},$$

where now $C'(z^{-1}) = 1 - z^{-1}$, and thus $C'(1) = 0$. To reconstruct e from y we consider the corresponding spectral representation yielding

$$d\zeta^e(\omega) = \frac{1}{1 - e^{-i\omega}} d\zeta^y(\omega).$$

The r.h.s. is well defined in frequency domain. To get a time domain representation of e in terms of y consider the approximating process $e_n(c)$ defined by

$$d\zeta^{e(c)}(\omega) = \frac{1}{1 - ce^{-i\omega}} d\zeta^y(\omega),$$

with $|c| < 1$. It is easy to see that

$$\lim_{c \nearrow 1} \frac{1}{1 - ce^{-i\omega}} = \frac{1}{1 - e^{-i\omega}}$$

in $L_2^c(|C(e^{-i\omega})|^2(d\omega))$, and also

$$\frac{1}{1 - ce^{-i\omega}} = \sum_{k=0}^{\infty} c^k e^{-ik\omega}$$

in $L_2^c(|C(e^{-i\omega})|^2(d\omega))$. It follows that

$$e_n = \lim_{c \nearrow 1} \sum_{k=0}^{\infty} c^k y_{n-k}.$$

Thus $e_n \in H_n^y$, and it follows that e is the innovation process of y. In particular, the one-step ahead predictor of y is given by

$$\widehat{y}_n = e_{n-1} = \lim_{c \nearrow 1} \sum_{k=0}^{\infty} c^k y_{n-1-k}.$$

Chapter 6

Multivariate time series

6.1 Vector valued wide sense stationary processes

Let us now consider the situation when we consider the price movements of several commodities simultaneously. Let the number of commodities be s, and let the $\mathbb{R}^s$-valued price-vector at time n be y_n. If there is an interaction between individual prices, which is often the case, then we expect to get better predictions for the individual price processes when treating them simultaneously. Therefore we need a theory of vector-valued wide sense stationary processes. The next definition is a straightforward extension of the definition for the scalar case.

Definition 6.1. *The $\mathbb{R}^s$-valued stochastic process (y_n), $-\infty < n < +\infty$ is called* wide sense stationary *if $\mathrm{E}|y_n|^2 < \infty$ for all n, $\mathrm{E}y_n = 0$ for all n, and the covariance matrix*

$$R(\tau) = \mathrm{E}(y_{n+\tau} y_n^\top)$$

is independent of n.

The matrix-valued function $(R(\tau))$ is called the *auto-covariance function* of (y_n). Obviously, we have

$$R(-\tau) = R(\tau)^\top.$$

As in the scalar case, the condition $\mathrm{E}y_n = 0$ can be replaced by the condition that $\mathrm{E}y_n = m$ with some fixed vector $m \in \mathbb{R}^s$ for all n.

The definition extends to $\mathbb{C}^s$-valued (complex) processes in a natural manner by requiring that

$$R(\tau) = \mathrm{E}(y_{n+\tau} \overline{y_n}^\top)$$

is independent of n. In this case we have

$$R(-\tau) = \overline{R}(\tau)^\top.$$

As in the scalar case, an eminent role is played by what are called wide sense stationary orthogonal processes.

Definition 6.2. *The $\mathbb{R}^s$-valued stochastic process (e_n) is called a* wide sense stationary orthogonal process *if it is wide sense stationary, and in particular*

$$\mathrm{E}e_{n+\tau}e_n^\top = 0 \quad \text{for} \quad \tau \neq 0 \quad \text{and} \quad \mathrm{E}e_n e_n^\top = \Sigma \quad \text{for all} \quad n,$$

where Σ is a fixed, $s \times s$ symmetric positive semi-definite matrix.

Note that Σ may be equal to any symmetric positive semi-definite matrix, i.e. (e_n) should not be normalized so that its covariance matrix is I.

Now all the results that we had for $\mathbb{R}$-valued or $\mathbb{C}$-valued wide sense stationary processes can be generalized to (real or complex) vector-valued, wide sense stationary processes. Let (y_n) be a $\mathbb{R}^s$-valued w.s.st. process.

If the components of y_n denote the prices of some commodities then the expectation of the next day value of a commodity, the price of which is correlated to the components of y_n can be reasonably expressed by taking a set of vectors $a_1, \dots a_p$ in $\mathbb{R}^s$, and defining

$$z_n = \sum_{k=1}^{p} a_k^\top y_{n-k}.$$

Exercise 6.1. *Show that (z_n) is an $\mathbb{R}^s$-valued wide sense stationary process and we have*

$$\mathrm{E}z_n^2 = \sum_{k=1}^{p}\sum_{l=1}^{p} a_k^\top R(k-l)a_l \geq 0. \tag{6.1}$$

Thus the *block matrix* R defined by the blocks

$$R_{k,l} = R(k-l), \qquad k,l = 1,\dots,p \tag{6.2}$$

is positive semi-definite. The size of R is $(ps) \times (ps)$.

Definition 6.3. *A $ps \times ps$ matrix R consisting of $m \times m$ blocks satisfying (6.2) is called a* block-Toeplitz matrix.

Note that a block-Toeplitz matrix is not necessarily Toeplitz in the usual sense, since already the diagonal $(1,1)$ block, equal to $R(0) = \mathrm{E}(y_n y_n^\top)$, is not Toeplitz in the usual sense.

Definition 6.4. *Let $R(\tau)$, $-\infty < \tau < +\infty$ be a sequence of $s \times s$ matrices such that $R(-\tau) = R(\tau)^\top$. Then $(R(\tau))$ is called* positive semi definite sequence, *if the associated block-Toeplitz matrix defined by (6.2) is positive semi-definite for all p.*

Thus we came to the following conclusion:

Proposition 6.5. *The auto-covariance matrices $R(\tau),\ -\infty < \tau < +\infty$ of a vector-valued wide sense stationary process (y_n) form a positive semi definite sequence.*

Exercise 6.2. *Prove the converse statement: let $R(\tau),\ -\infty < \tau < +\infty$ be a positive definite sequence of real-valued, $s \times s$ matrices. Then it is the auto-covariance sequence of an $\mathbb{R}^s$-valued, wide sense stationary Gaussian process.*

6.2 Prediction and the innovation process

Let now (y_n) be an $\mathbb{R}^s$-valued wide sense stationary process. To define the history of (y_n) up to time $n-1$ expressed via a Hilbert space we consider first the linear space of $\mathbb{R}$-valued (!) random variable -s

$$\mathcal{L}^y_{n-1} = \left\{ \sum_{k=1}^{p} a_k^\top y_{n-k},\ \text{for some } p,\ \text{and } a_k \in \mathbb{R}^s \right\}.$$

Thus $\mathcal{L}^y_{n-1}$ is a subspace of $L_2(\Omega, \mathcal{F}, P)$. We define H^y_{n-1} as the closure of $\mathcal{L}^y_{n-1}$ in $L_2(\Omega, \mathcal{F}, P)$. Note once again, that H^y_{n-1} is thus a Hilbert space of real-valued random variables. Defining the past of (y_n) this way is natural when thinking of linear prediction. Namely, it would be unnatural to define the past of y via the linear space consisting of $\mathbb{R}^s$-valued random variables $w = \sum_{k=1}^{p} a_k y_{n-k}$ for some p with scalar-valued a_k-s, and thus significantly restricting the range of available predictors.

Let $L^s_2(\Omega, \mathcal{F}, P)$ denote the Hilbert-space of $\mathbb{R}^s$-valued random variables x such that

$$\mathrm{E}|x|^2 = \mathrm{E}x^\top x < \infty.$$

The projection of a random variables x in $L^s_2(\Omega, \mathcal{F}, P)$ onto H^y_{n-1} will be defined componentwise:

$$\widehat{x} = (x|H^y_{n-1}) = \left((x_1|H^y_{n-1}), \ldots, (x_s|H^y_{n-1})\right)^\top.$$

Then for the error vector $\tilde{x}$ we have

$$\tilde{x} = x - \widehat{x} \perp H^y_{n-1}$$

where orthogonality is meant componentwise, i.e. we have for all $k = 1, ..., s$

$$\tilde{x}_k \perp H^y_{n-1}.$$

With this preparation we can now define innovation process (e_n) via

$$e_n = y_n - (y_n|H^y_{n-1}). \tag{6.3}$$

A vector valued process (y_n) is called *singular*, if its innovation process is an identically 0 process, or equivalently if

$$H^y_n = H^y_{n-1} \text{ for all } n. \tag{6.4}$$

Exercise 6.3. *Show that $e_n \equiv 0$ is indeed equivalent to (6.4).*

A novel phenomenon that we did not have in the scalar case is that the covariance-matrix of e_n, say

$$\Sigma = \mathrm{E} e_n e_n^\top$$

may be non-zero, but singular. If $\alpha \in \mathbb{R}^s$ is a non-zero vector such that $\Sigma\alpha = 0$, then

$$\alpha^\top \Sigma \alpha = \alpha^\top \mathrm{E} e_n e_n^\top \alpha = \mathrm{E}(\alpha^\top e_n)^2 = 0$$

implies that

$$\alpha^\top e_n = 0$$

w.p.1. It follows that the process

$$z_{\alpha,n} = \alpha^\top y_n$$

is singular. Assuming that, say, $\alpha_1 \neq 0$, we can express $y_{1,n}$ with arbitrary accuracy using its own strict past and the history of $y_2, ..., y_s$ up to time n.

6.3 Spectral theory

In this section the spectral theory for multivariate wide sense stationary processes will be discussed briefly, with the main steps of the proof. The first step is the appropriate representation of the auto-covariances $R(\tau)$ extending Herglotz's theorem.

Assume first that the auto-covariance function of (y_n) satisfies

$$\sum_{\tau=-\infty}^{\infty} \|R(\tau)\|^2 < \infty, \tag{6.5}$$

where $\|R\|$ denotes the operator norm of the matrix R, i.e.

$$\|R\| = \max_{x \neq 0} |Rx|/|x|.$$

Theorem 6.1. *Let (y_n) be an $\mathbb{R}^s$-valued wide sense stationary process, and let $R(\tau)$ be its auto-covariance function. Assume that $R(\tau)$ satisfies (6.5). Then we have*

$$R(\tau) = \frac{1}{2\pi} \int_0^{2\pi} e^{i\tau\omega} f(\omega) d\omega, \tag{6.6}$$

where $f(\omega)$ is a symmetric, positive semi-definite matrix-valued function in $L_2^{s\times s}(d\omega)$.

Proof. (Outline.) Let $\alpha \in \mathbb{R}^s$ and consider the scalar process

$$z_{\alpha,n} = \alpha^\top y_n.$$

If the components of y_n denote the prices of various stocks at time n, and the components of α denote the amounts of stocks held by an investor, (allowing negative components, i.e. allowing short positions), then $\alpha^\top y_n$ is the value of the portfolio at time n. The auto-covariance function of $z_{\alpha,n}$ is

$$r^\alpha(\tau) = \alpha^\top R(\tau)\alpha.$$

It is obvious by (6.5) that $\sum_{\tau=-\infty}^{\infty} (r^\alpha(\tau))^2 < +\infty$, hence, by the special case of the scalar version of Herglotz's theorem, we have

$$r^\alpha(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\omega\tau} f^\alpha(\omega) d\omega,$$

where $f^\alpha(\omega) \geq 0$ is the spectral density corresponding to $r^\alpha(\tau)$. We also know how to get $f^\alpha(\omega)$ from $r^\alpha(\omega)$ explicitly via

$$f^\alpha(\omega) = \sum_{\tau=-\infty}^{\infty} r^\alpha(\tau)\, e^{-i\tau\omega}.$$

Here convergence on the r.h.s. is meant in $L_2^c(d\omega) = L_2^c([0, 2\pi], d\omega)$. Substituting $r^\alpha(\tau) = \alpha^\top R(\tau)\alpha$ we get

$$f^\alpha(\omega) = \sum_{\tau=-\infty}^{\infty} \alpha^\top R(\tau)\alpha\, e^{-i\tau\omega}\,. \tag{6.7}$$

Taking finite truncations of the right hand side of (6.7) we get that

$$\sum_{\tau=-N}^{N} \alpha^\top R(\tau)\alpha e^{-i\tau\omega} = \alpha^\top (\sum_{\tau=-N}^{N} R(\tau)e^{-i\tau\omega})\alpha \tag{6.8}$$

converges in $L_2(d\omega)$ for any α. From here we would like to conclude that

$$f_N(\omega) = \sum_{\tau=-N}^{N} R(\tau)e^{-i\tau\omega}$$

itself converges in $L_2^{s\times s}(d\omega)$.

Exercise 6.4. *Prove that a quadratic form $\alpha^\top F\alpha$, with F symmetric, determines the bilinear form corresponding to F uniquely as*

$$\beta^\top F\gamma = \frac{1}{4}\left((\beta+\gamma)^\top F(\beta+\gamma) - (\beta-\gamma)^\top F(\beta-\gamma)\right). \tag{6.9}$$

Taking $F = f_N$, and taking any pair of unit vectors in $\mathbb{R}^s$, say, $\beta = e_k, \gamma = e_l$ we conclude that

$$\sum_{\tau=-\infty}^{\infty} R(\tau)e^{-i\tau\omega} =: f(\omega)$$

converges in $L_2(d\omega)$ componentwise, and thus also in $L_2^{s\times s}(d\omega)$. It follows that we have

$$f^\alpha(\omega) = \alpha^\top f(\omega)\alpha \quad \text{for} \quad \text{any} \quad \alpha \in \mathbb{R}^s.$$

Obviously $f(\omega)$ is symmetric, and $f^\alpha(\omega) \geq 0$ for any α implies that $f(\omega)$ is positive semi-definite, and this concludes the proof. □

In the general case we expect to get a representation of the form

$$R(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\tau\omega}dF(\omega), \tag{6.10}$$

where $F(\omega)$ is a matrix-valued function, which is monotone-nondecreasing in some sense. This is indeed the case, as stated in the next theorem:

Theorem 6.2. *Let (y_n) be an $\mathbb{R}^s$-valued wide sense stationary process, and let $R(\tau)$ be its autocovariance function. Then we have*

$$R(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\tau\omega}dF(\omega), \tag{6.11}$$

where $F(\omega)$ is a matrix-valued function such that the increments of $F(\cdot)$ are symmetric positive definite matrices. The elements of the matrix-valued function $F(\cdot)$ are functions of finite variations, and thus the above integral is defined as a Riemann-Stieltjes integral. We can also assume that $F(\cdot)$ is l.c. and that $F(0) = 0$.

Proof. The proof follows the logic of the proof for the scalar case. Consider the truncated sequences

$$R_N(\tau) = \begin{cases} R(\tau) & \text{for} \quad |\tau| \leq N \\ 0 & \text{otherwise.} \end{cases}$$

Then $R_N(\tau)$ is a positive definite sequence of $s \times s$ real matrices, for which the condition of Theorem 6.1, namely condition (6.5), is satisfied. Thus, by Theorem 6.1, we can write

$$R_N(\tau) = \frac{1}{2\pi}\int_0^{2\pi} e^{i\omega\tau} f_N(\omega)d\omega$$

with $f_N(\omega)$ symmetric, positive semidefinite. Also we have

$$R(0) = R_N(0) = \frac{1}{2\pi}\int_0^{2\pi} f_N(\omega)d\omega$$

for all N.

Now we would like to select a subsequence (N_k) such that the matrix-valued measures $f_{N_k}(\omega)d\omega$ converges weakly to some matrix-valued measure $dF(\omega)$. The simplest way to do this is to refer to weak convergence theory of measures. □

A more elementary argument, using Helly's theorem, is given below. For the sake of convenience we formulate the possibility of such a selection in the lemma below.

Lemma 6.6. *There exists a single subsequence (N_k) such that for all $j, l = 1, ..., s$ the measures $f_{N_k,j,l}(\omega)\, d\omega$ converge weakly to some measure $dF_{j,l}(\omega)$, formally written as*

$$f_{N_k,j,l}(\omega)\, d\omega \quad \Rightarrow \quad dF_{j,l}(\omega).$$

Here the matrix-valued function $F(\cdot) = (F_{j,l}(\cdot))$ is such that its increments are symmetric, positive semi-definite.

Proof. First, take any fixed $\alpha \in \mathbb{R}^s$, and consider the scalar-valued positive functions $\alpha^\top f_N(\omega)\alpha$. By the above equality we have for all N

$$\alpha^\top R(0)\alpha = \alpha^\top R_N(0)\alpha = \frac{1}{2\pi}\int_0^{2\pi} \alpha^\top f_N(\omega)\alpha d\omega.$$

Since the measures $\alpha^\top f_N(\omega)\alpha d\omega$ are concentrated on $[0, 2\pi]$ there exists a subsequence (N_k) such that the measures $\alpha^\top f_{N_k}(\cdot)\alpha\ d\omega$ converge weakly to some measure $dF_\alpha(\cdot)$, defined by a monotone non-decreasing function $F_\alpha(\cdot)$, i.e.

$$\alpha^\top f_{N_k}(\omega)\alpha\, d\omega \quad \Rightarrow \quad dF_\alpha(\omega).$$

Using the fact once again that a quadratic form uniquely determines the corresponding bilinear form as given in (6.9), it follows that for any fixed pair $\beta, \gamma \in \mathbb{R}^s$ there exists a subsequence (N_k) such that

$$\beta^\top f_{N_k}(\omega)\gamma\, d\omega \quad \Rightarrow \quad dF_{\beta,\gamma}(\omega),$$

where $F_{\beta,\gamma}(\omega)$ is now the difference of two monotone non-decreasing functions. Obviously, $F_{\beta,\gamma}(\cdot)$ is a function of finite variation.

Setting $\beta = e_j$, $\gamma = e_l$, with e_j, e_l denoting unit vectors, and letting j, l vary over $j, l = 1, \ldots, s$, we conclude that there exists a single subsequence (N_k) such that

$$f_{N_k,j,l}(\omega)\, d\omega \quad \Rightarrow \quad dF_{j,l}(\omega).$$

Let $F(\omega)$ denote the matrix with elements $F_{j,l}(\omega)$.

To prove that the increments of F are symmetric, positive semi-definite, note that with the above single subsequence N_k we have that for **any** $\alpha \in \mathbb{R}^s$

$$\alpha^\top f_{N_k}(\omega)\alpha d\omega = \sum_{j,\ l} \alpha_j \alpha_l f_{N_k,j,l}(\omega)\, d\omega \quad \Rightarrow \quad \sum_{j,l} \alpha_j \alpha_l dF_{j,\ l}(\omega).$$

Compactly written this reads as

$$\alpha^\top f_{N_k}(\omega)\alpha d\omega \quad \Rightarrow \quad \alpha^\top\ dF(\omega)\ \alpha.$$

Since the matrix $f_{N_k}(\omega)$ is symmetric, positive semi-definite for all ω, it follows that the increments of $\alpha^\top\ F(\omega)\ \alpha$ are non-negative for any $\alpha \in \mathbb{R}^s$. Thus the increments of $F(\cdot)$ are symmetric, positive semi-definite, and the proof of the lemma is complete. □

Now, weak convergence implies that the integrals of any bounded, continuous function converge, in particular for the function $e^{i\omega\tau}$ we have

$$\lim_k \int_0^{2\pi} e^{i\omega\tau} f_{N_k,j,l}(\omega)d\omega = \int_0^{2\pi} e^{i\omega\tau} dF_{j,l}(\omega).$$

In compact form we can write this as

$$\lim_k \int_0^{2\pi} e^{i\omega\tau} f_{N_k}(\omega) = \int_0^{2\pi} e^{i\omega\tau} dF(\omega),$$

Now, since the l.h.s. equals $R(\tau)$ for $N_k \geq \tau$, the required representation of $R(\tau)$ follows.

Finally, as in the case of the definition of a probability measure via a probability distribution, in defining the measure $dF_{j,l}(\omega)$ via $F_{j,l}(\omega)$ we have the freedom to choose $F_{j,l}(\cdot)$ l.c. (or r.c.). If we choose $F_{j,l}(\cdot)$ to be l.c. (corresponding to defining the probability distribution function as $P(\xi < x)$), then we may assume $F_{j,l}(0) = 0$. This completes the proof.

Remark. Note that if the $dF_{j,l}$-measure of the single point $\{0\}$ happens to be positive, then $F_{j,l}(\cdot)$ will be discontinuous at $\omega = 0$.

Exercise 6.5. * *Let $F(\cdot)$ be an $s \times s$ matrix-valued function on $[0, 2\pi]$ such that the increments $F(\cdot)$ are symmetric and positive semidefinite. Then for any $k, l = 1, \ldots, s$ the elements $F_{k,l}(\omega)$ are of finite variations.*

If the measure $dF(\omega)$ has a density, i.e. if

$$dF(\omega) = f(\omega)d\omega$$

then we have

$$R(\tau) = \frac{1}{2\pi} \int_0^{2\pi} e^{i\tau\omega} f(\omega) d\omega.$$

The function f is called *the spectral density*. A key property of $f(\omega)$ is that it is symmetric and positive semidefinite a.s., i.e.

$$f(\omega) \geq 0 \qquad \text{a.s.}$$

Exercise 6.6. *Show that for an $\mathbb{R}^s$-valued orthogonal wide sense stationary process (e_n) with covariance matrix $\Lambda = \mathrm{E} e_n e_n^\top$ we have*

$$f(\omega) = \Lambda \quad \text{for} \quad \forall \omega.$$

6.4 Filtering

Let us now consider the effect of filtering. Let (y_n) be an $\mathbb{R}^s$-valued wide sense stationary process and define

$$v_n = \sum_{k=0}^{p} h_k y_{n-k},$$

where the h_k-s are $r \times s$ matrices. Define the matrix-valued frequency response function

$$H(e^{-i\omega}) = \sum_{k=0}^{p} h_k e^{-i\omega k}.$$

Then we have the following result:

Proposition 6.7. *The spectral distribution of the process v is given by*

$$dF^v(\omega) = H(e^{-i\omega}) dF^y(\omega) H(e^{i\omega})^\top.$$

Exercise 6.7. *Prove Proposition 6.7.*

To extend the above result from FIR filters to the general case, i.e. to filters with infinite number of impulse responses we should ask ourselves: how do we associate a Hilbert-space with the matrix-valued measure dF? The natural choice is to take $\mathbb{R}^s$-valued or $\mathbb{C}^s$-valued measurable functions. Consider the set of $\mathbb{C}^s$-valued measurable functions $g(\omega)$ such that their squared norm defined as

$$\int_0^{2\pi} \overline{g(\omega)}^\top dF(\omega) g(\omega) d\omega = \int_0^{2\pi} g^\top(\omega) dF(\omega) \overline{g(\omega)}$$

is finite. The space of these functions will be denoted by $L_2^{c,s}(dF)$.

Extending the above definition, we may similarly define a Hilbert-space $L_2^{c,r\times s}(dF)$ of $r \times s$ matrices. Let $k(\omega)$ be a measurable function with its values being $r \times s$ matrices with *complex* entries, with $0 \le \omega \le 2\pi$. We say that $k(\omega)$ is in $L_2^{c,r\times s}(dF)$ if each row of $k(\omega)$ is in $L_2^{c,s}(dF)$, or equivalently, if

$$\int_0^{2\pi} \operatorname{tr}\, k(\omega) dF(\omega) \overline{k(\omega)}^\top < +\infty.$$

Having defined $L_2^{c,r\times s}(dF)$ we can now extend the previous result for filters with infinite number of impulse responses. So let us consider the linear filter of the form

$$v_n = \sum_{k=0}^{\infty} h_k y_{n-k}, \tag{6.12}$$

where the impulse responses h_k are $r \times s$ real matrices. Consider the associated (matrix-valued) frequency response function

$$H(e^{-i\omega}) = \sum_{k=0}^{\infty} h_k e^{-i\omega k}. \tag{6.13}$$

Then we have the following result:

Proposition 6.8. *Assume that the right hand side of (6.13) converges in $L_2^{c,r\times s}(dF)$. Then the process (v_n) under (6.12) is well-defined, i.e. the right hand side converges in $L_2^{c,r}(\Omega, \mathcal{F}, P)$, and the spectral distribution of (v_n) is given by*

$$dF^v(\omega) = H(e^{-i\omega}) dF^y(\omega) H^\top(e^{i\omega}).$$

Exercise 6.8. *Prove the above proposition following the proof for the scalar case.*

6.5 Multivariate random orthogonal measures

To describe the spectral representation of the process (y_n) itself we need the concept of $\mathbb{C}^s$-valued random orthogonal measures. Let $\zeta(\omega)$, $0 \le \omega \le 2\pi$ be a $\mathbb{C}^s$-valued, measurable* stochastic process such that for all ω we have $\zeta(\omega) \in L_2^{c,s}(\Omega, \mathcal{F}, P)$, or equivalently,

$$\mathrm{E}\zeta(\omega)^\top \overline{\zeta}(\omega) < \infty.$$

Assume that $\zeta(0) = 0$, and that $\zeta(\omega)$ is a zero-mean process, i.e. $\mathrm{E}\zeta(\omega) = 0$.

Definition 6.9. *The stochastic process $\zeta(\cdot)$ with the above properties is called a process with orthogonal increments if for any two non-overlapping intervals, defined via $0 \le a < b \le c < d \le 2\pi$, the covariance matrix of the increments is 0, i.e.*

$$\mathrm{E}(\zeta(b) - \zeta(a))(\overline{\zeta}(d) - \overline{\zeta}(c))^\top = 0 \in \mathbb{R}^{s\times s}.$$

The (matrix-valued) structure function corresponding to $\zeta(\omega)$ is defined as

$$F(\omega) = \mathrm{E}\zeta(\omega)\overline{\zeta(\omega)}^{\top}.$$

Integration with respect to a random orthogonal measure is defined by a straightforward extension of the scalar case. The most general problem of integration would be to ask ourselves, how to define integrals of the form

$$I(k) = \int_0^{2\pi} k(\omega) d\zeta(\omega)$$

where $k(\omega)$ is an $r \times s$ matrix. For a start we consider the simpler problem of integrating a vector-valued function g with values, say in C^s. We find that

$$I(g) = \int_0^{2\pi} g^{\top}(\omega) d\zeta(\omega)$$

is well-defined for any $g \in L_2^{c,s}(dF)$, defined as the set of measurable, C^s-valued functions such that

$$\int_0^{2\pi} g^{\top}(\omega) dF(\omega) \overline{g(\omega)} < \infty.$$

Then the following isometry property holds:

Theorem 6.3. *We have for any $g \in L_2^{c,s}(dF)$*

$$\mathrm{E}I(g)\overline{I(g)} = \mathrm{E}|I(g)|^2 = \int_0^{2\pi} g^{\top}(\omega) dF(\omega) \overline{g(\omega)}.$$

Thus I is an isometry from $L_2^{c,s}(dF)$ to $L_2^{c,s}(\Omega, \mathcal{F}, P)$.

Let now $k(\cdot)$ be an $r \times s$ matrix. We say that $k(\cdot)$ belongs to $L_2^{c,r\times s}(dF)$, if all rows of $k(\cdot)$ belong to $L_2^{c,s}(dF)$. Then the vector-valued stochastic integral

$$I(k) = \int_0^{2\pi} k(\omega) d\zeta(\omega)$$

is well defined, simply taking integration row-wise. However, the isometry property of stochastic integration now takes on a new interesting form.

Theorem 6.4. *Let $k(\omega)$ be an $r \times s$ matrix such that $k(\cdot) \in L_2^{c,r\times s}$. Then we have the matrix equality*

$$\mathrm{E}\ I(k)\ I(k)^* = \int_0^{2\pi} k(\omega) dF(\omega) \overline{k(\omega)}^{\top},$$

with * denoting transposition and conjugation.

Proof. Let $y = I(k)$, and let $\alpha \in \mathbb{R}^r$ be an arbitrary vector. Consider the random variable

$$\alpha^\top y = \int_0^{2\pi} \alpha^\top k(\omega) d\zeta(\omega).$$

Since for the vector-valued function $\alpha^\top k(\cdot)$ we have $\alpha^\top k(\cdot) \in L_2^{c,s}(dF)$, we have, by the isometry property of scalar-valued stochastic integration, given as Theorem 6.4,

$$\mathrm{E}|\alpha^\top y|^2 = \int_0^{2\pi} \alpha^\top k(\omega) dF(\omega) \overline{k(\omega)}^\top \alpha.$$

The left hand side can be written as $\alpha^\top \mathrm{E} y y^\top \alpha$. Since α is arbitrary, the claim follows.□

An interesting special case is the integration of a scalar-valued function, say g. The integral

$$I(g) = \int_0^{2\pi} g(\omega) d\zeta(\omega)$$

could be interpreted componentwise, if $g(\cdot) \in L_2^c(dF_{l,l})$ for all $l = 1, ..., s$, or, equivalently, if $g(\cdot) \in L_2^c(\mathrm{tr}\ dF)$. Thus we get a vector-valued, more exactly C^s-valued, integral $I(g)$. However, the interaction between the components of $d\zeta(\omega)$ and $I(g)$ is not reflected in this componentwise procedure.

An alternative, better option is to write the integral above as

$$I(g) = \int_0^{2\pi} g(\omega) \cdot I\ d\zeta(\omega),$$

where I is an $s \times s$ identity matrix. Then, setting $k(\omega) = g(\omega) \cdot I$, the integrability condition would read as follows: for each $l = 1, ..., s$ we should have $g(\cdot) e_l \in L_2^{c,s}(dF)$, (with e_l denoting the l-th unit vector). This is equivalent to saying that $g(\cdot) \in L_2^c(dF_{l,l})$ for each l, or briefly, $g(\cdot) \in L_2^c(\ \mathrm{tr} dF)$, just as above, which should not be a surprise. Now, taking into account the isometry property given as Theorem 6.4 we get the following result:

Theorem 6.5. . *Let $g(\cdot)$ be a C-valued function such that $g(\cdot) \in L_2^c(\mathrm{tr}\ dF)$. Then for the stochastic integral*

$$I(g) = \int_0^{2\pi} g(\omega) \cdot I\ d\zeta(\omega),$$

that can be interpreted as an integration componentwise, we have the matrix-equality

$$\mathrm{E}\ I(g)\ I(g)^* = \int_0^{2\pi} |g(\omega)|^2 dF(\omega).$$

All the above results stating various forms of isometry can be extended from quadratic forms to bilinear forms. Thus, e.g., the last result would yield: if $g(\cdot)$ and $h(\cdot)$ are C-valued functions such that $g(\cdot), h(\cdot) \in L_2^c(\text{tr } dF)$, then we have the matrix-equality

$$\text{E } I(g)\ I(h)^* = \int_0^{2\pi} g(\omega)\ \overline{h}(\omega) dF(\omega). \tag{6.14}$$

Especially when choosing $g = e^{in\omega}$ and $h = e^{im\omega}$, we get the following beautiful generalization of the corresponding scalar result:

Theorem 6.6. *Let $d\zeta(\omega)$ be a $\mathbb{C}^s$-valued random orthogonal measure, with structure function $dF(\omega)$. Then the $\mathbb{C}^s$-valued process*

$$y_n = \int_0^{2\pi} e^{in\omega} d\zeta(\omega)$$

is wide sense stationary, and its spectral distribution function is given by

$$dF^y(\omega) = 2\pi\, dF(\omega).$$

Exercise 6.9. *Prove Theorem 6.6.*

6.6 The spectral representation theorem

The centerpiece of spectral theory is the following spectral representation theorem, extending the corresponding result from scalar to the multivariate case:

Theorem 6.7. *Let (y_n) be an $\mathbb{R}^s$-valued ($\mathbb{C}^s$-valued) wide sense stationary process. Then there exists a unique random orthogonal measure $d\zeta(\omega)$ with values in $\mathbb{C}^s$ such that*

$$y_n = \int_0^{2\pi} e^{in\omega} d\zeta(\omega)$$

The idea of the proof is to reduce the problem to the scalar case by considering the processes $z_n^\alpha = \alpha^\top y_n$ with various α-s. Details will be given at the end of the chapter.

To understand the effect of linear filter on the spectral representation process we need to describe the multivariate change of measure formula. Let $d\zeta(\cdot)$ be a $\mathbb{C}^s$-valued random orthogonal measure with structure function $dF(\cdot)$. Let $k(\cdot)$ be an $r \times s$, complex matrix-valued function such that $k(\cdot) \in L_2^{c,r\times s}(dF)$.

Proposition 6.10. *Under the conditions above the change of measure formula*

$$d\eta(\omega) = k(\omega) d\zeta(\omega)$$

defines a $\mathbb{C}^r$-valued random orthogonal measure $d\eta(\omega)$ having the $r \times r$ structure function

$$dG(\omega) = k(\omega) dF(\omega) \overline{k(\omega)}^\top.$$

Integration with respect to the new measure $d\eta$ is reduced to integration with respect to $d\zeta$ in a straightforward manner:

Proposition 6.11. *Let $h(\cdot)$ be a $q \times r$ complex matrix-valued function in $L_2^{c,q\times r}(dG)$. Then*

$$\int_0^{2\pi} h(\omega)d\eta(\omega) = \int_0^{2\pi} h(\omega)k(\omega)d\zeta(\omega).$$

6.7 Linear filters

In this section we study the effect of linear filter on the spectral representation process. Let (y_n) be an $\mathbb{R}^s$-valued ($\mathbb{C}^s$-valued) wide sense stationary process with spectral representation process $d\zeta^y(\omega)$ and define

$$v_n = \sum_{k=0}^{p} h_k y_{n-k}$$

where the h_k-s are $r \times s$ real matrices. Let

$$H(e^{-i\omega}) = \sum_{k=0}^{p} h_k e^{-i\omega k}.$$

Exercise 6.10. *Show that the spectral representation process of v is given by*

$$d\zeta^v(\omega) = H(e^{-i\omega})d\zeta^y(\omega).$$

Let us now consider infinite linear combinations, i.e. let

$$v_n = \sum_{k=0}^{\infty} h_k y_{n-k}. \tag{6.15}$$

where the h_k-s are $r \times s$ matrices.

Exercise 6.11. *Assume, that the infinite series*

$$H(e^{-i\omega}) = \sum_{k=0}^{\infty} h_k e^{-i\omega k}.$$

converges in $L_2^{c,r\times s}(dF^y)$. Then the spectral representation process of (v_n) is

$$d\zeta^v(\omega) = H(e^{-i\omega})d\zeta^y(\omega).$$

(*Hint*: Take a finite truncation, and take the limit).

Exercise 6.12. *Re-derive the formula for the spectral distribution measure of v:*

$$dF^v(\omega) = H(e^{-i\omega})dF^y(\omega)H^\top(e^{i\omega})$$

using the exercise above.

6.8 Proof of the spectral representation theorem

The basic idea of the proof is very simple: consider the processes

$$z_n^\alpha = \alpha^\top y_n,$$

with α being an arbitrary vector in $\mathbb{R}^s$, and let the spectral representation of z_n^α be $\zeta^\alpha(\omega)$. I.e. let

$$z_n^\alpha = \int_0^{2\pi} e^{in\omega} d\zeta^\alpha(\omega).$$

We have seen that $\zeta^\alpha(\omega)$ can be obtained as the limit of linear transformations of z_n^α. It will be easily seen that these linear transformations are independent of α, and if applied to the process (y_n) itself then, after taking the limit, we shall obtain the required spectral representation measure of (y_n).

Recall that stochastic integration with respect to the random orthogonal measure $d\zeta^\alpha(\omega)$ is an isometry mapping $L_2^c(dF^\alpha)$, with $F^\alpha = \alpha^\top F\alpha$, onto H^{z^α}, the Hilbert-space spanned by $z_n^\alpha, -\infty < n < \infty$. Letting this isometry be denoted by I_α we have

$$I_\alpha(e^{in\omega}) = z_n^\alpha.$$

Moreover, for any characteristic function $\chi_{[0,a)}(\omega)$ with $0 \le a \le 2\pi$ we have

$$I_\alpha(\chi_{[0,a)}(\omega)) = \zeta^\alpha(a).$$

To express the spectral representation measure $\zeta^\alpha(a)$ via the observed process z_n^α we proceed to express $\chi_{[0,a)}(\omega)$ via $e^{in\omega}$, as the limit of trigonometric polynomials converging in $L_2^c(dF^\alpha)$.

Now observe that for $|\alpha| = 1$ the measure $dF^\alpha = \alpha^\top dF\alpha$ is majorized by the measure tr $dF = d$ tr$F$. Note that $(e^{in\omega})$, $-\infty < n < +\infty$ is dense in $L_2^c(\text{tr } dF)$. Thus if we represent $\chi_{[0,a)}(\omega)$ as the limit of trigonometric polynomials converging in $L_2^c(d\text{tr } F)$, say

$$\chi_{[0,a)}(\omega) = \lim_N \sum_{k=-N}^{+N} c_{N,k} e^{ik\omega}, \tag{6.16}$$

then the right hand side will converge also in $L_2^c(dF^\alpha)$ for **any** $|\alpha| = 1$. The expression of $\zeta^\alpha(a)$ via z_n^α is obtained by the isometry I_α, or equivalently, by stochastic integration w.r.t. z_n^α, giving

$$\zeta^\alpha(a) = \lim_N \sum_{k=-N}^{+N} c_{N,k} z_k^\alpha.$$

Here the right hand side converges in $L_2^c(\Omega, \mathcal{F}, P)$ for **any** α.

Now the sum on the right hand side can be written as

$$\alpha^\top \sum_{k=-N}^{+N} c_{N,k} y_k = \alpha^\top \zeta_N(a)$$

with

$$\zeta_N(a) = \sum_{k=-N}^{+N} c_{N,k} y_k.$$

Thus we can write that

$$\zeta^\alpha(a) = \lim_N \alpha^\top \zeta_N(a).$$

Since α can be an arbitrary unit vectors in $\mathbb{R}^s$, we conclude that

$$\lim_{N\to\infty} \zeta_N(a) = \zeta(a)$$

exists, with convergence meant in $L_2^{c,s}(\Omega, \mathcal{F}, P)$. It follows that we can write

$$\zeta^\alpha(a) = \alpha^\top \zeta(a).$$

Exercise 6.13. *Let $\zeta(a), 0 \le a < 2\pi$ be an C^s-valued stochastic process such that for any $\alpha \in \mathbb{R}^s$ the scalar-valued process $\alpha^\top \zeta(a), 0 \le a < 2\pi$ has orthogonal increments. Then the process $\zeta(\cdot)$ itself has orthogonal increments.*

By the exercise above $\zeta(\cdot)$ is a process with orthogonal increments, and obviously

$$\alpha^\top y_n = z_n^\alpha = \int_0^{2\pi} e^{in\omega} d\zeta^\alpha(\omega) = \alpha^\top \int_0^{2\pi} e^{in\omega} d\zeta(\omega).$$

Since α is arbitrary, the spectral representation for (y_n) follows.

Chapter 7

State-space representation

7.1 From multivariate AR(1) to state-space equations

We have seen so far that a key technical tool in handling wide sense stationary processes is their spectral representation. The purpose of this section is to extend our arsenal with another powerful method that is called *state space representation* of wide sense stationary processes.

To motivate our discussion recall that there was a simple - but not trivial - example in our previous investigations, for which a direct analysis, avoiding spectral representation, was possible. This was the example of a stable AR(1) process defined via the equation

$$y_n + ay_{n-1} = e_n,$$

with $|a| < 1$, and (e_n) a wide sense stationary orthogonal process.

Let us now consider a multivariate AR(1) process given by the equation

$$x_{n+1} = Ax_n + Bv_n \tag{7.1}$$

with $x_n \in \mathbb{R}^s$ and $v_n \in \mathbb{R}^t$, where (v_n) is assumed to be a wide sense stationary orthogonal process with $\mathrm{E}v_n v_n^T = \Sigma_{vv}$. The vector x_n in (7.1) is called a *state vector*, and (7.1) is called *state-space system*. More exactly we would call this a state-space system with full observation. Note here that, in contrast to the scalar case, the noise term v_n with index n enters the definition of the state-vector x_{n+1} rather than x_n. This discrepancy in notations is due to historical reasons, and is preserved in current literature.

We can now ask ourselves: under what condition does a wide sense stationary causal solution (x_n) exists. Following the arguments for the scalar case we get by iterating equation (7.1) $\tau \geq 1$ times

$$x_{n+1} = A^\tau x_{n+1-\tau} + \sum_{k=0}^{\tau-1} A^k B v_{n-k}. \tag{7.2}$$

In order to be able to transfer our arguments for AR(1) processes from the scalar case to the multivariate case we need to assume that $\lim_n \|A^n\| = 0$. Since

$$\lim_n \|A^n\|^{1/n} = \rho(A),$$

with $\rho(A)$ denoting the *spectral radius* of A, i.e.

$$\rho(A) = \max_{i=1,\dots s} |\lambda_i(A)|,$$

where $\lambda_i(A)$, $i = 1, \dots s$ denote the eigenvalues of A, we will have $\lim_n \|A^n\| = 0$ exactly if $\rho(A) < 1$.

Definition 7.1. *A square, $s \times s$ A matrix is stable (in discrete sense) if $\rho(A) < 1$. Equivalently, an $s \times s$ matrix A is stable (in the discrete sense) if all its eigenvalues are in the open unit disc $D = \{z \ : \ |z| < 1\}$ of the complex plane, i.e. all the roots of the polynomial equation*

$$|zI - A| = 0$$

are in D.

Now repeating the arguments given for scalar valued AR(1) processes we come to the following result:

Proposition 7.2. *Let us consider the multivariate linear stochastic equation (7.1). Let A be a stable $s \times s$ matrix. Then (7.1) has a unique wide sense stationary solution (x_n), given by*

$$x_{n+1} = \sum_{k=0}^{\infty} A^k B v_{n-k}. \tag{7.3}$$

It follows that x is a causal linear function of v, more exactly, for all n

$$H^x_{n+1} \subset H^v_n.$$

Remark. Note the shift in the time index, due to the way we write state-space equations.

Exercise 7.1. *Prove the above proposition.*

It may be of interest to see how a proof would go in the spectral domain. We can now ask ourselves: under what condition does a wide sense stationary causal solution exist? To answer this question we proceed as in the scalar case. Letting q^{-1} denote the backward shift operator equation (7.1) can be written as

$$(qI - A)x = Bv.$$

We need the following lemma:

Lemma 7.3. *Let A be an $s \times s$ stable real matrix. Then*

$$(e^{i\omega} I - A)^{-1} = \sum_{k=1}^{\infty} h_k e^{-ik\omega}, \qquad h_1 = I$$

with some sequences of $s \times s$ real matrices h_k, where convergence in the right hand side is understood in the sense of $L_2^{c,s\times s}(d\omega)$. In fact, convergence is also uniform *in ω.*

Exercise 7.2. *Prove the Lemma 7.3.*

Exercise 7.3. *Prove the Proposition 7.2 using Lemma 7.3.*

We note in passing that by the very same spectral methods we can also easily get an answer to the following question: under what condition does a wide sense stationary (not necessarily causal) solution of (7.1) exist?

Proposition 7.4. *Assume, that $e^{i\omega} I - A$ is not singular for all $\omega \in [0, 2\pi]$. Then (7.1) has a unique solution.*

Exercise 7.4. *Prove the above Proposition 7.4.*

As we have seen, state space equations provide a very convenient tool to model multivariate w.s.st. processes. To conclude this section we complete this discussion by noting that the above class of processes can be extended by allowing what is called **partial observation**. Mathematically speaking we consider the dynamics given by the set of equations

$$x_{n+1} = Ax_n + Bv_n \tag{7.4}$$
$$y_n = Cx_n + Dw_n, \tag{7.5}$$

where the dimension of the observed process y (simply called observation) is typically much smaller than the dimension of the state process x. The observation noise Dw_n is assumed to be such that the matrix D is square, and $\dim w = \dim y$.

Condition 7.5. *The joint noise process $(v_n, w_n), -\infty < n < \infty$ is a w.s.st. orthogonal process with covariance matrix*

$$\begin{pmatrix} \Sigma_{vv} & \Sigma_{vw} \\ \Sigma_{wv} & \Sigma_{ww} \end{pmatrix}. \tag{7.6}$$

The above set of equation for modelling a multivariate time series is called a state-space model or linear stochastic system. The foundations of the theory of linear stochastic systems has been laid down by the Kyoto prize laureate Hungarian scientist R. Kalman. This theory revolutionized the research in the theory of wide sense stationary processes, especially by allowing a very effective solution of the so-called filtering problem, to be discussed below.

7.2 Auto-covariances and the Lyapunov-equation

A remarkable property of stable state-space systems is that the covariance-matrix of the state-vector and the auto-covariance function of (x_n) or (y_n) are very easily computed. To get the covariance function of (x_n) take the dyadic product of (7.1) with itself:

$$x_{n+1}x_{n+1}^T = Ax_nx_n^TA^T + Bv_nv_n^TB^T + Ax_nv_n^TB^T + Bv_nx_n^TA^T. \tag{7.7}$$

Now the representation (7.3) implies that $x_n \in H_{n-1}^v$ and hence

$$x_n \perp v_n.$$

Then taking expectation on both sides of (7.7) we get the following result:

Proposition 7.6. *Let (x_n) be a w.s.st. process defined by the state-space equation (7.1) where A is stable. Then $P = \mathrm{E}x_nx_n^T$ satisfies the equation*

$$P = APA^T + B\Sigma_{vv}B^T. \tag{7.8}$$

The latter equation is called a (discrete-time) *Lyapunov equation.*

Exercise 7.5. *Show that P can be written as*

$$P = \sum_{k=0}^{\infty} A^kB\Sigma_{vv}B^T(A^T)^k. \tag{7.9}$$

Exercise 7.6. *Show directly, with purely algebraic arguments, that, if A is stable, the Lyapunov-equation (7.8) has a unique solution P, and show that it can be written in the form (7.9). Prove that the solution P, given by (7.9), is positive semi-definite.*

To compute the auto-covariance function of x_n note that iterating (7.1) forward in time τ times, with $\tau \geq 1$, we get

$$x_{n+\tau} = A^\tau x_n + \sum_{k=0}^{\tau-1} A^kBv_{n+\tau-1-k}.$$

Now note that $v_{n+\tau-1-k} \perp x_n$ for $0 \leq k \leq \tau - 1$. Thus taking the dyadic product of the above equation with itself and taking expectation we come to the following conclusion:

Proposition 7.7. *Let (x_n) be the wide sense stationary process defined by the state-space equation (7.1), with A stable. Then for the covariance function of (x_n) we have*

$$R(\tau) = \mathrm{E}x_{n+\tau}x_n^T = A^\tau P \qquad \text{for} \quad \tau \geq 0.$$

For $\tau \leq 0$ we have

$$R(\tau) = \mathrm{E}x_{n+\tau}x_n^T = (\mathrm{E}x_n x_{n+\tau}^T)^T = R(-\tau)^T,$$

and thus

$$R(\tau) = P(A^T)^\tau \qquad \text{for} \quad \tau \leq 0.$$

Exercise 7.7. *Consider two Lyapunov equations (7.8) with a* **common** *stable A and such that*

$$B_1\Sigma_{1,vv}B_1^T \leq B_2\Sigma_{2,vv}B_2^T.$$

Let the solutions be denoted by P_1 and P_2. Show that $P_1 \leq P_2$.

Let us now consider a general linear stochastic system given by

$$\begin{aligned} x_{n+1} &= Ax_n + Bv_n \\ y_n &= Cx_n + Dw_n, \end{aligned} \tag{7.10}$$

see (7.4)-(7.5). Then the autocovariance function of (y_n) can be directly obtained from the autocovariance function of (x_n) as follows:

Proposition 7.8. *Let (y_n) be the wide sense stationary process defined by the state-space equation (7.4)-(7.5), with A stable. Then the auto-covariance function of (y_n) is obtained by*

$$R^y(0) = \mathrm{E}(y_n y_n^T) = CPC^T + D\Sigma_{ww}D^T \qquad \text{and} \tag{7.11}$$

$$R^y(\tau) = \mathrm{E}(y_{n+\tau}y_n^T) = CA^\tau PC^T \qquad \text{for} \quad \tau \geq 1. \tag{7.12}$$

Exercise 7.8. *Prove that for $\tau < 0$ we have $R(\tau) = R(-\tau)^T$.*

Note that the state-space description of a w.s.st. process is far from being unique. First of all, the map from (v, w) can be realized in an infinite number of ways by allowing coordinate transformations of the state-space. Letting T be a non-singular linear transformation of the state space define a new state-vector by

$$x' = Tx.$$

Then we have

$$\begin{aligned} x'_{n+1} &= TAT^{-1}x'_n + TBv_n \\ y_n &= CT^{-1}x'_n + Dw_n. \end{aligned} \tag{7.13}$$

It follows that the two systems below are equivalent in the sense that they generate the same input-output mapping:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}. \tag{7.14}$$

is equivalent to

$$\begin{pmatrix} TAT^{-1} & TB \\ CT^{-1} & D \end{pmatrix}. \tag{7.15}$$

This transformation of the state-space systems is standard in the theory of linear systems.

Now, if we look for a representation of the process (y_n) only without specifying the driving noise process (v, w), then an additional degree of freedom is in the choice of the latter. This problem can be reformulated as the problem of realizing a given auto-covariance sequence $R^y(.)$ in the form given by Proposition 7.8. This is called the stochastic realization problem.

Initialization at time 0. Let us consider the situation when the state-space equation is initialized at $n = 0$, rather than assuming that $-\infty < n < +\infty$. This is the case when we process observed data using a time-invariant linear filter with the observations starting at 0. Let us assume that $\mathrm{E}x_0 = 0$ and it has a finite covariance, say

$$\mathrm{E}x_0 x_0^T = P_0.$$

Then, as it is easily seen, the covariance matrix of x_n, say $P_n = \mathrm{E}x_n x_n^T$, satisfies

$$P_{n+1} = AP_nA^T + BB^T \tag{7.16}$$

with initial condition P_0.

Exercise 7.9. *Prove the validity of the recursion (7.16) for P_n.*

Exercise 7.10. *Show that if A is stable, then P_n converges to the unique solution of (7.8).*

*C*ontrollability. Let us now consider the problem: under what condition is the state-covariance matrix non-singular, or equivalently, positive definite? This question is of practical interest. Namely, if the state-covariance matrix is singular, then the state-process lives on a proper linear subspace. In this case we may try to find an alternative description of our system using a state-vector of smaller dimension.

If the noise in the state equation is non-degenerated, i.e. if Σ_{vv} is non-singular, then we may assume $\mathrm{E}v_n v_n^T = I$ by simply redefining v_n as $\Sigma_{vv}^{-1/2} v_n$ and B as $B\Sigma_{vv}^{1/2}$. Let us introduce the "matrix"

$$\mathcal{C}_\infty = (B,\ AB,\ A^2B,\ \dots)$$

having s rows and an infinite number of columns. Then we can write P as

$$P = \mathcal{C}_\infty \mathcal{C}_\infty^T.$$

Now $\mathrm{rank}(P) = s$ exactly if $\mathrm{rank}(\mathcal{C}_\infty) = s$. Let us now focus on the column rank of $\mathcal{C}_\infty$. Since, by the Cayley-Hamilton theorem

$$\sum_{k=0}^{n} \alpha_k A^k = 0, \quad \alpha_n = 1,$$

where $\sum_{k=0}^{n} \alpha_k \lambda^k$ is the characteristic polynomial of A, we have that all the columns of A^n can be expressed via the columns of the so-called controllability matrix

$$\mathcal{C} = (B,\ AB,\ A^2B,\ \dots\ A^{n-1}B). \tag{7.17}$$

But then, by induction, the same is true for all powers of A, say A^m with $m \geq 0$. Thus it follows that

$$\text{rank}\mathcal{C}_\infty = \text{rank}\mathcal{C}.$$

Thus we come to the following conclusion:

Proposition 7.9. *Let (x_n) be a w.s.st. process defined by the state-space equation (7.1) where A is stable, and Σ_{vv} is nonsingular. Then $P = \mathrm{E}x_n x_n^T$ is non-singular exactly when the controllability matrix $\mathcal{C}$ has full rank.*

7.3 State space representation of ARMA processes

Linear stochastic systems given by a state-space system are not only simple and elegant construction. They also serve as powerful tools for analyzing processes of more complex structures, such as ARMA processes. Let us first consider a w.s.st. AR(p) process (y_n) defined by

$$A(q^{-1})y = e, \tag{7.18}$$

where $A(z^{-1}) \neq 0$ for $|z| \geq 1$, $a_0 = 1$ and $a_p \neq 0$. In other words, the polynomial A is stable, and the order of the AR-process is exactly p. Define the state vector

$$x_n = (y_{n-1}, \dots, y_{n-p}). \tag{7.19}$$

Note, that the shift in the time index (x_n vs y_{n-1}) is not accidental. This is the way tradition has established itself. Then the dynamics of x_n can be described by first noting that

$$x_{n+1,1} = y_n = -a_1 y_{n-1} \cdots - a_p y_{n-p} + e_n.$$

The remaining coordinates of x_{n+1} are obtained by shifting the coordinates of x_n one position down, e.g.

$$x_{n+1,2} = x_{n,1}.$$

To describe the state-space dynamics in matrix-vector notation define the matrix $\tilde{A}$ by

$$\tilde{A} = \begin{pmatrix} -a_1 & \dots & & -a_p \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}. \tag{7.20}$$

Definition 7.10. *The matrix $\tilde{A}$ is the companion matrix associated with the polynomial $A(z^{-1})$.*

Define the p dimensional vector

$$b = (1, 0, \dots 0)^T.$$

Then the above arguments lead to the following proposition:

Proposition 7.11. *The process (y_n) can be realized by the state-space system*

$$x_{n+1} = \tilde{A}x_n + be_n \tag{7.21}$$
$$y_n = b^T x_{n+1}, \tag{7.22}$$

where the state-vector process (x_n) is defined by (7.19), $\tilde{A}$ is the $p \times p$ companion matrix defined under (7.20) and $b = (1, 0, \dots 0)^T$.

Note that the observation equation is not quite in the standard form, we have x_{n+1} rather than x_n on the right hand side.

We assumed that $A(z^{-1})$ is a stable polynomial, hence e is the innovation process of y, thus $H_n^y = H_n^e$. It follows that $H_{n+1}^x = H_n^e$. Thus we may guess that the stability of $A(z^{-1})$ implies the stability of $\widetilde{A}$. To see this we need the following general simple result:

Proposition 7.12. *We have*

$$|zI - \widetilde{A}| = z^p A(z^{-1}).$$

Proof. The r.h.s. equal

$$z^p A(z^{-1}) = z^p + a_1 z^{p-1} + \cdots + a_p.$$

Let the left hand side be denoted by $\alpha_p(z)$, i.e. let

$$\alpha_p(z) = \begin{pmatrix} z + a_1 & a_2 & \dots & a_p \\ -1 & z & \dots & 0 \\ & \ddots & \ddots & \\ & & -1 & z \end{pmatrix}.$$

Obviously the proposition is true for $p = 1$. We use induction. Expanding the above determinant by the last column we get

$$\alpha_p(z) = z\alpha_{p-1}(z) + \alpha_p(-1)^{p+1}(-1)^{p-1} = z\alpha_{p-1}(z) + \alpha_p.$$

This is exactly the same recursion that $z^p A(z^{-1})$ satisfies, thus the proposition is true for all p. □

Corollary 7.13. *Let* $\deg A(z^{-1}) = p$ *and let* $a_p \neq 0$. *Then the eigenvalues of* $\widetilde{A}$ *are identical with the roots of* $A(z^{-1})$.

In particular, if $A(z^{-1})$ is stable, then $\widetilde{A}$ is also stable, and the above machinery developed for computing the auto-covariance function for state-space systems is applicable.

Exercise 7.11. *Prove that* R_N *is non-singular by taking a state-space representation of* y.

Chapter 8

Kalman filtering

8.1 The filtering problem

A major advance in the theory of w.s.st. processes is the explicit solution of the so-called filtering problem using state-space theory. The solution is the celebrated Kalman-filter. To formulate the problem let us consider a w.s.st. process (y_n) given by the state space equation, with $-\infty < n < \infty$,

$$\begin{aligned} x_{n+1} &= Ax_n + Bv_n \quad (8.1) \\ y_n &= Cx_n + Dw_n, \quad (8.2) \end{aligned}$$

where A is a stable matrix and (v_n, w_n) is a w.s.st orthogonal process with covariance matrix

$$\mathrm{E}\begin{pmatrix} v_n \\ w_n \end{pmatrix}(v_n^T, w_n^T) = \begin{pmatrix} R_{vv} & R_{vw} \\ R_{wv} & R_{ww} \end{pmatrix}.$$

Problem statement. We formulate three closely related problems. The first is the problem of prediction. To predict (y_n) we need to find a representation of y in terms of its innovation process

$$\nu_n = y_n - (y_n|H_{n-1}^y)$$

of the form

$$y = H(q^{-1})\nu \quad (8.3)$$

where $H(q^{-1})$ is a causal linear filter. This is called *the innovation representation of* y. Conditions under which the above filter is well-defined has been given in the chapter "Multivariate time series". The possibility of such a representation for multivariate w.s.st. processes given by a state-space systems will be proven below.

A closely related problem of practical relevance is to predict the hidden state x_n in terms of the past of the observation process y, i.e. to determine

$$\widehat{x}_n = (x_n|H_{n-1}^y).$$

This problem is called one-step ahead or *predictive filtering*. Finally, we may wish to estimate x_n in terms of values of y up to time n, i.e. to determine

$$\overline{x}_n = (x_n|H_n^y).$$

This last problem is called simply *filtering*. A key ingredient of Kalman-filtering is the discovery of an explicit, computable dynamics of the processes $\widehat{x}_n$ and $\overline{x}_n$.

From $\widehat{x}_n$ to $\overline{x}_n$: The first step in finding the dynamics of $\widehat{x}_n$ and $\overline{x}_n$ is the observation that

$$\nu_n = y_n - (y_n|H_{n-1}^y)$$

implies that

$$H_n^y = H_{n-1}^y \oplus \mathcal{L}(\nu_n) \tag{8.4}$$

where $\oplus$ indicates an orthogonal direct sum.

Exercise 8.1. *Provide an argument for the validity of (8.4).*

Now it follows that

$$\overline{x}_n = (x_n|H_n^y) = (x_n|H_{n-1}^y) + (x_n|\nu_n) = \widehat{x}_n + (x_n|\nu_n), \tag{8.5}$$

where $(x_n|\nu_n)$ is the shorthand notation for $(x_n|\mathcal{L}(\nu_n))$. Note that ν is a finite dimensional r.v., hence we can write

$$(x_n|\nu_n) = K\nu_n \tag{8.6}$$

with some matrix K of size $s \times m$, which – if not unique,– may be chosen so that it does not depend on n (due to the stationarity of $\begin{pmatrix} x_n \\ y_n \end{pmatrix}$). If the covariance matrix $\mathrm{E}\nu\nu^T$ is nonsingular, then K is unique. (See below.) We conclude that

$$\overline{x}_n = \widehat{x}_n + K\nu_n. \tag{8.7}$$

From $\overline{x}_n$ to $\widehat{x}_{n+1}$: The second observation is that by projecting both sides of the state space equation (8.1) to H_n^y, and taking into account that $v_n \perp H_n^y$, due to stability of A, gives

$$\widehat{x}_{n+1} = (x_{n+1}|H_n^y) = A\overline{x}_n. \tag{8.8}$$

The innovation. The third observation is that

$$(y_n|H_{n-1}^y) = (Cx_n + w_n|H_{n-1}^y) = C\widehat{x}_n.$$

This follows from the fact that

$$w_n \perp H_{n-1}^y.$$

Indeed, $H^y_{n-1} \subset H^x_{n-1} + H^w_{n-1}$, and $H^x_{n-1} \subset H^v_{n-2}$, due to the stability of A. Thus the orthogonality of the joint process (v, w) implies the claim. Thus

$$\nu_n = y_n - (y_n | H^y_{n-1}) = y_n - C\widehat{x}_n. \tag{8.9}$$

Now combining (8.7), (8.8) and (8.9) we have arrived, with minimal effort, at the following beautiful and important result:

Proposition 8.1. *Assume that A is stable. Then the filtered process $\widehat{x}_n = (x_n | H^y_{n-1})$ follows the state-space dynamics:*

$$\widehat{x}_{n+1} = A\widehat{x}_n + K\nu_n \tag{8.10}$$
$$\nu_n = y_n - C\widehat{x}_n. \tag{8.11}$$

with some fixed K. If $\mathrm{E}\nu\nu^T$ is nonsingular, then K is unique.

Note that with this result we have completed a major step in the program set forth by Kalman-filtering: namely, we have obtained a recursion for the predicted value of x. Before adding further details, we also note that rearranging (8.10) we get:

$$\widehat{x}_{n+1} = A\widehat{x}_n + K\nu_n \tag{8.12}$$
$$y_n = C\widehat{x}_n + \nu_n. \tag{8.13}$$

Thus we have reproduced y in terms of its own innovation process ν. The causal linear operator mapping ν to y is given by

$$H(q^{-1}) = C(qI - A)^{-1}K + I. \tag{8.14}$$

A rigorous interpretation of $H(q^{-1})$ can be given, as in the scalar case, using frequency domain representation of the relevant processes ν, $\widehat{x}$ and y.

8.2 The Kalman-gain matrix

.

The next step is to determine the matrix K which is called the *Kalman gain matrix*. Set $x = x_n$. If $\dim x = \dim \nu = 1$ then we can restrict ourselves to a 2-dimensional subspace spanned by x and ν. Then the projection of x on ν is obtained by elementary geometry as

$$\widehat{x} = \frac{\mathrm{E}(x\nu)}{\mathrm{E}(\nu\nu)} \cdot \nu.$$

Indeed, the projection of x on ν is $\widehat{x} = \lambda\nu$ with some $\lambda \in \mathbb{R}$, such that $x - \widehat{x} \perp \nu$. It means:

$$\mathrm{E}((x - \widehat{x})\nu) = \mathrm{E}((x - \lambda\nu)\nu) = \mathrm{E}(x\nu) - \lambda\mathrm{E}(\nu\nu) = 0.$$

From here we get $\lambda = \dfrac{\mathrm{E}(x\nu)}{\mathrm{E}(\nu\nu)}$.

This formula extends for the case when x and ν are vector-valued as

$$\widehat{x} = \mathrm{E}(x\nu^T)(\mathrm{E}(\nu\nu^T))^{-1}\nu \tag{8.15}$$

assuming that $\mathrm{E}\nu\nu^T$ is nonsingular.

Exercise 8.2. *Prove that the projection of the random vector $x \in L_2^s(\Omega, \mathcal{F}, P)$ onto the finite dimensional subspace of $L_2(\Omega, \mathcal{F}, P)$ spanned by the components of ν is given by (8.15).*

Specializing this result to our case we get that

$$K = \mathrm{E}(x\nu^T)\ (\mathrm{E}(\nu\nu^T))^{-1}. \tag{8.16}$$

The covariance matrix of ν. To compute $\mathrm{E}(\nu\nu^T)$ note that

$$\nu_n = y_n - C\widehat{x}_n = (Cx_n + Dw_n) - C\widehat{x}_n = C\widetilde{x}_n + Dw_n, \tag{8.17}$$

where

$$\widetilde{x}_n = x_n - \widehat{x}_n$$

is the state error process. Since $x_n \in H_{n-1}^v$ and $\widehat{x}_n \in H_{n-1}^y \subset H_{n-2}^v + H_{n-1}^w$, we have

$$\widetilde{x}_n \perp w_n.$$

Thus – with the notation $\mathrm{E}vv^T = R_{vv}$ etc. – we get from (8.17):

$$R_{\nu\nu} = CR_{\widetilde{x}\widetilde{x}}C^T + DR_{ww}D^T. \tag{8.18}$$

The covariance matrix of $\widetilde{x}$. The covariance matrix $R_{\widetilde{x}\widetilde{x}}$ can be obtained by noting that $x_n = \widehat{x}_n + \widetilde{x}_n$ and $\widehat{x}_n \perp \widetilde{x}_n$ imply

$$R_{xx} = R_{\widehat{x}\widehat{x}} + R_{\widetilde{x}\widetilde{x}}.$$

Now R_{xx} and $R_{\widehat{x}\widehat{x}}$ can be obtained from the Lyapunov equations:

$$\begin{aligned} R_{xx} &= AR_{xx}A^T + BR_{vv}B^T \\ R_{\widehat{x}\widehat{x}} &= AR_{\widehat{x}\widehat{x}}A^T + KR_{\nu\nu}K^T. \end{aligned}$$

Subtracting the second equation from the first one we get

$$R_{\widetilde{x}\widetilde{x}} = AR_{\widetilde{x}\widetilde{x}}A^T + BR_{vv}B^T - KR_{\nu\nu}K^T. \tag{8.19}$$

The covariance matrix $\mathrm{E}x\nu^T$**.** To compute $\mathrm{E}x\nu^T = \mathrm{E}(x_n\nu_n^T)$ write $x_n = \widehat{x}_n + \widetilde{x}_n$ and note that $\widehat{x}_n \in H_{n-1}^y$ implies

$$\nu_n \perp \widehat{x}_n.$$

Thus

$$\mathrm{E}x_n\nu_n^T = \mathrm{E}\widetilde{x}_n\nu_n^T = \mathrm{E}\widetilde{x}_n(C\widetilde{x}_n + Dw_n)^T.$$

Since $w_n \perp \widetilde{x}_n$ we get

$$\mathrm{E}x_n\nu_n^T = R_{\widetilde{x}\widetilde{x}}C^T. \tag{8.20}$$

Thus we get, after substitution in (8.16):

$$K = R_{\widetilde{x}\widetilde{x}}C^T R_{\nu\nu}^{-1}. \tag{8.21}$$

Now we have a set of circular expressions for $R_{\nu\nu}$, $R_{\widetilde{x}\widetilde{x}}$ and K given by (8.18), (8.19) and (8.21). Expressing $R_{\nu\nu}$ and K via $R_{\widetilde{x}\widetilde{x}}$ we get a single equation for the latter, which reads:

$$R_{\widetilde{x}\widetilde{x}} = AR_{\widetilde{x}\widetilde{x}}A^T + BR_{vv}B^T - R_{\widetilde{x}\widetilde{x}}C^T(CR_{\widetilde{x}\widetilde{x}}C^T + DR_{ww}D^T)^{-1}CR_{\widetilde{x}\widetilde{x}}. \tag{8.22}$$

The above matrix-equation is called an *algebraic Riccati equation*. Thus we arrived at the following conclusion:

Proposition 8.2. *Assume that the innovation process* (ν_n) *is non-degenerate, i.e.* $R_{\nu\nu} = \mathrm{E}\nu\nu^T$ *is non-singular. Then the Kalman-gain matrix* K *is uniquely determined and is given by*

$$K = R_{\widetilde{x}\widetilde{x}}C^T R_{\nu\nu}^{-1},$$

where $R_{\widetilde{x}\widetilde{x}}$ *is a symmetric positive definite solution of the algebraic Riccati equation (8.22), and* $R_{\nu\nu}$ *is readily expressed via* $R_{\widetilde{x}\widetilde{x}}$ *as given in (8.18).*

A simple condition that ensures that $R_{\nu\nu}$ is nonsingular is that D and R_{ww} are nonsingular.

Reconstruction of ν. The reconstruction of ν from y is formally straightforward. Setting $\nu_n = y_n - C\widehat{x}_n$ in (8.13) we get

$$\widehat{x}_{n+1} = A\widehat{x}_n + K(y_n - C\widehat{x}_n),$$

from which we get the inverse system:

$$\begin{aligned} \widehat{x}_{n+1} &= (A - KC)\widehat{x}_n + Ky_n \\ \nu_n &= y_n - C\widehat{x}_n. \end{aligned}$$

The corresponding operator, mapping y to ν is:

$$H^{-1}(q^{-1}) = I - C(qI - A + KC)^{-1}K. \tag{8.23}$$

Exercise 8.3. *Derive the above expression from (8.14) using the matrix inversion lemma.*

We would expect that $A - KC$ is stable, or at least does not have any eigenvalue outside the unit disc $\{z \ : \ |z| > 1\}$. However, the rigorous interpretation of the inverse $H^{-1}(q^{-1})$, when $A - KC$ has an eigenvalue of the unit circle, is beyond the scope of this course.

Chapter 9

Identification of AR processes

In the following two chapters we consider the problem of statistical analysis of a w.s.st. time series (y_n). Thus from now on we assume that we are given a sequence of observations $y_1, \ldots, y_N$, and we ask ourselves, how to infer structural properties of the complete process $y = (y_n)$. The first and obvious objective may be to estimate the auto-covariance function $r(\tau)$. A natural candidate for this is the sample covariance:

$$\widehat{r}(\tau) = \frac{1}{N-\tau} \sum_{n=1}^{N-\tau} y_{n+\tau} y_n$$

for $\tau \geq 0$. Note that the values $y_{n+\tau} y_n$ form a dependent sequence, therefore standard laws of large numbers (LLN) formulated for independent sequences are not applicable. Conditions under which $\widehat{r}(\tau)$ will converge to $r(\tau)$ will not be discussed in this course, rather we will simply assume that this convergence does take place. Note, however, that no matter how large N is, we will be able to estimate only a finite segment of the auto-covariance function. Certainly, no estimates will be available for $\tau \geq N$.

In order to get meaningful results we have to restrict ourselves to time series the structure of which can be perfectly described by a finite set of parameters. We will consider three classes of processes: AR, MA and ARMA processes.

9.1 Least Squares estimate of an AR process

Let (y_n) be w.s.st. stable AR(p) process defined by

$$y_n + a_1^* y_{n-1} + \cdots + a_p^* y_{n-p} = e_n. \tag{9.1}$$

The superscript * indicates that the corresponding parameters are "true parameters", as opposed to tentative values to be used later. Here (e_n) is, as usual, a w.s.st. orthogonal process. Due to the assumed stability of $A^*(z^{-1}) = \sum_{k=1}^{p} a_k z^{-k}$, the driving noise (e_n)

is the innovation process of (y_n). Our goal is to estimate θ^* using observations from time 1 to N. Introducing the notations

$$\varphi_n = (-y_{n-1} \cdots - y_{n-p})^T$$

and

$$\theta^* = (a_1^* \cdots a_p^*)^T$$

equation (9.1) can be rewritten as

$$y_n = \varphi_n^T \, \theta^* + e_n. \tag{9.2}$$

The advantage of this reformulation is that the original model is now rewritten as a linear regression model. More precisely, we get a (linear) stochastic regression, since the sequence of regressor vectors (φ_n) is not independent of the noise sequence (e_n).

To estimate θ^* using the observations $y_1, \ldots y_N$ a natural candidate is the least squares (LSQ) method. Let us fix a tentative value of $\theta \in \mathbb{R}^p$, and define the error process

$$\varepsilon_n(\theta) = y_n - \varphi_n^T \theta.$$

Here we should restrict n to be at least $p+1$, to ensure that $\varphi_n^T \theta$ is defined in terms of the observations for all n. Alternatively, we may assume that $y_0, y_{-1}, ..., y_{-p+1}$ are known. Following the tradition of the system identification literature, we shall use the latter option. Then, the LSQ estimation method amounts to minimizing the cost function defined as the sum of the squared errors:

$$V_N(\theta) = \frac{1}{2} \sum_{n=1}^{N} \varepsilon_n^2(\theta) = \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \varphi_n^T \theta \right)^2 .$$

Since $\varepsilon_n(\theta) = y_n - \varphi_n^T \theta$ is the best mean-squared prediction error when $\theta = \theta^*$, the LSQ estimate falls in the larger class of prediction error estimators, see the next chapter.

The above cost function is quadratic and convex in θ, therefore its minimum is attained. Moreover for any minimizing value of θ we have

$$\frac{\partial}{\partial \theta} V_N(\theta) = 0.$$

Differentiating $V_N(\theta)$ w.r.t. θ we get

$$\frac{\partial}{\partial \theta} V_N(\theta) = \sum_{n=1}^{N} \varepsilon_{\theta n}^T(\theta) \varepsilon_n(\theta),$$

where the subscript $_\theta$ denotes differentiation w.r.t. θ. Note that, following the convention of matrix analysis, the gradient of a scalar-valued function is represented as a row vector. Taking into account that

$$\varepsilon_{\theta n}(\theta) = -\varphi_n^T \tag{9.3}$$

we get the equation

$$\sum_{n=1}^{N} -\varphi_n^T \varepsilon_n(\theta) = 0.$$

From here we get, after substituting $\varepsilon_n(\theta)$,

$$\sum_{n=1}^{N} -\varphi_n^T (y_n - \varphi_n^T \theta) = 0.$$

This is a linear equation for θ, which certainly has a solution by the arguments above. After rearrangement we get the following result:

Proposition 9.1. *Let $\widehat{\theta}_N$ be a least squares estimator of the AR-parameter θ^* based on N samples. Then $\widehat{\theta}_N$ satisfies the following so-called normal equation:*

$$\left[\sum_{n=1}^{N} \varphi_n \varphi_n^T\right] \theta = \sum_{n=1}^{N} \varphi_n y_n. \tag{9.4}$$

The estimator $\widehat{\theta}_N$ is unique if the coefficient-matrix of the normal equation, i.e.

$$S_N = \sum_{n=1}^{N} \varphi_n \varphi_n^T$$

is non-singular. Equivalently, the estimator $\widehat{\theta}_N$ is unique if the normalized coefficient-matrix of the normal equation,

$$R_N = \frac{1}{N} \sum_{n=1}^{N} \varphi_n \varphi_n^T$$

is non-singular. Note that the elements of R_N are just empirical auto-covariances of (y_n) : say, the (k, l)-th element reads as:

$$\frac{1}{N} \sum_{n=1}^{N} y_{n-k} y_{n-l}. \tag{9.5}$$

To make use of this observation we impose the following assumption:

Condition 9.2. *Assume that the empirical auto-covariances of y converge to the theoretical auto-covariances almost surely, i.e. for any fixed k, l we have*

$$\lim_N \frac{1}{N} \sum_{n=1}^{N} y_{n-k} y_{n-l} = \mathrm{E}\ y_{n-k} y_{n-l} = r^y(l-k) \quad \text{a.s.} \tag{9.6}$$

The above condition simply states the validity of a strong law of large numbers for the dependent sequence $z_n = y_{n-k}y_{n-l}$. A standard way to ensure this is to prove some kind of mixing property of z_n. However, we do not have the space to discuss further details.

Proposition 9.3. *Let (y_n) be a w.s.st. stable AR(p) process defined by (9.1). Assume that Condition 9.2 is satisfied. Then the LSQ estimate $\hat{\theta}_N$ converges to the true system parameter vector θ^* almost surely.*

Proof. Under the above condition we have

$$\lim_N \frac{1}{N} S_N = \lim_N R_N = R^* \quad \text{a.s.}, \tag{9.7}$$

where R^* is the p-th order auto-covariance matrix. (Recall that R^* is a $p \times p$, symmetric, positive semi-definite Toeplitz matrix.)

Exercise 9.1. *Prove that R^* is non-singular.*

Exercise 9.2. *Prove that R^* is non-singular by taking a state-space representation of y.*

The r.h.s. of the normal equation, normalized by N, will be written as

$$-r_N = \frac{1}{N} \sum_{n=1}^{N} \varphi_n y_n.$$

Under Condition 9.2, we have

$$\lim_N (-r_N) = \lim_N \frac{1}{N} \sum_{n=1}^{N} \varphi_n y_n = \mathrm{E}\varphi_n y_n = -r^* \quad \text{a.s.},$$

where $r^* = (r^y(1), ..., r^y(p))^T$. Note that $\mathrm{E}\varphi_n y_n$ can also be written as

$$\mathrm{E}\varphi_n y_n = \mathrm{E}\varphi_n(\varphi_n^T \theta^* + e_n) = R^*\theta^*.$$

Thus we conclude that

$$\lim_N (-r_N) = -r^* = R^*\theta^*.$$

Now, rewrite the normal equation (9.4) as follows:

$$R_N \hat{\theta}_N + r_N = 0. \tag{9.8}$$

Note that for any fixed θ the l.h.s. of the equation converges:

$$\lim_N (R_N\theta + r_N) = (R^*\theta - R^*\theta^*) = R^*(\theta - \theta^*).$$

Since R^* is non-singular, the claim follows by standard arguments. □

Exercise 9.3. ** Show that Condition 9.2 implies that*

$$\lim_N \frac{1}{N}\sum_{n=1}^{N} \varphi_n e_n = \mathrm{E}\varphi_n e_n = 0 \quad \text{a.s.} \tag{9.9}$$

Remark. To conclude this subsection we note, that the normal equation can be simply obtained (and memorized) as follows: multiply (9.2) from the left by φ_n, sum it form 1 to N and omit the terms containing (e_n), in view of the fact that

$$\mathrm{E}\varphi_n \; e_n = 0$$

for all n. The beauty of this approach is that φ_n could be replaced by some other random vector ψ_n such that

$$\mathrm{E}\psi_n \; e_n = 0.$$

The vectors ψ_n are called *instrumental variables*. The the estimator of θ^* is obtained form the equation

$$\left[\sum_{n=1}^{N} \psi_n \varphi_n^T\right] \theta = \sum_{n=1}^{N} \psi_n y_n. \tag{9.10}$$

This method is called the instrumental variable (IV) method, that has been widely used in the early systems identification literature. The choice of an appropriate, convenient instrumental variable ensuring the non-singularity of the modified normal equation depends very much on the nature of the specific problem.

9.2 The asymptotic covariance matrix of the LSQ estimate

Next, we may ask ourselves about the quality of the estimator $\widehat{\theta}_N$, such as its bias and covariance matrix. Surprisingly (or not so surprisingly), the standard methods of regression analysis are not applicable in the present case. It is readily seen that the error $\tilde{\theta}_N = \hat{\theta}_N - \theta^*$ satisfies the equation

$$(\sum_{n=1}^{N} \varphi_n \varphi_n^T)\tilde{\theta}_N = \sum_{n=1}^{N} \varphi_n e_n, \tag{9.11}$$

and thus

$$\tilde{\theta}_N = (\sum_{n=1}^{N} \varphi_n \varphi_n^T)^{-1} \sum_{n=1}^{N} \varphi_n e_n.$$

As opposed to standard regression analysis, we can not conclude from here that $\hat{\theta}_N$ is unbiased, or equivalently that $\mathrm{E}\tilde{\theta}_N = 0$, due to the dependence of the regressor sequence (ϕ_n) and the noise sequence (e_n).

By the same reasoning, we can not compute the covariance matrix of $\hat{\theta}_N$ in a straightforward manner. In fact, it is not even guaranteed, that $\hat{\theta}_N$ has a finite covariance matrix (or finite second moments). As an example, consider an $AR(1)$-process

$$y_n + a^* y_{n-1} = e_n,$$

with $|a^*| < 1$. Then the error of the LSQ estimate of the single pole a^* is obtained as

$$\tilde{a}_N = (\sum_{n=1}^{N} y_{n-1} e_n)/(\sum_{n=1}^{N} y_{n-1} y_{n-1}). \tag{9.12}$$

Exercise 9.4. *Assume that (e_n) is Gaussian. Show that $(\sum_{n=1}^{N} y_{n-1} y_{n-1})^{-1}$ has no finite expectation.*

A simple remedy to the above difficulty is to consider an approximation of the error process $\tilde{\theta}_N$ by using the approximation

$$(\sum_{n=1}^{N} \varphi_n \varphi_n^T)^{-1} \cong \frac{1}{N}(R^*)^{-1},$$

and defining a new, approximating error process

$$\tilde{\tilde{\theta}}_N = \frac{1}{N}(R^*)^{-1} \sum_{n=1}^{N} \varphi_n e_n. \tag{9.13}$$

The (asymptotic) covariance matrix of $\tilde{\tilde{\theta}}_N$ is then completely determined by the (asymptotic) covariance matrix of

$$\rho_n = \frac{1}{N} \sum_{n=1}^{N} \varphi_n e_n.$$

To have a nice expression for this we need an additional, standard assumption:

Condition 9.4. . *Let $\{\mathcal{F}_n\}, -\infty < n < \infty$, be an increasing family of σ-algebras, such that e_n is $\mathcal{F}_n$-measurable for all n. It is assumed that*

$$\mathrm{E}(e_n|\mathcal{F}_{n-1}) = 0 \quad \text{and} \quad \mathrm{E}(e_n^2|\mathcal{F}_{n-1}) = \sigma^2 = \text{const.} \quad \text{for all} \quad n.$$

In other words, (e_n) is a martingale-difference sequence with constant conditional variance w.r.t. $\mathcal{F}_n$. Under the condition above, we have the following non-asymptotic result:

Lemma 9.5. *Under Condition 9.4 we have*

$$\mathrm{E}\rho_n\rho_n^T = \frac{1}{N}R^*\ \sigma^2.$$

Proof. We have

$$\mathrm{E}\rho_n\rho_n^T = \mathrm{E}\sum_{n,m=1}^{N}\varphi_n e_n \cdot e_m\varphi_m^T.$$

For a fixed pair $n < m$ we have

$$\mathrm{E}\varphi_n e_n \cdot e_m\varphi_m^T = \mathrm{E}\ [\mathrm{E}\ [\varphi_n e_n \cdot e_m\varphi_m^T|\mathcal{F}_{m-1}]] = \mathrm{E}\ (\varphi_n e_n \cdot \varphi_m^T)[\mathrm{E}\ [e_m|\mathcal{F}_{m-1}]] = 0.$$

Here we used the fact that φ_n, e_n and φ_m^T are $\mathcal{F}_{m-1}$-measurable, and that (e_n) is a martingale-difference sequence w.r.t. $(\mathcal{F}_n)$. On the other hand, for any fixed $n = m$ we have

$$\mathrm{E}[\varphi_n e_n \cdot e_n\varphi_n^T] = \mathrm{E}\ [\mathrm{E}\ [\varphi_n e_n \cdot e_n\varphi_n^T|\mathcal{F}_{n-1}]] = \mathrm{E}\ (\varphi_n \cdot \varphi_n^T)\ [\mathrm{E}\ [e_n e_n|\mathcal{F}_{n-1}]].$$

By Condition 9.4 the last expression can be written as

$$\mathrm{E}\ (\varphi_n \cdot \varphi_n^T)\sigma^2 = R^*\ \sigma^2,$$

which proves the claim. □

A direct consequence is the following proposition:

Proposition 9.6. *Under Condition 9.4 the approximating error process $\tilde{\tilde{\theta}}_N$ defined under (9.13) has the following covariance matrix:*

$$\mathrm{E}\tilde{\tilde{\theta}}_N\tilde{\tilde{\theta}}_N^T = \frac{1}{N}(R^*)^{-1}\sigma^2.$$

Note that this result is a mirror image of the corresponding result in the theory of linear regression.

The asymptotic covariance matrices of the LSQ estimators, assuming unit variance for the noise, are displayed for our three benchmark AR(4)-processes below:

The above result provides a guideline for the proof of an exact result. The first step in that direction may be the modification of the estimator itself by truncation to ensure finite second moments. One possible truncation is obtained as follows. Let K be a sufficiently large positive number such that $|\theta^*| < K$. Then define the truncated LSQ estimator as

$$\overline{\theta}_N = K\frac{\hat{\theta}_N}{|\hat{\theta}_N|} \quad \text{for} \quad |\hat{\theta}_N| > K \quad \text{and} \quad \overline{\theta}_N = \hat{\theta}_N \quad \text{otherwise.}$$

$$P^{-1} = \begin{bmatrix} 0.974 & -1.80 & 1.54 & -0.565 \\ -1.80 & 3.97 & -3.75 & 1.54 \\ 1.54 & -3.75 & 3.97 & -1.80 \\ -0.565 & 1.54 & -1.80 & 0.974 \end{bmatrix}$$

Figure 9.1: The asymptotic covariance matrix for an AR(4) process with two positive poles and two almost unstable complex pose whose real part is positive. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.3\pi$.

$$P^{-1} = \begin{bmatrix} 0.974 & -0.423 & 0.332 & -0.436 \\ -0.423 & 0.963 & -0.419 & 0.332 \\ 0.332 & -0.419 & 0.963 & -0.423 \\ -0.436 & 0.332 & -0.423 & 0.974 \end{bmatrix}$$

Figure 9.2: The asymptotic covariance matrix for an AR(4) process with two positive poles and two almost unstable complex pose whose real part is negative. The actual values: two real poles at 0.5, a pair of complex poles with length 0.8 and argument $\pm 0.6\pi$.

$$P^{-1} = \begin{bmatrix} 1.00 & 1.00 & 0.349 & 0.0476 \\ 1.00 & 2.00 & 1.33 & 0.349 \\ 0.349 & 1.33 & 2.00 & 1.00 \\ 0.0476 & 0.349 & 1.00 & 1.00 \end{bmatrix}$$

Figure 9.3: The asymptotic covariance matrix for an AR(4) process with four small negative poles. The actual values of the poles are $-0.1, -0.2, -0.3, -0.4$.

In trying to compute the asymptotic covariance matrix of this truncated estimator, we would need to estimate the probability of actual truncations. This indicates that additional technical analysis is needed, which is beyond the scope of the course. We simply note that under certain additional technical assumption, implying Condition 9.2, and also assuming Condition 9.4 we have the following result: the truncated LSQ estimate $\overline{\theta}_N$ is asymptotically unbiased and its asymptotic covariance matrix is exactly what we have obtained for the approximating error $\tilde{\tilde{\theta}}_N$:

$$\lim_N N\mathrm{E}(\overline{\theta}_N - \theta^*)(\overline{\theta}_N - \theta^*)^T = \sigma(e)^2 (R^*)^{-1}.$$

It is worth noting that the expression $\sigma(e)^2(R^*)^{-1}$ scales with e: multiplying e by a constant c the variance $\sigma(e)^2$ gets multiplied by c^2. On the other hand, the process y also gets multiplied by c, and hence R^* gets multiplied by c^2. Consequently, $\sigma(e)^2(R^*)^{-1}$ is unchanged. This is intuitively obvious from the fact that scaling e leaves the signal to noise ratio (SNR) unchanged.

As we see the (asymptotic) quality of the LSQ estimator is completely determined by the covariance matrix R^*. Recall that R^* is exactly the covariance matrix of the state-vector of the proposed state-space representation of (y_n), and thus it is easily found, at least in theory, as the solution of a Lyapunov-equation. Alternatively, we can use the Yule-Walker equations to find the auto-covariances of y. Consider the example of an $AR(1)$ process:

$$y_n + a^* y_{n-1} = e_n.$$

It is easily seen that $R^* = \sigma^2(e)(1 - (a^*)^2)^{-1}$, and thus the asymptotic variance of the LSQ estimator of a^* equals

$$1 - (a^*)^2.$$

It follows, that if as a^* is close to ± 1, then the asymptotic variance of the LSQ estimator is close to 0. This is again intuitively plausible: if a^* is close to ± 1 then the AR-system is nearly unstable, and hence the process (y_n) will take on very large values, leading to a very large SNR. We note in passing that AR processes with poles close to 1 are common in modeling economic time series.

9.3 The recursive LSQ method

Assume that

$$S_N = \sum_{n=1}^{N} \varphi_n \varphi_n^T$$

is nonsingular, and thus positive definite for some N. Then $S_{N'}$ will be nonsingular for any $N' > N$, and thus the LSQ estimator $\widehat{\theta}_{N'}$ is uniquely defined. Assume that $\widehat{\theta}_N$ is available. Suppose we get one more observation y_{N+1}. The question is then raised: do we need to recompute S_{N+1}^{-1} and $\widehat{\theta}_{N+1}$ from scratch or is there a way to compute S_{N+1}^{-1} and $\widehat{\theta}_{N+1}$ using S_N^{-1} and $\widehat{\theta}_N$? This question is partially answered in the following celebrated result:

Proposition 9.7. *(The matrix inversion lemma.) Let*

$$F = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

be a 2×2 block-matrix with A and D being square matrices. Assume that A, D are non-singular and so is $A - BD^{-1}C$. Then

$$\left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1}.$$

In particular, $D - CA^{-1}B$ is nonsingular.

Proof. Consider the equation for the inverse of F:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \mathbf{X} & Y \\ U & V \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

We will compute $\mathbf{X}$ via Gauss elimination in *two different* ways. We have

$$\begin{aligned} AX + BU &= I \qquad (9.14) \\ CX + DU &= 0. \qquad (9.15) \end{aligned}$$

From the second equation

$$U = -D^{-1}CX,$$

and thus we get from the first equation

$$\left(A - BD^{-1}C\right)X = I.$$

Thus

$$X = \left(A - BD^{-1}C\right)^{-1}.$$

It follows that (9.14)–(9.15) has a unique solution (X, U).

An alternative way of applying Gauss elimination is to start with the first equation. Then we get

$$X = A^{-1}(I - BU). \qquad (9.16)$$

Substituting into the second equation we get

$$CA^{-1}(I - BU) + DU = 0,$$

from which we get

$$(D - CA^{-1}B)U = -CA^{-1}.$$

Since U is uniquely determined, $D - CA^{-1}B$ must be nonsingular. Substituting the resulting U in (9.16) we get

$$X = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

and the lemma follows. Q.e.d. □

Remark. To memorize the matrix inversion lemma the following exercise may be useful:

Exercise 9.5. *Assume that A and D are non-singular. Then*

$$\frac{d}{d\varepsilon}\left(A-\varepsilon BD^{-1}C\right)^{-1}\Big|_{\varepsilon=0}=A^{-1}BD^{-1}CA^{-1}.$$

A special case of the matrix inversion lemma is the following result:

Proposition 9.8. *(The Sherman-Morrison lemma.) Let A be a square matrix, and let b, c be vectors of the same dimension. Assume that A is non-singular and so is $A+bc^T$. Then*

$$\left(A+bc^T\right)^{-1}=A^{-1}-\frac{1}{(1+c^TA^{-1}b)}A^{-1}bc^TA^{-1}.$$

In particular, $1+c^TA^{-1}b\neq 0$.

Exercise 9.6. *Prove the Sherman-Morrison lemma.*

A direct corollary of the above lemma is a recursion for the inverse of the coefficient matrix S_N of the normal equation. Noting that we have

$$S_{N+1}=S_N+\varphi_{N+1}\varphi_{N+1}^T,$$

and setting $A=S_N, b=c=\varphi_{N+1}$, we get the following recursion:

Proposition 9.9. *Let S_N denote the coefficient matrix of the normal equation. Assume that S_N is non-singular for some N. Then we have the following recursion:*

$$S_{N+1}^{-1}=S_N^{-1}-\frac{1}{1+\varphi_{N+1}^TS_N^{-1}\varphi_{N+1}}S_N^{-1}\varphi_{N+1}\varphi_{N+1}^TS_N^{-1}.$$

To get a recursion for $\widehat{\theta}_N$ let us consider the normal equation at time $N+1$:

$$S_{N+1}\widehat{\theta}_{N+1}=\sum_{n=1}^{N+1}\varphi_n y_n. \tag{9.17}$$

Write the right hand side as

$$\sum_{n=1}^{N}\varphi_n y_n+\varphi_{N+1}y_{N+1}=S_N\widehat{\theta}_N+\varphi_{N+1}y_{N+1}.$$

The trick is to express S_N via S_{N+1} as follows: $S_N=S_{N+1}-\varphi_{N+1}\varphi_{N+1}^T$. Substituting this expression into the equality above the normal equation at time $N+1$ becomes:

$$S_{N+1}\widehat{\theta}_{N+1}=(S_{N+1}-\varphi_{N+1}\varphi_{N+1}^T)\widehat{\theta}_N+\varphi_{N+1}y_{N+1}. \tag{9.18}$$

Multiplying by S_{N+1}^{-1}, taking out φ_{N+1}, and using the notation $R_N=S_N/N$, we get the following fundamental result, called the recursive least squares (RLSQ) method:

Proposition 9.10. *(The RLSQ method). Assume that S_N, and thus R_N, is non-singular, and let $\widehat{\theta}_N$ be the LSQ estimate of θ^*. Then $\widehat{\theta}_{N+1}$ and R_{N+1} can be computed via the recursion*

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N + \frac{1}{N+1} R_{N+1}^{-1} \varphi_{N+1}(y_{N+1} - \varphi_{N+1}^T \widehat{\theta}_N) \tag{9.19}$$

$$R_{N+1} = R_N + \frac{1}{N+1}(\varphi_{N+1}\varphi_{N+1}^T - R_N) \tag{9.20}$$

Note that the term $(y_{N+1} - \varphi_{N+1}^T \widehat{\theta}_N)$ is an approximation to $(y_{N+1} - \varphi_{N+1}^T \theta^*)$, which is just the innovation e_{N+1}. Also note that the expectation of the correction term $(y_{N+1} - \varphi_{N+1}^T \theta)$ is zero exactly when $\theta = \theta^*$.

Remark. Setting

$$\overline{R}_N = \frac{1}{N} R_N$$

we can write

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N + \frac{1}{N+1} \overline{R}_{N+1}^{-1} \varphi_{N+1}(y_{N+1} - \varphi_{N+1}^T \widehat{\theta}_N)$$

$$\overline{R}_{N+1} = \overline{R}_N + \frac{1}{N+1}(\varphi_{N+1}\varphi_{N+1}^T - \overline{R}_N)$$

The LSQ method and its recursive version is applicable for any wide sense stationary process to find *the best p-th order one step ahead predictor*, i.e. find the solution of the minimization problem

$$\mathrm{E}(y_n - \varphi_n^T \alpha) \quad \to \quad \min_{\alpha}.$$

The solution of it was found to be the solution of the linear equation in this form, see (9.4):

$$R\alpha = r.$$

Remark. Note that the recursive LSQ estimator above is just a recursive form for the off-line LSQ estimator. It follows that, under the conditions of Proposition 9.3 $\widehat{\theta}_N$ and R_N converge to θ^* and R^*, respectively. On the other hand, the RLSQ method stands on its own: taking any initial values $\widehat{\theta}_0$ and R_0, such that R_0 is positive definite, we can compute a sequence of estimators $\widehat{\theta}_N$ and R_N. If we take this point of view a standard choice for $\widehat{\theta}_0$ would be any a priori (experimental) estimate of θ^*, while R_0 would be $R_0 = \delta I$, with some $\delta > 0$. Surprisingly, the analysis of this modified RLSQ method is orders of magnitude harder, and requires a completely new arsenal of techniques. We are not much better off with the truncated version of the off-line LSQ method either, because it does not have a simple recursive form.

Chapter 10

Identification of MA and ARMA models

10.1 Identification of MA models

Let us now consider the problem of identifying an MA and ARMA model. Surprisingly, this is a much more demanding task than identifying an AR model. Thus, first consider an MA process (y_n) defined by

$$y = C^*(q^{-1})e,$$

with

$$\deg C^* = r, \qquad C^* = \sum_{l=0}^{r} c_l^* q^{-l}, \qquad c_0 = 1,$$

where (e_n) is a w.s.st. orthogonal process. We use the superscript * to indicate that the corresponding parameters are the true, but unknown parameters generating the data. We will use the notation

$$\theta^* = (c_1^*, \ldots, c_r^*)^T.$$

Condition 10.1. *We assume that the polynomial $C^*(z^{-1})$ is stable, i.e. all the roots of the equation $C^*(z^{-1}) = 0$ lie in the open unit disc of the complex plane.*

Note that under this condition e is the innovation process of y.

The key idea in identifying an MA process, widely used in other context as well in system identification, is the (attempted) reconstruction of the driving noise sequence $e_1, ..., e_N$ by inverting the system generating our observed data $y_1, ..., y_N$. Thus, let us take a polynomial $C(q^{-1})$ with

$$\deg C = r, \qquad C = \sum_{l=0}^{r} c_l q^{-l}, \qquad c_0 = 1,$$

and define an estimated driving noise process $\varepsilon = (\varepsilon_n)$ by

$$\varepsilon = C^{-1}(q^{-1})y. \tag{10.1}$$

According to our convention, this equality is to be understood on $-\infty < n < +\infty$. It is easily seen from previous results that if $C(z^{-1}) \neq 0$ for $|z| = 1$, then the process ε is well-defined. The latter equation can also be written as

$$C(q^{-1})\varepsilon = y. \tag{10.2}$$

Now, if data are available only for $n \geq 1$, then (10.2) can solved recursively for ε, assuming that the initial values of ε are given.

As an example, take the inverse of an MA(1) process:

$$\varepsilon_n + c_1\varepsilon_{n-1} = y_n, \qquad n \geq 1.$$

To generate ε_1 we would need to know ε_0 which is not available. In general, for the inversion of an MA(r) system we would need to know the values of ε_n for $-r+1 \leq n \leq 0$. The best we can do to circumvent this difficulty is to take arbitrary initial values for ε_n for $-r+1 \leq n \leq 0$. A standard choice is $\varepsilon_n = 0$ for $-r+1 \leq n \leq 0$. Then, we need to study the effect of the initial value on our estimation procedure.

Altogether, we need to introduce a dual definition of the estimated noise process (ε_n), depending on the time horizon in which we work. We will make this distinction explicit in what follows. Let us now introduce the notation

$$\theta = (c_1, \ldots, c_r)^T.$$

The w.s.st. process ε_n defined by (10.2) defined over $-\infty < n < +\infty$ will be denoted from now on as $\varepsilon_n^*(\theta)$. I.e. $\varepsilon_n^*(\theta)$ is defined by

$$C(q^{-1})\varepsilon^*(\theta) = y, \qquad -\infty < n < +\infty. \tag{10.3}$$

On the other hand, when (10.2) is solved for $n \geq 0$, with zero initial conditions, then the resulting process will be denoted by $(\varepsilon_n(\theta))$. I.e. $\varepsilon_n(\theta)$ is defined by

$$C(q^{-1})\varepsilon(\theta) = y, \qquad n \geq 0, \quad \varepsilon_0(\theta) = \cdots = \varepsilon_{-r+1}(\theta) = 0. \tag{10.4}$$

To ensure that the the choice of initial values does not affect the asymptotic behavior of the estimator we need the following condition:

Condition 10.2. *We assume that the polynomial $C(z^{-1})$ is stable, i.e. all the roots of the equation $C(z^{-1}) = 0$ lie in the open unit disc of the complex plane.*

To see why this condition is useful, note that a state-space realization of the system (10.4) with y as input and $\varepsilon(\theta)$ as output, is obtained by defining the state vector

$$x_n = (\varepsilon_{n-1}(\theta), \ldots, \varepsilon_{n-r}(\theta))^T.$$

Then we have for $n \geq 1$

$$\begin{aligned} x_{n+1}(\theta) &= \widetilde{C} x_n(\theta) + b y_n \qquad x_1(\theta) = 0 && (10.5) \\ \varepsilon_n(\theta) &= b^T \, x_{n+1}(\theta), && (10.6) \end{aligned}$$

where $\widetilde{C}$ is the companion matrix associated with $C(z^{-1})$, and $b = (1, 0, ..., 0)^T$ is a unit vector in $\mathbb{R}^r$. The parallel state-space system, defined over $-\infty < n < \infty$, is written as

$$\begin{aligned} x^*_{n+1}(\theta) &= \widetilde{C} x^*_n(\theta) + b y_n && (10.7) \\ \varepsilon^*_n(\theta) &= b^T \, x^*_{n+1}(\theta). && (10.8) \end{aligned}$$

Note that we have exactly the same dynamics, the two systems differ only in the initialization of the state-vectors. However, the effect of these initial values will asymptotically vanish as the next exercise states.

Exercise 10.1. *Prove that the stability of $C(z^{-1})$, implying the stability of $\widetilde{C}$, yields that*

$$\mathrm{E}\,|x^*_n(\theta) - x_n(\theta)|^2 = O(\gamma^n) \tag{10.9}$$

with any γ such that $\gamma > \varrho(\widetilde{C})$, with $\varrho(\widetilde{C})$ denoting the spectral radius of $\widetilde{C}$ (known to be less then 1).

It then follows that

$$\mathrm{E}\,|\varepsilon^*_n(\theta) - \varepsilon_n(\theta)|^2 = O(\gamma^n). \tag{10.10}$$

Now we are ready to estimate θ^* by considering, in the spirit of the LSQ estimator, the cost function

$$V_N(\theta) = \frac{1}{2} \sum_{n=0}^{N-1} \varepsilon_n^2(\theta).$$

Then, define the estimator of θ^* as the solution of the minimization problem

$$\min_\theta V_N(\theta). \tag{10.11}$$

The range of θ over which minimization is performed is the set

$$D = \{\theta \in \mathbb{R}^r \ : \ C(z^{-1}) = 1 + \sum_{l=1}^{r} c_r z^{-l} \text{ is stable}\}.$$

The resulting estimator $\widehat{\theta}_N$ is called a prediction error (PE) estimator.

When talking about "the" solution of the minimization problem (10.11) we may have been too ambitious, namely the function $V_N(\theta)$ is not known to be convex in θ, hence finding the global minimum of $V_N(\theta)$ over D may be too hard. Therefore we relax our definition as follows:

Definition 10.3. *The prediction error estimator $\widehat{\theta}_N$ of the MA parameter θ^* is a D-valued r.v. variable such that*

$$\frac{\partial}{\partial\theta}V_N(\widehat{\theta}_N) = 0 \tag{10.12}$$

if a solution of $\frac{\partial}{\partial\theta}V_N(\theta) = 0$ exists at all, allowing multiple solutions.

Remark. This definition of $\widehat{\theta}_N$ is still not completely satisfactory, since it implicitly assumes that if there exists a solution, then we can actually find it. Also note that the existence of $\widehat{\theta}_N$ as a random variable in face of multiple solutions is not obvious. In fact, we need to use the so-called measurable selection theorem of Filippov.

Exercise 10.2. *Provide an expression of the coefficients c_r in terms of the roots, say γ_r, and express $\dfrac{\partial}{\partial\gamma}$ via $\dfrac{\partial}{\partial\theta}$.*

After all, let us settle with (10.12) and let us see, how we can compute the left hand side. Obviously, we have

$$\frac{\partial}{\partial\theta}V_N(\theta) = V_{\theta N}(\theta) = \sum_{n=1}^{N}\varepsilon_{\theta n}(\theta)\varepsilon_n(\theta), \tag{10.13}$$

where the subscript θ denotes differentiation w.r.t. θ. To get $\varepsilon_{\theta n}(\theta)$ note that the process $\varepsilon_n(\theta)$, as defined by (10.4), is obtained by a finite recursion starting at time $n = 1$. Therefore, we can differentiate this set of equations without any additional consideration to get

$$\frac{\partial}{\partial\theta}C(q^{-1})\varepsilon(\theta) + C(q^{-1})\frac{\partial}{\partial\theta}\varepsilon(\theta) = 0. \tag{10.14}$$

Obviously, the initial values for $\frac{\partial}{\partial\theta}\varepsilon_n(\theta) = \varepsilon_{\theta n}(\theta)$ will be 0 for $n \leq 0$. Now

$$\frac{\partial}{\partial\theta_l}C(q^{-1}) = q^{-l},$$

and thus

$$\frac{\partial}{\partial\theta}C(q^{-1}) = (q^{-1}, \ldots, q^{-r})^T.$$

The action of the r.h.s on the sequence $\varepsilon_n(\theta)$ results in $(\varepsilon_{n-1}(\theta), ..., \varepsilon_{n-r}(\theta))^T$. Introducing the notation

$$\phi_n(\theta) = (\varepsilon_{n-1}(\theta), ..., \varepsilon_{n-r}(\theta))^T,$$

substituting this into (10.14), and rearranging it we come to the following conclusion:

Lemma 10.4. *The gradient process $\varepsilon_\theta(\theta)$ satisfies, with zero initial conditions,*

$$C(q^{-1})\varepsilon_\theta(\theta) = -\phi(\theta). \tag{10.15}$$

From the above arguments it readily follows that the equation defining the PE estimator, i.e. equation (10.12) is non-linear in θ. Therefore, the asymptotic analysis of the estimator requires a lot of technicalities even on a heuristic level.

10.2 The asymptotic covariance matrix of $\widehat{\theta}_N$

In this section we shall give an outline for the computation of the asymptotic covariance matrix of $\widehat{\theta}_N$ only. Consider the equation (10.12)

$$\frac{\partial}{\partial\theta} V_N(\widehat{\theta}_N) = 0,$$

and make a Taylor-series expansion around $\widehat{\theta}_N$, and evaluate $V_{\theta N}(\widehat{\theta})$ for $\theta = \theta^*$:

$$V_{\theta N}(\theta^*) = V_{\theta N}(\widehat{\theta}_N) + \int_0^1 V_{\theta\theta N}(\overline{\theta})(\lambda)\; d\lambda \cdot (\theta^* - \widehat{\theta}), \tag{10.16}$$

where $\overline{\theta}(\lambda) = \lambda\theta^* + (1-\lambda)\widehat{\theta}_N$. Now the Hessian under the integral will be approximated so that we replace $\overline{\theta}(\lambda)$ by θ^*, and then using (10.13) we write

$$V_{\theta\theta N}(\theta^*) = \sum_{n=1}^{N} \Big(\varepsilon_{\theta\theta n}(\theta^*)\varepsilon_n(\theta^*) + \varepsilon_{\theta n}(\theta^*)\varepsilon_{\theta n}^T(\theta^*) \Big).$$

In the next step of the approximation, we replace the computable values of $\varepsilon_n(\theta)$ and their derivatives by their stationary variants initiated at $-\infty$. To be more specific, consider (10.14) defining $\varepsilon_\theta(\theta)$. On its r.h.s. replace $\varepsilon_n(\theta)$ by its stationary variant $\varepsilon_n^*(\theta)$, define $\phi_n^*(\theta)$ accordingly, and consider (10.15) defined for $-\infty < n < \infty$. Then we get a w.s.st. process $\varepsilon_\theta^*(\theta)$ defined by

$$C(q^{-1})\varepsilon_\theta^*(\theta) = -\phi^*(\theta) \qquad -\infty < n < \infty \tag{10.17}$$

such that, in analogy with (10.9), we have

$$\mathrm{E}\,|\varepsilon_{\theta n}^*(\theta) - \varepsilon_{\theta n}(\theta)|^2 = O(\gamma^n).$$

We can proceed with the second derivatives similarly. (Note that have not claimed that $\varepsilon_{\theta n}^*(\theta)$ is the derivative of $\varepsilon_n^*(\theta)$ in any sense, although the latter is indeed the case in an appropriate sense). Setting $\theta = \theta^*$ we get $\varepsilon_n^*(\theta^*) = e_n$. Finally, assuming that a strong law of large number holds, we get that

$$\lim_N \frac{1}{N} V_{\theta\theta N}(\theta^*) = \mathrm{E}\Big(\varepsilon_{\theta\theta n}^*(\theta^*)e_n + \varepsilon_{\theta n}^*(\theta^*)\varepsilon_{\theta n}^{*T}(\theta^*) \Big).$$

Exercise 10.3. *Show that the first term on the r.h.s. of the above equality is zero.*

Introducing the notation

$$R^* = \mathrm{E}\varepsilon^*_{\theta n}(\theta^*)\varepsilon^{*T}_{\theta n}(\theta^*),$$

and approximating the l.h.s of (10.16) using stationary variants of $\varepsilon_n(\theta)$ and their derivatives, and taking into account that $V_{\theta N}(\widehat{\theta}_N) = 0$, we get an approximation for the error $(\widehat{\theta}_N - \theta^*)$ called $\tilde{\tilde{\theta}}_N$ defined by the equation

$$\frac{1}{N}\sum_{n=1}^{N}\varepsilon^*_{\theta n}(\theta^*)e_n = -R^*\tilde{\tilde{\theta}}_N.$$

From here we get

$$\tilde{\tilde{\theta}}_N = -(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\varepsilon^*_{\theta n}(\theta^*)e_n. \tag{10.18}$$

Now for the covariance matrix of $\tilde{\tilde{\theta}}_N$ we get a mirror image of the corresponding result for AR-processes, given as Proposition 9.6 in Chapter 9:

Proposition 10.5. *Assume that $C^*(z^{-1})$ is stable, and that the driving noise sequence (e_n) satisfies Condition 9.4. Then the approximating error process $\tilde{\tilde{\theta}}_N$ defined under (10.18) has the following covariance matrix:*

$$\mathrm{E}\tilde{\tilde{\theta}}_N\tilde{\tilde{\theta}}^T_N = \frac{1}{N}(R^*)^{-1}\sigma^2(e).$$

Just like in the AR case, the above result provides a guideline for the proof of an exact result. Thus we get, that using an appropriate truncation procedure we can define a new prediction error estimator $\overline{\theta}_N$ for which we have, under additional technical conditions,the following result: the truncated prediction error estimate $\overline{\theta}_N$ is asymptotically unbiased and its asymptotic covariance matrix is exactly what we have obtained for the approximating error $\tilde{\tilde{\theta}}_N$:

$$\lim_N N\mathrm{E}(\overline{\theta}_N - \theta^*)(\overline{\theta}_N - \theta^*)^T = \sigma(e)^2(R^*)^{-1}.$$

To interpret R^* note that for $\theta = \theta^*$ we have $\phi^*_n(\theta^*) = (e_{n-1}, ..., e_{n-r})$, and so we get

$$C^*(q^{-1})\varepsilon^*_\theta(\theta^*) = -(e_{n-1}, ..., e_{n-r}).$$

It follows that the gradient process $\varepsilon^*_\theta(\theta^*)$ is identical with the state process of an AR(r)-process defined by

$$C^*(q^{-1})v = -e.$$

A remarkable feature of the above result is that it implies that the asymptotic covariance matrix of the PE estimator of the parameters of the MA system

$$y = C^*(q^{-1})e$$

is the same as the asymptotic covariance matrix of the LSQ estimator of the parameters of the AR system

$$C^*(q^{-1})y = e$$

Consider the example of an MA(1) process:

$$y_n = e_n + c^* e_{n-1}.$$

We have seen that $R^* = \sigma^2(e)(1-(c^*)^2)^{-1}$, and thus the asymptotic variance of the PE estimator of c^* equals

$$1-(c^*)^2.$$

It follows, that if as c^* is close to ± 1, then the asymptotic variance of the PE estimator is close to 0. In contrast to the AR case, there is no direct evidence for this phenomenon, in fact it is quite a surprise.

Remark. The above observation can be generalized to saying that a transfer function $H(e^{-i\omega}, \theta^*)$, depending on a parameter θ^*, and its inverse $H^{-1}(e^{-i\omega}, \theta^*)$ can be equally accurately estimated, at least asymptotically.

To outline the proof assume that θ^* is a scalar. Then, it is easily seen that

$$y = H(q^{-1}, \theta^*)e$$

implies

$$\varepsilon^*_\theta(\theta^*) = H^{-1}(q^{-1}, \theta^*)H_\theta(q^{-1}, \theta^*)e.$$

The latter can be written, at least *formally*, as

$$\varepsilon^*_\theta(\theta^*) = \frac{\partial}{\partial\theta}\left(\log\ H(q^{-1}, \theta)\right)_{|\theta=\theta^*} e.$$

Switching $H(q^{-1}, \theta)$ for its inverse will change only the sign of $\varepsilon^*_\theta(\theta^*)$, thus R^* will be unaffected.

The question arises, how to proceed when $C(z^{-1})$ is not stable. Here we need the following observation. If (y_n) is a w.s.st. process that is observed for $1 \le n < +\infty$ then we may be able to reconstruct its auto-covariance function $r(\tau)$, and hence its spectral density given by

$$f(\omega) = |C(e^{-i\omega})|^2\sigma^2(e).$$

But there seems no way to reconstruct the spectral factor $C(e^{-i\omega})$ itself, unless we specify that we are looking for a spectral factor with additional specification such as stability. Therefore, may redefine our identification problem by saying that we are looking for an MA representation of (y_n) such that $C(z^{-1})$ is stable. Such a reformulation of the problem is feasible whenever the original polynomial $C(z^{-1})$ does not have a zero on the unit circle, or equivalently, whenever $f(\omega) \neq 0$ for $0 \le \omega \le 2\pi$.

10.3 Identification of ARMA models

Let us now consider the problem of identifying an ARMA model. Thus, consider an ARMA process (y_n) defined by

$$A^*(q^{-1})y = C^*(q^{-1})e,$$

with $\deg A^* = p$ and $\deg C^* = r$,

$$A^* = \sum_{k=0}^{p} a_k^* q^{-k}, \qquad C^* = \sum_{l=0}^{r} c_l^* q^{-l}, \qquad a_0^* = c_0^* = 1,$$

where (e_n) is a w.s.st. orthogonal process. As always, we use the superscript * to indicate that the corresponding parameters are the true, but unknown parameters generating the data. We will use the notation

$$\theta^* = (a_1^*, \dots, a_p^*, c_1^*, \dots, c_r^*)^T.$$

We need the following condition:

Condition 10.6. *We assume that the polynomials $A^*(z^{-1})$ and $C^*(z^{-1})$ are stable.*

Note that under this condition e is the innovation process of y.

A new feature of the problem of identifying an ARMA model is that the observed data determine only the spectral density, which is $|C^*(e^{-i\omega})/A^*(e^{-i\omega})|^2$, and thus if $A^*(z^{-1})$ and $C^*(z^{-1})$ have a common factor then this will not be identifiable. Therefore we impose the following condition:

Condition 10.7. *We assume that the polynomials $A^*(z^{-1})$ and $C^*(z^{-1})$ are relative prime, i.e. they do not have any non-trivial common factor.*

We could try to use the Least Squares method that was appropriate for the AR case. Rearrange the ARMA equation as

$$y_n = -a_1^* y_{n-1} - ... - a_p^* y_{n-p} + e_n + c_1^* e_{n-1} + \cdots + c_r^* e_{n-r} = -a_1^* y_{n-1} - ... - a_p^* y_{n-p} + f_n. \quad (10.19)$$

Let us try to identify the parameters a_k^*. Define

$$\varphi_n = (-y_{n-1}, \dots, -y_{n-p})^T.$$

Multiplying the above equation by φ_n^T from the left, and taking expectation, unfortunately in general

$$E\varphi_n f_n \neq 0,$$

hence the instrumental variable interpretation of the LSQ method does not work.

Following the method for identifying an MA process, we attempt to reconstruct the driving noise sequence $e_1, ..., e_N$ by inverting the system generating our observed data $y_1, ..., y_N$. Thus, let us take a polynomials $A(q^{-1})$ and $C(q^{-1})$ with with $\deg A = p$ and $\deg C = r$,

$$A = \sum_{k=0}^{p} a_k q^{-k}, \qquad C = \sum_{l=0}^{r} c_l q^{-l}, \qquad a_0 = c_0 = 1,$$

and define an estimated driving noise process $\varepsilon = (\varepsilon_n)$ by

$$C(q^{-1})\ \varepsilon = A(q^{-1})\ y. \tag{10.20}$$

According to our convention, this equality is to be understood on $-\infty < n < +\infty$. It is easily seen from previous results that if $C(z^{-1}) \neq 0$ for $|z| = 1$, then the process ε is well-defined. Now, if data are available only for $n \geq 1$, then (10.2) can be solved recursively for ε, assuming that the initial values of y and ε are given. In this case, we set $y_n = \varepsilon_n = 0$ for $n \leq 0$. Altogether, we need to introduce a dual definition of the estimated noise process (ε_n), depending on the time horizon in which we work. We will make this distinction explicit in what follows.

Let us now introduce the notation

$$\theta = (a_1, \ldots, a_p, c_1, \ldots, c_r)^T.$$

The w.s.st. process (ε_n) defined by (10.20) defined over $-\infty < n < +\infty$ will be denoted from now on as $(\varepsilon_n^*(\theta))$, i.e. $(\varepsilon_n^*(\theta))$ is defined by

$$C(q^{-1})\ \varepsilon^*(\theta) = A(q^{-1})\ y, \tag{10.21}$$

$-\infty < n < +\infty$. On the other hand, when (10.2) is solved for $n \geq 0$, with zero initial conditions, then the resulting process will be denoted by $(\varepsilon_n(\theta))$ i.e. $(\varepsilon_n(\theta))$ is defined by

$$C(q^{-1})\ \varepsilon(\theta) = A(q^{-1})\ y, \qquad n \geq 0, \tag{10.22}$$

with $y_n = \varepsilon_n(\theta) = 0 \quad \text{for} \quad n \leq 0$. To ensure that the choice of initial values does not affect the asymptotic behavior of the estimator we need the following condition:

Condition 10.8. *We assume that the polynomial $C(z^{-1})$ is stable, i.e. all the roots of the equation $C(z^{-1}) = 0$ lie in the open unit disc of the complex plane.*

It then follows, just like in the MA case, that

$$\mathrm{E}\,|\varepsilon_n^*(\theta) - \varepsilon_n(\theta)|^2 = O(\gamma^n),$$

with some $0 < \gamma < 1$, and similar approximations hold for the derivatives of $\varepsilon_n(\theta)$. Now we are ready to estimate θ^* by considering the cost function

$$V_N(\theta) = \frac{1}{2} \sum_{n=0}^{N-1} \varepsilon_n^2(\theta).$$

Then, define the estimator of θ^* as the solution of the minimization problem

$$\min_{\theta} V_N(\theta).$$

The range of θ over which minimization is performed is the set

$$D = \{\theta \in \mathbb{R}^r \ : \ A(z^{-1}),\ C(z^{-1}) \text{ are stable and relative prime.}\}.$$

Remark. . A more transparent parametrization can be given in terms of poles and zeros, however, due to the non-linear dependence of the coefficients of $A(z^{-1})$ and $C(z^{-1})$ on the respective roots, the computations that follow become more complicated.

The resulting estimator $\widehat{\theta}_N$ is called a prediction error (PE) estimator. Repeating the arguments given in the MA case, we redefine the notion of "the" solution as follows:

Definition 10.9. *The prediction error estimator $\widehat{\theta}_N$ of the ARMA parameter θ^* is a D-valued r.v. variable such that*

$$\frac{\partial}{\partial\theta} V_N(\widehat{\theta}_N) = 0 \tag{10.23}$$

if a solution of $\frac{\partial}{\partial\theta} V_N(\theta) = 0$ exists at all, allowing multiple solutions.

After all, let us settle with (10.23) and let us now see, how can we compute the left hand side of (10.23). Obviously, we have

$$\frac{\partial}{\partial\theta} V_N(\theta) = V_{\theta N}(\theta) = \sum_{n=1}^{N} \varepsilon_{\theta n}(\theta)\varepsilon_n(\theta), \tag{10.24}$$

where the subscript θ denotes differentiation w.r.t. θ. To get $\varepsilon_{\theta n}(\theta)$ note that the process $\varepsilon_n(\theta)$, as defined by (10.22), is obtained by a finite recursion starting at time $n = 1$. Therefore, we can differentiate this set of equations without any additional consideration w.r.t. any coordinates , say η of $\theta = (a, c)$ as follows:

$$\frac{\partial}{\partial\eta} C(q^{-1})\ \varepsilon(\theta) + C(q^{-1})\ \frac{\partial}{\partial\eta}\varepsilon(\theta) = \frac{\partial}{\partial\eta} A(q^{-1})\ y.$$

Setting $\eta = a_k$ and $\eta = c_l$, respectively, we get

$$C(q^{-1})\ \frac{\partial}{\partial a}\varepsilon(\theta) \ = \ \frac{\partial}{\partial a} A(q^{-1})\ y = (q^{-1}, ..., q^{-p})y, \tag{10.25}$$

$$C(q^{-1})\ \frac{\partial}{\partial c}\varepsilon(\theta) \ = \ -(q^{-1}, ..., q^{-r})\ \varepsilon(\theta). \tag{10.26}$$

The initial values for $\frac{\partial}{\partial\theta}\varepsilon_n(\theta) = \varepsilon_{\theta n}(\theta)$ will be 0 for $n \leq 0$. Introducing the notation

$$\phi_n(\theta) = (-y_{n-1}, ..., -y_{n-p}, \varepsilon_{n-1}(\theta), ..., \varepsilon_{n-r}(\theta))^T,$$

substituting this into (10.26), and rearranging it we come to the following conclusion:

Lemma 10.10. *The gradient process $\varepsilon_\theta(\theta)$ satisfies, with zero initial conditions,*

$$C(q^{-1})\ \varepsilon_\theta(\theta) = -\phi(\theta).$$

Using this result we will derive a neat formula for the asymptotic covariance matrix of the estimator $\widehat{\theta}_N$. Following the arguments given for MA processes replace the computable values of $\varepsilon_n(\theta)$ and their derivatives by their stationary variants initiated at $-\infty$. Thus we get the processes $\varepsilon^*(\theta)$, $\phi^*(\theta)$ and $\varepsilon^*_\theta(\theta)$, the latter defined by

$$C(q^{-1})\varepsilon^*_\theta(\theta) = -\phi^*(\theta), \qquad -\infty < n < \infty.$$

Setting $\theta = \theta^*$ we get $\varepsilon^*_n(\theta^*) = e_n$. Introducing the notation

$$R^* = \mathrm{E}\varepsilon^*_{\theta n}(\theta^*)\varepsilon^{*T}_{\theta n}(\theta^*),$$

we get an approximation for the error $(\widehat{\theta}_N - \theta^*)$ given by

$$\tilde{\tilde{\theta}}_N = -(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\varepsilon^*_{\theta n}(\theta^*)e_n. \tag{10.27}$$

Now for the covariance matrix of $\tilde{\tilde{\theta}}_N$ we get a mirror image of the corresponding result for AR and MA processes, see Proposition 9.6 and 10.27:

Proposition 10.11. *Assume that $A^*(z^{-1})$ and $C^*(z^{-1})$ are stable, and that the driving noise sequence (e_n) satisfies Condition 9.4. Then the approximating error process $\tilde{\tilde{\theta}}_N$ defined under (10.18) has the covariance matrix:*

$$\mathrm{E}\tilde{\tilde{\theta}}_N\tilde{\tilde{\theta}}_N^T = \frac{1}{N}(R^*)^{-1}\sigma^2(e).$$

Exercise 10.4. *Show that if $A^*(z^{-1})$ and $C^*(z^{-1})$ have a common factor then R^* is singular.*

The asymptotic covariance matrices of the PE estimators, assuming unit variance for the noise, are displayed for our three benchmark ARMA$(2,2)$-processes below:

Just like in the AR and MA case, the above result provides a guideline for the proof of an exact result. Thus we get, that using an appropriate truncation procedure we can define a new prediction error estimator $\overline{\theta}_N$ for which we have, under additional technical conditions,the following result: the truncated prediction error estimate $\overline{\theta}_N$ is asymptotically unbiased and its asymptotic covariance matrix is exactly what we have obtained for the approximating error $\tilde{\tilde{\theta}}_N$:

$$\lim_N N\mathrm{E}(\overline{\theta}_N - \theta^*)(\overline{\theta}_N - \theta^*)^T = \sigma(e)^2(R^*)^{-1}.$$

$$(R^*)^{-1} = \begin{bmatrix} 1.35 & -0.880 & -0.703 & -0.109 \\ -0.880 & 1.20 & 0.674 & -0.263 \\ -0.703 & 0.674 & 0.784 & -0.164 \\ -0.109 & -0.263 & -0.164 & 0.530 \end{bmatrix}$$

Figure 10.1: The asymptotic covariance matrix of the PE estimator for an ARMA$(2,2)$ process with similar AR poles and MA zeros. The actual poles: length 0.8, arguments $\pm 0.3\pi$, the actual zeros: length 0.9, arguments $\pm 0.4\pi$.

$$(R^*)^{-1} = \begin{bmatrix} 1.02 & 0.0684 & -0.768 & -0.767 \\ 0.0684 & 0.631 & 0.212 & 0.190 \\ -0.768 & 0.212 & 1.40 & 1.37 \\ -0.767 & 0.190 & 1.37 & 1.38 \end{bmatrix}$$

Figure 10.2: The asymptotic covariance matrix of the PE estimator for an ARMA$(2,2)$ process with complex AR poles with small negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.6\pi$, the actual zeros: $-0.6, -0.9$.

$$(R^*)^{-1} = \begin{bmatrix} 10.3 & 8.68 & -11.3 & -10.9 \\ 8.68 & 7.50 & -9.36 & -9.14 \\ -11.3 & -9.36 & 13.2 & 12.7 \\ -10.9 & -9.14 & 12.7 & 12.2 \end{bmatrix}$$

Figure 10.3: The asymptotic covariance matrix of the PE estimator for an ARMA$(2,2)$ process with complex AR poles with large negative real part combined with two negative MA zeros. The actual poles: length 0.8, arguments $\pm 0.9\pi$, the actual zeros: $-0.6, -0.9$.

Exercise 10.5. *Compute the gradient process for the following models: MA*(1)*, AR*(1)*, ARMA*$(1,1)$*.*

Exercise 10.6. *Show that for* $\theta = \theta^*$ *we have*

$$\varepsilon_\theta(\theta)_{|\theta=\theta^*} = (-\frac{1}{A^*}\left[q^{-1}\dots q^{-p}\right] e, \frac{1}{C^*}\left[q^{-1}\dots q^{-r}\right] e)^T.$$

Chapter 11

Non-stationary models

Many economic time series exhibit non-stationary behavior due to such factors as trends or seasonality. The purpose of this chapter is to give a brief summary of some of the basic ideas in the theory of non-stationary processes. Our main reference is [21].

A classical approach in analyzing a non-stationary time series would be to remove non-stationarity by some appropriate procedure, such as taking the difference process in the case of the presence of trend. Then we could use the theory of stationary time series for the residual process. As we have seen, the theory of stationary time series is well developed. We mention here only one additional source, the book of Box and Jenkins [12]. For the more mathematically skilled reader a very useful, although not easily readable book, is the book of Hannan and Deistler [28],

The structure of the chapter is the following. In the first section we discuss the notion of integrated processes. In its simplest form it is just a random walk. We show that the LSQ estimation of the pole $\alpha = 1$, responsible for the integrating effect, converges with a rate faster than the usual $N^{-1/2}$. In the next section we consider a special class of integrated *v*ector processes, the individual components of which have an integrator effect, but there exists a nontrivial linear combinations of these components, or simply said a projection of the vector process which is is stationary. Then we give the maximum likelihood (ML) estimation of the projection subspace, as computed first in [32]. Finally, in the last section we consider fractionally integrated processes exhibiting a certain long-memory behavior, originally introduced in the physical sciences.

11.1 Integrated models

Definition 11.1. *A stochastic process $y = (y_n)$, $-\infty < n < +\infty$, is called integrated of order 1 if y is non-stationary, but the difference process $[1 - q^{-1}]y$, defined by*

$$\big((1 - q^{-1})y\big)_n = y_n - y_{n-1} \tag{11.1}$$

is wide sense stationary, not necessarily of zero mean.

Here we shall consider the case when $(1-q^{-1})y$ is a w.s.st. ARMA process. Thus we consider processes defined by the dynamics

$$(1-q^{-1})A(q^{-1})y = C(q^{-1})e. \tag{11.2}$$

Example. Consider the special case with $A = C = 1$, and define a process y by the dynamics

$$y_n = y_{n-1} + c + \varepsilon_n,$$

where ε is a white nose process, i.e a w.s.st. orthogonal process. Then (y_n) has a linear trend both in the mean and the variance, while $y_n - y_{n-1} = c + \varepsilon_n$ is stationary.

In estimating the parameters of an integrated process we may formally proceed by applying a prediction error method. In the case of an integrated AR process this boils down to a LSQ method. Taking $C = 1$, and writing (11.2) in the form

$$A_1(q^{-1})A_2(q^{-1})y = e$$

and pretending that the processes y_1 and y_2 defined by

$$A_1(q^{-1})y_1 = e \quad \text{and} \quad A_2(q^{-1})y_2 = e,$$

are known, we can study the estimation problem of $A_1(q^{-1})$ and $A_2(q^{-1})$ separately.

Exercise 11.1. *Express the asymptotic covariance matrix of the LSQ estimator of the parameters of $A_1(q^{-1})$ in terms of A_2 and the auto-covariance matrix of y.*

Omitting further details let us see how the LSQ method work for the simplest integrated process. A remarkable fact is that the rate of convergence of the LSQ estimates will be of the order faster than $N^{-1/2}$.

Thus we consider a scalar AR(1) process

$$y_n = \rho y_{n-1} + \varepsilon_n, \qquad y_0 \text{ given},$$

where (ε_n) is white noise with variance σ^2. Then the LSQ estimator of ρ is given by

$$\widehat{\rho}_N = \frac{\sum_{n=1}^{N} y_n y_{n-1}}{\sum_{n=1}^{N} y_n^2}. \tag{11.3}$$

It is well-known, that if $|\rho| < 1$, then $\widehat{\rho}_N$ is a consistent estimator of ρ and

$$\sqrt{N}(\widehat{\rho}_N - \rho) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0,\ 1-\rho^2).$$

Let us now consider the limiting, non-stationary case when $\rho = 1$. Assume that $y_0 = 0$, then

$$y_n = \varepsilon_1 + \cdots + \varepsilon_n.$$

Define the piecewise linear process X_N by

$$X_N(r) = \begin{cases} ry_1 & 0 \le r \le \frac{1}{N} \\ \vdots & \\ \frac{y_k}{N} + (r - \frac{k}{N})\frac{y_{k+1} - y_k}{N} & \frac{k}{N} \le r \le \frac{(k+1)}{N} \\ \vdots & \\ \frac{y_N}{N} & r = 1 \end{cases}$$

Then we have by Donsker's functional central limit theorem that the process $X_N(.)$ converges weakly to the standard Brownian motion $W(\cdot)$:

$$\frac{\sqrt{N}X_N(\cdot)}{\sigma} \quad \xrightarrow{\mathcal{L}} \quad W(\cdot).$$

For the LSQ estimator we have

$$\widehat{\rho}_N - 1 = \frac{\sum_{n=1}^{N} y_n y_{n-1}}{\sum_{n=1}^{N} y_n^2} - \frac{\sum_{n=1}^{N} y_{n-1}^2}{\sum_{n=1}^{N} y_{n-1}^2},$$

thus, replacing $\sum_{n=1}^{N} y_n^2$ by $\sum_{n=1}^{N} y_{n-1}^2$ in the first term on the r.h.s. we get

$$N(\widehat{\rho}_N - 1) \cong \frac{\frac{1}{N}\sum_{n=1}^{N} y_{n-1}\varepsilon_n}{\frac{1}{N^2}\sum_{n=1}^{N} y_{n-1}^2}. \tag{11.4}$$

For the numerator of this fraction we have

$$\frac{1}{N}\sum_{n=1}^{N} y_{n-1}\varepsilon_n = \frac{1}{N}\sum_{n=1}^{N} \varepsilon_n(\sum_{s=1}^{n-1} \varepsilon_s) = \frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{n-1} \varepsilon_n\varepsilon_s =$$

$$= \frac{1}{2}\left(\frac{(\sum_{n=1}^{N} \varepsilon_n)^2}{n} - \frac{\sum_{n=1}^{N} \varepsilon_n^2}{N}\right) = \frac{1}{2}\big(NX_N^2(1) - \frac{\sum_{n=1}^{N} \varepsilon_n^2}{N}\big).$$

As $NX_N^2(1) \to \sigma^2 W^2(1)$ in law, and $\sum_{n=1}^N \varepsilon_n^2/N \to \sigma^2$ w.p.1, we have

$$\frac{1}{N}\sum_{n=1}^{N} y_{n-1}\varepsilon_n \quad \xrightarrow{\mathcal{L}} \quad \frac{1}{2}(W^2(1)-1)\sigma^2.$$

To compute the denumerator of (11.4), first note that we have

$$\int_0^1 X_N^2(r)dr = \frac{\sum_{n=1}^N y_{n-1}^2}{N^3}.$$

Thus due to the continuous mapping theorem we get

$$\frac{1}{N^2}\sum_{n=1}^{N} y_{n-1}^2 = \int_0^1 NX_N^2(r)dr \quad \xrightarrow{\mathcal{L}} \quad \int_0^1 \sigma^2 W^2(r)dr,$$

since $NX_N^2(\cdot) \quad \xrightarrow{\mathcal{L}} \quad \sigma^2 W^2(\cdot)$.

Finally we get the following result

Proposition 11.2. *We have*

$$N(\widehat{\rho}_N - 1) \quad \xrightarrow{\mathcal{L}} \quad \frac{\frac{1}{2}(W^2(1)-1)}{\int_0^1 W^2(r)dr}.$$

Thus in the case $\rho = 1$ the convergence of LSQ estimator is faster than in the stationary case. It is N-consistent rather then $\sqrt{N}$-consistent.

A number of results for approximately integrated processes has been obtained by Hungarian mathematicians, in particular by Gy. Pap, and his co-workers, see: http://www.math.u-szeged.hu/ papgy/. For an early work of his we refer to [47].

11.2 Co-integrated models

Definition 11.3. *An integrated $\mathbb{R}^m$-valued vector-process (y_n) is called* co-integrated, *if $\exists \alpha \in \mathbb{R}^m$, $\alpha \neq 0$ such that (αy_n) is stationary. The number of linearly independent α-s such that (αy_n) is stationary is called the* co-integrity rank.

A possible structure of a co-integrated vector process can be described as follows. First, let (y_n) be simply an $\mathbb{R}^m$-valued integrated vector-process, i.e. let

$$(1 - q^{-1})y = u,$$

where u is a w.s.st. $\mathbb{R}^m$-valued vector-process. Assume that u is completely regular, i.e.

$$u = k(q^{-1})\varepsilon,$$

where ε is a w.s.st. $\mathbb{R}^m$-valued, orthogonal vector-process, being the innovation process of u, and $k(q^{-1})$ is a $\mathbb{R}^{m\times m}$ dimensional matrix-valued operator such that

$$k(z^{-1}) = \sum_0^\infty k_j z^{-j}$$

converges in $L_2^{m\times m}[0, 2\pi]$ for $z = e^{2\pi\omega}$. It follows that $k(z^{-1})$ is analytic for $|z| > 1$. Write

$$k(z^{-1}) = k(1) + \widetilde{k}(z^{-1}),$$

where $\widetilde{k}(1) = 0$, and thus

$$\widetilde{k}(z^{-1}) = (1 - z^{-1})l(z^{-1}),$$

where $l(z^{-1})$ is analytic for $|z| > 1$. Then we can write

$$y_n = (1 - q^{-1})k(1)\varepsilon_n + l(q^{-1})\varepsilon_n. \tag{11.5}$$

In the above equation the second term on the right hand side, $l(q^{-1})\varepsilon_n$ is w.s.st. Thus (y_n) is co-integrated, if $k(1)$ is singular, but not zero.

11.3 Long memory models

Considering the auto-correlation function (ρ_k) of an ARMA process it is easily seen that it decays exponentially fast to 0. On the other hand, the auto-correlation function of the simplest integrated process (random walk) is a constant, namely $(\rho_k) = 1$ for all k. In analyzing financial data the need for models with slowly decaying auto-correlation functions arises. Thus we come to the concept of with *long range dependence*. A special class of such processes will be now discussed.

Definition 11.4. *The process (y_n) is called fractionally integrated (of order d) or ARFIMA, if*

$$(1 - q^{-1})^d y = u, \tag{11.6}$$

where $0 < d < 1$, and u is a stationary ARMA process.

In the analysis of an ARFIMA model we estimate both the parameter d, and the ARMA parameters of (u_n). We can represent (y_n) formally by multiplying (11.6) by the inverse of

$$(1 - q^{-1})^d = 1 - dq^{-1} + d(d-1)\frac{q^{-2}}{2!} - ...$$

In this way we get an infinite MA representation of the process y in terms of u, which in turm can be expressed via its innovation process, say e. It turns out, that for $d \in (0,\ 0.5)$ the process y is well-defined as a w.s.st. process. An explanation for this can be given using spectral theory. Let us write the spectral density of (u_n) in the form

$$f_u(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{b(e^{-i\lambda})}{a(e^{-i\lambda})}\right|^2,$$

Then, assuming that y is stationary, its spectral density equals:

$$f_y(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{b(e^{-i\lambda})}{a(e^{-i\lambda})}\right|^2 |1-e^{-i\lambda}|^{-2d}.$$

It is clearly seen that $f_y(\lambda)$ is indeed integrable when $d \in (0,\ 0.5)$ and thus (y_n) is well-defined. Note that

$$f_y(0) = \infty,$$

indicating that low-frequencies have a dominant role in the construction of y, which is a nice explanation for long memory.

To estimate d we first estimate the spectral density itself. Computing the logarithm of $f_y(\lambda)$, and using the fact, that

$$|1-e^{-i\lambda}|^2 = |1-\cos\lambda + i\sin\lambda|^2 = 4\sin^2\frac{\lambda}{2},$$

we get

$$\log f_y(\lambda) = \log\frac{f_u(\lambda)}{f_u(0)} + \log f_u(0) - d\big(\log 4\sin^2\frac{\lambda}{2}\big).$$

Based on this formula we can estimate parameter d using a linear regression (cf. [27]).

Chapter 12

Stochastic volatility: ARCH and GARCH models

An important class of technical models are the so-called ARCH and GARCH models. Their simplicity and capability to reproduce some important features of financial time series lead to their unprecedented popularity. In addition, mathematically tractable technical models, such as GARCH models, are getting more attention recently with increasing interest in automated trading.

At this point we briefly collect a few important features of time series of financial data. These features, called "stylized facts", are crucial in building a model for financial time series.

12.1 Some stylized facts of asset returns

Consider a time series of "raw financial data" given as the time series of prices P_n, $n = 1, \ldots, N$, of a certain asset, such as the stock of a company, a stock index or the price of a foreign currency. The observations are assumed to have been taken at equidistant moments. Normally prices tend to increase. A number of empirical studies, including those of the father modern financial mathematics, Louis Bachelier, lead to the (slightly arguable) assumption that the price process is a stochastic process with stationary, independent, Gaussian increments.

This would imply that prices can take on negative values. This anomaly has been rectified by the economist Paul Samuelson by formulating the alternative hypothesis that the infinitesimal returns rather than the price process itself, or equivalently, the logarithm of the price process is a stochastic process with stationary, independent, Gaussian increments. Therefore we rather focus on *log-returns* defined as

$$y_n = \log P_n - \log P_{n-1} = \log\left(1 + \frac{P_n - P_{n-1}}{P_{n-1}}\right).$$

For small values log-returns are close to the relative returns defined as

$$y'_n = \frac{P_n - P_{n-1}}{P_{n-1}},$$

which describes the relative change of the price process over time.

Figure 12.1 displays the daily closing prices and log-returns of the Standard and Poor's 500 Composite Stock Price Index (S&P 500) over the period January 1, 1950 through January 28, 2011.[1]

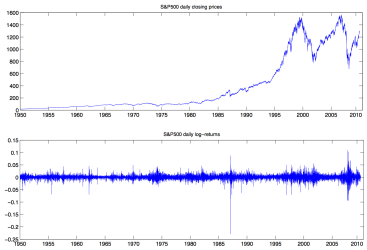

Figure 12.1: S&P 500 daily closing prices and log-returns from January 1, 1950 to January 28, 2011

The study of statistical properties of financial time series revealed a wealth of stylized facts which seem to be common in a wide range of financial time series (see e.g. Cont [18] and Pagan [46]). We mention a few of them. First, in many cases the unconditional distribution of the log-returns has a *heavy tail*, meaning that the tail probabilities are sub-exponential: for any $\lambda > 0$ we have

$$\lim_{x\to\infty} e^{\lambda x} P(|X| > x) = \infty.$$

Second, the distribution of the log-returns has a positive excess kurtosis.

Then, it is found that the *autocorrelations of returns are often insignificant*. We should mention though that insignificant autocorrelation is a world apart from zero autocorrelation, or even more from the assumption of independent log-returns, which has been a prevailing hypothesis since the works of Paul Samuelson, improving the earlier, even more unrealistic assumption of Louis Bachelier on independent increments of prices. In fact, the assumption of independent log-returns would bring us to the conclusion that the variance of price processes, even if discounted to take care of the effect of inflation, tend to infinity, which is a conclusion certainly against our common sense and experience.

[1] Information about the composition of this index and historical data can be found at the address http://www.standardandpoors.com/indices.

Finally, there is the phenomenon of *volatility clustering* phenomena, meaning that long periods of low volatility are followed by short periods of high volatility. To understand the mechanism behind this phenomenon we can argue that price volatility is due to the arrival of new information. Thus volatility clustering is the same as clustering of information arrivals, which corresponds to the simple statement that news are clustered in time. Another economic explanation may be that the behavior of major market agents may switch from fundamentalist to chartists, or the other way round. For further details see in Cont [19].

Volatility clustering can be captured mathematically as strong autocorrelations of the time series of absolute log-returns, see the next figure.

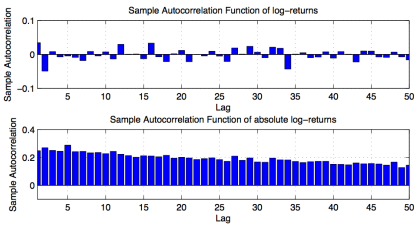

Figure 12.2: Autocorrelations of S&P 500 log-returns and absolute log-returns

For further details on many other stylized facts related to financial time series we refer the reader to Bollerslev et al. [9].

12.2 Stochastic volatility models

In modelling of financial time series one of the indicator of the quality of a model is its capability to reproduce some of the stylized facts detailed above. In classical time series analysis attention was focused on modelling the internal dependence structure of time series using second order properties. This has lead to the development of the theory of linear processes, such as AR or ARMA processes. They are attractive partially due to the fact that a beautiful mathematical theory has been developed to understand the properties of ARMA processes and to perform statistical analysis of data using ARMA models. A widely used reference on ARMA models is Box and Jenkins [12].

However ARMA models have one serious shortcoming when used for modelling return data: the conditional variance, a potential measure of volatility, is constant over time, a fact that is not supported by real financial data. In an ARCH process volatility is modeled as the output of a linear finite impulse response (FIR) system, combined with static non-linearities, driven by observed log-returns. In turn, log-returns are assumed to be defined as an i.i.d. process multiplied by the current volatility. Thus we get a stochastic non-linear feedback system, driven by an i.i.d. process.

One of the key problems in the application of GARCH models is fitting the model to real data, i.e. the estimation of the parameters. We will describe the so-called off-line quasi-maximum likelihood method, and outline the proofs of some of its fundamental properties. Potential alternatives to ARCH models, requiring a number of similar mathematical techniques, are the so-called bilinear models, see Terdik [52].

The material of this chapter is partially based on the PhD thesis of Zs. Orlovits, [45].

The simplest alternative to the classic random walk model is obtained by considering the model

$$y_n = \sigma_n \varepsilon_n, \quad -\infty < n < +\infty \tag{12.1}$$

where (ε_n) is an exogenous noise source, modelled as a sequence of i.i.d. r.v.-s. This noise source may be due a variety of factors such as variations in the agents' behavior or fluctuations of interest rate. The variable σ_n is called the *volatility* of y_n. It is to assumed that the current volatility (or current level of market activity) is uniquely determined by past returns, reflecting a certain feedback mechanism in the market. This is expressed by the condition that

$$\mathcal{F}_n^{\sigma} \subset \mathcal{F}_{n-1}^{y} \quad \text{for all } n,$$

where $\mathcal{F}_n^y = \sigma\{y_k : k \leq n\}$, and similarly for σ. In this case we say that (σ_n) is $\mathcal{F}^y$-predictable. More precisely, we require that

$$\sigma_n = F(y_{n-1}, y_{n-2}, ...), \tag{12.2}$$

where F may depend on the complete past of y up to time $n-1$, as in the case of linear systems with poles, but it is independent of n. The two most common specific forms of this feedback mechanism will be given below.

It is also assumed that (y_n) itself is a casual function of the exogenous noise source (ε_n), i.e.

$$\mathcal{F}_n^{y} \subset \mathcal{F}_n^{\varepsilon} \quad \text{for all } n. \tag{12.3}$$

Now $\varepsilon_n = y_n/\sigma_n$ implies the converse inclusion $\mathcal{F}_n^{\varepsilon} \subset \mathcal{F}_n^{y}$, leading to

$$\mathcal{F}_n^{y} = \mathcal{F}_n^{\varepsilon} \quad \text{for all } n. \tag{12.4}$$

To summarize and further specify the above speculations, we introduce the following definition:

Definition 12.1. *The return process (y_n) is defined by a stochastic volatility model if there exists an i.i.d. sequence (ε_n), and a stochastic volatility process (σ_n) such that (y_n) is strictly stationary, moreover (12.1), (12.2) and (12.3) are satisfied with a fixed, time-invariant F.*

Exercise 12.1. *Show that that the triplet $(\varepsilon_n, y_n, \sigma_n)$ itself is jointly strictly stationary.*

Note that at this point it is not clear that a return process (y_n) defined by a stochastic volatility model exists at all. A technical difficulty with the above definition is that y_n is defined in terms of σ_n (and of ε_n) while σ_n is defined in terms of the past of (y_n). This circular definition is typical for feedback systems.

To simplify the discussion let us now assume that that e_n has zero mean and finite second moments:

$$\mathrm{E}e_n = 0 \quad \text{and} \quad \mathrm{E}e_n^2 = 1. \tag{12.5}$$

Furthermore assume that

$$\mathrm{E}y_n^2 < \infty \quad \text{and} \quad \mathrm{E}\sigma_n^2 < \infty. \tag{12.6}$$

Exercise 12.2. *Show that under the conditions above*

$$\mathrm{E}[y_n|\mathcal{F}_{n-1}^y] = 0 \quad \text{a.s.}$$

In other words (y_n) is a martingale difference process.

Exercise 12.3. *Show that under the conditions above*

$$\mathrm{E}[y_n^2|\mathcal{F}_{n-1}^y] = \sigma_n^2 \quad \text{a.s.}$$

Taking expectation on both sides of the above equality we get that under the conditions above we have

$$\mathrm{E}\sigma_n^2 = \mathrm{E}y_n^2.$$

The fact that the conditional variance of y_n is random is referred to by the terminology that the process exhibits *conditional heteroscedasticity*. (For the origin of the terminology we note that "dispersion" in Greek is "skedasis").

Exercise 12.4. *Show that under the conditions above (y_n) is a w.s.st. orthogonal process.*

This is in line with our intuition or expectations about returns. Note however that the returns y_n are far from being independent, in fact the absolute value process $|y_n|$ has a very slowly decaying autocorrelation function, see Figure

Remark. We should note here that (ε_n) is not the innovation process of (y_n) as defined in the theory of linear processes. This may seem counter-intuitive in light of the identity $\mathcal{F}_n^y = \mathcal{F}_n^\varepsilon$, see (12.4). However, note that this identity postulates the existence of a *non-linear* function mapping the past of y up to time n onto the past of ε up to time n, and vice versa.

12.3 ARCH and GARCH models

The next step in building our model is to specify a suitable feedback mechanism defining σ_n that would ensure that extreme values of the returns generate more activity in the market, expressed in higher volatility, which in turn would explain the phenomenon of volatility clustering. The first widely accepted feedback mechanism was developed by Engle [22], leading to the celebrated ARCH model. A return process (y_n) defined by a stochastic volatility model is called an ARCH process of order r, or briefly an ARCH(r)-process if its volatility is defined via a feedback mechanism

$$(\sigma_n^2 - \gamma^*) = \sum_{i=1}^{r} \alpha_i^* (y_{n-i}^2 - \gamma^*), \quad n \in \mathbb{Z} \tag{12.7}$$

with $\alpha_i^* \geq 0$, $i = 1, \ldots, r$, and $\gamma^* = \mathrm{E}y_n^2 = \mathrm{E}\sigma_n^2 > 0$. The upper indices * indicate that we are talking about the true, but unknown parameters of the model, as opposed to tentative values that are chosen in fitting a model to data, see below. The feedback path given by (12.7) does indeed reflect the fact that that extreme values of the (squared) returns that are far apart from their expectations would generate volatilities (the squares of) which are far from their respective expectation. It is still not clear under what conditions is the overall feedback system well-defined in the sense that there exists a unique strictly stationary solution (e_n, y_n, σ_n).

Remark 12.2. Although the overall mapping from (ε_n) into (y_n) is non-linear, it has a simple structure: namely the current volatility is obtained as the output of a linear finite impulse response (FIR) system, cascaded with static, non-linear functions of observed returns. In short, the feedback path from returns to volatilities is defined in terms of a so-called *Hammerstein-system*. On the forward path returns are simply obtained by static nonlinearity as the product of the current volatility and of an exogenous i.i.d. source.

The parametrization that we use for ARCH models is mathematically appealing in understanding the role of the parameters. A disadvantage of this parametrization is that parameters show up in a non-linear fashion. Historically, the original parametrization of Engle [22] was different. It is obtained by merging all nonlinear terms into a single constant

$$\alpha_0^* = \gamma^* \left(1 - \sum_{i=1}^{r} \alpha_i^*\right).$$

Thus our model becomes linear in its parameters:

$$\sigma_n^2 = \alpha_0^* + \sum_{i=1}^{r} \alpha_i^* y_{n-i}^2, \quad n \in \mathbb{Z}. \tag{12.8}$$

Now, assuming that the returns in the feedback path, $y_{n-i}, i = 1, ..., r$ can take on any small values, the conditional variance σ_n^2 is positive only if $\alpha_0^* > 0$. Using the definition of α_0^* and the fact that $\gamma^* > 0$ we must have

$$\alpha_0^* > 0 \quad \text{and} \quad \sum_{i=1}^{r} \alpha_i^* < 1. \tag{12.9}$$

This is in fact a necessary and sufficient condition for the existence of a unique strictly stationary solution (e_n, y_n, σ_n) with finite second moments. We will prove sufficiency below.

A major feature of this model class is that it captures certain stylized facts such as volatility clustering. On the other hand it is mathematically simple enough for further theoretical investigations and statistical analysis of real data. In particular the estimation of the coefficients or weights α_i^* from historical data can be carried out relatively easily, in spite of the fact that the actual volatilities σ_n are not observed. This important advance in modelling financial data was recognized by a shared Nobel Prize in Economics in 2003.

A weak point in using ARCH models is that ARCH(r) processes may not fit log-returns very well unless one chooses the order of r very large. A natural extension is obtained by adding a moving average of $\sigma_n^2 - \gamma^*$ of order, say, s on the right hand side of the feedback path in defining an ARCH process, see (12.7). Thus we arrive at a so-called generalized ARCH model of order (r, s), briefly referred to as GARCH(r, s), which was independently introduced by Bollerslev [7] and Taylor [51] in 1986.

A GARCH(r, s) model is thus defined via the multiplicative model (12.1) for the returns, together with the specification of the feedback path defining the squared conditional variance process σ_n^2 as

$$(\sigma_n^2 - \gamma^*) = \sum_{i=1}^{r} \alpha_i^* (y_{n-i}^2 - \gamma^*) + \sum_{j=1}^{s} \beta_j^* (\sigma_{n-j}^2 - \gamma^*), \quad n \in \mathbb{Z}, \tag{12.10}$$

where $\gamma^* = \mathrm{E}y_{n-i}^2 = \mathrm{E}\sigma_{n-j}^2 > 0$ and $\alpha_i^*, \beta_j^* \geq 0$, $i = 1, \ldots, r$, $j = 1, \ldots, s$.

Remark 12.3. In the original definition of Bollerslev [7] used an alternative standard parametrization of equation (12.10) which can be obtained by collecting the constant terms into a single term

$$\alpha_0^* = \gamma^* \left(1 - \sum_{i=1}^{r} \alpha_i^* - \sum_{j=1}^{s} \beta_j^* \right).$$

Thus we get the defining equation

$$\sigma_n^2 = \alpha_0^* + \sum_{i=1}^{r} \alpha_i^* y_{n-i}^2 + \sum_{j=1}^{s} \beta_j^* \sigma_{n-j}^2. \quad n \in \mathbb{Z}, \tag{12.11}$$

We can argue as in the case of ARCH processes that a likely necessary condition for the existence of a strictly stationary solution with finite second order moments is that

$$\alpha_0^* > 0 \quad \text{and} \quad \sum_{i=1}^{r} \alpha_i^* + \sum_{j=1}^{s} \beta_j^* < 1. \tag{12.12}$$

The sufficiency of this condition will be proven below. The positive number

$$1 - \sum_{i=1}^{r} \alpha_i^* - \sum_{j=1}^{s} \beta_j^* < 1$$

is called the stability margin of the process. When fitting a GARCH model to real data the stability margin is typically below the threshold 0.2.

The scaling properties of a GARCH process are very simple. If (y_n) is a GARCH(r, s) process with parameters α_i^*, β_j^* and variance γ^*, then for any $\lambda > 0$ the process $(\sqrt{\lambda}\, y_n)$ is a GARCH(r, s) process with identical parameters α_i^*, β_j^* and variance $\lambda\gamma^*$. The conditional variance process becomes $\lambda\sigma_n^2$, while the driving noise process (e_n) remains the same.

The graph of an almost unstable simulated GARCH$(1, 1)$ process is displayed on Figure 12.3:

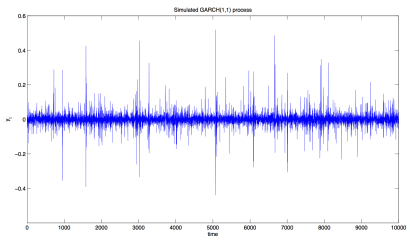

Figure 12.3: Simulated GARCH$(1, 1)$ process for 10000 observations with parameters $\alpha_0 = 0.0002$, $\alpha_1 = 0.955$, $\beta_1 = 0.0023$.

Three further examples of generated GARCH$(2, 2)$ processes are displayed on the figures below. The first two figures display the graphs of two almost unstable processes with stability margins 0.05. The first model is in fact an ARCH(2) model, while the second model is in a sense its opposite. The stability margin for the fourth model is 0.2.

It is interesting to note that the model exhibiting the phenomenon of volatility clustering in the most convincing manner is the ARCH(2) model.

A weak point of the GARCH model is that the volatility is insensitive to the sign of the return. Now if we look we look at the prices and the corresponding volatilities

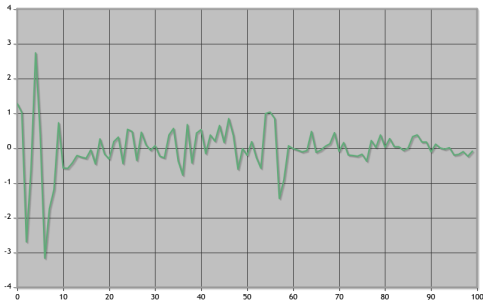

Figure 12.4: Simulated almost unstable ARCH(2) process with $\alpha_1^* = 0.3$, $\alpha_2^* = 0.65$. and $\gamma^* = 0.5$.

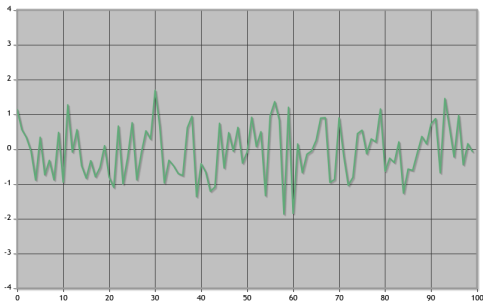

Figure 12.5: Simulated almost unstable GARCH$(2, 2)$ process with $\alpha_1^* = \alpha_2^* = 0$, and $\beta_1^* = 0.3$, $\beta_2^* = 0.65$ and $\gamma^* = 0.5$.

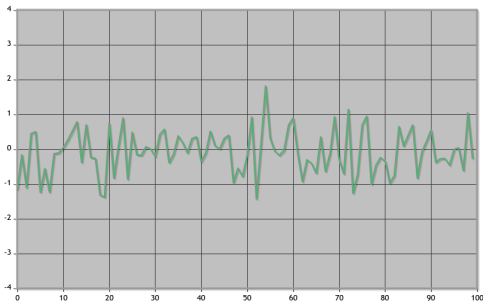

Figure 12.6: Simulated GARCH$(2, 2)$ process far from the stability margin with $\alpha_1^* = \alpha_2^* = \beta_1^* = \beta_2^* = 0.2$ and $\gamma^* = 0.5$.

on Figure 12.1 it can be observed that the volatility is higher when prices are falling. This implies that bad news on the market, i.e. negative shocks, tends to have a larger impact on volatility than good news, i.e. positive shocks. This asymmetry on volatility is called the leverage effect, first noted by Black [5]. For further details see Engle [23]. This leverage effect can be incorporated into the GARCH model by choosing different static non-linearities in the feedback path.

12.4 State space representation

In this chapter we summarize some of the basic mathematical tools found to be useful in studying the GARCH processes. Above all, we present a state space representation of GARCH processes, and show that this representation leads to very simple computational procedures. A novel, non-standard feature of this state space system is that its system matrices are not constant, but rather they form an i.i.d. sequence. Conditions for the existence of a strictly stationary, causal solution are therefore very different from what we discussed in the context of time-invariant linear stochastic systems.

Let (y_n), $n \in \mathbb{Z}$ be a GARCH(r, s) process having finite second moments, defined via the non-linear stochastic feedback system

$$y_n = \sigma_n \varepsilon_n, \tag{12.13}$$

$$(\sigma_n^2 - \gamma^*) = \sum_{i=1}^{r} \alpha_i^* (y_{n-i}^2 - \gamma^*) + \sum_{j=1}^{s} \beta_j^* (\sigma_{n-j}^2 - \gamma^*). \tag{12.14}$$

Here the exogenous i.i.d. noise process (ε_n), $n \in \mathbb{Z}$ has zero mean and unit variance, and $\gamma^* = \mathrm{E} y_{n-i}^2 = \mathrm{E}\sigma_{n-j}^2 > 0$ and $\alpha_i^*, \beta_j^* \geq 0$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ denote the true, but possibly unknown parameters of the model. Defining the polynomials

$$A^*(q^{-1}) = \sum_{i=1}^{r} \alpha_i^* q^{-i}, \quad B^*(q^{-1}) = 1 - \sum_{j=1}^{s} \beta_j^* q^{-j}, \tag{12.15}$$

equation (12.14) can be written in a compact form as

$$B^*(q^{-1})(\sigma^2 - \gamma^*) = A^*(q^{-1})(y^2 - \gamma^*), \tag{12.16}$$

where q^{-1} is the backward shift operator.

Let us define the random, $(r + s)$-dimensional state vector X_n^* as

$$X_n^* = (y_n^2, \ldots y_{n-r+1}^2, \sigma_n^2, \ldots, \sigma_{n-s+1}^2)^T. \tag{12.17}$$

Note that, in contrast to state-vectors defined in the context of linear stochastic systems, X_n^* is defined in terms of past values of y^2 and σ^2 starting at present time n, rather than $n - 1$.

Then it is easy to verify that X_n^* satisfies a first order, random coefficient linear stochastic difference equation

$$X_{n+1}^* = A_{n+1}^* X_n^* + u_{n+1}^*, \quad n \in \mathbb{Z}, \tag{12.18}$$

with the fairly sizeable random state-matrices $A_n^* \in \mathbb{R}^{(r+s)\times(r+s)}$ defined as

$$\left(\begin{array}{cccc|c|cccc|c}
\alpha_1^*\varepsilon_n^2 & \alpha_2^*\varepsilon_n^2 & \dots & \alpha_{r-1}^*\varepsilon_n^2 & \alpha_r^*\varepsilon_n^2 & \beta_1^*\varepsilon_n^2 & \beta_2^*\varepsilon_n^2 & \dots & \beta_{s-1}^*\varepsilon_n^2 & \beta_s^*\varepsilon_n^2 \\
\hline
1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\
\hline
\alpha_1^* & \alpha_2^* & \dots & \alpha_{r-1}^* & \alpha_r^* & \beta_1^* & \beta_2^* & \dots & \beta_{s-1}^* & \beta_s^* \\
\hline
0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 \\
0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0
\end{array}\right) \tag{12.19}$$

and $u_n^* \in \mathbb{R}^{r+s}$ defined as

$$u_n^* = (\alpha_0^*\varepsilon_n^2, 0, \dots, 0, \alpha_0^*, 0, \dots, 0)^T.$$

Note that the random state-matrices $A_n^* \in \mathbb{R}^{(r+s)\times(r+s)}$ are highly structured, in particular $\mathrm{E}A_n$ has two identical rows, (namely the block row 1 and 3). In addition, the sequence (A_n^*) is i.i.d.

The next natural question to ask is this: under what conditions does a unique, strictly stationary solution of (12.13) and (12.14) exist satisfying the causality conditions $\sigma_n \in \mathcal{F}_{n-1}^y$ and $y_n \in \mathcal{F}_n^\varepsilon$, such that $\mathrm{E}y_n^2 = \mathrm{E}\sigma_n^2 = \gamma^* < \infty$. The answer to this question is given in the following theorem, see Bollerslev [7].

Theorem 12.1. *The non-linear closed loop system given by (12.13) and (12.14), defining a GARCH(r,s) process, has a unique, causal, strictly stationary solution solution satisfying $\mathrm{E}y_n^2 = \mathrm{E}\sigma_n^2 = \gamma^* < \infty$ if and only if*

$$\sum_{i=1}^{r} \alpha_i^* + \sum_{j=1}^{s} \beta_j^* < 1. \tag{12.20}$$

Proof. To prove sufficiency note that iterating equation (12.18) infinitely many times, and replacing $n+1$ by n we get the formal expansion of the assumed solution as follows:

$$X_n^* = u_n + \sum_{k=1}^{\infty} A_n A_{n-1} \dots A_{n-k+1} u_{n-k}. \tag{12.21}$$

□

We show that the infinite sum on the right hand side of equation (12.21) converges in L_1. Indeed, the assumed independence of the sequence (ε_n) implies that

$$\mathrm{E}[A_n A_{n-1} \dots A_{n-k+1} u_{n-k}] = \mathrm{E}A_n \cdot \mathrm{E}A_{n-1} \dots \mathrm{E}A_{n-k+1} \cdot \mathrm{E}u_{n-k}.$$

Now, the stability condition (12.20) implies that $\mathrm{E}A_0^*$ is *sub row-stochastic*, i.e. $\mathrm{E}A_0^*$ is a matrix with non-negative elements such that its row-sums are less than or equal to one. In addition, there is at least 1 row (in fact two rows) with row-sum strictly less than 1. Hence, by a well-known theorem of linear algebra of non-negative matrices, the Perron-Frobenius theorem it follows, that the all eigenvalues have modulus (absolute value) strictly less than 1. This is equivalent to saying that the spectral radius of $\overline{A} = \mathrm{E}A_0^*$, denoted by $\rho(\mathrm{E}A_0^*) = \rho(\overline{A})$ satisfies $\rho(\overline{A}) < 1$.

Now, we have $\mathrm{E}A_l = \overline{A}$ for any l, and thus

$$||\mathrm{E}A_n \cdot \mathrm{E}A_{n-1} \ldots \mathrm{E}A_{n-k+1} \cdot \mathrm{E}u_{n-k}|| = ||\overline{A}^{(k-1)} \cdot \mathrm{E}u_{n-k}||$$

with some $C > 0$ and $\rho < 1$. It follows that the partial sums of the infinite series on right hand side of (12.21) form a Cauchy-sequence in L_1, hence this infinite series converges in L_1, and thus X_n is well-defined.

Exercise 12.5. *Show that X_n defined by the right hand side of (12.21) satisfies the state equation (12.18) corresponding to the GARCH process.*

Obviously, (X_n) is $\mathcal{F}_n^{\varepsilon}$-measurable, i.e. it is a causal function of the process ε. It is also obvious that (X_n) is strictly stationary. Finally, $\mathrm{E}|\mathrm{X_n}| < \infty$ implies that $\mathrm{E}y_n^2 = \mathrm{E}\sigma_n^2 = \gamma^* < \infty$. To complete the proof solve the following exercise:

Exercise 12.6. *Prove uniqueness as stated in the theorem.*

The identification of GARCH models can be carried out along the lines of the identification of ARMA processes: we invert the system to reconstruct the noise, and pretending that the noise is Gaussian, we apply a maximum likelihood method. The resulting estimator is called a quasi-maximum likelihood estimator, or QMLE for short. Here we present the Fisher information matrices of the estimators, and their eigenvalues, assuming unit variance for the noise, for two of our benchmark models:

12.5 Existence of a strictly stationary solution

The condition $\mathrm{E}e_n^2 < \infty$ or $\mathrm{E}y_n^2 = \mathrm{E}\sigma_n^2 = \gamma^* < \infty$ may be not quite appropriate in all circumstances, and it is a challenging problem to see what can we achieve without using second order techniques. Thus let us consider the non-linear closed loop dynamics described by (12.13) and (12.14), and let us ask ourselves: under what conditions does a unique, causal, strictly stationary solution exist, without additional constraints or expectations on moments of y_n or σ_n^2.

The main tool for tackling this problem is to consider the state space equation (12.18) and analyze the resulting stability properties of the resulting random coefficient linear

stochastic system. Recall first that a process $(X_n), n \in \mathbb{Z}$ is strictly stationary if for all $n, m \in \mathbb{Z}$, the law of $(X_n, X_{n+1}, \ldots, X_{n+m})$ is independent of n. Using the state-space representation of GARCH processes it can be readily seen that the GARCH equations (12.13) and (12.14) have a unique, strictly stationary, causal solution if and only if the linear stochastic system (12.18) has a unique, strictly stationary, causal solution with non-negative coordinates. Let us now consider a generalization of (12.18) as follows:

$$X_{n+1} = A_{n+1}X_n + u_{n+1}, \qquad n \in \mathbb{Z}, \tag{12.22}$$

where $X_n \in \mathbb{R}^d$, A_n is a random matrix in $\mathbb{R}^{d\times d}$ and u_n is a random vector in $\mathbb{R}^d$. Assume that the following condition holds:

Condition 12.4. *(A_n, u_n) is a jointly strictly stationary, ergodic sequence of $d \times (d+1)$ random matrices over some probability space $(\Omega, \mathcal{F}, P)$.*

A strictly stationary solution (X_n) is called casual if X_{n+1} is measurable with respect to the σ-field $\mathcal{F}_n = \sigma\{A_i, u_i, i \leq n\}$. Both necessary and sufficient conditions for the existence of a strictly stationary casual solution of (12.22) have been given by Bougerol and Picard in [11]. To formulate these results we need the concept of a Lyapunov-exponent. Let $|\cdot|$ be any vector norm in $\mathbb{R}^d$ and define an operator norm on the set of $d \times d$ real matrices by

$$\|M\| := \sup_{x\in\mathbb{R}^d, x\neq 0} \frac{|Mx|}{|x|}$$

for $M \in \mathbb{R}^{d\times d}$. Let $\mathcal{A} = (A_n)$ be as above such that

$$\mathrm{E}\log^+ \|A_n\| < +\infty, \tag{12.23}$$

where $\log^+ x$ denotes the positive part of $\log x$. Then we have the following result:

Theorem 12.2. *Under the conditions above the limit*

$$\lambda = \lim_{n\to\infty} \frac{1}{n}\mathrm{E}\log\|A_n \ldots A_1\| \tag{12.24}$$

exists, where $-\infty \leq \lambda < +\infty$. Moreover

$$\lambda = \inf_{n>0} \frac{1}{n}\mathrm{E}\log\|A_n \ldots A_1\|. \tag{12.25}$$

The number λ is called the *t*op-Lyapunov exponent of $\mathcal{A}$, and is denoted by $\lambda(\mathcal{A})$. If $A_n = A$ for all n then $\lambda(\mathcal{A})$ is simply the spectral radius of A.

Proof. Consider the sequence $u_n = \mathrm{E}\log\|A_n \ldots A_1\|$. □

Exercise 12.7. *Show that (u_n) is sub-additive, i.e. we have for any $n, m > 0$ the inequality $u_{n+m} \leq u_n + u_m$.*

Now for subadditive sequences the following general result holds:

Lemma 12.5. *Let (u_n) be a sub-additive sequence. Then the limit*

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \mathrm{E} \log u_n \tag{12.26}$$

exists, where $-\infty \leq \lambda < +\infty$, moreover

$$\lambda = \inf_{n\to\infty} \frac{1}{n} \mathrm{E} \log u_n. \tag{12.27}$$

Exercise 12.8. *Prove the above lemma.*

(*Hint*: take $\varepsilon > 0$ and choose a k such that $u_k/k < \lambda + \varepsilon$. Show that then for any integer n we have $u_{nk}/(nk) < \lambda + \varepsilon$.

Now a major result of the theory of random matrices is the theorem of Fürstenberg and Kesten as follows, see [25]:

Theorem 12.3. *Assume that (A_n) is ergodic, and that (12.23) holds. Then we the top-Lyapunov exponent can be represented as the almost sure limit*

$$\lambda(\mathcal{A}) = \lim_{n\to\infty} \frac{1}{n} \log \|A_n \dots A_1\| \quad \text{a.s.} \tag{12.28}$$

Then we have the following simple consequence of the previous theorem, a result due to Bougerol and Picard [11]:

Theorem 12.4. *Assume that (A_n, u_n) satisfies Condition 12.4, (12.23) holds, and*

$$\mathrm{E} \log^+ |u_n| < +\infty. \tag{12.29}$$

Then $\lambda(\mathcal{A}) < 0$ implies that (12.22) has a unique strictly stationary and causal solution given by

$$X_n^* = u_n + \sum_{k=1}^{\infty} A_n A_{n-1} \dots A_{n-k+1} u_{n-k}. \tag{12.30}$$

Outline of proof: The first step in the proof is to solve the following exercise.

Exercise 12.9. *Prove that for any $\varepsilon > 0$ there exist finite r.v.-s $C_n(\omega, \varepsilon)$ such that for any n we have*

$$||A_n A_{n-1} \dots A_{n-k+1}|| \leq C_n(\omega, \varepsilon) e^{(\lambda(\mathcal{A})+\varepsilon)(n-k)}.$$

Show that $C_n(\omega, \varepsilon)$ can be assumed to be a stationary sequence.

The second step in the proof is formulated in the next exercise:

Exercise 12.10. *Prove that for any sequence of r.v.-s* (u_n) *the condition*

$$\sup_n \mathrm{E} \log^+ |u_n| < +\infty$$

implies that for any $\varepsilon > 0$ *there exists a r.v.* $C(\omega, \varepsilon)$ *such that*

$$|u_n| \le C(\omega, \varepsilon) e^{\varepsilon n}.$$

Combining the results of the two exercises yield that the r.h.s. of (12.30) converges a.s. It is trivially seen that the resulting r.v.-s do indeed satisfy the state-equation (12.22).

Remark 12.6. A remarkable deep result due to Bougerol and Picard [11] is that the condition $\lambda(\mathcal{A}) < 0$ is also a *necessary* condition for the existence of a strictly stationary causal solution of (12.22) when (A_n, u_n) is an i.i.d. sequence, assuming that (A_n, u_n) is controllable in the sense that there is no proper subspace $V \subset \mathbb{R}^d$, such that

$$A_0 V + u_0 \subset V \qquad \text{w.p.1.}$$

Remark 12.7. *An simple way to verify that* $\lambda(\mathcal{A}) < 0$ *we may use the following observation: if for some* $m \ge 1$ *we have*

$$\mathrm{E} \|A_m \dots A_1\| < 1, \tag{12.31}$$

then $\lambda(\mathcal{A}) < 0$. *This follows from the definition of the Lyapunov-exponent given in (12.25), and Jensen's inequality.*

$$\Sigma = \begin{bmatrix} -3102 & -2885 & -1928 & -1809 \\ -2885 & -3124 & -2029 & -1853 \\ -1928 & -2029 & -1694 & -1561 \\ -1809 & -1853 & -1561 & -1973 \end{bmatrix}$$

Figure 12.7: The asymptotic covariance matrix the QMLE, and its eigenvalues for an ARCH(2) process with parameters, $\alpha_1^* = 0.3$ and $\alpha_2^* = 0.5$, and $\gamma^* = 0.5$

$$\Sigma = \begin{bmatrix} -3.681e+4 & -3.685e+4 & -3.316e+4 & -3.272e+4 \\ -3.685e+4 & -3.740e+4 & -3.351e+4 & -3.309e+4 \\ -3.316e+4 & -3.351e+4 & -3.083e+4 & -3.000e+4 \\ -3.272e+4 & -3.309e+4 & -3.000e+4 & -2.736e+4 \end{bmatrix}$$

Figure 12.8: The asymptotic covariance matrix the QMLE, and its eigenvalues for an ARCH(2) process with parameters, $\alpha_1^* = 0.3$ and $\alpha_2^* = 0.65$, and $\gamma^* = 0.5$

$$\Sigma = \begin{bmatrix} -7483 & -6344 & -1131 & -1200 \\ -6344 & -6925 & -1183 & -1238 \\ -1131 & -1183 & 0.000 & 0.000 \\ -1200 & -1238 & 0.000 & 0.000 \end{bmatrix}$$

Figure 12.9: The asymptotic covariance matrix the QMLE, and its eigenvalues for a GARCH(2,2) process with $\alpha_i^* = 0$ and high β_i^*, ($\gamma^* = 0.5$)

$$\Sigma = \begin{bmatrix} -1784 & -1444 & -828.5 & -793.6 \\ -1444 & -1658 & -877.1 & -808.1 \\ -828.5 & -877.1 & -561.7 & -582.7 \\ -793.6 & -808.1 & -582.7 & -638.8 \end{bmatrix}$$

Figure 12.10: The asymptotic covariance matrix the QMLE, and its eigenvalues for a GARCH(2,2) process with medium parameters, $\alpha_i^* = \beta_i^* = 0.2$ and $\gamma^* = 0.5$

Chapter 13

High-frequency data. Poisson processes

13.1 Motivation

When dealing with high-frequency data it is natural to build models in continuous time. A classical model for modelling market dynamics in continuous time is geometric Brownian motion, proposed by Paul Samuelson, modifying an earlier model of Louis Bacehelier. This model is still the accepted core model despite the fact that empirical studies revealed that its assumptions are not realistic. For example, since price movements are induced by transactions which can be unevenly distributed in real time, it would be more natural to use a time changed geometric Brownian motion to model price dynamics.

If the time change is defined by a so-called gamma process, which is a non-negative strictly increasing process with independent and stationary increments, (see below), we obtain the so-called VG (shorthand for Variance Gamma) process. VG processes reproduce a number of stylized facts of real price processes, such as fat tails and large kurtosis.

Another shortcoming of the geometric Brownian motion is that it is unsuitable to model shocks or jumps in the price process, which are indeed observed. Therefore in modelling high-frequency data an alternative approach is to use a general class of continuous-time processes with independent and stationary increments, allowing jumps. Thus we come to the notion of Lévy processes, which are obtained as the sum of a Wiener process and another independent process modelling jumps.

Lévy processes are widely used to model phenomena arising in natural sciences, economics, financial mathematics, queueing theory and telecommunication [3],[44],[20]. It can be shown that the above time changed Brownian process itself is a Lévy process. Extending the above construction novel price dynamics have been proposed by a variety of authors, called the geometric Lévy processes obtained by exponentiating a Lévy

process. Lévy processes have become a widely used tool in modelling price processes of financial instruments, such as stock prices or indices [44].

In this section we provide the basic mathematical technology, the elements of the theory of Poisson point processes, through which the construction of Lévy process can be conveniently described, see the next chapter.

13.2 Basic properties of the Poisson distribution

A stochastic process modelling jumps could be described by defining a point process, i.e. an increasing sequence of random times $0 < T_1 < T_2 < \dots$ indicating the time points when a jump occurs. Then for each T_i we should take a r.v. X_i giving the size of jumps. A more convenient and more powerful approach is to work in the (t, x) domain (i.e. in the product space $\mathbb{R}^+ \times \mathbb{R}$), and considering random point (T_α, X_α) with index α belonging to a denumerable set. A novel feature of this approach that the time instants T_α-s are not ordered in an increasing sequence. In fact it may happen that in any finite interval $[t_0, t_1]$ an infinite number of T_α-s occur, indicating an infinite number of transactions, as an idealization. We start by extending the idea of stochastic processes with independent increments for random sets of points in $\mathbb{R}^+ \times \mathbb{R}$, or more generally in a measurable space $(S, \mathcal{G})$. Our presentation is based on [36].

Let $(S, \mathcal{G})$ be a measurable space such that $|S| \geq \infty$. For example, for $S = \mathbb{R}^+ \times \mathbb{R}$ we take $\mathcal{G} = \mathbf{B}(\mathbb{R}^+ \times \mathbb{R})$, the set of Borel-sets of $\mathbb{R}^+ \times \mathbb{R}$. Let Π be a random, finite set in S. For any $\mathcal{G}$-measurable test set $A \subset S$ define

$$N(A) = |\Pi \cap A|,$$

i.e. $N(A)$ denotes the number of random points lying in A. An awkward looking technical condition we need is the following:

Condition 13.1. *We assume that the diagonal*

$$D = \{(x, x) : x \in S\}$$

is measurable in $(S \times S, \mathcal{G} \times \mathcal{G})$.

Obviously, this condition is satisfied when $\mathcal{G}$ is the set of Borel-subsets of S, which is a Borel set of $\mathbb{R}^n$.

To fix the notations recall that a r.v. X has Poisson distribution $P(\mu)$ with $0 < \mu < \infty$ if X takes its values in the set of non-negative integers, i.e. X is $\mathbb{Z}^+$-valued, and

$$P(X = n) = \pi_n(\mu) = \mu^n e^{-\mu}/n!$$

for $n \geq 0$, with the usual convention that $0! = 1$.

The above, standard definition of a Poisson-distribution can be extended to cover the extreme case when $\mu = 0$ or $\mu = \infty$. For $\mu = 0$ we define $P(\mu)$ by

$$P(X = 0) = 1.$$

For $\mu = \infty$ we set:

$$P(X = \infty) = 1.$$

Proposition 13.2. *Let X have a Poisson distribution $P(\mu)$. Then for $|z| \leq 1$ we have*

$$E(z^X) = e^{-\mu(1-z)}.$$

Proof.

$$E(z^X) = e^{-\mu} \sum_{n=0}^{\infty} \frac{\mu^n}{n!} z^n = e^{-\mu} e^{\mu z} = e^{-\mu(1-z)}.$$

Using this result one can obtain formulas, which are easy to remember, for the moments of X. Differentiating w.r.t. z and setting $z = 1$ we get

$$E\ [X] = \mu$$

$$E\ [X(X-1)] = \mu^2$$

$$E\ [X(X-1)(X-2)] = \mu^3,$$

and so on. □

Let us now consider the problem of adding two a more independent Poisson r.v.-s. For a start we recall the following elementary result:

Proposition 13.3. *If X and Y are independent r.v.-s with*

$$X \stackrel{\mathcal{L}}{=} P(\lambda) \qquad Y \stackrel{\mathcal{L}}{=} P(\mu)$$

then

$$X + Y \stackrel{\mathcal{L}}{=} P(\lambda + \mu).$$

The proof can be easily obtained by direct calculation.

To extend the above result for countably many r.v.-s let X_j, $j = 1, 2, ...$ be independent r.v.-s with

$$X_j \stackrel{\mathcal{L}}{=} P(\mu_j).$$

Proposition 13.4. *If*

$$\sigma = \sum_{j=1}^{\infty} \mu_j \tag{13.1}$$

converges, then

$$S = \sum_{j=1}^{\infty} X_j$$

converges with probability 1, and

$$S \stackrel{\mathcal{L}}{=} P(\sigma).$$

If, on the other hand, if (13.1) diverges, then S diverges w.p.1.

Proof. (Outline) By induction on n we have:

$$S_n = \sum_{j=1}^{n} X_j \stackrel{\mathcal{L}}{=} P(\sigma_n),$$

where

$$\sigma_n = \sum_{j=1}^{n} \mu_j.$$

But

$$P(S \leq r) = \lim_{n \to \infty} P(S_n \leq r).$$

Using the continuity of $\pi_\mu(k)$ in μ gives the claim. □

The next natural problem we consider is this: what is the conditional distribution of the individual X_i-s under the condition that their sum is given ? The answer is that this conditional distribution is multinomial, as given in the following proposition:

Proposition 13.5. *Let $X_1, X_2, ...X_n$ be independent, $X_j \stackrel{\mathcal{L}}{=} P(\mu_j)$, and set*

$$S = X_1 + ... + X_n.$$

Then the conditional distribution of $(X_1, X_2, ...X_n)$ given $S = s$ is:

$$P(X_1 = r_1, ..., X_n = r_n | S = s) = \frac{s!}{r_1!...r_n!} (\frac{\mu_1}{\sigma})^{r_1} ... (\frac{\mu_n}{\sigma})^{r_n}.$$

with $\sigma = \mu_1 + ... + \mu_n$.

For $n = 2$ the previous result gives that the above conditional distribution is binomial:

Corollary 13.6. *Let X and Y be independent Poisson r.v.-s. Then*

$$P(X|X+Y=n) \stackrel{\mathcal{L}}{=} \mathcal{B}(n,p)$$

with

$$p = \mathrm{E}(X)/(\mathrm{E}(X)+\mathrm{E}(Y)).$$

Recall that $\mathcal{B}(n,p)$ is the binomial distribution for n trials with success probability p. Thus

$$P(X=r|X+Y=n) = b(n,p;r) = \binom{n}{r} p^r (1-p)^r.$$

13.3 Poisson point processes on a general state space

Now we define and study Poisson point processes in a general setting. It turns out that the proposed general point of view is exceptionally fruitful in understanding a number of features of Poisson point processes. Let $(S,\mathcal{G})$ be a measurable space s.t. $|S| \geq \infty$.

Definition 13.7. *. A random, countable subset Π of S is a Poisson process if*
(i) for any $A_1, ..., A_n \subset S$ measurable, disjoint sets the counts

$$N(A_i) = |\Pi \cap A_i|$$

are independent.
(ii) for any $A \in \mathcal{G}$ we have

$$N(A) \stackrel{\mathcal{L}}{=} P(\mu(A)),$$

where μ is a measure on $(S,\mathcal{G})$. It is called the mean measure.

Exercise 13.1. *Show that μ is non-atomic, i.e. for all $x \in S$ we have $\mu(\{x\}) = 0$.*

(*Hint*: Otherwise we would have $P(N(\{x\}) \geq 2) > 0$!)

Exercise 13.2. *Show that for any $A_1, A_2 \in \mathcal{G}$ with $\mu(A_i) < \infty$ we have*

$$\mathrm{cov}(N(A_1), N(A_2)) = \mu(A_1 \cap A_2).$$

Definition 13.8. *Let $S = \mathbb{R}^d$. If μ is given in terms of a measurable function λ on S by*

$$\mu(A) = \int_A \lambda(x)dx,$$

then $\lambda(x)$ is called the intensity.

Note that is not assumed that $\mu(A) < +\infty$. If $\lambda(x) = \lambda$ constant, then Π is called a homogeneous Poisson point process.

The next theorem is the so-called *superposition theorem*. In the language of finance superposition amounts to considering the union of two (a more) independent markets as a single market.

Theorem 13.1. *Let $\Pi_1, \Pi_2, ...$ be independent Poisson processes on S, and let Π_n have mean measure μ_n. Then their superposition*

$$\Pi = \bigcup_{n=1}^{\infty} \Pi_n$$

is a Poisson process with mean measure

$$\mu = \sum_{n=1}^{\infty} \mu_n.$$

Proof. Let

$$N_n(A) = \#\{\Pi_n \cap A\} \quad \text{and} \quad N(A) = \#\{\Pi \cap A\}.$$

If

$$N(A) = \sum_{n=1}^{\infty} N_n(A), \quad w.p.1 \tag{13.2}$$

then we are ready (by the countable additivity theorem). □

To prove (13.2) we use the following disjointness lemma, which is applicable if $\mu_n(A) < +\infty$ for all n:

Lemma 13.9. *Let Π_1 and Π_2 be independent Poisson processes on S and let $A \subset S$ be measurable with*

$$\mu_1(A), \mu_2(A) < +\infty.$$

Then

$$P\{\Pi_1 \cap \Pi_2 \cap A = \varnothing\} = 1. \tag{13.3}$$

A very simple proof of this lemma can be obtained by using the construction of the Poisson process via so-called Bernoulli processes, see below.

The next three theorem fall into the category of *m*apping theorems. We begin with the simplest possible mapping: restriction.

Theorem 13.2. *Let Π be a Poisson process with mean measure μ on S and let $S_1 \subset S$ be measurable. Then*

$$\Pi_1 = \Pi \cap S_1$$

is a Poisson process on S with mean measure

$$\mu_1(A) = \mu(A \cap S_1).$$

Exercise 13.3. *Prove the above theorem.*

A much more exiting and challenging problem is the following. Let Π be a Poisson process on S, having mean measure μ, and let

$$f : S \to T$$

be a mapping into a measurable space T satisfying Condition D. Assume that f is measurable and let the induced measure be μ^*:

$$\mu^*(B) = \mu(f^{-1}(B))$$

for $B \subset T$ measurable. What can we say about the induced measure ? The following theorem gives the answer:

Theorem 13.3. *Assume that μ is σ-finite, and the induced measure μ^* has no atoms. Then*

$$f(\Pi)$$

is a Poisson process on T with mean measure μ^*.

Outline of the proof: We show that for $B \subset T$ measurable we have for the counts

$$N^*(B) = \{X \in \Pi,\ f(X) \in B\} = N(f^{-1}(B)) \quad \text{w.p.1.}$$

A *non-trivial* technical tool: we have w.p.1

$$x, y \in \Pi, x \neq y \Rightarrow f(x) \neq f(y).$$

Proving this will be an exercise later on.

A simple application of the mapping theorem is the following projection theorem:

Theorem 13.4. *Let Π be a Poisson process in $\mathbb{R}^D$ with rate function $\lambda(x_1, ..., x_D)$. Let $d < D$ and let*

$$\Pi_d$$

be the projection of Π on the first d coordinates. Then Π_d is a Poisson-process with rate function

$$\lambda^*(x_1, ..., x_d) = \int ... \int \lambda(x_1, ..., x_D) dx_{d+1} ... dx_D.$$

Exercise 13.4. *Prove the above projection theorem.*

13.4 Construction of Poisson processes

So far it has not been discussed how to generate a Poisson point process. In the case of $S = \mathbb{R}^1$, i.e. in time domain a standard procedure is to generate the times of events occurring T_k simply adding i.i.d, exponentially distributed random variables. In our general setup this approach is not applicable.

Let us work backward. Let Π be a Poisson process on S with $\mu(S) < +\infty$. We ask the question: what is the distribution of Π under the condition that $N(S)$ is given ? Let $A_0, A_1, ..., A_k \subset S$ be a partition of S, and let $\sum_{i=0}^{k} n_i = n$. Then

$$P(N(A_0) = n_0, ...N(A_k) = n_k \mid N(S) = n)$$

$$= \frac{n!}{n_0!...n_k!}\left(\frac{\mu(A_0)}{\mu(S)}\right)^{n_0}...\left(\frac{\mu(A_k)}{\mu(S)}\right)^{n_k}.$$

The idea is then to construct first a conditional Poisson point process under the condition that $N(S) = n$. This is a countable random set of interest of its own:

Definition 13.10. *The random finite set $\Pi \subset S$ with $|\Pi| = n$ and*

$$P(N(A_0) = n_0, ..., N(A_k) = n_k) = \frac{n!}{n_0!...n_k!}\eta(A_0)^{n_0}...\eta(A_k)^{n_k} \tag{13.4}$$

where η is a probability measure on S, is called a Bernoulli process. The measure $n\ \eta(.)$ is called the mean measure.

Proposition 13.11. *Let $X_1, ..., X_n$ be i.i.d. on S according to the probability measure η. Assume that η has no atom. Then*

$$\Pi = \{X_1, .., X_n\}$$

is a Bernoulli process.

Exercise 13.5. *Prove the above proposition.*

Theorem 13.5. *Let μ be a non-atomic measure on S such that*

$$\mu = \sum_{n=1}^{\infty} \mu_n, \qquad \mu_n(S) < \infty. \tag{13.5}$$

Then there exists a Poisson process having μ as its mean measure.

Note that (13.5) is slightly weaker than σ-finiteness.
Outline of the proof: Assume $\mu_n(S) > 0$ for all n. Let for $n = 1, 2, ...$

$$N_n \stackrel{\mathcal{L}}{=} P(\mu_n(S))$$

be a sequence of independent r.v.-s. For all n let

$$X_n = (X_{nr}) \qquad r = 1, ..., N_n$$

be an i.i.d. sequence of S-valued r.v.-s with distributions

$$P_n(\cdot) = \mu_n(\cdot) \diagup \mu(S).$$

Assume that

$$(X_{nr}) \quad \text{and} \quad N = (N_n), \quad n, r = 1, 2, ...$$

are mutually independent. Write

$$\Pi_n = \{X_{n1}, ..., X_{nN_n}\}$$

and set

$$\Pi = \bigcup_{n=1}^{\infty} \Pi_n.$$

This Π will do.

Exercise 13.6. *Work out the details of the proof.*

13.5 Sums and integrals over Poisson processes

Let f be a real-valued function over the state space S. The object to be discussed in this section is the sum:

$$\Sigma = \Sigma_f = \sum_{x \in \Pi} f(x).$$

This problem will lead us to the definition of integrals with respect to random point processes, in close analogy to stochastic integrals w.r.t. random orthogonal measures.
Example. The shot effect. Let $0 < T_1 < T_2 < ...$ be a Poisson process on $(0, \infty)$. Let an impulse response be ϕ. Then the total response is

$$\sum_j \phi(t - T_j).$$

First we define the integral of simple, i.e. piecewise constant functions. Let f take on a finite number of values:

$$f(x) = f_i \quad \text{on} \quad A_i$$

$i = 1, ..., k$, where the sets A_i are disjoint, and $\mu(A_i) < +\infty$. Then

$$\Sigma = \Sigma_f = \sum_{x \in \Pi} f(x) = \sum_{j=1}^{k} f_j N_j$$

where $N_j = N(A_j)$.

Compute the *moment generating function* of Σ. For any real or complex θ we have

$$\mathrm{E}(e^{\theta\Sigma}) = \prod_{j=1}^{k} \mathrm{E}(e^{\theta f_j N_j}).$$

The j-th term is $\exp\{(e^{\theta f_j} - 1)m_j\}$ with $m_j = \mu(A_j)$. Thus we get

$$\mathrm{E}(e^{\theta\Sigma}) = \exp\{\sum_{j=1}^{k}(e^{\theta f_j} - 1)\mu(A_j)\} = \exp\{\int_S (e^{\theta f(x)} - 1)\mu(dx)\}.$$

Thus we arrive at the following *master equation:*

Theorem 13.6. *Let f be a real valued function taking on a finite number of values. Then we have for any real or complex θ*

$$E(e^{\theta\Sigma}) = \exp\{\int_S (e^{\theta f(x)} - 1)\mu(dx)\}. \tag{13.6}$$

For the means and covariances of Σ-s we get:

$$\mathrm{E}(\Sigma) = \int_S f(x)\mu(dx).$$

and, if

$$\Sigma_1 = \sum_{x \in \Pi} f_1(x), \qquad \Sigma_2 = \sum_{x \in \Pi} f_2(x),$$

then

$$\mathrm{Cov}(\Sigma_1, \Sigma_2) = \int_S f_1(x) f_2(x)\mu(dx).$$

Exercise 13.7. *Prove the above two identities.*

Clearly, the extension of the concept of integrals may not be possible for any function f. First of all we introduce some technical conditions to ensure the integrability of the right hand side of (13.6). Assuming that θ is purely imaginary, say $\theta = i\omega$, with ω real, we can estimate the absolute value of the r.h.s. of (13.6) by splitting the region of integration into two, according to whether $|f(x)| > d$ or $|f(x)| \leq d$, and taking a Taylor-series approximation of the integrand on the latter region. Thus we come to the following conditions:

Condition 13.12. *For all $d > 0$ we have with $B_d = \{x : |f(x)| > d\}$*

$$\mu(B_d) < \infty.$$

Condition 13.13.

$$\int_{|f(x)|\leq d} |f(x)|\mu(dx) < \infty. \tag{13.7}$$

The next condition merges the previous two conditions with $d = 1$:

Condition 13.14.

$$\int_S \min(|f(x)|, 1)\mu(dx) < \infty. \tag{13.8}$$

This is a familiar condition that frequently shows up in the theory of Lévy processes. Under the integrability condition (13.8) it follows that the r.h.s. of (13.6) is well-defined and is finite. Note that we can write

$$\sum_{x\in\Pi} f(x) = \int_S f(x)N(dx),$$

where $N(dx)$ is a random counting measure. The following result, known as (*Campbell's theorem* gives a necessary and sufficient condition for the existence of an integral over a Poisson point process, and in addition it gives a novel version of the master equation.

Theorem 13.7. *Let f be a real-valued measurable function. Then*

$$\Sigma = \sum_{x\in\Pi} f(x)$$

is absolutely convergent w.p.1 if and only if (13.8) holds. Under (13.8):

$$\mathrm{E}(e^{\theta\Sigma}) = \exp\{\int_S (e^{\theta f(x)} - 1)\mu(dx)\}. \tag{13.9}$$

for $\theta = it$, t real. If $f \geq 0$ then (13.9) holds for any complex θ with

$$\Re\ \theta \leq 0.$$

Outline of the proof: Let $f \geq 0$. Let (f_j) be a sequence of simple functions such that

$$f_j \nearrow f \quad \mu \text{ a.s.}$$

Then for *any* real θ, we have by the monotone convergence theorem

$$\mathrm{E}(e^{\theta\Sigma}) = \lim_j \mathrm{E}(e^{\theta\Sigma}{}_j).$$

Note that both sides can be $+\infty$! Using the master equation (13.6) and monotone convergence we get (13.9) for *any* real θ.

Now let $\theta < 0$. If the integrability condition (13.8) holds, then

$$\mathrm{E}(e^{\theta\Sigma}) = \int_S (e^{\theta f(x)} - 1)\mu(dx) = I(\theta).$$

Moreover the r.h.s. is finite and tends to 0 as $\theta \to 0$. Hence Σ is finite w.p.1.

If (13.8) *does not* hold, then $I(\theta) = +\infty$ for any $\theta > 0$. Thus

$$\mathrm{E}(e^{\theta\Sigma}) = \infty$$

for any $\theta > 0$, and hence $\Sigma = \infty$ w.p.1.

If the integrability condition (13.8) holds, then

$$\mathrm{E}(e^{\theta\Sigma}) \qquad \text{and} \qquad I(\theta) = \exp\{\int_S (e^{\theta f(x)} - 1)\mu(dx)\}$$

are well-defined for $\Re\theta \leq 0$, and are *analytic* for $\Re\theta < 0$. Since they agree for θ real, $\theta < 0$, we have

$$\mathrm{E}(e^{\theta\Sigma}) = I(\theta)$$

for all $\Re\theta \leq 0$, in particular for $\theta = it$. In the general case we write

$$f = f^+ - f^-$$

with

$$f^+ = \max(f, 0) \quad , \quad f^- = \max(-f, 0),$$

and proceed using straightforward arguments.

The next result gives the mean and variance of Σ.

Theorem 13.8. *Under the integrability condition (13.8) we have*

$$\mathrm{E}(\Sigma) = \int_S f(x)\mu(dx) \tag{13.10}$$

in the sense that the l.h.s. exists if the r.h.s. converges. Furthermore, if (13.10) converges, then

$$\mathrm{var}(\Sigma) = \int_S f^2(x)\mu(dx)$$

finite or infinite.

Theorem 13.9. *Let $f_{1,}f_2$ satisfy (13.8), and let*

$$\int_S f_j(x)\mu(dx)$$

converge, and let

$$\int_S f_j^2(x)\mu(dx) < +\infty.$$

Then

$$\mathrm{cov}(\Sigma_1, \Sigma_2) = \int_S f_1(x)f_2(x)\mu(dx).$$

Exercise 13.8. *Provide a formal proof for the above expression of* $\mathrm{var}(\Sigma)$ *by differentiating the master equation*

$$\mathrm{E}(e^{it\Sigma}) = \exp\{\int_S (e^{itf(x)} - 1)\mu(dx)\}$$

w.r.t. t once and twice, and setting $t = 0$.

Exercise 13.9. *Provide a formal proof for the above expression of* $\mathrm{cov}(\Sigma_1, \Sigma_2)$ *by considering the master equation*

$$\mathrm{E}(e^{it_1\Sigma_1 + it_2\Sigma_2}) = \exp\{\int_S (e^{it_1 f_1(x) + it_2 f_2(x)} - 1)\mu(dx)\}.$$

and taking mixed second order partial derivatives

$$\frac{\partial^2}{\partial t_1 \partial t_2},$$

and setting $t_1 = t_2 = 0$.

Chapter 14

High-frequency data. Lévy Processes

14.1 Motivation and basic properties

To model shocks or jumps in the price process the simplest starting point is the *compound Poisson process*. Classically it is defined as a Poisson process with random i.i.d. jumps. An alternative representation can be obtained as follows. Let $N(dt, dx)$ be a time-homogeneous, space-time Poisson point process, counting the number of jumps of size x at time dt. The intensity of $N(dt, dx)$ is defined by

$$\mathrm{E}N(dt, dx) = dt \cdot \nu(dx),$$

where $\nu(dx)$ is called the *Lévy-measure*. Intuitively, $\nu(x)$ can be interpreted as the rate of jumps with size of x. Consider now the process defined by

$$L_t = \int_0^t \int_{\mathbf{R}^1} xN(ds, dx), \tag{14.1}$$

assuming that for any finite interval $[t_0, t_1]$ the number of random points (t, x) with t falling in the selected interval is finite. Then we get a piecewise constant process with a finite number of jumps in any finite interval. This is called a compound Poisson process.

Now let us allow an infinite number of random points (t, x) in any finite interval. If t denotes the time of transaction then we have a model with infinite activity. The integral representation (14.1) still makes sense under the technical conditions given in Campbell's theorem. In particular, if

$$\int_{\mathbf{R}^1} \min(|x|, 1)\nu(dx) < \infty,$$

then the above representation is mathematically certainly rigorous. Under this condition the sample paths of (L_t) are of *finite variation*, a property supported by empirical evidence for most indices.

We note in passing that using an appropriate limiting procedure in L_2 the integral given on the r.h.s. of (14.1) can be interpreted under the more general condition:

$$\int_{\mathbf{R}^1 \setminus 0} \min(|x|^2, 1)\nu(dx) < \infty.$$

Lévy processes have become a widely used tool in modeling price processes of financial instruments, such as stock prices or indices [44]. A Lévy process (L_t) is much like a Wiener process: a process with stationary an independent increments, but discontinuities or jumps are allowed, hence, they can be used to model shocks in financial markets. For an excellent introduction to see [31]. The relevance of Lévy processes can be justified by visual inspection of historical data, obviously exhibiting shocks and jumps. We present three time series in the figures below: stock prices for IBM, Coke and Microsoft.

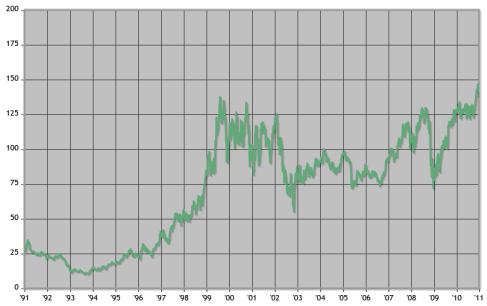

Figure 14.1: Historic daily closing prices of IBM stocks, 1991-2011

Figure 14.2: Historic daily closing prices of Coke stocks, 1991-2011

Let us now give a brief summary of basic definitions related to Lévy processes. The material of this chapter is partially based on the PhD thesis of M. Mánfay, [40].

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Figure 14.3: Historic daily closing prices of Microsoft stocks, 1991-2011

Definition 14.1. *We say that* $(L_t), t \geq 0$ *is a Lévy process if* $L_0 = 0$, *and*

1. *for any given* $0 \leq t_1 < \ldots < t_n$, *the random variables* $L_{t_2} - L_{t_1}, L_{t_3} - L_{t_2}, \ldots,$ $L_{t_n} - L_{t_{n-1}}$ *are independent.*
2. *for any* $0 \leq s < t$, *and* $\tau \geq 0$ *the distribution of* $L_{t+\tau} - L_{s+\tau}$ *is independent of* τ.

Exercise 14.1. *Show that the characteristic function of* L_t *can be written in the form*

$$\mathrm{E}e^{iuL_t} = e^{t\psi(u)}.$$

Here $\psi(u)$ is called the characteristic exponent.

Note that the logarithm of the characteristic function is linear in t, which is implied by the fact that (L_t) has independent and stationary increments. The characteristic function plays a key role in the study of Lévy processes, because unlike the density function of L_t it typically has a closed form. The c.f. of a Lévy process is given by the following celebrated Lévy-Khintchine formula:

Theorem 14.1. *Let* (L_t) *be a Lévy process. Then there exist a triplet* (b, c, ν), *with* $b, c \in \mathbf{R}, c \geq 0$, *and* ν *being a Lévy measure satisfying* $\nu(0) = 0$ *and* $\int_{\mathbf{R}^1} \min(|x|^2, 1)\nu(dx) < \infty$, *such that*

$$\mathrm{E}\left[e^{iuL_t}\right] = \exp\left[t\left(ibu - \frac{u^2 c}{2} + \int_{\mathbf{R}^1}\left(e^{iux} - 1 - iux\mathbf{1}_{|x|<1}\right)\nu(dx)\right)\right].$$

For pure-jump Lévy processes with finite variation trajectories defined by 14.1 we have the following simplified form of the Lévy-Khintchine formula:

Theorem 14.2. *Let* (L_t) *be a pure-jump Lévy process (having no Brownian motion component), with finite variation trajectories, defined by (14.1). Then*

$$\mathrm{E}\left[e^{iuL_t}\right] = \exp\left[t\left(ibu + \int_{\mathbf{R}^1}\left(e^{iux} - 1\right)\nu(dx)\right)\right].$$

Exercise 14.2. *Provide a derivation of the above simplifies version of the Lévy-Khintchine formula using Campbell's theorem.*

14.2 Lévy processes in finance

A wide range of geometric Lévy processes has been proposed by a variety of authors. One of the early models is the so-called Variance Gamma (VG)-process, which is a time changed Brownian motion with drift. The time change is realized by a so-called gamma process, which is essentially the continuous time extension of the inverse of a Poisson process. In particular, a gamma-process has independent increments having gamma distributions. A γ process is characterized by two parameters μ and ν, the mean-rate and the variance rate, respectively. For details see the Appendix of this chapter. The Lévy measure of a γ-process $\gamma(t;\mu,\nu)$ is given by:

$$\nu(dx) = \mu x^{-1} e^{-\nu x} dx$$

To define the VG process let $B(t;\theta,\sigma)$ be a Brownian motion with drift, given by

$$B(t;\theta,\sigma) = \theta t + \sigma B(t),$$

where $(B(t))$ is a standard Brownian motion. Let $\gamma(t;\mu,\nu)$ be a gamma process with mean rate μ and variance rate ν,. Then the VG process $(X(t;\sigma,\nu,\theta))$ is defined as

$$X(t;\sigma,\nu,\theta) = B(\gamma(t;1,\nu);\theta,\sigma).$$

Note that the mean-rate of the γ-process is fixed as 1. This is due to our freedom to fix one of the scaling factors μ or σ. Thus VG-processes form a three-parameter class of processes.

The Lévy measure of a VG process can be obtained by first computing its characteristic function and then applying Lévy-Khintchine's formula in the inverse direction. Thus we get:

$$\nu(dx) = \begin{array}{ll} \frac{\mu_n^2}{\nu_n} \frac{\exp(-\frac{\mu_n}{\nu_n}|x|)}{|x|} dx & \text{if } x < 0 \\ \frac{\mu_p^2}{\nu_p} \frac{\exp(-\frac{\mu_p}{\nu_p} x)}{x} dx & \text{if } x > 0, \end{array}$$

where the parameters $\mu_p, \nu_p, \mu_n, \nu_n$ are obtained in terms of the original parameters as follows:

$$\begin{aligned} \mu_p &= \frac{1}{2}\sqrt{\theta^2 + \frac{2\sigma^2}{\nu}} + \frac{\theta}{2} \\ \mu_n &= \frac{1}{2}\sqrt{\theta^2 + \frac{2\sigma^2}{\nu}} - \frac{\theta}{2} \\ \nu_p &= \mu_p^2 \nu \\ \nu_n &= \mu_n^2 \nu \end{aligned}$$

From here we get the following remarkable property of VG processes: a VG process $X_t(\sigma, \nu, \theta)$ can be written as the difference of two gamma processes:

$$X_t(\sigma, \nu, \theta) = \gamma_p(t, \mu_p, \nu_p) - \gamma_1(t, \mu_n, \nu_n).$$

In particular, it follows that a VG process is of finite variation.

Another early model in finance, proposed by Mandelbrot to model cotton prices, is the symmetric α-stable process, with $0 < \alpha < 2$, is defined via the Lévy measure

$$\nu(dx) = C|x|^{-1-\alpha}dx.$$

A recently widely studied class of Lévy processes is the CGMY process, due to Carr, Geman, Madan and Yor [14]. It is obtained by multiplying the Lévy-density of a stable process with a decreasing exponential on each half of the real axis. Its Lévy-measure, using standard parametrization, is of the form:

$$\nu(dx) = \frac{Ce^{-G|x|}}{|x|^{1+Y}}\mathbf{1}_{x<0}dx + \frac{Ce^{-Mx}}{|x|^{1+Y}}\mathbf{1}_{x>0}dx,$$

where $C, G, M > 0$, and $0 < Y < 2$. Intuitively, C controls the level of activity, G and M together control skewness. Y controls the density of small jumps, i.e. the fine structure. For $Y < 1$ the corresponding Lévy process is of finite variation. Allowing $Y = 0$ yields the Variance Gamma process.

The characteristic exponent of the CGMY process is given by

$$\psi(u) = C\Gamma(-Y)\left((M - iu)^Y - M^Y + (G + iu)^Y - G^Y\right),$$

where Γ denotes the gamma-function.

In the following three figures we present the trajectories of compound Poisson approximations of three CGMY processes, with cut-off $\epsilon = 10^{-5}$, meaning that jumps with absolute value less than 1 are nullified.

On Figures bellow we see a varieties of CGMY processes: with significant drift and local noise on the fine structure due to the large value of Y, a symmetric CGMY process with weak tempering, exhibiting some clustering phenomenon. Finally, on the last Figure the tempering effect is stronger, and the local noise is weaker, resulting in an almost constant process.

Although the geometric CGMY model is widely used in finance, it can not always be validated on real data. Surprisingly, even the assumed independence of daily log-returns may not always be validated on historical data.

14.3 The empirical characteristic function method

An important practical problem is the estimation of the parameters of an assumed model class of Lévy processes from historical data. Suppose that we are given a sample of N independent and identically distributed observations obtained as the increments of a Lévy

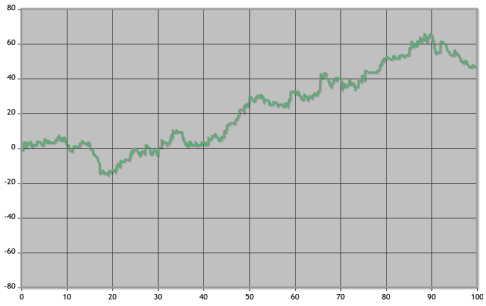

Figure 14.4: A CGMY process with visible drift and local noise, due to a high value of $Y = 1.95$. The further parameters are $C = 1, G = 0.1$ and $M = 0.2$.

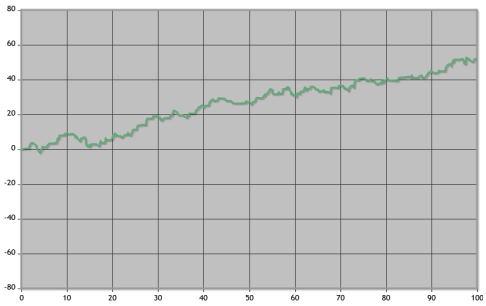

Figure 14.5: A CGMY process with visible drift, but less volatile local noise, due to a medium value of $Y = 1.2$. The further parameters are $C = 2, G = 1.2$ and $M = 1.5$

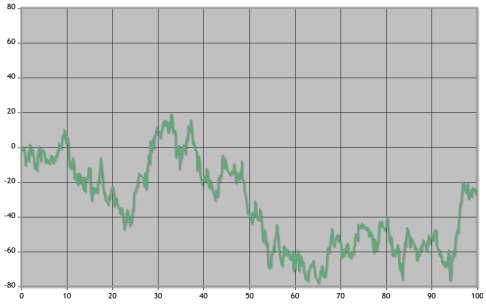

Figure 14.6: A symmetric CGMY process with $Y = 0.5$. The further parameters are $C = 2$ and $G = M = 0.3$

process. If we knew the probability density function of these random variables then we could apply an ML (Maximum Likelihood) estimation method. The challenge of the present problem is that it is the characteristic function of the noise that is explicitly given, rather than the density function. Namely, it is typical for a Lévy process (X_t) that the probability density function of X_t does not have a closed form, while its characteristic function is known up to some unknown parameters. A natural approach to solve

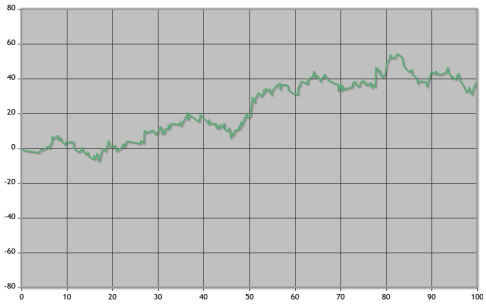

Figure 14.7: A CGMY process with a low value of $Y = 0.3$. The further parameters are $C = 1, G = 0.7$ and $M = 0.8$.

this problem is therefore to apply the so-called empirical characteristic function (ECF) method. In fact, the ECF method is widely used in finance.

We briefly describe the method for i.i.d. samples, see [17]. Let $(r_1, r_2, \dots r_N)$ be i.i.d. observations, and let a closed form of the characteristic function $\varphi(u, \eta)$ be known, with η being a p-dimensional parameter vector, and $u \in \mathbf{R}$. The true value of the parameter will be denoted by η^*.

The idea is to estimate η^* by a value of η for which the characteristic function (cf) best matches the empirical characteristic function (ecf). The error for any fixed u is defined as

$$\overline{h}_N(u, \eta) = \frac{1}{N} \sum_{k=1}^{N} h_k(u, \eta),$$

where $h_k(u, \theta)$ is the generalized, normalized moment function, or score:

$$h_k(u, \eta) = e^{iur_k} - \varphi(u, \eta).$$

An important, obvious property of the score is that

$$\mathrm{E}h_k(u, \eta^*) = 0, \quad \text{for all} \quad u,$$

where η^* denotes the true parameter. Take a finite set of moments, evaluated, say at $u_1, ..., u_m$, with $m > p$, and set

$$\overline{h}_N(\eta) = (\overline{h}_N(u_1, \eta), ..., \overline{h}_N(u_m, \eta))^T.$$

We would the estimate η^* by minimizing the weighted quadratic cumulative error

$$V_N(\eta) = |K^{-1/2}\overline{h}_N(\eta)|^2, \tag{14.2}$$

where K is an appropriate, $m \times m$ weighing matrix. The resulting estimator will be denoted by $\hat{\eta}_N$.

To get the asymptotic covariance matrix of $\hat{\eta}_N$, we proceed along standard arguments (see also the analysis of the PE estimator for MA or ARMA processes). We will provide only an outline here. First, let the expectations of the scores be denoted by

$$g_k(\eta) = \mathrm{E} h_{k,n}(\eta).$$

Let the complex conjugate of a vector or matrix M be denoted by M^*. Then the gradient equation (p equations) reads as:

$$\overline{h}_\eta^*(\eta) K^{-1} \overline{h}(\eta) = 0.$$

The left hand side can be considered as a new set of exactly p scores. The corresponding asymptotic score can be written as

$$g_\eta^*(\eta) K^{-1} g(\eta),$$

while its derivative at η^*, (the Hessian of the asymptotic cost) is

$$R = g_\eta^*(\eta^*) K^{-1} g_\eta(\eta^*).$$

Let us define the $M \times p$ matrix

$$G = g_\eta(\eta^*).$$

Then the Hessian of the asymptotic cost is

$$T = G^* K^{-1} G.$$

To get the asymptotic covariance of the new set of scores define the $M \times M$ covariance matrix by

$$C_{k,l} = \mathrm{E}\ h_{k,n}^*(\eta^*) h_{l,n}(\eta^*).$$

Note that we have

$$C_{k,l} = \varphi(u_k - u_l, \eta^*) - \varphi(u_k, \eta^*)\varphi(-u_l, \eta^*).$$

Set

$$C = (C_{k,l}).$$

Thus the asymptotic covariance of the new set of scores is

$$S = G^* K^{-1} C K^{-1} G.$$

The asymptotic covariance of the estimator $\hat{\eta}_N$ is then

$$T^{-1} S T^{-1},$$

or equivalently,

$$(G^*K^{-1}G)^{-1}\ G^*K^{-1}CK^{-1}G\ (G^*K^{-1}G)^{-1}.$$

It is relatively easy to see (using simple matrix inequalities) that the optimal value of K is

$$K = C.$$

Thus we finally arrive at the following conclusion:

Proposition 14.2. *The asymptotic covariance matrix of the ECF estimator $\hat{\eta}_N$ for i.i.d. data, using $K = C$, is*

$$\Sigma = (G^*C^{-1}G)^{-1}.$$

It can be shown that the method gives a consistent estimate of η^*, and that the distribution of the estimation error $\hat{\eta}_N - \eta^*$ is normal.

A fascinating feature of the ECF method is that it can as efficient as the ML method, when taking a continuum of u-s. To conclude this section we provide a heuristics behind the latter claim. Let the data be $\{y_n\}_{n=1}^N$, and let the empirical distribution be $F_n(x)$. Let the family of distributions is $F(x,\eta)$ having a density $f_\eta(y_n,\eta)$. The likelihood equation is then

$$\sum_{n=1}^{N} \frac{f_\eta(y_n,\eta)}{f(y_n,\eta)} = 0.$$

Rewrite this as

$$\int_{-\infty}^{\infty} \frac{f_\eta(x,\eta)}{f(x,\eta)}(dF_n(x) - dF(x,\eta)) = 0.$$

Using Parseval's theorem (formally) yields

$$\int_{-\infty}^{\infty} w(t,\eta)(\varphi_n(t) - \varphi(t,\eta))dt = 0,$$

where

$$w(t,\eta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\partial \log f(x,\eta)}{\partial \eta} e^{-itx}dx, \tag{14.3}$$

and $\varphi_n(t)$ is the empirical c.f. This reasoning shows that the likelihood equation of the ML method is equivalent to an empirical characteristic function method using a weighted integral of the scores $(\varphi_n(t) - \varphi(t,\eta))$ as given by the equation (14.3). Since the ML estimates are consistent, it is reasonable to argue that the weights $w(t,\eta)$ can be replaced by their asymptotic value

$$w(t,\eta^*) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left.\frac{\partial \log f(x,\eta)}{\partial \eta}\right|_{\eta=\eta^*} e^{-itx}dx$$

without effecting the asymptotic properties of the ECF estimator. Now, it can be shown that this variant of the ECF method is equivalent to choosing $K = C$, as described above, using a *continuum* of u-s. Thus we finally conclude that the ECF method with a continuum of u-s can be as efficient as the ML method.

In the figures below we present the Hessian of the asymptotic cost function for our benchmark examples defined above:

$$H = \begin{bmatrix} 6.272 & -17.70 & 30.25 & 111.3 \\ -17.70 & 159.5 & 0.4964 & -323.4 \\ 30.25 & 0.4964 & 294.9 & 511.7 \\ 111.3 & -323.4 & 511.7 & 1981 \end{bmatrix}$$

Figure 14.8: The Hesse matrix of the asymptotic cost function of the ECF method, and its eigenvalues for a CGMY process with visible drift and local noise, due to a high value of $Y = 1.95$. The further parameters are $C = 1, G = 0.1$ and $M = 0.2$. We use 100 equidistant u-s on $[-10, 10]$.

$$H = \begin{bmatrix} 182.0 & 339.6 & 450.9 & -279.7 \\ 339.6 & 4140 & 2586 & -1354 \\ 450.9 & 2586 & 2029 & -1167 \\ -279.7 & -1354 & -1167 & 713.9 \end{bmatrix}$$

Figure 14.9: The Hesse matrix of the asymptotic cost function of the ECF method, and its eigenvalues for a CGMY process with visible drift and local noise, due to a high value of $Y = 1.5$. The further parameters are $C = 1.5, G = 1.2$ and $M = 0.5$. We use 100 equidistant u-s on $[-10, 10]$.

$$H = \begin{bmatrix} 5.237e+16 & 6.517e+16 & 7.284e+16 & -4.592e+17 \\ 6.517e+16 & 1.078e+17 & 1.108e+17 & -6.000e+17 \\ 7.284e+16 & 1.108e+17 & 1.165e+17 & -6.604e+17 \\ -4.592e+17 & -6.000e+17 & -6.604e+17 & 4.060e+18 \end{bmatrix}$$

Figure 14.10: The Hesse matrix of the asymptotic cost function of the ECF method, and its eigenvalues for a CGMY process with visible drift, but less volatile local noise, due to a medium value of $Y = 1.2$. The further parameters are $C = 2, G = 1.2$ and $M = 1.5$ We use 100 equidistant u-s on $[-10, 10]$.

$$H = \begin{bmatrix} 0.1468 & 0.03579 & 0.02700 & -0.2609 \\ 0.03579 & 0.01155 & 0.007205 & -0.06422 \\ 0.02700 & 0.007205 & 0.01012 & -0.07268 \\ -0.2609 & -0.06422 & -0.07268 & 0.6234 \end{bmatrix}$$

Figure 14.11: The Hesse matrix of the asymptotic cost function of the ECF method, and its eigenvalues for a CGMY process with a low value of $Y = 0.3$. The further parameters are $C = 1, G = 0.7$ and $M = 0.8$. The further parameters are $C2, G = 1.2$ and $M = 1.5$. We use 100 equidistant u-s on $[-10, 10]$.

14.4 Appendix: the gamma process

In this section we provide a brief introduction to gamma-processes that play a major role in VG modelling. To construct a gamma process let ξ_n be an i.i.d. sequence of random variables with exponential distribution having density $\lambda e^{-\lambda x}$ for $x > 0$. Let $s_k = \xi_1 + ... + \xi_k$. The probability density function and characteristic function of s_k are given by

$$\lambda^k x^{k-1} e^{-\lambda x}/(k-1)! \tag{14.4}$$

and

$$\mathrm{E}\exp(ius_k) = \left(\frac{\lambda}{\lambda - iu}\right)^k = \left(\frac{1}{1 - \frac{iu}{\lambda}}\right)^k. \tag{14.5}$$

For the means and the variances of s_k we have:

$$\mathrm{E}s_k = k/\lambda \quad \text{and} \quad \sigma^2(s_k) = k/\lambda^2.$$

We can look upon s_k as a stochastic process defined over the positive integers, with independent and identically distributed increments. Let us re-parametrize the density above by introducing the new variables

$$\nu = 1/\lambda \quad \text{and} \quad t = k/\lambda = k\nu. \tag{14.6}$$

Here $\nu = 1/\lambda$ is the *mean life-time*. Then define

$$\gamma_t = s_k = s_{t/\nu}.$$

Remember that t/ν is the number of exponential terms. The probability density function of γ_t can be written as

$$f_t(x) = \left(\frac{1}{\nu}\right)^{t/\nu} \frac{x^{t/\nu - 1} e^{-x/\nu}}{\Gamma(t/\nu)}. \tag{14.7}$$

The characteristic function of γ_t is given by

$$\mathrm{E}\exp(iu\gamma_t) = \left(\frac{1}{1-i\nu u}\right)^{t/\nu}. \tag{14.8}$$

Finally, the means and the variances of γ_t are

$$\mathrm{E}\gamma_t = t \quad \text{and} \quad \sigma^2(\gamma_t) = t\nu. \tag{14.9}$$

Now it can be shown that $f_t(x)$ as defined above, is a density function for *any* real $t \geq 0$. This is called a gamma-density. The corresponding characteristic function is given by (14.8) for *any* real $t \geq 0$. Obviously, thus set of characteristic functions is closed under multiplication. Thus gamma-densities are closed under convolution. Consequently, we can construct a stochastic process γ_t, with $t \geq 0$ real, with stationary independent increments, so that the the density function of $\gamma_{t+h} - \gamma_t$ is f_h. This is called a gamma-process. Obviously, the means and variances of γ_t are obtained as in (14.9) for *any* real t. Therefore we say that the *mean rate* of γ_t is 1, and its *variance rate* is ν.

Finally, we can re-scale the process by setting, with some $\mu' > 0$,

$$t' = t/\mu' \quad \text{and} \quad \gamma'_{t'} = \gamma_t. \tag{14.10}$$

Then

$$\mathrm{E}\gamma'_{t'} = \mu' t' \quad \text{and} \quad \sigma^2(\gamma'_{t'}) = (\mu' t')\nu = (\mu'\nu)t'. \tag{14.11}$$

Correspondingly, we say that the *mean-rate* of the re-scaled process is μ', and the *variance rate* of the re-scaled process is $\nu' = \mu'\nu$. We can express the old variables in terms of the new variables by the the following scaling equations:

$$t = \mu' t' \quad \text{and} \quad \nu = \nu'/\mu'.$$

Expressing the density function of $\gamma'_{t'} = \gamma_t$ in terms of these parameters, and changing the roles of parameters with and without superscripts, and correspondingly making the replacements

$$t \to \mu t \quad \text{and} \quad \nu \to \nu/\mu$$

we get

$$f_t(x;\mu,\nu) = \left(\frac{\mu}{\nu}\right)^{\mu^2\frac{t}{\nu}} \frac{x^{\mu^2\frac{t}{\nu}-1}e^{-\mu\frac{x}{\nu}}}{\Gamma(\mu^2\frac{t}{\nu})}. \tag{14.12}$$

Note that the following scaling property holds: for any $c > 0$

$$f_t(x;\mu,\nu) = f_{ct}(x;\mu/c,\nu/c). \tag{14.13}$$

A random variable with this distribution will be denoted by $\gamma_t = \gamma_t(\mu, \nu)$, with μ denoting the mean rate and ν denoting the variance rate. Its characteristic function is given by

$$\phi_t(u; \mu, \nu) = \left(\frac{1}{1 - i\frac{\nu}{\mu}u} \right)^{\mu^2 \frac{t}{\nu}}.$$

Similarly, a stochastic process γ_t with stationary independent increments, so that the the density function of $\gamma_{t+h} - \gamma_t$ is $f_h(x; \mu, \nu)$ will be denoted by $\gamma_t(\mu, \nu)$. This is then a Lévy process, the Lévy density of which can be explicitly determined, see [49].

$$k(x; \mu, \nu) = \frac{\mu^2 e^{-\mu \frac{x}{\nu}}}{x} \mathbf{1}_{x>0}. \tag{14.14}$$

Bibliography

[1] T.W. Anderson. *An introduction to multivariate statistical analysis.* Wiley, New York, 1984.

[2] L. Bachelier Théorie de la spéculation, in Annales Scientifiques de l' École Normale Supérieure, 3 (17), pp. 21-86, 1900.

[3] F.E. Benth, S.J. Benth, Dynamic pricing of wind futures. Energy Economics, 31, pp. 16-24, 2009.

[4] A. Benveniste, M. Métivier and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer Verlag, Berlin, 1990.

[5] F. Black, Studies of stock price volatility changes, In Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section 177-181, 1976.

[6] F. Black and M. Scholes, The valuation of option contracts and a test of market efficiency, Journal of Finance, 27(2) 399-417, 1972.

[7] T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, Journal of Econometrics, 31 307-327, 1986.

[8] T. Bollerslev, R. Y. Chou, and K.F. Kroner. ARCH modelling in Finance. *Journal of Econometrics*, 52:5–59, 1992.

[9] T. Bollerslev, R.F. Engle and D.B. Nelson, ARCH models. In R.F. Engle and D.L. McFadden, editors, Handbook of Econometrics Vol.4, Elsevier Science B.V., Amsterdam, 1994.

[10] P. Bougerol, N. Picard. Stationarity of GARCH processes and of some nonnegative time series, Journal of Econometrics, 52 115-127, 1992.

[11] P. Bougerol, N. Picard. Strict stationarity of generalized autoregressive processes, The Annals of Probability, 20(4) 1714-1730, 1992.

[12] G.E.P. Box and G.M. Jenkins, Time Series Analysis, Forecasting and Control, Wiley, 1976.

[13] A. Brandt, The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients, Adv. Appl. Prob. 18 211-220, 1986.

[14] P.Carr, H.Geman, D.Madan, M.Yor, The fine structure of asset returns: an empirical investigation. Journal of Business, 75 (2), pp. 305-332, 2000.

[15] M. Carrasco, X. Chen, Mixing and moment properties of various GARCH and stochastic volatility models. Econometric Theory 18(1) 17-39, 2002.

[16] M. Carrasco, M. Chernov, J.-P. Florens, E. Ghysels, *Efficient estimation of general dynamic models with a continuum of moment conditions*, in Journal of Econometrics, 140 (2), pp. 529–573, 2007.

[17] M.Carrasco, J.-P.Florens, Efficient GMM estimation using the empirical characteristic function. Idei working papers, 140. 2002.

[18] R. Cont, Empirical properties of asset returns: Stylized facts and statistical issues, Quantitative Finance 1 1-14, 2001.

[19] R. Cont, Volatility clustering in financial markets: empirical facts and agent-based models, In: Long memory in economics (eds.: A. Kirman and G. Teyssiere), 2005

[20] R. Cont, P. Tankov, Financial Modelling with Jump Processes. Journal of the American Statistical Association, 101, pp. 1315-1316, 2006.

[21] M. Deistler. System identification in econometrics, Manuscript. 1996.

[22] R.F. Engle. Autoregressive Conditional Heteroskedasticity with Estimates of The Variance of the UK Inflation. *Econometrica*, 50:987–1008, 1982.

[23] R.F. Engle, Risk and volatility: econometrics models and financial practice, The American Economic Review 94(3) 405-420, 2004.

[24] P.D. Feigin and R.L. Tweedie, Random coefficient autoregressive processes: a Markov chain analysis of stationarity and finiteness of moments, Journal of Time Series Analysis, 6(1) 1-14, 1985.

[25] H. Furstenberg and H. Kesten, Products of random matrices, Ann. Math. Statist. 31 457-469, 1960.

[26] L. Gerencsér and Zs. Orlovits, L_q-stability of products of block-triangular stationary random matrices, Acta Scientiarum Mathematicarum (Szeged) 74 927-944, 2008.

[27] J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series models. Journal of Time Series Analysis, 4 221–238, 1983.

[28] E.J. Hannan and M. Deistler. The statistical theory of linear systems. Wiley, 1988.

[29] R.Z. Hasminskii, Stochastic stability of differential equations, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.

[30] C. He and T. Terasvirta, Properties of moments of a family of GARCH processes, Journal of Econometrics 92 173-192, 1999.

[31] J. Jacod, A.N. Shiryaev, Limit theorems for stochastic processes (2. ed.). Springer, 2002.

[32] S. Johansen. Statistical analysis of cointegration vectors. Journal of Economics Dynamics and Control, 12:231–254, 1988.

[33] M. Karanasos, The second moment and the autocovariance function of the squared errors of the GARCH model, Journal of Econometrics 90, 63-76, 1999.

[34] H.A. Karlsen, Existence of moments in a stationary stochastic difference equation, Adv. Appl. Prob. 22 129-146, 1990.

[35] H. Kesten and F. Spitzer, Convergence in distribution of products of random matrices, Zeitschrift für Wahrscheinlichkeit und verwandete Gebiete, Vol. 67 pp. 363-386, 1984.

[36] J.F.C. Kingman: Poisson Processes, Oxford Science Publications, Clarendon Press, Oxford, 1995.

[37] S. Ling, On the probabilistic properties of a double threshold ARMA conditional heteroscedasticity model, Journal of Applied Probability 36 688-705, 1999.

[38] S. Ling, M. McAleer, Necessary and sufficient moment conditions for the GARCH(p, q) and asymmetric power GARCH(p, q) models, Econometric Theory 18 722-729, 2002.

[39] B. Madan, P. Carr, C. Chang, The Variance Gamma Process and Option Pricing. European Finance Review, 2, pp. 79–105, 1998.

[40] M. Mánfay, Identification of Financial Time Series Driven by Lévy Processes Ph.D. Thesis (to be submitted), Central European University, Doctoral Program in Mathematics and its Applications. 2014.

[41] H.M. Markowitz, Portfolio Selection, Journal of Finance 7(1) 77-91, 1952.

[42] R.C. Merton Theory of rational options pricing, Bell Journal of Economics and Management Science 4(1) 141-183, 1973.

[43] A. Milhøj, The moment structure of ARCH processes, Scandinavian Journal of Statistics 12, 281-292, 1985.

[44] Y. Miyahara and A. Novikov, Geometric Lévy Process Pricing Model, in Research Paper Series 66, Quantitative Finance Research Centre, University of Technology, Sydney, 2001.

[45] Zs. Orlovits, Statistical Analysis of Stochastic Volatility Models. Ph.D. Thesis, Eötvös Loránd University, Faculty of Sciences, Doctoral School of Mathematics, 2011.

[46] A. Pagan, The econometrics of financial markets, Journal of Empirical Finance 3 15-102, 1986.

[47] Gy. Pap, M. Ispány, M.van Zuijlen, Asymptotic inference for nearly unstable INAR(1) models, J. Appl. Probab. 40(3), 750-765. 2003.

[48] D.T. Pham, The mixing property of bilinear and generalised random coefficient autoregressive models, Stochastic Processes and their Applications 23 291-300, 1986.

[49] D. Revuz, M. Yor, Continuous martingales and Brownian motion. Springer, 293. 1999.

[50] W. Sharp, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, Journal of Finance 19(3) 425-442, 1964.

[51] S. Taylor, Modelling financial time series, New York, Wiley, 1986.

[52] Gy. Terdik, Bilinear Stochastic Models and Related Problems of Nonlinear Time Series Analysis, volume 142 of Lecture Notes in Statistics, Springer Verlag, New York 1999.

[53] W. Vervaat, On a stochastic difference equation and a representation of nonnegative infinitely devisible random variables, Adv. Appl. Prob. 11 750-783, 1979.

[54] P.A. Zadrozny, Necessary and sufficient restrictions for existence of a unique fourth moment of a univariate GARCH(p, q) process, Advances in Econometrics 20(1) 365-379, 2006.