# Stochastic Machine Learning
# Chapter 03 - Time series and LSTM


Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de


SS 2024

# Universal approximation theorems

We mention finally the two papers

- Kratsios (2021)
- Benth, Detering, and Galimberti (2023)

# Time series

Dynamic phenomena are the classical frame for **time series**.

- A **time series** $X = (X_t)_{t\in\mathcal{T}}$ with at most countable $\mathcal{T}$ is a family of random variables.
- The state space can be quite general - for example $\mathbb{R}$, $\mathbb{R}^d$ or even a Hilbert/ Banach-space (functional time series).
- If $\mathcal{T}$ is not finite then we need to require that the state space is[1] **Polish**. Then the extension theorem of Kolmogorov says that the distribution of $Y$ is already determined by all finite-dimensional marginal distributions (fidis), i.e. the distribution of

$$\big(Y_{t_1}, \dots, Y_{t_n}\big), \quad t_1, \dots, t_n \subset \mathcal{T}, n \in \mathbb{N}.$$

- If the random variables have densities we have that

$$f(y_1, \dots, y_n) = f(y_n|y_{n-1}, \dots, y_1) \cdots f(y_2|y_1) \cdot f(y_1),$$

  i.e. the joint distribution is determined by the conditional probabilities. This is an important step for developing dynamic evolutions.

[1] A Polish space is a separable, completely metrizable space. If the space is not Polish, many things may go wrong since product sets become very badly behaved, see for example https://mathoverflow.net/questions/20919/polish-spaces-in-probability.

- If all fidis are Gaussian, then the process is called **Gaussian** or a **Gauss-process**.
- (In which spaces can we define Gauss distributions?)
- In this case, it is sufficient to study mean- and covariance function, i.e.

  $$m(t) = E[Y_t]$$

  and

  $$\gamma(t,h) = \mathrm{Cov}(Y_t, Y_t + h) = E\big[(Y_t - m(t)) \cdot (Y_{t+h} - m(t+h))\big].$$
- $\gamma$ is called the **auto-covariance function**.
- The process $Y$ is called **strictly stationary**, if all distributions $Y_{t+1}, \ldots, Y_{t+n}$ and $Y_1, \ldots, Y_n$ are identical for all $t \in \mathcal{T}$ and all $n \in \mathbb{N}$ (whenever this makes sense - this can also be formulated for general $\mathcal{T}$, but we stay a bit simpler).

# Examples

- $Y_0, Y_1, \ldots$ are i.i.d. Then they are also stationary. If the first two moments exist, we have that
$$m(t) = m(0) = m, \qquad \gamma(t, h) = 0$$
for any $h > 0$.
- Random walk: For $(X_i)$ i.i.d., we define
$$Y_t = X_1 + \cdots + X_t,$$
$Y_0 = 0$. If $X_i \in \{-1, 1\}$ this is the **binomial tree**

# Properties

- Consider for simplicity $\mathcal{T} = \{0, 1, 2, \dots\}$. Assume that $f \in \mathcal{F}$ is a distribution determing class (for example Fourier transforms, or all continuous and bounded functions)
- The process $Y$ is called **Markovian**, if
$$E[f(Y_t)|Y_{t-1}, \dots, Y_0] = E[f(Y_t)|Y_{t-1}]$$
for all $f \in \mathcal{F}$ and all $t \geq 1$.
- Intuitively, the distribution of $Y_t$ does not depend on the full past but only on the previous value $Y_t$.
- This is often very useful and simplifies the setting - but it is also very often not true. Think of financial time series, etc.
- By enlarging the state space we can always make a process Markovian - how?

# Autoregressive and moving average

- A process is called **autoregressive** of order $k$, if for all $t \geq 1$,
$$E[Y_t|Y_{t-1},\dots,Y_1] = E[Y_t|Y_{t-1},\dots,Y_{t-k}].$$
- Here the expectation only depends on the last $k$ values - however in any form. Typically one uses affine dependence only -
- The process is called **linear autoregressive**, if
$$E[Y_t|Y_{t-1},\dots,Y_1] = a_0 + a_1 Y_{t-1} + \dots + a_k Y_{t-k}.$$
We denote the process class by **AR(k)**.
- We consider $(Z_t)$ as i.i.d. with mean zero and variance $\sigma^2$ and would consider
$$Y_t = \frac{1}{k}\sum_{i=1}^{k} Z_{t-k}$$
as a moving average. This inspires the following definition.

## Definition

A real-valued process $Y$ is called **ARMA(p,q)** if it is strictly stationary and can be represented as
$$Y_t - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p} = c + Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}.$$
Here, $c \in \mathbb{R}$ and $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p)$ and $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^p$ have no common factors.

We have the compact representation
$$\phi(L)Y_t = c + \theta(L)Z_t,$$
where $L$ is the lag-operator.

## Example (MA(1))

In this case we have

$$Y_t = c + Z_t + \theta_1 Z_{t-1}.$$

Moreover,

$$\gamma(t,h) = \begin{cases} \sigma^2 & h = 0 \\ \theta_1 \sigma^2 & h = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then, if $Y_0$ is distributed as $Z_1$ and independent of $(Z)$, then $Y$ is stationary.

## Example (AR(1))

In this case, (for simplicity $c = 0$)

$$Y_t = \phi_1 Y_{t-1} + Z_t.$$

Then,

$$\gamma(t,h) = \begin{cases} \gamma(t,0) & h = 0 \\ \mathbb{E}(Y_t Y_{t+1}) = \mathbb{E}(Y_t^2 \phi_1) = \phi_1 \gamma(t,0) & h = 1 \\ \phi_1^h \gamma(t,0) & \text{otherwise.} \end{cases}$$

# AR(1)

- Does there exist $Y = (Y_t)_{t\in\mathbb{Z}}$ such that $Y$ is AR(1)?
- Let us look at the following equation

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + Z_t \\ &= Z_t + \phi_1 Z_{t-1} + \cdots + \phi_1^k Z_{t-k} + \phi_1^{k+1} Y_{t-k-1} \\ &\overset{?}{=} \sum_{j=0}^{\infty} \phi_1^j Z_{t-j}. \end{aligned} \tag{1}$$

In which sense do we have to understand the last equality? The answer resides on the following result (Brockwell and Davis (1991), Prop 3.1.1.):

## Satz

*Consider $(Z_t)_{t\in\mathbb{Z}}$ such that $\sup_{t\in\mathbb{Z}} \mathbb{E}(Z_t^2) < \infty$ and $(\psi_i)_{i\in\mathbb{Z}}$ such that $\sum_{i\in\mathbb{Z}} |\psi_i| < \infty$. The the series*

$$\psi(L) Z_t := \sum_{i\in\mathbb{Z}} \psi_i L^i Z_t$$

*converges absolutely with probability one.*

Since $Y$ is stationary, $\gamma(t,0) = \gamma(0)$ and

$$\gamma(0) = \mathrm{Cov}(Y_t, Y_t) = \mathrm{Cov}(\phi_1 Y_{t-1} + Z_t, \phi_1 Y_{t-1} + Z_t) = \phi_1^2 \gamma(0) + \sigma^2.$$

We have the nice formula

$$\gamma(0) = \frac{\sigma^2}{1-\phi_1^2}.$$

# Existence

## Proposition

A stationary solution $Y$ for an ARMA(p,q)-process exists if and only if

$$\phi(z) \neq 0, \qquad \text{for all } z : |z| = 1.$$

It is interesting to have a causal solution which can be obtained if $\phi(z) \neq 0$ for all $z : |z| \leq 1$.

## Proof.

With $\phi(z) \neq 0 \quad \forall |z| \leq 1$ we have

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \zeta_j z^j =: \zeta(z) \qquad |z| \leq 1$$

Then

$$Y_t = \zeta(L)\theta(L)Z_t$$

and $Y$ is causal. □

# Estimation

- A weakly stationary process is determined by $m$ and $\gamma$. We estimate these by moment methods.
$$\hat{m}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i.$$
which is consistent and asymptotically normal.
- The estimator
$$\hat{\gamma}_n(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}(Y_{t+|h|} - \hat{m}_n)(Y_t - \hat{m}_n)$$
however has a number of difficulties, in particular when the sample size is to small. One estimates typically using the Yule-Walker equations which exploit better the properties of the time series.

# Financial time series

- Financial time series have characteristics which differ from most other time series.
- Typically we have clusters of high and low volatility.
- It was the idea of Robert Engle to put forward a model which realizes this in a time series structure.
- **Homescedasticity** denotes a homogeneous variance, heteroscedasticity a time-varying variance.
- **Conditional heteroscdasticity**, i.e. a time varying variance conditional on the past observations is the key to model financial markets.

# ARCH

- The simplest ARCH specification looks as follows: we consider $(Z_t)$ iid with $E[Z_t] = 0$ and $\mathrm{Var}(Z_t) = 1$. For this we say $Z$ is strict white noise and write $Z_t \sim SWN(0, 1)$.
- $Y$ is called ARCH(1) if

$$Y_t = \sigma_t Z_t \tag{2}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2, \quad \alpha_0, \alpha_1 > 0. \tag{3}$$

- We have obviously conditional heteroscedasticity: clearly, $E[Y_t|Y_{t-1}] = 0$ and

$$\mathrm{Var}(Y_t|Y_{t-1}) = \sigma_t^2.$$

- Engle (1982) assumed additionally that $Z_t \sim \mathcal{N}(0, 1)$. Then $Y_t$ given $Y_{t-1}$ is normally distributed.
- The unconditional distribution is obtained as mixture

$$\begin{aligned} &E\left[P(Y_t \le x|Y_{t-1})\right] \\ &= E\left[\Phi\left(\frac{x}{\sigma_t}\right)\right] = \int \Phi\left(\frac{x}{y}\right) f_{\sigma_t}(y)dy, \end{aligned}$$

  where $f_{\sigma_t}$ is the unconditional distribution of $\sigma_t$. The mixture distribution often has much fatter tails than the normal distribution, matching the observations in financial markets.

- Weak stationarity is equivalent to $\alpha_1 < 1$.

## Definition

Let $Z$ be strict white noise. Then $Y$ is a **GARCH(p,q)** process if

(i) $Y$ iis strictly stationary

(ii) $\forall\ t \in \mathbb{Z}$ and appropriate $\alpha_0 > 0, \alpha_i, \beta_j \geq 0 \quad i = 1, \ldots, p\ j = 1, \ldots, q$ it holds that

$$Y_t = \sigma_t Z_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 + \sum_{j=1}^{q} \beta_i \sigma_{t-i}^2.$$

# Volatility

- A GARCH(p,0) process is an ARCH(p) process, of course
- The generalization is similar to AR $\rightarrow$ ARMA, but of course not the same.
- As expected $Y^2$ is an ARMA process:

$$Y_t^2 = E[Y_t^2|Y_{t-1}, Y_{t-2}, \dots] + \underbrace{Y_t^2 - E[Y_t^2|Y_{t-1}, Y_{t-2}, \dots]}_{=:V_t}$$

$$= \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}^2 + \sum_{j=1}^{q} \beta_i \sigma_{t-i}^2.$$

- Here, $V_t = \sigma_t^2(Z_t^2 - 1) = Y_t^2 - \sigma_t^2$. We obtain

$$Y_t^2 = \alpha_0 + \sum_{i=1}^{max(p,q)} (\alpha_i + \beta_i) Y_{t-i}^2 - \sum_{j=1}^{q} \beta_i V_{t-i}^2 + V_t,$$

  with $\alpha_i = 0$ and $\beta_j = 0$ for $i > p$ and $j > q$.
- The GARCH model leads hence to an ARMA model in the squares.
- We have stationarity of GARCH(1,1) if $\alpha_1 + \beta_1 < 1$.

# Existence and Asymptotics

- Nelson (1990) showed that for GARCH(1,1), a unique stationary solution exists if and only if $E[\log(\beta_1 + \alpha_1 Z_0^2)] < 0$.
- Bougerol and Picard in 1992 found conditions for the more general case, which however are also a bit more complicated (and skipped here)
- Consistency and asymptotic normality can be obtained under fairly general conditions, see Berkes, Horváth, and Kokoszka (2003)
- For example, if a bit higher then second moments on $Z_0$ exist and $Z_0$ satisfies

$$\lim_{x \to 0} \frac{P(Z_0^2 \leq x)}{x^c} = 0$$

  for some $c > 0$, then the QMLE is consistent, i.e. $\hat{\theta}_n \to \theta$ a.s.
- With a bit more than 4th moments, also asymptotic normality follows, i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \Sigma)$$

# ARMA-GARCH

- For complex model we will mix the two approaches as follows.

## Definition

Consider $Z$ as $SWN(0,\sigma^2)$. A process $Y$ is called ARMA $(p_1,q_1)$-process with GARCH $(p_2,q_2)$-errors, if

$$
\begin{aligned}
Y_t &= \mu_t + \sigma_t Z_t \\
\mu_t &= \mu + \sum_{i=1}^{p_1} \phi_i (Y_{t-i} - \mu) + \sum_{j=1}^{q_1} \beta_j \sigma_{t-j} Z_{t-j} \\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p_2} \alpha_i (Y_{t-i} - \mu)^2 + \sum_{j=1}^{q_2} \beta_j \sigma_{t-j}^2 .
\end{aligned}
$$

Here is $\alpha_0 > 0, \alpha_i, \beta_i \geq 0$ and $\sum \alpha_i + \beta_j < 1$.

There are many more approaches, asymmetric approaches and so on. We first discuss the estimation of the GARCH models.

# Maximum-Likelihood

Define the **Likelihoodfunction** (conditional on $Y_0$) through

$$L(\boldsymbol{y};\theta) = \prod_{t=1}^{T} f_t(y_t|y_{t-1},\dots,y_0;\theta).$$

As shorthand for the probably more precise notation

$$L_{Y_1,\dots,Y_T|Y_0}(y_1,\dots,y_T|y_0;\theta) = \prod_{t=1}^{T} f_t(y_t|y_{t-1},\dots,y_0;\theta).$$

The ML estimator for $\theta$ is given by

$$\hat{\theta}_T := \arg\max_{\theta} L(\boldsymbol{Y};\theta) = \arg\max_{\theta} \log L(\boldsymbol{Y};\theta).$$

If the true conditional density $f$ is replaced by a normal density, one calls the approach a **quasi-maximum likelihood estimator**.

- Let us be more precise: we additionally match the first two moments.
- Hence, we replace the density of $Y_t|Y_{t-1},\ldots$ by a normal density with mean
$$E[Y_t|Y_{t-1},\ldots]$$
and variance
$$\mathrm{Var}(Y_t|Y_{t-1},\ldots).$$
- Let us denote this density by $\tilde{f}_t(y_t|y_{t-1},\ldots;\theta)$ and define the QMLE by
$$\tilde{L}(y;\theta) := \prod_{t=1}^{T} \tilde{f}_t(y_t|y_{t-1},\ldots,y_0;\theta)$$
through
$$\tilde{\theta}_T := \arg\max_{\theta} \tilde{L}(\boldsymbol{Y};\theta).$$
- Under suitable regularity assumptions on the true distribution of $Y$ this estimator is consistent and asymptotically normal.

# Higher dimensions

- How do we come to higher dimensions ?
- We can build factor models
- We can assume component-wise GARCH and assume that the errors have a joint normal distribution or a more general distribution.
- Several multivariate extensions have been proposed.
- We therefore shortly dive into the concept of multivariate random variables.

Benth, Fred Espen, Nils Detering, and Luca Galimberti (2023). „Neural networks in Fréchet spaces“. In: **Annals of Mathematics and Artificial Intelligence** 91.1, pp. 75–103.

Berkes, I., Lajos Horváth, and Piotr Kokoszka (2003). „GARCH processes: structure and estimation“. In: **Bernoulli** 9.2, pp. 201–227.

Brockwell, Peter J. and Richard A. Davis (1991). **Time Series: Theory and Methods**. 2nd. Springer.

Kratsios, Anastasis (2021). „The universal approximation property: characterization, construction, representation, and existence“. In: **Annals of Mathematics and Artificial Intelligence** 89.5, pp. 435–469.