Time series analysis
Notes for the course HS 2022

Nicolai Meinshausen
meinshausen@stat.math.ethz.ch

September 22, 2022

These notes are intended to just give a quick summary of what we discussed in the course. Some parts of this script are reused from an earlier script of Prof. Künsch. For examples and illustrations of the concepts and methods, you should look at the *R*-demonstrations which are on the course web page and the examples in the book Shumway & Stoffer. There are bound to be errors – I would appreciate if you point them out to me so I can get a corrected version to everybody.

Objectives of time series analysis

Goals of time series analysis can be classified in one of the following

i) Compact description of data as $X_t = T_t + S_t + Y_t$, where X_t is the observed time-series, T_t a trend, S_t a seasonal component and Y_y stationary noise. This can aid with interpretation for example by seasonal adjustment of unemployment figures.

ii) Hypothesis testing. We might for example want to test whether the trend component T_t vanishes for summer rainfall figures in Zurich over the last 10 years.

iii) Prediction. Examples are: predict unemployment data/ strength of El Nino / airlines passenger numbers or next word in a text. Might sometimes only be possible via simulation, as when trying to forecast hurricane intensity for the next decade at a specific location.

iv) Control/Causality/Reinforcement learning. One example is impact of monetary policy (interest rates) on inflation, where causal impact is quite different (possibly even different sign) to pure observational correlation. Or optimal filling and draining of lakes for energy storage.

Overview of the course

1. Characteristics of Time-Series (mostly Chapter 1 in book)
   (a) Stationarity
   (b) Auto-correlation function
   (c) Transformations
2. Time domain methods (mostly Chapter 3)
   (a) AR/MA/ARMA processes
   (b) Forecasting
   (c) Parameter estimation
   (d) ARIMA models
3. Spectral analysis (mostly Chapter 4.1-4.5)
   (a) Periodogram
   (b) Spectral density
   (c) Spectral estimation
4. State-space models (parts of Chapter 6)
   (a) Hidden Markov models (HMM)
   (b) Inference for HMM
   (c) Forward-backward algorithms
   (d) Kalman filter
   (e) Particle filters
5. Additional topics (possibly independent component analysis, recurrent neural networks, reinforcement learning)

1 Characteristics of Time Series

1.1 Stochastic Process

A stochastic process is a mathematical model for a time series.

Stochastic process = Collection of random variables $(X_t(\omega); t \in T)$. Alternative view: Stochastic process as a random function from T to $\mathbb{R}$.

A basic distinction is between continuous and discrete equispaced time T. Models in continuous time are prefered for irregular observation points. In this course we will restrict ourselves mostly to discrete equispaced time and, if not stated otherwise, use $T = \mathbb{Z}$.

In all interesting cases, there is dependence between the random variables at different times. Hence need to consider joint distributions, not only marginals. Gaussian stochastic processes have joint Gaussian distribution for any number of time points.

A stochastic process describes how different time series (when different ω's are drawn) could look like. In most cases, we observe only one realization $x_t(\omega)$ of the stochastic process (a single ω). Hence it is clear that we need additional assumptions, if we want to draw conclusions about the joint distributions (which involves many ω's) from a single realization. The most common such assumption is stationarity.

Stationarity means the same behavior of the observed time series in different time windows. Mathematically, it is formulated as invariance of (joint) distributions when time is shifted. Stationarity justifies taking of averages (mathematically, one needs ergodicity in addition).

Some examples:

a) White noise $X_t = W_t$, where $W_t \sim WN(0, \sigma^2)$, that is $W_t \sim F$ iid for some distribution F with mean 0 and variance σ^2. Special case is Gaussian White noise, where $F = \Phi$.

b) Harmonic oscillations plus (white) noise,

$$X_t = \sum_{k=1}^{K} \alpha_k \cos(\lambda_k t + \phi_k) + W_t,$$

where $W_t \sim WN(0, \sigma^2)$ as above a white noise process and K, α, λ, ϕ unknown parameters.

c) Moving averages. For example

$$X_t = \frac{1}{3}(W_t + W_{t-1} + W_{t-2}),$$

where $W_t \sim WN(0, \sigma^2)$.

d) Auto-regressive processes. For example

$$X_t = 0.9X_{t-1} + W_t,$$

plus initial conditions.

e) Random Walk (special case of an auto-regressive process)

$$X_t = X_{t-1} + W_t$$

or, with drift,

$$X_t = X_{t-1} + 0.2 + W_t.$$

f) Auto-regressive conditional heteroscedastic models

$$X_t = \sqrt{1 + 0.9X_{t-1}^2} \cdot W_t,$$

where again $W_t \sim WN(0, \sigma^2)$.

1.2 Measures of dependence

We want to summarize the distribution of a stochastic process (X_t) by the first two moments.

1. The **mean function** of process (X_t) is defined as

$$\mu_t := E(X_t) = \int_{-\infty}^{\infty} x f_t(x) dx,$$

where f_t is the density of X_t (if it exists).

2. The **auto-covariance function** is for all $s, t \in \mathbb{Z}$ defined as

$$\gamma(s,t) := \mathrm{Cov}(X_s, X_t) = E((X_s - \mu_s)(X_t - \mu_t)).$$

3. The **auto-correlation function** (ACF) is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

4. The **cross-covariance** for two time series (X_t) and (Y_t) is defined as

$$\gamma_{X,Y}(s,t) = \mathrm{Cov}(X_s, Y_t).$$

We can also use a subscript X for the first three definition but usually drop it for notational simplicity.

In vector notation, we can thus write for a collection of n time-points that the vector

$$(X_1, \dots, X_n)$$

has mean

$$(\mu_1, \dots, \mu_n)$$

and covariance matrix

$$\begin{pmatrix} \gamma(1,1) & \gamma(1,2) & \gamma(1,3) & \dots & \gamma(1,n) \\ \gamma(2,1) & \gamma(2,2) & \dots & & \\ \dots & & & & \\ \dots & & & & \\ \dots & & & & \\ \gamma(n,1) & \dots & & & \gamma(n,n) \end{pmatrix}.$$

The first two moments of *any* collection of n random variables can be described as above. For time-series, we would like to see translational invariance in time, which will be called stationarity.

1.3 Stationarity

Definition: A time-series (X_t) is **strictly stationary** iff the distribution of $(X_{t_1}, \ldots, X_{t_k})$ is identical to the distribution of $(X_{t_1+h}, \ldots, X_{t_k+h})$ for all $k \in \mathbb{N}^+$, time-points $t_1, \ldots, t_k \in \mathbb{Z}$ and shifts $h \in \mathbb{Z}$.

If (X_t) is strictly stationary, then

(i) $\exists \mu \in \mathbb{R}$ such that $\mu_t = \mu$ for all $t \in \mathbb{Z}$, that is the mean is constant.

(ii) $\gamma(s,t) = \gamma(s+h, t+h)$ for all $s, t, h \in \mathbb{Z}$, that is the covariance is invariant under time-shifts and we write γ without the second argument as

$$\gamma(k) := \gamma(k, 0) \qquad \forall k \in \mathbb{Z}.$$

We can also use just the last two properties about the first two moments (which are implied by strict stationarity) to define weak stationarity.

Definition: A time-series (X_t) is **weakly stationary** (or just stationary henceforth) iff (X_t) has finite variance and

(i) the mean function μ_t does not depend on $t \in \mathbb{Z}$

(ii) the autocovariance $\gamma(s,t)$ depends on s, t only through $|s-t|$ and we use again the notation

$$\gamma(k) := \gamma(k, 0) \qquad \forall k \in \mathbb{Z}.$$

The mean and covariance of a collection of n consecutive observations, for example $(X_1, \ldots, X_n)$, are now

$$(\mu, \ldots, \mu)$$

and covariance matrix

$$\left(\begin{array}{ccccc} \gamma(0) & \gamma(1) & \gamma(2) & \ldots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \ldots & \\ \gamma(2) & \ldots & & & \\ \ldots & & & & \\ \ldots & & & & \\ \ldots & & & & \\ \gamma(n-1) & \ldots & & & \gamma(0) \end{array}\right), \tag{1}$$

in contrast to the general case discussed above. The invariance under time shifts is now easily visible.

Note that a strictly stationary time-series is always also weakly stationary.

Moreover, if the distribution of $(X_{t_1}, \ldots, X_{t_k})$ is multivariate Gaussian for all $k \in \mathbb{N}$ and $t_1, \ldots, t_k \in \mathbb{Z}$, then weak stationarity implies strict stationarity (the proof is trivial since a multivariate Gaussian distribution is uniquely identified by its mean and covariance).

Examples. We look at the same examples as above and see whether they are (weakly) stationary.

a) White noise $X_t = W_t$ with $W_t \sim WN(0, \sigma^2)$, that is W_t is iid with distribution F with mean 0 and variance σ^2. The expected value is

$$E(X_t) = 0 \forall t \in \mathbb{Z}.$$

The variance is given by $E(X_t^2) = \sigma^2$ for all t and the auto-covariance is

$$\gamma(t+h, t) = \begin{cases} 0 & \text{if } h \neq 0 \\ \text{Var}(X_t) = \sigma^2 & \text{if } h = 0 \end{cases}$$

We can thus write $\gamma(t+h, t) = \gamma(h)$ for all t and white noise is (weakly) stationary. The ACF is given by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 0 & \text{if } h \neq 0 \\ 1 & \text{if } h = 0 \end{cases}$$

b) The harmonic oscillator is not stationary as the mean function μ_t is not constant in time.

c) Moving average. Say

$$X_t = W_t + \theta W_{t-1},$$

where $W_t \sim WN(0, \sigma^2)$. Then

$$E(X_t) = 0$$

and

$$\gamma(t+h, t) = \text{Cov}(X_t, X_{t+h}) = E(X_t X_{t+h}) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0 \\ \sigma^2\theta & \text{if } h \in \{-1, 1\} \\ 0 & \text{if } |h| > 1 \end{cases},$$

and we can again write $\gamma(h) = \gamma(t+h, t)$ for all t, and the process is weakly stationary. The ACF is then

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\theta}{1+\theta^2} & \text{if } h \in \{-1, 1\} \\ 0 & \text{if } |h| > 1 \end{cases}.$$

d) AR(1)-process. Will be discussed in the second Chapter in more detail.

e) Random Walk

$$X_t = \sum_{j=1}^{t} W_j \qquad \text{and } W_j \sim WN(0, \sigma^2).$$

First,

$$E(X_t) = 0 \quad \forall t \in \mathbb{Z},$$

and second

$$\gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{Cov}(\sum_{j=1}^{s} W_j, \sum_{j=1}^{t} W_j) = E(\sum_{j=1}^{\min\{s,t\}} W_j^2) = \min\{s,t\}\sigma^2,$$

and the Random Walk is thus not stationary.

f) ARCH model

$$X_t = \sqrt{1 + \phi X_{t-1}^2} W_t \qquad \text{and } W_t \sim WN(0, \sigma^2).$$

First,

$$E(X_t) = 0 \quad \forall t \in \mathbb{Z}.$$

Second, for $0 \leq \phi\sigma^2 < 1$, weakly stationary since

$$\gamma(t, t+h) = \mathrm{Cov}(X_t, X_{t+h}) = 0 \quad \text{if } h \neq 0,$$

and the variance $\gamma(t,t)$ is time-invariant with

$$E(X_t^2) = E(1 + \phi X_{t-1}^2)\sigma^2 = \frac{\sigma^2}{1 - \phi\sigma^2}.$$

The ACF is hence $\rho(h) = 1\{h = 0\}$, just as for a white noise process. Note, however, that while X_{t-1} and X_t are uncorrelated, they are not independent. For example, $|X_{t-1}|$ and $|X_t|$ or X_{t-1}^2 and X_t^2 will in general have a positive correlation in this model.

While the stationarity above refers to weak stationarity, all weakly stationary examples above are also strongly stationary.

1.4 Properties of the autocovariance for stationary time-series

In general, for a stationary time-series,

(i) The variance is given by $\gamma(0) = E((X_t - \mu)^2) \geq 0$.

(ii) $|\gamma(h)| \leq \gamma(0)$ for all $h \in \mathbb{Z}$. This follows by Cauchy-Schwarz as

$$\begin{aligned} |\gamma(h)| &= |E((X_t - \mu)(X_{t+h} - \mu))| \\ &\leq \left[E((X_t - \mu)^2)E((X_{t+h} - \mu)^2)\right]^{1/2} \\ &= [\gamma(0)^2]^{1/2} = \gamma(0). \end{aligned}$$

(iii) $\gamma(-h) = \gamma(h)$ (follows trivially).

(iv) γ is positive semi-definite, that is for all $a \in \mathbb{R}^n$ (and any choice of $n \in \mathbb{N}$),

$$\sum_{i,j=1}^{n} a_i \gamma(i-j) a_j \geq 0.$$

As a proof of (iv), consider the variance of $(X_1, \ldots, X_n)a = \sum_{i=1}^n a_i X_i$, where $a \in \mathbb{R}^n$ is a column-vector:

$$0 \leq \operatorname{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i,j=1}^n a_i a_j \operatorname{Cov}(X_i, X_j) = \sum_{i,j=1}^n a_i \gamma(i-j) a_j,$$

which completes the proof.

1.5 Estimating the auto-covariance

For observations $x_1, \ldots, x_n$ of a stationary time-series, estimate the mean, auto-covariance and auto-correlation as follows

(i) Sample mean $\hat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

(ii) Sample auto-covariance function is, for $-n \leq h \leq n$

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}),$$

and set to 0 otherwise.

(iii) Sample auto-correlation is given by

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Note that $\hat{\gamma}(h)$ is identical to the sample covariance of $(X_1, X_{1+h}), \ldots, (X_{n-h}, X_n)$, except that we normalize by n instead of $n - h$ to keep $\hat{\gamma}$ positive semi-definite (see below).

Properties of the sample ACF The four properties of the ACF are also true for the sample ACF:

(iii) $\hat{\gamma}(-h) = \hat{\gamma}(h)$ holds trivially.

(iv) $\hat{\gamma}$ is positive semi-definite (proof below).

(i)+(ii) $\hat{\gamma}(0) \geq 0$ and $|\hat{\gamma}(h)| \leq \hat{\gamma}(0)$ follows from property (iv).

Proof of (iv): We can write

$$\hat{\Gamma}_n = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \hat{\gamma}(2) & \ldots & \hat{\gamma}(n-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \hat{\gamma}(1) & \ldots & \\ \hat{\gamma}(2) & \ldots & & & \\ \ldots & & & & \\ \ldots & & & & \\ \ldots & & & & \\ \hat{\gamma}(n-1) & \ldots & & & \hat{\gamma}(0) \end{pmatrix} = \frac{1}{n} M M^T,$$

where the $n \times (2n-1)$-dimensional matrix M is given by

$$M := \begin{pmatrix} 0 & \dots & & 0 & \tilde{X}_1 & \tilde{X}_2 & \tilde{X}_3 & \dots & \tilde{X}_{n-1} & \tilde{X}_n \\ 0 & \dots & 0 & \tilde{X}_1 & \tilde{X}_2 & \tilde{X}_3 & & \dots & \tilde{X}_n & 0 \\ \dots & & & & & & & & & \\ 0 & \tilde{X}_1 & \tilde{X}_2 & \dots & & \tilde{X}_n & 0 & \dots & 0 & 0 \\ \tilde{X}_1 & \tilde{X}_2 & \dots & & \tilde{X}_n & 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

where $\tilde{X}_t := X_t - \hat{\mu}$. Hence, for any $a \in \mathbb{R}^n$,

$$a^T \hat{\Gamma}_n a = \frac{1}{n}(a^T M)(M^T a) = \frac{1}{n}\|M^T a\|_2^2 \geq 0,$$

which completes the proof.

1.6 Transforming to stationarity

Several steps/strategies, not always in the same order

1 Plot the time series: look for trends, seasonal components, step changes, outliers etc.

2 Transform data so that residuals are stationary.

  (a) Estimate and subtract trend T_t and seasonal components S_t

  (b) Differencing

  (c) Nonlinear transformations (log, $\sqrt{\cdot}$ etc.).

3 Fit a stationary model to residuals. This yields then an overall model for the data.

For 2(a), we can use non-parametric estimation (with large bandwidth) to get trend T_t and smoothing (with medium bandwidth) to get seasonal component. Seasonal component can also be estimated as empirical average of detrended data in, for example, each given month (if its yearly data).

For 2(b), define **lag-1 difference operator** via

$$(\nabla X)_t = (1-B)X_t = X_t - X_{t-1},$$

where B is the **backshift operator** defined via

$$(BX)_t = X_{t-1}.$$

(i) For a linear trend, that is if

$$X_t = \mu + \beta t + N_t,$$

with N_t the noise process, we have

$$(1-B)X_t = \beta + (1-B)N_t.$$

If differenced noise $(1-B)N_t$ is stationary, we can estimate slope β from data as the mean of the differenced time-series ∇X.

(ii) For a polynomial trends+noise, that is if

$$X_t = \sum_{j=1}^{k} \beta_j t^j + N_t,$$

difference k times to get

$$\nabla^k X_t = (1-B)^k X_t = k!\beta_k + (1-B)^k N_t.$$

If k-times differenced noise $(1-B)^k N_t$ is stationary, can estimate highest-order term as the mean of the k-times differenced time-series.

(iii) For a seasonal variation of length s, define lag-s differencing as

$$(1-B^s)X_t = X_t - B^s X_t = X_t - X_{t-s},$$

where B^s is the backshift operator applied s times. If

$$X_t = T_t + S_t + N_t,$$

and S_t has period s, then

$$(1-B^s)X_t = T_t - T_{t-s} + (1-B^s)N_t,$$

and the seasonal component has been removed and we can then proceed as in (i) or (ii), depending on the nature of the trend.

1.7 Cointegration

We say $(X_t)_{t\in\mathbb{Z}}$ is integrated of order d if the d times differenced time-series

$$(1-B)^d X_t \text{ is stationary}$$

but $(1-B)^{d'} X_t$ is not stationary for all $d' < d$. Let $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ be two time-series and integrated of order $d = 1$. The two time-series are said to be co-integrated if there exists a linear combination of both that is integrated of order 0 (that is the linear combination is stationary), that is there exists a $\beta \in \mathbb{R}^2$ such that

$$U_t = X_t\beta_1 + Y_t\beta_2$$

is stationary. The β is then clearly not unique and one can for example set $\beta_1 = 1$ wlog. Examples are stock prices of Apple and Google (where the difference is perhaps stationary) or economic data on money supply, income, prices and interest rates. If cointegration hold, we can model the difference as a stationary process. Cointegration means intuitively that while the processes marginally can drift they cannot drift far apart from each other .

1.8 Regression with correlated errors

We can try to estimate a linear trend in two different models

(i) A trend-stationary model, where

$$X_t = \mu + \beta t + N_t,$$

where N_t is stationary noise.

(ii) A difference-stationary model with

$$X_t = \mu + \beta t + N_t,$$

where ∇N_t is stationary noise.

Our goal is to estimate the trend β and give a confidence interval for this parameter.

Note: if model (i) is correct, the model (ii) is also correct, but we lose efficiency in the estimation if we proceed with differencing.

1.8.1 Trend-stationary models and pre-whitening

If model (i) is correct, we would like to use least-squares estimation to estimate the slope β and derive a confidence interval as in standard least-squares regression. We can write in vector notation

$$X = Z\theta + N, \tag{2}$$

where

$$\begin{aligned} X &= (X_1, \dots, X_n)^T, \\ N &= (N_1, \dots, N_n)^T, \\ Z &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \dots & \\ 1 & n \end{pmatrix}, \\ \theta &= \begin{pmatrix} \mu \\ \beta \end{pmatrix}. \end{aligned}$$

Note that we have chosen the special case $(t_1, \dots, t_n) = (1, \dots, n)$ in this example to simplify notation.

Naive least-squares estimation would then use an estimator

$$\hat{\theta} = \mathrm{argmin}_{\theta'} \|X - Z\theta'\|_2^2.$$

The least-squares estimator is motivated as maximum-likelihood estimator if the noise contributions are independent and have a Gaussian distribution. In a time-series context, the first assumption will be violated. Assume, for simplicity, though that the Gaussian assumption is correct and $N \sim \mathcal{N}(0, \Sigma)$ for a $n \times n$ full-rank matrix Σ of the general form of stationary time-series discussed before. Then, to get the maximum-likelihood estimate

$$\text{argmin}_{\theta'} (X - Z\theta')^T \Sigma^{-1} (X - Z\theta),$$

observe that (2) is equivalent to

$$AX = AZ\theta + AN, \tag{3}$$

for any full-rank matrix $A \in \mathbb{R}^{n \times n}$. Now choose A such that $AN \sim \mathcal{N}(0, 1_n)$, where 1_n is the identity matrix in n dimensions. If $\Sigma = C^T C$ is the Cholesky decomposition of Σ (and C invertible since we assumed that Σ has full rank), then such a pre-whitening matrix A is given for example by

$$A = C^{-T},$$

since then

$$\text{Var}(AN) = E(ANN^T A^T) = A \underbrace{E(NN^T)}_{=\Sigma} A^T = AC^T CA^T = 1_n.$$

Alternatively, if $\Sigma = \Phi \Lambda \Phi^T$ is the eigenvalue decomposition of Σ with orthogonal Φ (specifically, $\Phi^T \Phi = 1_n$) and a diagonal $\Lambda \in \mathbb{R}^{n \times n}$, then we can equivalently choose

$$A = \Lambda^{-1/2} \Phi^T$$

as a pre-whitening matrix, where $\Lambda_{ij}^{-1/2} = 0$ if $i \neq j$ and $\Lambda_{ii}^{-1/2} = 1/\sqrt{\Lambda_{ii}}$ for $i = 1, \ldots, n$ since then

$$\text{Var}(AN) = A\Sigma A^T = A\Phi\Lambda\Phi^T A^T = \Lambda^{-1/2} \underbrace{\Phi^T \Phi}_{=1_n} \Lambda \underbrace{\Phi^T \Phi}_{=1_n} \Lambda^{-1/2} = 1_n.$$

Once we have such a “pre-whitening” matrix A, then the maximum-likelihood estimator for θ is

$$\hat{\theta} = \text{argmin}_{\theta'} \|AX - AZ\theta'\|_2^2 = ((AZ)^T (AZ))^{-1} (AZ)^T (AX).$$

The point-estimate for β is thus the second entry in $\hat{\theta}$,

$$\hat{\beta} = \big(((AZ)^T (AZ))^{-1} (AZ)^T (AX)\big)_2.$$

To get a confidence interval for β, we need to know the variance of $\hat{\beta}$. A confidence interval is easiest to derive if $\hat{\beta}$ if unbiased (that is $E(\hat{\beta}) = \beta$), which is true for the estimator above) and it has a Gaussian distribution (which it has under the assumption made above that

$N \sim \mathcal{N}(0, \Sigma)$). Otherwise some modifications are necessary. The distribution of $\hat{\theta}$ under the Gaussian distribution for the noise is given by

$$\hat{\theta} \sim \mathcal{N}(\theta, ((AZ)^T(AZ))^{-1}).$$

A brief argument for this: note that the least squares estimator is given by

$$\hat{\theta} = ((AZ)^T(AZ))^{-1}(AZ)^T(AX).$$

The expected value is hence θ as $AX = AZ\theta + AN$. The variance of $\hat{\theta}$ is thus given by

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\Big(((AZ)^T(AZ))^{-1}(AZ)^T AN\Big),$$

where $AN \sim \mathcal{N}(0, 1_n)$. Thus

$$\mathrm{Var}(\hat{\theta}) = ((AZ)^T(AZ))^{-1}(AZ)^T \underbrace{E\big((AN)(AN)^T\big)}_{=1_n}(AZ)((AZ)^T(AZ))^{-T} = ((AZ)^T(AZ))^{-1}.$$

The distribution is furthermore joint normal (as its a linear combination of normal random variables) which completes the argument.

Remember that $\hat{\beta}$ is the second component in $\hat{\theta}$. We thus know that

$$\frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}(\hat{\beta})}} \sim \mathcal{N}(0, 1),$$

where $\mathrm{Var}(\hat{\beta}) = (((AZ)^T(AZ))^{-1})_{2,2}$ in our example. Thus

$$P\Big(-q \leq \frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}(\hat{\beta})}} \leq q\Big) \geq 1 - \alpha,$$

where $q = \Phi^{-1}(1 - \alpha/2)$ the $1 - \alpha/2$ quantile of a standard normal distribution (and for example $q \approx 1.96 \approx 2$ for $\alpha = 0.05$). A $(1 - \alpha)$-confidence interval for β is then given by

$$\Big[\hat{\beta} - q\sqrt{\mathrm{Var}(\hat{\beta})}, \hat{\beta} + q\sqrt{\mathrm{Var}(\hat{\beta})}\Big].$$

If unsure about any of this this, please consult a textbook on regression or introductory statistics. If we estimate Σ from the data, we will have to modify the confidence intervals accordingly (they tend to get wider) but this is beyond the scope here.

1.8.2 Difference-stationary models

For model (ii), let $Z_t = \nabla X_t$ be the differenced time-series. The slope β in the model can be estimated as the mean of the differenced time-series, that is

$$\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} Z_t.$$

To get a confidence interval, we need to know again the variance of $\hat{\beta}$. If we do know the variance (and assume a Gaussian distribution of $\hat{\beta}$ for simplicity), then a confidence interval is given again by

$$\left[\hat{\beta} - q\sqrt{\mathrm{Var}(\hat{\beta})}, \hat{\beta} + q\sqrt{\mathrm{Var}(\hat{\beta})}\right],$$

where the quantile q of the standard normal distribution is defined just as above. Now, if Z_t were independent (and Z_t stationary), then

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\frac{1}{n}\sum_{t=1}^{n} Z_t) = \mathrm{Var}(Z_1)/n.$$

More generally, if we allow correlations and Z_t has autocovariance γ, then

(a)

$$\mathrm{Var}(\frac{1}{n}\sum_{t=1}^{n} Z_t) = \frac{1}{n}\sum_{k=-n+1}^{n-1}(1 - \frac{|k|}{n})\gamma(k)$$

(b) If $\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$, then, as $n \to \infty$,

$$\begin{aligned} n\mathrm{Var}(\frac{1}{n}\sum_{t=1}^{n} Z_t) \to \sum_{k=-\infty}^{\infty}\gamma(k) &= \mathrm{Var}(Z_1)(\sum_{k=-\infty}^{\infty}\rho(k)) \\ &= \mathrm{Var}(Z_1)(1 + \sum_{k=-\infty, k\neq 0}^{\infty}\rho(k)) \quad = \mathrm{Var}(Z_1)(1 + 2\sum_{k=1}^{\infty}\rho(k)) \end{aligned}$$

and the asymptotic variance is thus inflated (or deflated) compared to the independence case by the factor $(1 + 2\sum_{k=1}^{\infty}\rho(k))$, which is typically larger than 1 but can also be less than 1 when negative auto-correlations appear.

Proof of (a):

$$\begin{aligned}
\mathrm{Var}(\sum_{t=1}^{n} Z_t) &= \sum_{t=1}^{n}\sum_{s=1}^{n} \mathrm{Cov}(Z_t, Z_s) \\
&= \sum_{t=1}^{n}\sum_{s=1}^{n} \gamma(t-s) \\
&= \sum_{k=-n+1}^{n-1} \gamma(k) \cdot \underbrace{(\text{number of pairs } (t,s) \text{ with } t-s=k)}_{=n-|k|} \\
&= \sum_{k=-n+1}^{n-1} \gamma(k) \cdot (n-|k|)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\frac{1}{n}\sum_{t=1}^{n} Z_t) &= \frac{1}{n^2}\mathrm{Var}(\sum_{t=1}^{n} Z_t) \\
&= \frac{1}{n}\sum_{k=-n+1}^{n-1} (1-\frac{|k|}{n})\gamma(k).
\end{aligned}$$

Proof of (b): Using (a),

$$\sum_{k=-n+1}^{n-1} (1-\frac{|k|}{n})\gamma(k) = \sum_{k=-\infty}^{\infty} \underbrace{\max\{0, 1-\frac{|k|}{n}\}\gamma(k)}_{\to\gamma(k) \text{ as } n\to\infty},$$

and the claim follows by dominated convergence.

2 Time-Domain models

2.1 Causal and stationary autoregressions

A stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called a Markovian autoregressive process of order p if

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + W_t$$

where W_t is independent of all X_s, $s < t$ (this is the "causal" condition; see discussion next Section) and $\phi_p \neq 0$. If the process is stationary, we call this a causal, stationary AR(p) process.

The variable W_t is called the innovation at time t. In operator notation,

$$\Phi(B)X_t = W_t,$$

where B is again the backshift operator and

$$\Phi(B) := 1 - \phi_1 B^1 - \ldots - \phi_p B^p.$$

For a Markovian autoregression, $\phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + E(W_t)$ is the best prediction of X_t from the past. Furthermore, the innovations at different times are independent: For $t > s$ W_t is independent of $X_s - \phi_1 X_{s-1} - \ldots - \phi_p X_{s-p} = W_s$.

When is a Markovian autoregression stationary? First it is clear that under stationarity, the innovations are not only independent, but also identically distributed. For an AR(1)-process, we obtain by iteration

$$X_t = \sum_{j=0}^{t-1} \phi^j W_{t-j} + \phi^t X_0.$$

Hence if second moments exist and if (X_t) is stationary, then

$$\gamma(0) = \mathrm{Var}(W) \sum_{j=0}^{t-1} \phi^{2j} + \phi^{2t} \gamma(0)$$

since by assumption all terms on the right are independent. Clearly this implies that $|\phi| < 1$ as the equation cannot be fulfilled if $|\phi| > 1$ on the hand (the right hand side will always be larger than the left hand side for example) and it has a solution for $|\phi| < 1$:

$$\gamma(0) = \mathrm{Var}(W) \frac{\sum_{j=0}^{t-1} \phi^{2j}}{1 - \phi^{2t}} = \mathrm{Var}(W) \frac{(1 - \phi^{2t})/(1 - \phi^2)}{1 - \phi^{2t}} = \frac{\mathrm{Var}(W)}{1 - \phi^2}.$$

The general case is covered by the next theorem.

Theorem 1. *A stationary Markovian autoregression with finite second moments exists iff all zeroes of the polynomial*

$$\Phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$$

are outside the unit disc $\{z; |z| \leq 1\}$. In that case the process has the representation

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

where the coefficients ψ_j are the solution of the recursion

$$\psi_j = \phi_1 \psi_{j-1} + \ldots + \phi_p \psi_{j-p} \quad (j \geq 1) \tag{4}$$

with initial conditions $\psi_0 = 1$, $\psi_{-1} = \ldots \psi_{1-p} = 0$ and thus converge to zero exponentially fast. Moreover, if we define for $t > 0$

$$X_t^* = \phi_1 X_{t-1}^* + \ldots \phi_p X_{t-p}^* + W_t$$

with arbitrary initial conditions $X_0^*, X_{-1}^*, \ldots, X_{1-p}^*$, $X_t - X_t^* \to 0$ almost surely and in L_1.

Proof. We write the autoregression of order p as a vector autoregression of order 1: If we set $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})^T$, $\eta_t = (W_t, 0, \ldots, 0)^T$ and

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 & \ldots & \phi_{p-1} & \phi_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix}$$

(with I_{p-1} the identity matrix in dimension $p-1$), then

$$Z_t = \mathbf{\Phi} Z_{t-1} + \eta_t.$$

Iterating this autoregression, we obtain

$$Z_t = \sum_{j=0}^{t-1} \mathbf{\Phi}^j \eta_{t-j} + \mathbf{\Phi}^t Z_0.$$

If Z_t is stationary with finite variance, then $\mathbf{\Phi}^t$ must converge to zero, and this is known to be equivalent to the condition that all eigenvalues of $\mathbf{\Phi}$ are smaller than one in absolute value[1]. The characteristic polynomial of $\mathbf{\Phi}$ is however nothing else than the polynomial $z^p\Phi(1/z) = z^p - \phi_1 z^{p-1} - \ldots - \phi_p$. In more detail: for an eigenvalue λ we need that $\mathbf{\Phi} v = \lambda v$, where $v \in \mathbb{C}^p$ is the eigenvector. This last equation is the same as

$$\begin{pmatrix} \phi_1 & \ldots & \phi_{p-1} & \phi_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_p \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_p \end{pmatrix},$$

which is equivalent (using simple algebra) to the condition that $\lambda^p - \phi_1 \lambda^{p-1} - \ldots - \phi_p = 0$ and the roots of $\Phi(z)$ are thus outside the unit circle iff all eigenvalues of $\mathbf{\Phi}$ are inside the unit circle. The above argument shows that if there are eigenvalues of $\mathbf{\Phi}$ with absolute value larger or equal to 1 (or, equivalently, if there are roots of the polynomial $\Phi(z)$ that lie inside the unit circle), then the process can not be stationary.

If on the other hand, we assume that all eigenvalues of $\mathbf{\Phi}$ are indeed smaller than 1, we can take the limit in the above recursion and obtain

$$Z_t = \sum_{j=0}^{\infty} \mathbf{\Phi}^j \eta_{t-j}. \tag{5}$$

[1] since –if any eigenvalue of $\mathbf{\Phi}$ is larger than 1– the term $E(\|\mathbf{\Phi}^t Z_0\|_2^2)$ would increase exponentially with increasing t and $E(\|Z_t\|_2^2)$ can hence not be constant, as required for a stationary process. To see this, note that $\mathbf{\Phi}^t$ is identical to $\mathbf{U}\Lambda^t\mathbf{U}^{-1}$ if $\mathbf{U}\Lambda\mathbf{U}^{-1}$ with diagonal Λ is the eigenvalue decomposition of $\mathbf{\Phi}$

This process has a finite variance now and is clearly stationary (since the η_t contributions are iid). Hence both parts of the first claim (stationary if and only if roots of Φ are outside the unit circle).

We still need to show the recursion (4) for the coefficients ψ_j, $j \geq 1$. From the definition of Z_t iand η_t and (5), it follows that

$$\psi_j = (\mathbf{\Phi}^j)_{1,1},$$

that is the coefficient Ψ_j is equal to the $(1,1)$ components of the matrix $\mathbf{\Phi}^j$ for all $j \geq 0$. For $\mathbf{\Phi}^j$ we have the recursion

$$\mathbf{\Phi}^j = \mathbf{\Phi}\mathbf{\Phi}^{j-1}$$

Hence the first column of $\mathbf{\Phi}^j$, say $c^{(j)}$, satisfies the recursion

$$c^{(j)} = ((\phi_1, \ldots, \phi_p)^T c^{(j-1)}, c_1^{(j-1)}, \ldots, c_{p-1}^{(j-1)}).$$

Note that $\psi_j = (\mathbf{\Phi}^j)_{1,1} = c_1^{(j)}$. The first components of the recursion reads

$$\psi_j = c_1^{(j)} = \phi_1 c_1^{(j-1)} + \phi_2 c_2^{(j-1)} + \ldots + \phi_p c_p^{(j-1)}.$$

But we can see from the recursion that $c_2^{(j-1)} = c_1^{(j-2)} = \psi_{j-2}$ and $c_k^{(j-1)} = c_1^{(j-k)} = \psi_{j-k}$ which shows the claimed recursion for ψ, namely that

$$\psi_j = \phi_1\psi_{j-1} + \phi_2\psi_{j-2} + \ldots + \phi_p\psi_{j-p}.$$

□

2.2 Some more discussion of term "causality"

The theorem above shows that a stationary Markovian autoregression, where W_t is independent of past values X_s, $s < t$, can be written as a linear combination of past innovations. Without the condition that W_t is independent of past values X_s, $s < t$, the theorem is false. To see why, take any $|\phi| > 1$ and set

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} W_{t+j}.$$

Clearly this is stationary if the W_t are i.i.d. Moreover,

$$\phi X_{t-1} = -W_t + X_t,$$

so the recursion is satisfied. However, W_t contributes to the sum defining X_s for $s < t$, and thus the two variables are dependent.

As a side remark: We called the condition that W_t is independent of all preceding values X_s, $s < t$ the "causal" condition" as it implies conversely that X_t is not a function of future values of W_s, $s > t$. The book Shumway & Stoffer defines an AR(p) process without the last "causal" condition that W_t is independent of all previous X_s, $s < t$. Which condition is called causal and which condition leads to stationarity is inconsistent in the literature, unfortunately. To make this a bit more transparent, we have the corollary that if the process $(X_t)_{t\in\mathbb{Z}}$ is of the Markovian autoregressive form

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + W_t$$

and stationary with finite variance, then the following three conditions are equivalent:

(i) W_t is independent of all X_s, $s < t$ [first possible definition of a "causal" process].

(ii) X_t can be written as a one-sided linear process:

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$ [second possible definition of a "causal" process].

(iii) All zeroes of the polynomial $\Phi(z)$ are outside the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$ [third possible definition].

The first and second make arguably the most sense for defining "causal" but it is a matter of taste. Note that if we assume (i), as we have done here, then (iii) is a condition for stationarity; in other words, (i) can be true and (iii) violated, but the process is then not stationary any longer. Conversely, the process given above is an example where (iii) is violated and the process stationary, but then condition (i) is violated (the book discusses more, see "every explosion has a cause" and related parts).

2.3 Skeletons of AR(p) processes and impulse-response functions.

Suppose we fix initial conditions $X_0, \ldots, X_{p-1}$ and compute the solutions for the homogeneous equation

$$\Phi(B)X_t = 0 \qquad \forall t \geq p. \tag{6}$$

The solution X_t is then an "impulse-response" (the response to the initial impulse of the initial condition) and the solutions to the general case are superpositions of such impulse-response functions since we have a linear system. The solutions to (6) are linear combinations of the solutions to the homogenous equations

$$u_t = \sum_{j=1}^{p} \phi_j u_{t-j} \quad \text{or, equivalently,} \quad \Phi(B)u_t = 0 \qquad \forall t \in \mathbb{Z}. \tag{7}$$

Theorem 2. *The set of sequences (u_t) that satisfy the above difference equation (7) is a vector space of dimension p. A basis is given by the sequences of the form*

$$u_t = t^j z_0^{-t}$$

where z_0 is a root of the polynomial

$$\Phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p$$

with multiplicity m and $0 \leq j < m$.

Proof. It is clear that a linear combination of two solutions is again a solution. Moreover, if p consecutive values $u_{k+1}, \ldots, u_{k+p}$ of a solution (u_t) are given, then the solution is unique: Values u_t for $t > k + p$ follow by forward iteration, those for $t \leq k$ follow by backward iteration

$$u_{t-p} = \frac{u_t - \phi_1 u_{t-1} - \ldots - \phi_{p-1} u_{t-p+1}}{\phi_p}.$$

Therefore the dimension of the vector space is p.

Next, we show that the above sequences are indeed solutions. First we take $j = 0$:

$$z_0^{-t} - \phi_1 z_0^{-(t-1)} - \ldots \phi_p z_0^{-(t-p)} = z_0^{-t} \Phi(z_0) \equiv 0.$$

Similarly, for $j = 1$ we have

$$t z_0^{-t} - \phi_1 (t-1) z_0^{-(t-1)} - \ldots \phi_p (t-p) z_0^{-(t-p)} = z_0^{-t} t \Phi(z_0) - z_0^{-(t-1)} \Phi'(z_0) \equiv 0.$$

The general case follows because

$$\Phi^{(j)}(z_0) = -z_0^j \sum_{k=j}^{p} \phi_k k(k-1) \cdots (k-j+1) z_0^k = 0 \quad (j < m)$$

implies that also

$$\sum_{k=1}^{p} \phi_k k^j z_0^k = 0 \quad (j < m).$$

The proof will be completed if we can show that the above solutions are linearly independent since the number of zeroes of a polynomial of degree p counted with their multiplicity is equal to p. For a proof of the linear independence, see Brockwell and Davies, Theorem 3.6.2. □

For real-valued processes, the solutions are thus given by the basis vectors

(i) single real root $z_0 \in \mathbb{R}$ yields basis vector

$$u_t = z_0^{-t}$$

(ii) complex roots $z_0 = r \exp(i\mu), \bar{z}_0 = r \exp(-i\mu)$ yield the two vectors

$$u_t = \cos(\mu t) r^{-t}$$
$$u_t = \sin(\mu t) r^{-t}$$

(iii) real root $z_0 \in \mathbb{R}$ with multiplicity k yields basis vectors

$$u_t = z_o^{-t} t^j \qquad \text{for } j = 0, \ldots, k-1.$$

(iv) Complex roots $z_0 = r\exp(i\mu) \in \mathbb{C}$ with multiplicity k yields

$$\begin{aligned} u_t &= \cos(\mu t) r^{-t} t^j \\ u_t &= \sin(\mu t) r^{-t} t^j \qquad \text{for } j = 0, \ldots, k-1. \end{aligned}$$

The coefficients in the basis are chosen to satisfy initial conditions. The solutions converge to 0 as $t \to \infty$ if and only if all roots are outside the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$ (and if initial conditions are not chosen precisely to cancel out the corresponding exponential growth terms if they exist).

Examples:

1) **AR(1).** $u_t = \phi_1 u_{t-1}$ with $\phi_1 \neq 0$ has root $z_0 = 1/\phi_1$. Exponential growth if $|\phi_1| > 1$. Critical behaviour (random-walk type) if $\phi_1 = 1$ and convergent to 0 from all initial conditions if $|\phi_1| < 1$.

2) **AR(2).** The roots of $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2$ are

$$z_{1,2} = -\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2\phi_2}$$

One can verify that z_1 and z_2 are both outside the unit circle iff

$$-1 < \phi_2 < 1, \quad \phi_2 < 1 - |\phi_1|$$

Hence the set of parameters which correspond to a stationary Markovian autoregression is a triangle. The roots are complex (and the solutions then show oscillatory behaviour) for $\phi_2 < -\frac{1}{4}\phi_1^2$.

The above discussion shows (again) that the roots of $\Phi(z)$ are critical for stationarity of the process (here in the noiseless case) and determine whether there is oscillatory behaviour.

2.4 Invertible moving averages

A linear moving average of order q

$$X_t = W_t + \theta_1 W_{t-1} + \ldots \theta_q W_{t-q}$$

with W_t i.i.d. is always stationary. In operator notation, we write:

$$X_t = \Theta(B) W_t,$$

where

$$\Theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q.$$

W_t is always independent of X_s for $s < t$ for such a process. However we cannot call W_t the innovation of the process unless the other terms $\theta_1 W_{t-1} + \ldots \theta_q W_{t-q}$ on the right hand side can be expressed with the values X_s for $s < t$. If they can be expressed in this way, W_t is independent of all X_s with $s < t$ and a function of only X_s, $s \leq t$.

We therefore call a moving average invertible if there are coefficients (π_j) with $\sum |\pi_j| < \infty$ such that

$$W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

(An autoregression is always invertible, just set $\pi_0 = 1$, $\pi_j = -\phi_j$ for $1 \leq j \leq p$ and $\pi_j = 0$ for $j > p$). Invertibility helps with the uniqueness of the representation, see for example 3.4 and 3.5 in the book, where it is shown that the two processes have the identical distribution under Gaussian noise

$$\begin{aligned} X_t &= W_t + \frac{1}{5} W_{t-1} \qquad \text{with } W_t \text{ iid } \mathcal{N}(0, 25) \\ X_t &= W_t + 5 W_{t-1} \qquad \text{with } W_t \text{ iid } \mathcal{N}(0, 1) \end{aligned}$$

We prefer the first representation as it is invertible. Specifically, using the same recursion idea as before,

$$\begin{aligned} W_t &= X_t - \frac{1}{5} W_{t-1} \\ &= X_t - \frac{1}{5}(X_{t-1} - \frac{1}{5} W_{t-2}) \\ &= X_t - \frac{1}{5}(X_{t-1} - \frac{1}{5}(X_{t-2} - \frac{1}{5} W_{t-3}) \\ &= \dots \\ &= \sum_{j=0}^{\infty} \pi_j X_{t-j} \qquad \text{with } \pi_j = (-\frac{1}{5})^j \end{aligned}$$

Theorem 3. *A moving average is invertible iff all zeroes of the polynomial*

$$\Theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$$

are outside the unit circle $\{z; |z| \leq 1\}$. In that case the coefficients π_j are the solution of the recursion

$$\pi_j = -\theta_1 \pi_{j-1} - \dots - \theta_q \pi_{j-q}$$

with initial conditions $\pi_0 = 1$, $\pi_{-1} = \dots \pi_{1-q} = 0$ and thus converge to zero exponentially fast.

Proof. We use the same proof idea as in the theorem about causality, reversing the roles of W_t and X_t. For $q = 1$, we simply iterate the equation $X_t = W_t + \theta W_{t-1}$ (which is $W_t = X_t - \theta W_{t-1}$). For $q > 1$, we write the process as a vector moving average of order 1. □

2.5 ARMA-Processes

An autoregressive moving average process of order (p, q), called ARMA(p, q), combines the properties of the two previous models. The recursion is

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j W_{t-j} + W_t.$$

For a reasonable model W_t should again be independent of X_s for $s < t$ and W_t should depend only on past values $X_s, s \leq t$, i.e. the model should be invertible, that is there exists summable coefficients π_j, $j = 0, 1 \ldots$, such that

$$W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

Again it then follows that the variables W_t are independent for different times t, and if (X_t) is stationary, the W_t are even i.i.d.

We also want the condition (that is sometimes refereed to as a condition for stationarity and sometimes as a causal condition, see Section 2.2): There are summable coefficients ψ_j such that

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}.$$

If one wants to generalize the arguments for the autoregressive case one sees, however, that a problem occurs: For instance for any ϕ

$$X_t = \phi X_{t-1} + W_t - \phi W_{t-1}$$

has the stationary solution $X_t = W_t$ which is also invertible and W_t is independent of X_s for $s < t$. The reason for this problem is that Φ and Θ have common zeroes.

If we assume that Φ and Θ have no common zeroes, then the conditions that all zeroes of Φ and Θ are outside of the unit circle are again necessary and sufficient for the existence of a stationary ARMA model which is invertible and causal.

The recursion of the ARMA process can then be written as

$$\Phi(B) X_t = \Theta(B) W_t.$$

Formally, we can thus write

$$X_t = \Phi(B)^{-1} \Theta(B) W_t, \quad W_t = \Theta(B)^{-1} \Phi(B) X_t.$$

If $\Phi(z)$ has no zeroes in $\{z; |z| \leq 1\}$, the Taylor series

$$\frac{\Theta(z)}{\Phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j$$

converges on $\{z; |z| \leq 1\}$ and thus we can define

$$\Phi(B)^{-1}\Theta(B) = \sum_{j=1}^{\infty} \psi_j B^j.$$

From the equality

$$\Theta(z) = \Phi(z) \cdot \sum_{j=0}^{\infty} \psi_j z^j,$$

we obtain by comparing the coefficient of z^j on both sides the equations

$$\psi_j - \sum_{k=1}^{\min(p,j)} \phi_k \psi_{j-k} = \begin{cases} \theta_j & 0 \leq j \leq q \\ 0 & j > q \end{cases} \tag{8}$$

(we set $\theta_0 = 1$). This is a generalisation of the corresponding recursion (4) for AR-processes (where $\theta_0 = 1$ and all $\theta_j = 0$ for $j > 0$) and the most convenient way to compute ψ_j numerically. A similar argument applies for the coefficients in the invertibility representation $W_t = \Theta(B)^{-1}\Phi(B)X_t$. In particular, ψ_j again satisfies a difference equation except for an initial part of length q. This generalizes the previous recursive way of computing the MA-style ψ-coefficients for an AR-model. Note that the coefficients ψ decay exponentially fast again for a causal stationary process and we can in practice stop the recursion at some point when the coefficients have become very small.

2.6 Autocorrelation

Let (X_t) be a stationary, causal and invertible ARMA(p,q) process.

We compute the autocovariance $\gamma(h)$ of the process. There are two options. The first one uses the causal MA(∞)-representation

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

to get

$$\begin{aligned} \gamma(h) = \text{Cov}(X_h, X_0) &= \text{Cov}(\sum_{j=1}^{\infty} \psi_j W_{t+h-j}, \sum_{j'=0}^{\infty} \psi_{j'} W_{t-j'}) \\ &= \sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \psi_j \psi_{j'} \underbrace{E(W_{t+h-j} W_{t-j'})}_{=\text{Var}(W)1\{h-j=-j'\}} \\ &= \sum_{j=0}^{\infty} \psi_j \psi_{j+h}. \end{aligned}$$

The second possibility is as follows. Because the covariance is linear in both arguments, we obtain

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(X_h, X_0) = \sum_{j=1}^{p} \phi_j \mathrm{Cov}(X_{h-j}, X_0) + \sum_{k=0}^{q} \theta_k \mathrm{Cov}(W_{h-k}, X_0) \\ &= \sum_{j=1}^{p} \phi_j \gamma(h-j) + \sum_{k=h}^{q} \theta_k \mathrm{Cov}(W_{h-k}, X_0).\end{aligned}$$

In the last equality, we have used the property that W_t is independent and thus uncorrelated with X_0 for $t > 0$.

For $h > q$, the second sum on the right runs over an empty set and is thus zero. Therefore, we have shown that for $h \geq \max(p, q+1)$ the autocovariance function satisfies the difference equation

$$\gamma(h) = \sum_{j=1}^{p} \phi_j \gamma(h-j).$$

In particular, it decays to zero exponentially fast. Moreover, the properties are closely linked to properties of the zeroes of the polynomial Φ. If Φ has two zeroes $r \exp(\pm i\nu)$ with r close to one, then the covariance is (approximately) a damped harmonic with period $2\pi/\nu$.

In order to compute the values $\gamma(h)$ for $h < \max(p, q+1)$, we need $\mathrm{Cov}(W_s, X_0)$ for $s \leq 0$. These covariances can be computed from the causal representation:

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} \Rightarrow \mathrm{Cov}(W_s, X_0) = \mathrm{Var}(W)\psi_{-s} \quad (s \leq 0).$$

Example: Autoregressions. For autoregressions, we only need $\mathrm{Cov}(W_0, X_0)$ which is equal to $\mathrm{Var}(W)$. The autocovariances $\gamma(h)$ for $0 \leq h \leq p$ are then obtained from the equations

$$\begin{aligned}\gamma(0) - \sum_{j=1}^{p} \phi_j \gamma(j) &= \mathrm{Var}(W) \\ \gamma(h) - \sum_{j=1}^{p} \phi_j \gamma(h-j) &= 0 \quad (1 \leq h \leq p).\end{aligned}$$

These equations are called Yule-Walker equations and can be written in matrix form as

$$\Gamma_p \phi = \gamma_p \tag{9}$$

$$\mathrm{Var}(W) = \sigma^2 = \gamma(0) - \phi^t \gamma_p, \tag{10}$$

where

$$\Gamma_p = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \\ \gamma(2) & \dots & & & \\ \dots & & & & \\ \dots & & & & \\ \dots & & & & \\ \gamma(p-1) & \dots & & & \gamma(0) \end{pmatrix}, \ \phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \dots \\ \\ \phi_p \end{pmatrix}, \ \gamma_p = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(3) \\ \dots \\ \\ \gamma(p) \end{pmatrix}$$

Replacing Γ_p and γ_p by their empirical counterparts $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ will yield a possible estimator (the Yule-Walker estimator) for the coefficients ϕ in an AR(p)-model and the variance $\mathrm{Var}(W)$ of the innovations (remember that $\hat{\Gamma}_p$ is positive semi-definite).

Example: ARMA(1,1). From

$$X_t = \phi X_{t-1} + W_t + \theta W_{t-1}$$

we obtain

$$X_t = \phi^2 X_{t-2} + W_t + (\phi + \theta) W_{t-1} + \phi\theta W_{t-2}$$

and therefore $\psi_0 = 1$, $\psi_1 = (\phi + \theta)$. Hence the autocovariances $\gamma(0)$ and $\gamma(1)$ can be found by solving the equations

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \mathrm{Var}(W)(1 + \theta(\phi + \theta)) \\ \gamma(1) &= \phi\gamma(0) + \mathrm{Var}(W)\theta. \end{aligned}$$

This gives the variance

$$\gamma(0) = \mathrm{Var}(W) \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}$$

and the autocorrelations

$$\rho(1) = \phi + \frac{\theta(1-\phi^2)}{1 + 2\theta\phi + \theta^2}, \quad \rho(h) = \phi^{h-1}\rho(1) \ (h > 1).$$

2.7 Linear prediction and partial autocorrelations

The Yule-Walker equations from above also appear in the context of linear forecasting.

The best linear prediction of X_t based on $(X_r, X_{r+1}, \dots X_s)$ for $r \le s < t$ or $t < r \le s$ is the linear combination

$$\widehat{X}_t^{r:s} = \alpha + \sum_{k=0}^{s-r} \beta(k) X_{s-k}$$

which minimizes the mean square error of prediction:

$$\widehat{X}_t^{r:s} = \ \mathrm{argmin}_x \ E((X_t - x)^2 | X_r, \dots, X_s).$$

Schematically, we try to do the following:

$$\ldots, X_0, \ldots, X_{r-1}, \quad \underbrace{X_r, X_{r+1}, \ldots, X_s}_{\text{use these to predict}} \;, \ldots, \qquad \underbrace{X_t}_{\text{this target value}} \qquad , \ldots .$$

$\widehat{X}_t^{r:s}$ is determined (using the projection theorem) by a system of linear equations which involve the mean and autocovariance of X_t only:

$$\begin{aligned} E(X_t - \widehat{X}_t^{r:s}) &= 0 \\ E((X_t - \widehat{X}_t^{r:s})X_u) &= 0 \quad (r \leq u \leq s) \end{aligned}$$

Assume for simplicity of notation that $\mu = 0$ and a stationary process. Due to stationarity the problem is translational invariant (if we shift r, s, t by an amount $h \in \mathbb{Z}$, nothing changes). The equations above for the optimal linear coefficients are then equal to (the first one is automatically fulfilled if the mean is a constant 0 and we use the order $u = s$ as first equation, $u = s - 1$ as second until $u = r$ as last equation):

$$\Gamma_t^{r:s} \beta_t^{r:s} = \gamma_t^{r:s}, \tag{11}$$

where

$$\beta_t^{r:s} = \text{argmin}_\beta \; E((X_t - \sum_{k=0}^{s-r} \beta(k) X_{s-k})^2 | X_r, \ldots, X_s).$$

where (analogous to the Yule-Walker equations in (9) for the coefficients ϕ, which are for an AR(p) process also the best linear predictors β), which we recover if $s = t-1$ and $r = t-p$,

$$\Gamma_t^{r:s} = \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \ldots & \gamma(s-r) \\ \gamma(1) & \gamma(0) & \gamma(1) & \ldots & \\ \gamma(2) & \ldots & & & \\ \ldots & & & & \\ \ldots & & & & \\ \ldots & & & & \\ \gamma(s-r) & \ldots & & & \gamma(0) \end{pmatrix},$$

$$\beta_t^{r:s} = \begin{pmatrix} \beta_t^{r:s}(0) \\ \beta_t^{r:s}(1) \\ \beta_t^{r:s}(2) \\ \ldots \\ \\ \\ \beta_t^{r:s}(s-r) \end{pmatrix}, \; \gamma_t^{r:s} = \begin{pmatrix} \gamma(t-s) \\ \gamma(t-s+1) \\ \gamma(t-s+2) \\ \ldots \\ \\ \\ \gamma(t-r) \end{pmatrix}$$

We can estimate β by using $\hat{\Gamma}_t^{r:s}$ and $\hat{\gamma}_t^{r:s}$ instead of $\Gamma_t^{r:s}$ and $\gamma_t^{r:s}$ inverting the matrix $\hat{\Gamma}_t^{r:s}$ (which is positive semi-definite; need to regularize if one or more singular values are exactly zero). Note that for stationary time-series,

$$\beta_t^{r:s} = \beta_{t+h}^{r+h:s+h}$$

for any $h \in \mathbb{Z}$, that is the optimal linear prediction coefficients are invariant under a time-shift (as the same is true for the first two moments of the time-series itself).

Prediction intervals: We can also compute the variance of the residual

$$\begin{aligned}
(\sigma_t^{r:s})^2 = E((X_t - \widehat{X}_t^{r:s})^2) &= \mathrm{Var}(X_t - \sum_{k=0}^{s-r} \beta(k) X_{s-k}) \\
&= \mathrm{Var}(X_t) - 2\sum_{k=0}^{s-r} \beta(k)\mathrm{Cov}(X_t, X_{s-k}) + \sum_{k,k'=0}^{s-r} \beta(k)\beta(k')\mathrm{Cov}(X_{s-k}, X_{s-k'}) \\
&= \gamma(0) - 2\sum_{k=0}^{s-r} \beta(k)\gamma(t-s+k) + \sum_{k,k'=0}^{s-r} \beta(k)\beta(k')\gamma(k-k') \\
&= \gamma(0) - 2(\beta_t^{r:s})^t \gamma_t^{r:s} + (\beta_t^{r:s})^t \Gamma_t^{r:s} \beta_t^{r:s}.
\end{aligned}$$

For Gaussian noise, a $(1-\alpha)$-prediction interval for X_t is thus (under non-Gaussian noise, this holds approximately in most cases),

$$P(X_t \in \widehat{X}_t^{r:s} \pm \Phi^{-1}(1-\alpha/2)\sigma_t^{r:s}) = 1 - \alpha.$$

As $t - s \to \infty$, we have for stationary processes that $\widehat{X}_t^{r:s} \to 0$ (or the mean μ if $\mu \neq 0$ as assumed above) and $(\sigma_t^{r:s})^2 \to \gamma(0) = \mathrm{Var}(X_t)$.

Example: One-step-ahead prediction. Let $r = s = t - 1$ (and we can introduce again a possibly non-zero mean μ). One easily verifies that

$$\widehat{X}_t^{t-1:t-1} = \mu + \rho(1)(X_{t-1} - \mu), \quad E((X_t - \widehat{X}_t^{t-1:t-1})^2) = \gamma(0)(1 - \rho(1)^2).$$

For $r < s$, specifically $r = 0$ and $s = k - 1$ for $t = k$, we can either use the linear system of equations above. Alternatively, the Durbin-Levinson algorithm allows to compute the coefficients of the linear predictions

$$\widehat{X}_k^{0:k-1} = \alpha_k^{0:k-1} + \sum_{j=1}^{k} \beta_k^{0:k-1}(j) \cdot X_{k-j}$$

and the mean square errors $(\sigma_k)^2 = E((X_k - \widehat{X}_k^{0:k-1})^2)$ recursively. We start with $(\sigma_0)^2 = \gamma(0)$ (as we do not have any values to base the forecast on and the predicted value is just the mean μ) and $\alpha_0 = \mu$. Then we have

$$\begin{aligned}
\beta_k^{0:k-1}(j) &= \beta_k^{0:k-2}(j) + \tau(k)\beta_j^{0:k-2}(k-j) \quad (1 \le j < k), \\
\beta_k^{0:k-1}(k) &= \tau(k), \\
\alpha_k &= \mu(1 - \sum_{j=1}^{k} \beta_p^{0:p-1}(j)), \\
\sigma_k^2 &= \sigma_{k-1}^2(1 - \tau(k)^2)
\end{aligned}$$

where

$$\tau(k) = \frac{\gamma(k) - \sum_{j=1}^{k-1} \beta_k^{0:k-2}(j)\gamma(k-j)}{\sigma_{k-1}^2}$$

where $\tau(k)$ is the so-called partial autocorrelation of lag k. According to the above formula, τ_k is the coefficient of X_0 in $\widehat{X}_k^{0:k-1}$, and $1 - \tau_k^2$ gives the reduction in mean square error if one more observation from the past becomes available.

The distinction between the autocorrelation $\rho(\cdot)$ and partial autocorrelation $\tau(\cdot)$ can be characterized as follows:

$$\begin{aligned} \text{autocorrelation } \rho(k) &= \mathrm{Cor}(X_0, X_k) \\ \text{partial autocorrelation } \tau(k) &= \mathrm{Cor}(X_0 - \widehat{X}_0^{1:k-1}, X_k - \widehat{X}_k^{1:k-1}) \end{aligned}$$

that is the partial autocorrelation $\tau(k)$ is the correlation between X_0 and X_k once the linear effects of the intermediate time-points $X_1, \ldots, X_{k-1}$ have been removed:

$$\ldots, X_{-1}, \underline{X_0}, \underbrace{X_1, X_2, \ldots, X_{k-1}}_{\text{effects removed}}, \underline{X_k}, \ldots$$

For a derivation of these formulae, see for instance Brockwell and Davies, Chapter 2.5.

Examples:

(i) For an AR(1)-process we have $\gamma(h) = \gamma(0)\phi^{|h|}$ and therefore

$$\tau(2) = \frac{\gamma(2) - \phi \cdot \gamma(1)}{\sigma_1^2} = 0.$$

This means that if we know X_{t-1}, then there is no additional information in X_{t-2} that can be used for predicting X_t. This holds because $X_t = \phi X_{t-1} + W_t$ and W_t is independent of all past values.

(ii) For general AR(p)-processes $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + W_t$, the autocorrelation decays exponentially (as discussed before). For the partial autocorrelation, observe that the best prediction for $k > p$ is

$$\widehat{X}_k^{1:k-1} = \sum_{j=1}^{p} \phi_j X_{k-j}.$$

Thus $X_k - \widehat{X}_k^{1:k-1} = W_k$, which is uncorrelated with any past values of X_t, $t < k$. Hence $\tau(k) = 0$ for $k > p$ for an AR(p)-process.

(iii) For a MA(q)-process, the partial autocorrelation decays exponentially but the autocorrelation is zero after q lags.

In summary, we have

	AR(p)-process	MA(q)-process	ARMA-process
Autocorrelation ρ	decays	zero after lag q	decays
Partial Autocorrelation τ	zero after lag p	decays	decays

This distinction can be used for model identification.

Prediction from the infinite past: We assume that both X_t and W_t have mean zero (i.e. we have subtracted the mean). For a causal and invertible ARMA model

$$\widehat{X}_t^{-\infty:t-1} = \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j W_{t-j}. \tag{12}$$

is then the best prediction of X_t based on the infinite past $(X_s, s < t)$, and W_t is the prediction error. In order to compute it, one can either express W_{t-j} with past observations by computing the coefficients π_j according to the formula given above, or one can set $W_{s-1} = \cdots = W_{s-q} = 0$ for a time point $s \ll t$ and then iterate the relation

$$W_u = X_u - \sum_{j=1}^{p} \phi_p X_{u-j} - \sum_{j=1}^{q} \theta_j W_{u-j}$$

for $u = s, s+1, \ldots t$. The error due to assuming $W_{s-1} = \cdots = W_{s-q} = 0$ decays exponentially as $t - s \to \infty$.

Predictions from the infinite past for more than one time step ahead can be made as follows: Because the best prediction is linear, we obtain for $k > 0$

$$\widehat{X}_{t+k}^{-\infty:t-1} = \sum_{j=1}^{\min(p,k-1)} \phi_j \widehat{X}_{t+k-j}^{-\infty:t-1} + \sum_{j=k}^{p} \phi_j X_{t+k-j} + \sum_{j=k}^{q} \theta_j W_{t+k-j}.$$

Hence we see that the predictions for different lead times satisfy the difference equation associated with the AR part, except for finitely many lead times at the beginning. In particular, as the lead time increases, the predictions tend to zero, the mean of X_t.

Use of exponentially weighted moving averages. Take again the one-step forecast $k = 1$. In practice, many people regress X_t onto the $m + r$ predictor variables

$$(X_{t-1}, X_{t-2}, \ldots, X_{t-m}, E_{t-1}^{\lambda_1}, E_{t-1}^{\lambda_2}, \ldots, E_{t-1}^{\lambda_r}),$$

using m past values of X_t (which might for example be a return time-series of an asset) and the exponentially weighted moving averages are defined for $\lambda \in (1, \infty]$ as

$$E_t^{\lambda} = \sum_{j=0}^{\infty} \lambda^{-j} X_{t-j}.$$

Note that this allows an easy recursion and updating as

$$E_t^\lambda = X_t + \frac{1}{\lambda} E_{t-1}^\lambda.$$

In practice, if observing the time-series at n time-points $t_1, \ldots, t_n$ and trying to regress it onto the $m + r$ variables defined above, one can then define the target variable $Y \in \mathbb{R}^n$ and the design matrix $Z \in \mathbb{R}^{n \times p}$ as

$$Z = \left(\begin{array}{llllllll} X_{t_1-1} & X_{t_1-2} & \ldots & X_{t_1-m} & E_{t_1-1}^{\lambda_1} & E_{t_1-1}^{\lambda_2} & \ldots & E_{t_1-1}^{\lambda_r} \\ X_{t_2-1} & X_{t_2-2} & \ldots & X_{t_2-m} & E_{t_2-1}^{\lambda_1} & E_{t_2-1}^{\lambda_2} & \ldots & E_{t_2-1}^{\lambda_r} \\ \ldots & & & & & & & \\ X_{t_n-1} & X_{t_n-2} & \ldots & X_{t_n-m} & E_{t_n-1}^{\lambda_1} & E_{t_n-1}^{\lambda_2} & \ldots & E_{t_n-1}^{\lambda_r} \end{array} \right), \tag{13}$$

$$Y = \left(\begin{array}{l} X_{t_1} \\ X_{t_2} \\ \ldots \\ X_{t_n} \end{array} \right).$$

The least squares estimator is then

$$\hat{\beta} = \mathrm{argmin}_\beta \|Y - Z\beta\|_2^2 = (Z^t Z)^{-1} Z^t Y,$$

assuming that $Z^t Z$ is invertible. Combining the $m + r$ predictor variables with the coefficient $\hat{\beta}$ will then yield optimal linear forecasts (given the set of predictor variables).

Is this model useful if the underlying process is an ARMA model? Let us compare this model to an optimal forecast for an invertible stationary ARMA model. Let π be the coefficients for expressing W_t as a function of past X_t-values:

$$W_t = \sum_{j=0}^{p} \pi_j X_{t-j},$$

assuming again that X_t is invertible. From Theorem 3, we do know that we can write π_j as

$$\pi_j = \sum_{k=1}^{q} \alpha_k z_k^{-j} + \tilde{\pi}_j,$$

where the values $z_k, k = 1, \ldots, q$ are the roots of the polynomial $\theta(z)$ (outside the unit circle as X_t is invertible and assumed to be single roots for simplicity here), α_k are the linear coefficients for each of the solutions and the coefficients $\tilde{\pi}_j$ (where $\tilde{\pi}_j = 0$ for all $j > p$) account for the initial solutions up to $j = p$ where the inhomogenous solutions lead to deviations from the exponentially decaying solutions (compare with (8) to validate, reversing the role of the AR and MA part).

Plugging this into (12),

$$\begin{aligned}
\widehat{X}_t^{-\infty:t-1} &= \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j W_{t-j} \\
&= \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \sum_{j'=0}^{\infty} \pi_{j'} X_{t-j-j'} \\
&= \sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \sum_{j'=0}^{\infty} \underbrace{\Big(\tilde{\pi}_{j'} + \sum_{k=1}^{q} \alpha_k z_k^{-j'}\Big)}_{\pi_{j'}} X_{t-j-j'} \\
&= \sum_{j=1}^{p} \big(\phi_j + \sum_{l=1}^{\min\{j,q\}} \theta_l \tilde{\pi}_{j-l}\big) X_{t-j} + \sum_{k=1}^{q} \alpha_k \sum_{j''=1}^{\infty} \big(\sum_{l=1}^{q} \theta_l z_k^{l-1}\big) z_k^{-j''+1} X_{t-j''} \\
&= \sum_{j=1}^{p} \underbrace{\big(\phi_j + \sum_{l=1}^{\min\{j,q\}} \theta_l \tilde{\pi}_{j-l}\big)}_{=:\gamma_j} X_{t-j} + \sum_{k=1}^{q} \underbrace{\alpha_k \big(\sum_{l=1}^{q} \theta_l z_k^{l-1}\big)}_{=:\delta_k} \qquad \underbrace{\sum_{j''=1}^{\infty} z_k^{-j''+1} X_{t-j''}}_{=E_{t-1}^{z_k}} \text{ by definition of EWMA's} \\
&= \sum_{j=1}^{p} \gamma_j X_{t-j} + \sum_{k=1}^{q} \delta_k E_{t-1}^{z_k}, \qquad (14)
\end{aligned}$$

where

$$\begin{aligned}
\gamma_j &:= \phi_j + \sum_{l=1}^{\min\{j,q\}} \theta_l \tilde{\pi}_{j-l} \quad 1 \le j \le p \text{ (and 0 for } j > p) \\
\delta_k &:= \alpha_k \sum_{l=1}^{q} \big(\theta_l z_k^{l-1}\big) \quad k = 1, \dots, q.
\end{aligned}$$

We see from (14) that the optimal forecast can indeed be written as a linear combination of the past $m = p$ values $X_{t-1}, X_{t-2}, \dots, X_{t-p}$ plus a linear combination of the r EWMA where the parameters λ_k, $k = 1, \dots, r$ need to contain the roots z_k, $k = 1, \dots, q$ of the characteristic polynomial $\theta(z)$. For example, if $\lambda_k = z_k$ for all k and $r = q$ and $m = p$, setting $\beta \in \mathbb{R}^{m+r}$ to be

$$\beta_j = \gamma_j \text{ if } j \le p \text{ and } \beta_j = \delta_{j-p} \text{ if } p < j \le (p+q),$$

we get that the optimal linear forecast at time-points $t_1, \dots, t_n$ is indeed given by $Z\beta$, where Z is given as in (13).

In practice, the coefficients ϕ and θ that define the optimal basis vectors (via p, q and the roots z_k, $k = 1, \dots, q$) are unknown but we can estimate the optimal coefficients β directly from data by least-squares regression or similar, using a basis λ_k, $k = 1, \dots, r$ for the EWMA with very large r that should allow a good approximation to the unknown roots of the polynomial $\theta(z)$.

2.8 Statistical inference for ARMA models

2.8.1 Estimation of coefficients

Estimation of the unknown parameters ϕ_j, θ_k and $\sigma_W^2 = \mathrm{Var}(W_t)$ is usually done with exact or approximate Gaussian maximum likelihood (MLE). An unknown mean is usually estimated first by the arithmetic mean of the data and then subtracted.

We have the following general formula for the density of $X_1, \ldots, X_n$

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2|x_1)\cdots f(x_n|x_1, \ldots, , x_{n-1}).$$

In the Gaussian case, the conditional densities $f(x_t|x_1, \ldots, , x_{t-1})$ are again Gaussian with mean equal to the best linear prediction $\widehat{X}_{t|1:t-1}$ and variance equal to $\mathrm{Var}(X_t - \widehat{X}_t^{1:t-1})$. For the exact MLE, one computes these means and variances exactly as a function of the unknown parameters. We can restrict ourselves to mean-zero process (estimating the mean as the empirical mean of $X_1, \ldots, X_n$ and subtracting it from the data). The covariance of $X_1, \ldots, X_n$ for an ARMA(p,q) process with coefficients $\phi = (\phi_1, \ldots, \phi_p)$ and $\theta = (\theta_1, \ldots, \theta_q)$ and the noise variance $\sigma_W^2 = \mathrm{Var}(W_t)$ is then the matrix

$$\Gamma_n(\phi, \theta, \sigma_W^2) = \begin{pmatrix} \gamma_{\phi,\theta,\sigma_W^2}(0) & \gamma_{\phi,\theta,\sigma_W^2}(1) & \gamma_{\phi,\theta,\sigma_W^2}(2) & \cdots & \gamma_{\phi,\theta,\sigma_W^2}(n-1) \\ \gamma_{\phi,\theta,\sigma_W^2}(1) & \gamma_{\phi,\theta,\sigma_W^2}(0) & \gamma_{\phi,\theta,\sigma_W^2}(1) & \cdots & \\ \gamma_{\phi,\theta,\sigma_W^2}(2) & \cdots & & & \\ \cdots & & & & \\ \cdots & & & & \\ \cdots & & & & \\ \gamma_{\phi,\theta,\sigma_W^2}(n-1) & \cdots & & & \gamma_{\phi,\theta,\sigma_W^2}(0) \end{pmatrix},$$

where $\gamma = \gamma_{\phi,\theta,\sigma_W^2}$ is the auto-covariance under parameters ϕ, θ, σ_W^2. For a given ϕ, θ, σ_W^2, we can compute γ and hence $\Gamma_n(\phi, \theta, \sigma_W^2)$ and compute the likelihood

$$L(\phi, \theta, \sigma_W^2) := \frac{1}{\sqrt{(2\pi)^n |\Gamma_n(\phi, \theta, \sigma_W^2)|}} \exp\big(-\frac{1}{2}(X_1, \ldots, X_n)\Gamma_n(\phi, \theta, \sigma_W^2)^{-1}(X_1, \ldots, X_n)^t\big).$$

We would like to choose ϕ, θ, σ_W^2 such that we maximize the likelihood $L(\phi, \theta, \sigma_W^2)$ or minimize the negative log-likelihood

$$\begin{aligned} -2\ell(\phi, \theta, \sigma_W^2) &= -2\log(L(\phi, \theta, \sigma_W^2)) \\ &= \text{constant} + \log(|\Gamma_n(\phi, \theta, \sigma_W^2)|) + (X_1, \ldots, X_n)\Gamma_n(\phi, \theta, \sigma_W^2)^{-1}(X_1, \ldots, X_n)^t. \end{aligned}$$

This is a difficult optimisation problem as the likelihood and the log-likelihood are in general not concave or convex functions of their parameters.

An approximate likelihood uses $\widehat{X}_t^{-\infty:t-1}$ and $\mathrm{Var}(W_t)$ instead where $\widehat{X}_t^{-\infty:t-1}$ is computed recursively starting with $W_0 = \cdots = W_{1-q} = 0$. In order to reduce the effect of these artificial starting values, one typically omits the first $r = \max(p, q+1)$ factors in the likelihood, that is one takes

$$f(x_{r+1}, \dots, x_n | x_1, \dots, x_r) = \prod_{t=r+1}^{n} f(x_t | x_1, \dots x_{t-1}).$$

For the AR(p) model and Gaussian innovations, this reduces to the least squares estimator

$$\arg\min \sum_{t=p+1}^{n} (x_t - \sum_{j=1}^{p} \phi_j x_{t-j})^2$$

which is particularly simple to compute. The design matrix X and response in the traditional regression setting are given by

$$X = \begin{pmatrix} x_p & x_{p-1} & \dots & x_1 \\ x_{p+1} & x_p & \dots & x_2 \\ \dots & & & \\ x_{n-1} & x_{n-2} & \dots & x_{n-p} \end{pmatrix}, \qquad Y = \begin{pmatrix} x_{p+1} \\ x_{p+2} \\ \dots \\ x_n \end{pmatrix}$$

and we try to minimize

$$\|X\phi - Y\|_2^2, \qquad \text{where } \phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \dots \\ \phi_p \end{pmatrix},$$

yielding the solution

$$\hat{\phi} = (X^t X)^{-1} X^t Y.$$

The Yule-Walker estimator determines the unknown ϕ_j and σ_W^2 from the Yule-Walker equations with estimated covariances $\widehat{\rho}(h)$ $(0 \leq h \leq p)$ as

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p,$$

where $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ are of the form discussed above, using the estimated auto-covariance function. Note that for large n (and large $n-p$),

$$\hat{\Gamma}_p \approx X^t X \text{ and } \hat{\gamma}_p \approx X^t Y.$$

The first are estimates for Γ_p and the latter are for γ_p and it is easy to see thus that the Yule-Walker and least-squares estimators for AR(p) processes are converging to the same value as n is growing.

The Burg estimator proceeds recursively with respect to p, that is, it estimates the partial autocorrelations, and it does this by minimizing forward and backward prediction errors.

For long series, all the different versions give similar estimates, but for shorter series and parameters close to the boundary of the causality and invertibility region, the choice of the estimator can matter. Usually one prefers the exact MLE or the Burg estimator.

2.8.2 Order selection

A simple technique is to identify the orders p and q from the plot of the autocorrelations and partial autocorrelations. For an MA(q) process, all autocorrelations $\rho(h) = 0$ for $h > q$ whereas the partial autocorrelations $\tau(h)$ decay exponentially or like a damped harmonic as $h \to \infty$. For an AR(p) process, the partial autocorrelations $\tau(h)$ are zero for $h > p$ and the autocorrelations decay exponentially or like a damped harmonic as $h \to \infty$. For an ARMA(p,q) process with $p > 0$ and $q > 0$ both $\tau(h)$ and $\rho(h)$ decay exponentially or like a damped harmonic.

It is also possible to fit ARMA(p,q) models for all $p \leq p_0$ and $q \leq q_0$ and to choose the one with the best fit afterwards. The most popular methods to choose the order are then the selection critera AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). They are defined as follows:

$$-2 \sup \ell(\phi, \theta, \sigma_W^2) + C(p+q)$$

where ℓ is the log likelihood function and $C = 2$ in case of the AIC and $C = \log(n)$ in case of the BIC. If the order increases, the first term always decreases because the supremum is taken over a larger set. The second term is a penalty for the complexity of the model. If the estimates $\widehat{\phi}$ and $\widehat{\theta}$ are based on an approximate likelihood, then one uses this approximate likelihood in the AIC or BIC instead of the exact likelihood ℓ. We then search for the model which minimizes the AIC criterion.

Validation on independent data is an alternative to penalized likelihood.

I do not discuss here the justification of these criteria, but just mention two results: 1) The AIC is an unbiased estimate of a distance between the fitted and the true model. 2) The AIC favors complex models and does not provide a consistent estimate of the true order if the true order is finite.

Goodness of fit Once a model has been fitted (that is both the orders and the parameters have been estimated), one should check whether the fit is adequate. As a minimum, one should look at the time series plot and the acf of the residuals $\widehat{W_t} = X_t - \widehat{X}_t^{-\infty:t-1}$ which approximate the innovations W_t and thus should be approximately i.i.d..

The auto-correlation-function $\rho_{res}(h)$ of the residuals should vanish at all lags $h > 0$. To test, this, one can for example compute the test statistic

$$Q(H) = n(n+2) \sum_{k=1}^{H} \frac{\hat{\rho}_{res}(h)^2}{n-k},$$

where $\hat{\rho}_{res}(h)$ is the empirical auto-correlation at lag h of the residuals. If the residuals are really i.i.d., then $Q(H)$ will have approximately a χ^2_{H-K}-distribution, where K is the number of fitted parameters (so $K = p + q$ for an ARMA(p,q)-model). The corresponding p-value of the test (as a function of H, even though $H = 10$ or $H = 20$ is an often-made choice) are shown as Ljung-box p-values in the residual plot diagnostics.

It is also a good idea to simulate from the fitted model and compare the plot of a simulated series with the plot of the original series. Ideally, the two plots should be visually indistinguishable. One can also look for nonlinear dependence among the residuals, by using for instance lag plots ($\widehat{W}_{t+h}$ versus $\widehat{W}_t$) or the acf of the squared residuals, or for non-Gaussianity with a normal plot of the residuals.

ARIMA-Models

So far all ARMA models were stationary. One way to analyze nonstationary data is to take differences, see 1.2. This can be included in the ARMA model:

$$\Phi(B)(1-B)^d X_t = \Phi^*(B)X_t = \Theta(B)W_t.$$

The polynomial $\Phi^*(z)$ has degree $d+p$ and it has a root at $z=1$ of multiplicity d and p roots outside of the unit circle. Such a model is called an ARIMA(p, d, q) model (autoregressive integrated moving average). Note that an ARIMA model is not unique: If (X_t, W_t) satisfies the above recursion, then so does $(X_t + A_0 + \ldots A_{d-1}t^{d-1}, W_t)$ for arbitrary coefficients $A_0, \ldots, A_{d-1}$. In other words, the ARIMA model only specifies the conditional distribution of $X_1, X_2, \ldots$ given the initial values $X_0, X_{-1}, \ldots, X_{d-1}$, and not the distribution of these initial values. Also note that if $E(W_t) \neq 0$, then $E(X_t)$ contains a term ct^d with $c \neq 0$. Because of this, one usually assumes that $E(W_t) = 0$ if $d > 0$.

Whether we should choose $d > 0$ usually becomes clear from the inspection of the time series plot (slowly changing level or slowly changing slope of the series) and of the acf (behaviour of $\widehat{\rho}(h) \sim 1 - \text{const.}h$ with a small value of const.). Identifying p and q and estimating the coefficients is then done based on the differenced series $Y_t = (1-B)^d X_t$.

For forecasting, one usually assumes that the initial values $X_0, X_{-1}, \ldots, X_{d-1}$ are independent of the differenced series $Y_t = (1-B)^d X_t$. Then the same formula can be used for recursive computation of the forecast k steps ahead as in the stationary case.

If a series contains a seasonal component, then we often need to take also seasonal differences to achieve stationarity. This means that we use a model of the form

$$\Phi(B)(1-B)^d(1-B^M)^D X_t = \Theta(B)W_t$$

where M is the number of observations in one seasonal cycle. Moreover, empirically the seasonal behavior also shows up in the structure of the polynomials Φ and Θ. For instance in the autoregressive case, X_t depends usually on X_{t-1}, X_{t-M} and maybe X_{t-M-1}. This leads to the so-called seasonal ARIMA(p, d, q, P, D, Q) model:

$$\Phi(B)\Phi_M(B^M)(1-B)^d(1-B^M)^D X_t = \Theta(B)\Theta_M(B^M)W_t.$$

In general, one would expect that $d + D \leq 2$ and either P or Q is equal to 0 (so the seasonal component is either of AR or MA but not general ARMA form).

An example is given by the so-called airline model (because it fits the data on airline passengers, one of the standard data sets in R, well):

$$(1-B)(1-B^M)X_t = (1+\theta_1 B)(1+\theta_{M,1}B^M)W_t,$$

which is an ARMA(0,1,1,0,1,1) model.

There are many other possible seasonal models but the main point is that the stationary ARMA models yield a good modelling basis and allow to derive many interesting non-stationary models for a given dataset.

3 Spectral methods

3.1 The spectral representation

3.1.1 Some results from deterministic spectral analysis

Fourier theory is concerned with the representation of signals g as a superposition of harmonics with different frequencies and amplitudes. If g is a signal in continuous time t with finite energy

$$\int_{-\infty}^{\infty} g(t)^2 dt < \infty,$$

then it can be represented as

$$g(t) = \int_{-\infty}^{\infty} G(\nu)\exp(i2\pi\nu t)d\nu \tag{15}$$

where

$$G(\nu) = \int_{-\infty}^{\infty} g(t)\exp(-i2\pi\nu t)dt. \tag{16}$$

Hence g is a superposition of harmonics with continuous frequencies ν. If we write $G(\nu)$ in polar coordinates $G(\nu) = |G(\nu)|\exp(i\phi(\nu))$, we see that the harmonic $G(\nu)\exp(i2\pi\nu t)$ has amplitude $|G(\nu)|$ and phase $\phi(\nu)$. Since g is real, $G(-\nu) = \overline{G(\nu)}$ and we also have a representation in terms of sine and cosine functions with frequencies $\nu > 0$. Moreover, Parseval's theorem says that

$$\int_{-\infty}^{\infty} g(t)^2 dt = \int_{-\infty}^{\infty} |G(\nu)|^2 d\nu,$$

i.e. the energy is the integral of squared amplitudes.

Next, we consider a signal (g_t) observed at time points $t\Delta$ with $t = 0, \pm 1, \ldots$ with finite energy $\sum_{t=-\infty}^{\infty} g_t^2 < \infty$. If we replace the integrand in (16) by a function which is constant on intervals of length Δ, then we obtain

$$G_p(\nu) = \Delta \sum_{t=-\infty}^{\infty} g_t \exp(-i2\pi t\Delta\nu), \tag{17}$$

and we can represent the signal with G_p:

$$g_t = \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(i2\pi t\Delta\nu) d\nu. \tag{18}$$

Hence again g is a superposition of harmonics, but the continuous frequencies ν are now restricted to $|\nu| \leq 1/(2\Delta)$. The reason for this is that in discrete time we cannot distinguish between harmonics at frequencies ν, $\nu \pm 1/\Delta$, $\nu \pm 2/\Delta$ etc. . This is called aliasing, and $1/(2\Delta)$ is called the Nyquist frequency.

If we consider G_p as a function of arbitrary ν, then it is periodic with period $1/\Delta$ (this is the reason for the subscript p). Note that $G_p(\nu) \neq G(\nu)$ for $|\nu| \leq 1/\Delta$, but rather

$$G_p(\nu) = \sum_{k=-\infty}^{\infty} G(\nu + k/\Delta) = G(\nu) + \sum_{k=1}^{\infty} (G(\nu + k/\Delta) + \overline{G(-\nu + k/\Delta)}).$$

This means that we add up the amplitudes at all frequencies we cannot distinguish. Finally, for a discrete time signal, Parseval's theorem says that

$$\Delta \sum_{t=-\infty}^{\infty} g_t^2 = \int_{-1/(2\Delta)}^{1/(2\Delta)} |G_p(\nu)|^2 d\nu.$$

In the last step, we consider a signal g observed at finitely many discrete time points $t\Delta$ with $t = 0, 1, \ldots n-1$. By replacing the integrand in (18) by a function which is constant on intervals of length $1/(n\Delta)$, we obtain the representation

$$g_t = \frac{1}{n\Delta} \sum_{k=0}^{n-1} G_k \exp(i2\pi t\Delta \frac{k}{n\Delta}) = \frac{1}{n\Delta} \sum_{k=0}^{n-1} G_k \exp(i2\pi tk/n) \tag{19}$$

whose inversion is

$$G_k = \Delta \sum_{t=0}^{n-1} g_t \exp(-i2\pi tk/n). \tag{20}$$

Hence the signal is now a superposition of harmonics with a finite number of frequencies $\nu_k = k/(\Delta n)$, the so-called Fourier frequencies. Again Parseval's theorem holds

$$\Delta \sum_{t=0}^{n-1} g_t^2 = \frac{1}{n\Delta} \sum_{k=0}^{n-1} |G_k|^2.$$

If we use (19) or (20) to define g_t for any $t \in \mathbb{Z}$ or G_k for any $k \in \mathbb{Z}$, we obtain periodic sequences. If we restrict an infinite sequence with Fourier representation

$$g_t = \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(i2\pi t\Delta\nu) d\nu,$$

to $0 \leq t < n$, then the relation between $G_p(\nu)$ and the discrete amplitudes G_k is

$$G_k = n\Delta \int_{-1/(2\Delta)}^{1/(2\Delta)} G_p(\nu) \exp(-i\pi(n-1)(\nu_k - \nu)\Delta) D_n(|\nu - \nu_k|\Delta) d\nu$$

where D_n is the so-called Dirichlet kernel

$$D_n(\nu) = \frac{\sin(n\pi\nu)}{n \sin(\pi\nu)}.$$

This means that the amplitude G_k in the discrete representation is a weighted average of the amplitudes $G_p(\nu)$ for ν around ν_k. The phase shift in the above formula occurs because the time points are not symmetric around the origin. The proof of this formula uses the the summation formula of a geometric series

$$\sum_{t=0}^{n-1} e^{i\lambda t} = \frac{e^{i\lambda n} - 1}{e^{i\lambda} - 1} = e^{i(n-1)\lambda/2} \frac{e^{in\lambda/2} - e^{-in\lambda/2}}{e^{i\lambda/2} - e^{-i\lambda/2}} = e^{i(n-1)\lambda/2} \frac{\sin(n\lambda/2)}{\sin(\lambda/2)}.$$

The discrete Fourier transform $(g_t) \to (G_k)$ can be computed by the Fast Fourier Transform (FFT) with $O(n\log_2(n))$ operations instead of $O(n^2)$ operations in a naive implementation. This algorithm is crucial for the widespread use of Fourier methods in many applications.

The spectral representation of stationary stochastic processes

For a stationary stochastic process $(X_t; t \in \mathbb{Z})$, the energy $\sum_{t=-\infty}^{\infty} X_t^2$ is infinite, but if second moments exist, the power (energy per time unit) converges to a finite value

$$\frac{1}{2T+1} \sum_{t=-T}^{T} X_t^2 \to E(X_t^2).$$

Hence we cannot expect to have a representation of the form

$$X_t(\omega) = \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(\nu, \omega) d\nu.$$

However, a deep result says that we have the representation

$$X_t(\omega) = E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(d\nu, \omega)$$

where Z is a (complex) stochastic process with uncorrelated increments:

1. $Z(-\nu) - Z(-\nu - h) = \overline{Z(\nu + h) - Z(\nu)}$ for all ν, h.
2. $E(Z(\nu + h) - Z(\nu)) = 0$ for all ν, h.

3. $E|Z(\nu+h)-Z(\nu)|^2 = S(\nu+h)-S(\nu)$ where S is the spectral distribution function $S(\nu) = S([-1/2,\nu])$.
4. For $\nu < \nu+h < \nu' < \nu'+h'$, $E((Z(\nu+h)-Z(\nu))\overline{(Z(\nu'+h')-Z(\nu'))}) = 0$.

Here the integral is defined as the limit of

$$\sum_j \exp(i2\pi\nu_j t)(Z(\nu_j,\omega)-Z(\nu_{j-1},\omega))$$

as the partition $\nu_0 = -1/2 < \nu_1 < \ldots < \nu_J = 1/2$ becomes finer. Hence intuitively, the process is a superposition of harmonics with uncorrelated mean zero amplitudes, and the variance of the amplitudes are given by the increments of the spectrum. In other words, the spectrum spectrum says how strongly the different frequencies are represented in the process. If the spectral density exists, $E(|Z(\nu_j,\omega)-Z(\nu_{j-1},\omega)|^2)$ is of the order $\nu_j - \nu_{j-1}$ and therefore $|Z(\nu_j,\omega)-Z(\nu_{j-1},\omega)|$ is typically of the order $\sqrt{\nu_j-\nu_{j-1}} > \nu_j - \nu_{j-1}$. This is the crucial difference between the spectral representation here and the representations in the previous subsection.

Formally, we can write the properties of Z as

$$E(Z(d\nu)\overline{Z(d\nu')}) = \delta_{\nu,\nu'}S(d\nu)$$

where $\delta_{\nu,\nu'} = 0$ for $\nu \neq \nu'$ and $\delta_{\nu,\nu} = 1$ (the Kronecker delta). We then obtain the spectral representation of the autocovariances (Herglotz's Theorem)

$$\begin{aligned}\gamma(k) &= \operatorname{Cov}(X_{t+k}(\omega),X_t(\omega)) = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \exp(i2\pi\nu(t+k))\exp(-i2\pi\nu' t)E(Z(d\nu,\omega)\overline{Z(d\nu',\omega)}) \\ &= \int_{-1/2}^{1/2} \exp(i2\pi\nu k)S(d\nu).\end{aligned}$$

In particular,

$$E((X_t - E(X_t)^2) = \int_{-1/2}^{1/2} S(d\nu)$$

which is the analogue of Parseval's theorem.

3.1.2 Linear filters

A (time invariant) linear filter is a transformation of an input time series (X_t) into an output time series (Y_t) of the following form

$$Y_t = \sum_k a_k X_{t-k}$$

The input or output can be either deterministic or stochastic. Usually one assumes that $\sum_k |a_k| < \infty$ or some other condition in order that the right hand side is well defined.

If the input is an impulse at time zero, $X_t = \delta_{t0}$, then the output is equal to $Y_t = a_t$. Because of this, the coefficients a_k are called the impulse response coefficients. If the input is a harmonic with frequency ν, $X_t = G \exp(i2\pi\nu t)$ then the output is again a harmonic with the same frequency

$$Y_t = GA(\nu)\exp(i2\pi\nu t), \quad A(\nu) = \sum_k a_k \exp(-i2\pi\nu k).$$

There is however a change in amplitude by $|A(\nu)|$ and also a phase shift unless the coefficients are symmetric ($a_{-k} = a_k$). $A(\nu)$ is called the transfer function.

By linearity of the linear filter, a superposition of harmonic oscillations is transformed into another superposition of harmonics where the amplitudes and phases are changed by the transfer function.

Stationary stochastic processes are superpositions of harmonic oscillations:

$$X_t = E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) Z(d\nu).$$

If the coefficients (a_k) are summable,

$$Y_t = A(0)E(X_t) + \int_{-1/2}^{1/2} \exp(i2\pi\nu t) A(\nu) Z(d\nu).$$

Therefore the spectral increment process of (Y_t) is $A(\nu)Z(d\nu)$ and we have the following relation between the spectral measures of (X_t) and (Y_t):

$$S_Y(d\nu) = |A(\nu)|^2 S_X(d\nu).$$

3.2 The periodogram

The periodogram of a time series of length n with sampling interval Δ is defined as

$$I_n(\nu) = \frac{\Delta}{n}\left|\sum_{t=1}^{n}(X_t - \overline{X})\exp(-i2\pi\nu t\Delta)\right|^2.$$

In words, we compute the absolute value squared of the Fourier transform of the sample, that is we consider the squared amplitude and ignore the phase.

Note that I_n is periodic with period $1/\Delta$ and that $I_n(0) = 0$ because we have centered the observations at the mean. The centering has no effect for Fourier frequencies $\nu = k/(n\Delta)$, $k \neq 0$.

By mutliplying out the absolute value squared on the right, we obtain

$$I_n(\nu) = \frac{\Delta}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \overline{X})(X_s - \overline{X})\exp(-i2\pi\nu(t-s)\Delta) = \Delta \sum_{h=-n+1}^{n-1} \widehat{\gamma}(h)\exp(-i2\pi\nu h).$$

Hence the periodogram is nothing else than the Fourier transform of the estimated acf. In the following, we assume that $\Delta = 1$ in order to simplify the formula (although for applications the value of Δ in the original time scale matters for the interpretation of frequencies).

By the above result, the periodogram seems to be the natural estimator of the spectral density

$$s(\nu) = \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-i2\pi\nu h).$$

However, a closer inspection shows that the periodogram has two serious shortcomings: It has large random fluctuations, and also a bias which can be large.

We first consider the bias. Using the spectral representation, we see that up to a term which involves $E(X_t) - \overline{X}$

$$I_n(\nu) = \frac{1}{n} \left| \int \sum_{t=1}^{n} e^{-i2\pi(\nu-\nu')t} Z(d\nu') \right|^2 = n \left| \int e^{-i\pi(n+1)(\nu-\nu')} D_n(\nu-\nu') Z(d\nu') \right|^2 .$$

Taking the expectation on both sides and using the properties of Z, we obtain

$$E(I_n(\nu)) = n \int D_n(\nu-\nu')^2 s(\nu') d\nu'.$$

In order to gain insight from this formula, we need to understand the behaviour of the Dirchlet kernel D_n and the so-called Fejér kernel

$$F_n(\nu) = n D_n(\nu)^2.$$

It can be checked that $F_n(0) = n$, $F_n(\nu) \to 0$ as $n \to \infty$ for all $0 < |\nu| \leq 1/2$ and $\int_{-1/2}^{1/2} F_n(\nu) = 1$ for all n. Hence F_n approximates the Dirac delta function and for a continuous density we obtain $E(I_n(\nu)) \to s(\nu)$ for any $\nu \neq 0$.

Still, for some applications, the bias of the periodogram can be substantial. In such cases the bias is reduced if we use a so-called taper. This is a set of weights $h_1, h_2, \ldots, h_n$ which are one for t close to $n/2$ and decay smoothly to zero for t near 1 and n. With these weights, we compute the tapered periodogram as follows

$$I_n^h(\nu) = \frac{1}{\sum_{t=1}^{n} h_t^2} \left| \sum_{t=1}^{n} h_t (X_t - \overline{X}) \exp(-i2\pi\nu t) \right|^2 .$$

If we use a taper, then we obtain

$$E(I_n^h(\nu)) = \int H_n(\nu-\nu') s(\nu') d\nu'$$

where

$$H_n(\nu) = \frac{1}{\sum_{t=1}^{n} h_t^2} \left| \sum_{t=1}^{n} h_t \exp(-2\pi i\nu t) \right|^2 .$$

If h_t is as described above, $H_n(\nu)$ has smaller sidelobes than the Fejér kernel.

The variances and covariances of the periodogram depend in principle on the fourth moments of the process. However, for many processes a Central Limit Theorem applies for the Fourier transform and thus the real and imaginary part of $\sum h_t(X_t - \overline{X}) \exp(2\pi i \nu t)$ have asymptotically a normal distribution with mean zero and variance $s(\nu)/2$ for $\nu \neq 0, 1/2$. Because of this

$$\frac{I_n^h(\nu)}{s(\nu)} \text{ approximately } \sim \text{Exp}(1) \qquad (= \frac{\chi_2^2}{2}).$$

In particular, the periodogram is an asymptotically unbiased, but not consistent estimator for the spectral density, and

$$\left[\frac{I_n^h(\nu)}{-\log(0.025)}, \frac{I_n^h(\nu)}{-\log(0.975)}\right]$$

is an approximate 95% confidence interval for $s(\nu)$. On the logarithmic scale, this interval has constant width.

For two different frequencies $\nu \neq \nu'$, the periodogram values are asymptotically independent, in particular the covariance tends to zero. This explains the irregular behaviour of the periodogram as a function of frequency. Because of this and because of the inconsistency, the periodogram is of limited value.

For two frequencies close together, we have the following approximation

$$\text{Cov}(I_n^h(\nu), I_n^h(\nu')) \approx \frac{s(\nu)s(\nu')}{\sum_{t=1}^n h_t^2} \left| \sum_{t=1}^n h_t^2 \exp(-2\pi i(\nu - \nu')t) \right|^2 .$$

Without a taper, i.e. for $h_t \equiv 1$, the periodogram values at two Fourier frequencies j/n and j'/n are thus approximately uncorrelated. This does not hold if we use a taper.

I refer to the literature for exact statements and proofs of these results.

3.3 Smoothing the periodogram

The reason why the periodogram is not consistent is that as the length n of the time series increases , we obtain independent estimates of the spectral density at an increasingly dense set of Fourier frequencies $\nu_k = k/n$. If the spectral density is smooth, we can therefore pool the information from nearby frequencies.

The tapered and smoothed spectral estimate is

$$\hat{s}^{(ts)}(k/n) = \sum_{j=-J}^{J} w_j I_n^h((k-j)/n),$$

where the w_j's are weights with the following properties

$$w_j > 0,\ w_j = w_{-j}\ (-J \leq j \leq J),\ \sum_{j=-J}^{J} w_j = 1.$$

If $k \leq J$, the smoothing includes the periodogram at the origin which is equal or very close to zero if the mean μ is estimated. In this case, we exclude $j = k$ from the sum and renormalize the weights.

The properties of this estimator can be derived by the same arguments that are used for kernel smoothers in nonparametric regression. If we neglect the bias of the tapered periodogram, the bias of $\hat{s}^{(ts)}$ is approximately

$$\frac{s''(k/n)}{2}\frac{1}{n^2}\sum_{j=-J}^{J} j^2 w_j.$$

The variance of $\hat{s}^{(ts)}(k/n)$ depends on whether or not a taper is used. Without a taper the summands are approximately uncorrelated, and we obtain for $k \neq 0, n/2$

$$\mathrm{Var}(\hat{s}^{(ts)}(k/n)) \approx s(k/n)^2 \sum_{j=-J}^{J} w_j^2.$$

With a taper, we have to take the correlation of the summands into account. We skip the details and just state that in this case the variance is increased by the factor

$$M(h) = \frac{\frac{1}{n}\sum_{t=1}^{n} h_t^4}{(\frac{1}{n}\sum_{t=1}^{n} h_t^2)^2}.$$

By Cauchy-Schwarz, $M(h)$ is strictly greater than one unless h_t is constant, and thus asymptotically tapering entails some loss of precision. However, this is often more than compensated by a reduction in bias.

The choice of J, that is the number of frequencies involved in the smoothed estimate, is difficult. Small values of J give a small bias, but a large variance, and vice versa. Asymptotically, the optimal choice is $J = O(n^{4/5})$, but the constants involve both s and s'' which are unknown. In practice, one often looks at the estimate for different values of J and then makes a subjective choice.

The above results imply that to a first approximation

$$E\left(\frac{\hat{s}^{(ts)}(k/n)}{s(k/n)}\right) = 1, \quad \mathrm{Var}\left(\frac{\hat{s}^{(ts)}(k/n)}{s(k/n)}\right) = \sum_{j=-J}^{J} w_j^2\ M(h).$$

Because the periodogram values have asymptotically an exponential distribution and the sum of m independent exponential random variables is distributed as $1/2$ times a chi-squared random variable with $2m$ degrees of freedom, one approximates the distribution of $\hat{s}^{(ts)}(k/n)/s(k/n)$ by Z_d/d where $Z_d \sim \chi^2_d$ and the degrees of freedom d are chosen to match the variance given above. This then leads to the following confidence interval for $s(k/n)$

$$\left[\frac{\hat{s}^{(ts)}(k/n)\ d}{\chi^2_{d,1-\alpha/2}}, \frac{\hat{s}^{(ts)}(k/n)\ d}{\chi^2_{d,\alpha/2}}\right] \quad \text{where } d = \frac{2}{\sum_{j=-J}^{J} w_j^2\ M(h)}.$$

3.4 Alternative estimators of the spectrum

So far, we have averaged over the values of the periodogram at the Fourier frequencies k/n because they are approximately independent in the case of no taper and because the fast Fourier transform can be used for computation.

We can also use a different grid k/n' with $n' > n$ (we then have to set $X_t = \bar{X}$ for $n < t \leq n'$ in order to use the fast Fourier transform). In the limit we then have a continuous average

$$\widehat{s}^{(lw)}(\nu) = \int W(\nu - \nu') I_n^h(\nu') d\nu'.$$

This can be shown to be equal to

$$\sum_{k=-n+1}^{n-1} w_k \widehat{\gamma}^h(k) \exp(-2\pi i \nu k)$$

where

$$w_k = \int W(\nu) \exp(2\pi i \nu k) d\nu$$

and

$$\widehat{\gamma}^h(k) = \frac{1}{\sum_{t=1}^{n} h_t^2} \sum_{t=1}^{n-|k|} h_t(X_t - \bar{X}) h_{t+|k|}(X_{t+|k|} - \bar{X})$$

are the autocovariances of the tapered series. In other words, smoothing of the periodogram is equivalent to downweighting the estimated autocovariances in the inversion formula

$$s(\nu) = \sum_{k=-\infty}^{\infty} \gamma(k) \exp(-2\pi i k \nu).$$

This estimator is therefore called a lag weight estimator (which explains the superscript lw). For computational reasons, $\widehat{s}^{(ts)}$ is usually preferred.

A different approach consists in averaging the periodograms for segments of $m < n$ consecutive observations:

$$\widehat{s}^{(os)}(\nu) = \frac{1}{J \sum_{t=1}^{m} h_t^2} \sum_{j=0}^{J-1} \left| \sum_{t=1}^{m} h_t (X_{t+jd} - \bar{X}) e^{-2\pi i \nu t} \right|^2$$

where J is the integer part of $(n-m)/d$. The parameter d regulates how much the segments overlap: For $d = 1$ we have maximal overlap whereas for $d = m$ there is no overlap (*os* stands for overlapping segments). It can be shown that in case of maximal overlap, this is essentially a lag weight estimator. It has however the advantage that it gives also information about changes in the periodogram over time. It is thus the first step towards a time-frequency analysis where one wants to analyze how strongly different frequencies are present at different times. This is however an ill-posed question since a high resolution in time entails a low resolution in frequency and vice versa.

Yet a different approach to spectral estimation consists in using the spectral density of a fitted autoregressive model. Usually, one chooses the order of the autoregression by AIC. This usually gives very smooth estimates, but sometimes details are lost that can be detected by $\widehat{s}^{(ts)}$. A combination of both methods fits an autoregression, usually of low order without assuming that the innovations

$$W_t = X_t - \sum_{k=1}^{p} \phi_k X_{t-k}$$

are exactly white noise. In any case, the general formula

$$s_X(\nu) = \frac{s_W(\nu)}{|1 - \sum \phi_k \exp(-2\pi i \nu k)|^2}.$$

holds, and one estimates $s_W(\nu)$ by smoothing the periodogram of the residuals. Even when $s_W(\nu)$ is not exactly constant, it is at least much flatter than $s_X(\nu)$ and thus the problems with the bias are less serious. This approach is called prewhitening.

Wavelets in time series analysis

Wavelets are suitable both for smoothing time series and for a time-frequency analysis. We can only give a very brief introduction. The discrete wavelet transform decomposes an equispaced time series of length n as follows:

$$X_t = \sum_{j=1}^{J} \sum_{k=0}^{2^{-j}n-1} d_{j,k} 2^{-j/2} \psi(2^{-j}t - k) + \sum_{k=0}^{2^{-J}n-1} a_{J,k} 2^{-J/2} \phi(2^{-J}t - k) \quad (t = 0, 1, \ldots, n-1)$$

where ψ is the so-called mother wavelet – a small wave located near zero – and ϕ is the so-called father wavelet or scaling function which represents a smooth part. Hence we have a decomposition into oscillations with frequencies 2^{-j} located at times $k2^j$ for $j = 1, 2, \ldots, J$ and a part which contains the lower frequencies. The simplest example is the Haar wavelet where

$$\psi(t) = 1_{[0,1/2)}(t) - 1_{[1/2,1)}(t), \quad \phi(t) = 1_{[0,1)}(t).$$

For other cases, ψ and ϕ are defined through a limiting operation and thus have to be calculated numerically.

The amplitudes $d_{j,k}$ and $a_{J,k}$ are computed from the original series by iterative application of an orthogonal transformation. We start with $a_{0,t} = X_t$ and set for $j = 1, 2, \ldots, J \leq \log_2(n)$

$$a_{j,t} = \sum_{\ell=0}^{L-1} g_\ell \; a_{j-1,2t+1-\ell}, \quad d_{j,t} = \sum_{\ell=0}^{L-1} h_\ell \; a_{j-1,2t+1-\ell} \;\; (t = 0, 1, \ldots, 2^{-j}n - 1).$$

(all indices are extended periodically). In words, we take the coefficients $a_{j-1,t}$ for odd times and apply to them two linear filters with impulse response coefficients g_ℓ and $h_\ell = (-1)^\ell g_{L-\ell-1}$, respectively. The coefficients g_ℓ are defined through the father wavelet (details omitted). They can be chosen arbitrarily subject to the constraints that L must be even and

$$\sum_{\ell=0}^{L-1} g_\ell = \sqrt{2}, \qquad \sum_{\ell=0}^{L-1-2n} g_\ell g_{\ell+2n} = \delta_{n,0} \;\; (n = 0, 1, \ldots, L/2 - 1).$$

For $L = 2, 4$ there is essentially only one solution, e.g. for $L = 2$ we have $g_0 = g_1 = 1/\sqrt{2}$. For $L \geq 6$, there are several solutions.

Because the discrete wavelet transform is a product of orthogonal linear transformations and thus is again linear and orthogonal, the computation of the inverse is easy. For smoothing, one typically sets $d_{j,k}$ and $a_{J,k}$ equal to zero if their absolute value is small and then applies the inverse transform. This retains features in the data which are not smooth in a conventional sense.

In the maximal overlap discrete wavelet transform, one uses the above recursions without omitting coefficients $a_{j-1,t}$ for even t:

$$\tilde{a}_{j,t} = 2^{-j/2} \sum_{\ell=0}^{L-1} g_\ell \; \tilde{a}_{j-1,t-2^{j-1}\ell}, \quad \tilde{d}_{j,t} = 2^{-j/2} \sum_{\ell=0}^{L-1} h_\ell \; \tilde{a}_{j-1,t-2^{j-1}\ell} \;\; (t = 0, 1, \ldots, n-1).$$

This creates redundancies, but is sometimes easier for a time-frequency interpretation.

If X_t is a stochastic process, the amplitudes $a_{j,t}$ and $d_{j,t}$ are random variables, and one can study their distributions. Because the wavelet transform is orthogonal, these amplitudes are again i.i.d. for Gaussian white noise. It turns out that also under dependence they become approximately independent like the periodogram values. In addition, the average of the $a_{j,t}^2$ for fixed j is essentially an estimate of the spectrum integrated over the frequency interval $[2^{-j-1}, 2^{-j}]$. A key difference is however that this holds also for integrated processes: We only need that $(1-B)^d X_t$ is stationary for some $d < L/2$.

4 State space models

The following is meant just as a quick overview of state space models[2]

[2] most parts here are in analogy (but simplified) from a book chapter “State Space and Hidden Markov Models” by Prof. Hansruedi Künsch

4.1 General state space models/ Hidden Markov models

General state space models (or Hidden Markov models - HMM) consist of

(i) An unobserved (latent) state process (X_t) with Markovian dependence

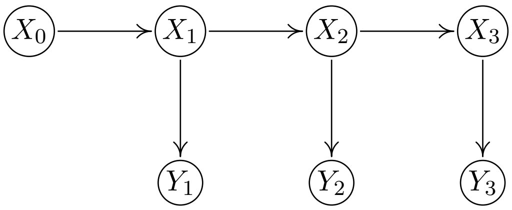
(ii) Observations (Y_t) which are derived from X_t.

Concretely, this means

(i) $X_0, X_1, X_2, \dots$ is a Markov chain

(ii) Conditionally on (X_t), all Y_t are independent and depend only on X_t.

As a graphical model, can use a directed acyclic graph:



The graphical model is a convenient way of writing down assumptions (i) and (ii). For a general joint distribution of $(X_0, X_1, \dots, X_T, Y_1, \dots Y_T)$ we can always write the density as

$$f_0(X_0)\prod_{t=1}^{T} f_t(X_t|X_0, X_1, \dots, X_{t-1}, Y_1, \dots, Y_{t-1})g_t(Y_t|X_0, X_1, \dots, X_{t-1}, X_t, Y_1, \dots, Y_{t-1})$$

for appropriate conditional densities f_t, $t = 0, \dots, T$ and g_t, $t = 1, \dots, T$. The model above is equivalent to the joint density having a simpler factorization

$$f_0(X_0)\prod_{t=1}^{T} f_t(X_t|X_{t-1})g_t(Y_t|X_t).$$

Note that (X_t) is a Markov chain and also (Z_t) is a Markov chain if we concatenate $Z_t = (X_t, Y_t)$. However, the observations (Y_t) on their own are not a Markov chain and exhibit more complex time-dependencies. The HMM allows, in other words, to model such more complex time-dependencies in the observations by a simple Markov model.

Goals of HMM analysis include

(a) Given observations $y_1, \dots, y_T$ and **known** transition densities/probabilities $f_t(X_t|X_{t-1})$ and $g_t(Y_t|X_t)$, provide inference for the underlying state vector $x_0, \dots, x_T$ (there are several forms of different inference which we return to).

(b) Given observations $y_1, \dots, y_T$ and **unknown** transition densities/probabilities $f_t(X_t|X_{t-1})$ and $g_t(Y_t|X_t)$, provide inference for the underlying state vector $x_0, \dots, x_T$ and for f_t and g_t simultaneously, either in a Bayesian or frequentist form

In the following, will first assume we are in setting (a), that is the transition densities are assumed to be known.

Examples for state space models:

(i) **Linear state space model** for $X_t \in \mathbb{R}^p$ and $Y_t \in \mathbb{R}^q$:

$$\begin{aligned} X_t &= \mathbf{G}_t X_{t-1} + V_t \\ Y_t &= \mathbf{H}_t X_t + W_t, \end{aligned}$$

where $\mathbf{G}_t \in \mathbb{R}^{p\times p}$ and $\mathbf{H}_t \in \mathbb{R}^{q\times p}$. The state vector X_t could for example be position and velocity of a moving object and Y_t are noisy measurements of the objects location. Goal is to infer X_t as accurately as possible. In case of Gaussian error terms there is an explicit solution (Kalman filter).

(ii) **ARMA models.** Let (Y_t) be a causal and invertible ARMA(p,q) process. Can be written as an HMM. Take as an example the previously discussed case of an AR(p) model:

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + W_t.$$

Define X_t as the collection of the past p observations $X_t := (Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_{t-p+1})^t$. Then

$$\begin{aligned} X_t &= \mathbf{\Phi}_t X_{t-1} + \eta_t \\ Y_t &= \mathbf{H}_t X_t + \varepsilon_t, \end{aligned}$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 & \dots & \phi_{p-1} & \phi_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix}, \ \eta_t = \begin{pmatrix} W_t \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} \ \mathbf{H_t} = (1, 0, 0, \dots, 0), \text{ and } \varepsilon_t \equiv 0.$$

Advantages of writing an ARMA process as a state space model include the ability to deal easily with missing data (for example by setting $H_t = (0, 0, \dots, 0)$ and $Y - t = 0$ at times t where we have missing data) and the ability to introduce a different type of outlier in the noise distribution (using η_t and ε_t respectively).

(iii) **Speech recognition.** The sequence (X_t) can be seen as the hidden sequence of words a speaker is trying to say and Y_t are the observed sound measurements.

(iv) **Biological examples** include ion-channel analysis (determining whether an ion gate is on or off–the underlying state $X_t \in \{0, 1\}$) based on noisy measurements Y_t of the current flowing through the gate. Includes also DNA analysis where the index of time is taken by the index of position along the chromosome and we can try to infer for example regions with heightened copynumbers of so-called CG-islands (areas where the acids C and G appear more often than A and T).

(v) **Physics/meteorology.** State X_t includes all relevant atmospheric variables at time t. Transition dynamics of X_t are given by a underlying physics (and approximated by meteorological models). The observations Y_t can be satellite measurements, wind and rainfall sensors etc. that help to infer the true underlying state.

4.2 Discrete state space models

Inference is easier (also notationally) for discrete state space models, where without limitation of generality

$$\begin{aligned} X_t &\in \{1, \ldots, \ell\} \\ Y_t &\in \{1, \ldots, m\}. \end{aligned}$$

Joint density factorizes again as

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_T, Y_1 = y_1, Y_2 = y_2, \ldots Y_T = y_T) = \\ &P(X_0 = x_0) \cdot \prod_{t=1}^{T} \Big[P(X_t = x_t | X_{t-1} = x_{t-1}) \cdot P(Y_t = y_t | X_t = x_t) \Big] \end{aligned}$$

Or, taking the logarithm,

$$\begin{aligned} &\log P(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_T, Y_1 = y_1, Y_2 = y_2, \ldots Y_T = y_T) = \\ &\log P(X_0 = x_0) + \sum_{t=1}^{T} \Big[\log P(X_t = x_t | X_{t-1} = x_{t-1}) + \log P(Y_t = y_t | X_t = x_t) \Big] \end{aligned} \tag{21}$$

Let matrices $\mathbf{A} \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ describe the transition probabilities $X_{t-1} \to X_t$ and $X_t \to Y_t$ in the sense that

$$\begin{aligned} P(X_t = j' | X_{t-1} = j) &= \mathbf{A}_{j',j} \\ P(Y_t = o | X_t = j) &= \mathbf{B}_{o,j} \end{aligned}$$

(Note that often people work with $\mathbf{A}^t$ instead of $\mathbf{A}$ and $\mathbf{B}^t$ instead of $\mathbf{B}$ but for our purposes it is more convenient to define the matrices as above.) We have that (using the fact that conditional probabilities have to be positive and sum to 1):

$$\begin{aligned} \sum_{j'=1}^{\ell} \mathbf{A}_{j',j} = 1 \;\; &\text{for all } j \in \{1, \ldots, \ell\} \text{ (matrix } \mathbf{A} \text{ is column-normalised)} \\ \text{and } \sum_{o=1}^{m} \mathbf{B}_{o,j} = 1 \;\; &\text{for all } j \in \{1, \ldots, \ell\} \text{ (matrix } \mathbf{B} \text{ is column-normalised)} \\ \text{and } \mathbf{A}_{j',j} \geq 0 \text{ and } \mathbf{B}_{o,j} \geq 0 &\text{ for all } j, j' \in \{1, \ldots, \ell\} \text{ and } o \in \{1, \ldots, m\}. \end{aligned}$$

4.3 Filtering, smoothing and prediction

Let $\pi^0 \in \mathbb{R}^\ell$ be the initial/prior distribution of X_0:

$$\pi_j^0 := P(X_0 = j) \text{ for all } j = 1, \dots, \ell.$$

Let $y_s^t = (y_s, \dots, y_t)$ be a vector of observations.

Goal: find conditional distribution $P(X_{t+k}|y_s^t)$. This is called

i) **Prediction** if $k > 0$

ii) **Filtering** if $k = 0$

iii) **Smoothing** if $k < 0$.

Will here look mostly at filtering and prediction.

Prediction. The prediction problem can be solved iteratively and hence reduced to filtering.

Let $\pi^{t+k|t}$ be the conditional distribution of X_{t+k}, given y_1^t in the sense that for all $j \in \{1, \dots, \ell\}$,

$$\pi_j^{t+k|t} := P(X_{t+k} = j|y_1^t).$$

We can now get a recursion for $\pi^{t+k|t}$ (a recursion in k) by conditioning on X_{t+k-1} and using the conditional independence between X_{t+k} and Y_1^t given X_{t+k-1}:

$$\begin{aligned}
\pi_j^{t+k|t} &= P(X_{t+k} = j|y_1^t) \\
&= \sum_{j'=1}^{\ell} P(X_{t+k} = j|X_{t+k-1} = j', y_1^t) \cdot P(X_{t+k-1} = j'|y_1^t) \\
&= \sum_{j'=1}^{\ell} P(X_{t+k} = j|X_{t+k-1} = j') \cdot P(X_{t+k-1} = j'|y_1^t) \\
&= \sum_{j'=1}^{\ell} \mathbf{A}_{j,j'} \cdot \pi_{j'}^{t+k-1|t}.
\end{aligned}$$

Or, in vector form,

$$\pi^{t+k|t} = \mathbf{A}\ \pi^{t+k-1|t}.$$

Reiterating back to time t, we get

$$\pi^{t+k|t} = \mathbf{A}^k\ \pi^{t|t}.$$

and we have thus reduced it to a filtering problem since $\pi^{t|t}$ is the conditional distribution of X_t, given $(y_1, \dots, y_t)$.

The distribution of Y_{t+k}, given y_1^t, follows by conditioning on X_{t+k} in similar form. If we set

$$p_o^{t+k|t} := P(Y_{t+k} = o|y_1^t),$$

then, by conditioning,

$$\begin{aligned} p_o^{t+k|t} &= P(Y_{t+k} = o|y_1^t) \\ &= \sum_{j=1}^{\ell} P(Y_{t+k} = o|X_{t+k} = j, y_1^t) \cdot P(X_{t+k} = j|y_1^t) \\ &= \sum_{j=1}^{\ell} P(Y_{t+k} = o|X_{t+k} = j) \cdot P(X_{t+k} = j|y_1^t) \end{aligned}$$

In vector form,

$$p^{t+k|t} = \mathbf{B} \cdot \pi^{t+k|t}.$$

Substituting from the result above , we can also write it as

$$p^{t+k|t} = \mathbf{B} \cdot \mathbf{A}^k \cdot \pi^{t|t}.$$

and we are going to look at the filtering distribution $\pi^{t|t}$ next.

Filtering. We want a recursion for the filtering density $\pi_j^{t|t} = P(X_t = j|y_1^t)$, which is also used in the prediction tasks. We can again use conditional independence of Y_1^t and Y_{t+1}, given X_{t+1}, and Bayes formula to get the desired recursion.

Recall that for two events A, B Bayes formula derives from

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A) = P(A, B),$$

and can be written as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

We can furthermore condition on yet another event C:

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}.$$

Setting

$$\begin{aligned} A &= \{X_{t+1} = j\} \\ B &= \{Y_{t+1} = y_{t+1}\} \\ C &= \{Y_1^t = y_1^t\}, \end{aligned}$$

we get for the desired filtering density

$$\begin{aligned}\pi_j^{t+1|t+1} &= P(X_{t+1} = j|y_1^{t+1}) \\ &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j, Y_1^t = y_1^t)P(X_{t+1} = j|Y_1^t = y_1^t)}{P(Y_{t+1} = y_{t+1}|Y_1^t = y_1^t)}.\end{aligned}$$

Using the conditional independencies at this point, we can simplify to

$$\begin{aligned}\pi_j^{t+1|t+1} &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j)P(X_{t+1} = j|Y_1^t = y_1^t)}{P(Y_{t+1} = y_{t+1}|Y_1^t = y_1^t)} \\ &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j)P(X_{t+1} = j|Y_1^t = y_1^t)}{\sum_{j'=1}^{\ell} P(Y_{t+1} = y_{t+1}|X_{t+1} = j')P(X_{t+1} = j'|Y_1^t = y_1^t)} \\ &= \frac{\pi_j^{t+1|t}\mathbf{B}_{y_{t+1},j}}{\sum_{j'=1}^{\ell} \pi_{j'}^{t+1|t}\mathbf{B}_{y_{t+1},j'}}\end{aligned}$$

The recursion works thus schematically in computation as

$$\pi^{t|t} \to \pi^{t+1|t} \to \pi^{t+1|t+1} \to \ldots,$$

where the second step $\pi^{t+1|t} \to \pi^{t+1|t+1}$ requires the new observation y_{t+1} at time $t+1$.

Using from the prediction task the recursion for $\pi^{t+1|t}$, we can directly write the recursion for $\pi^{t|t}$ without going via the prediction density as

$$\pi_j^{t+1|t+1} = \frac{(\mathbf{A}\pi^{t|t})_j\mathbf{B}_{y_{t+1},j}}{\sum_{j'=1}^{\ell}(\mathbf{A}\pi^{t|t})_{j'}\mathbf{B}_{y_{t+1},j'}}.$$

The denominator can be seen as a normalisation that ensures

$$\sum_{j=1}^{\ell} \pi_j^{t|t} = 1 \text{ for all times } t$$

(and can conveniently by implemented by such a normalisation without having to compute the denominator explicitely).

4.4 Posterior mode, viterbi and forward-backward algorithms and dynamic programming

The filtering approach yields posterior densities of, say, X_T, given $y_1, \ldots, y_T$. It does not and cannot answer questions about the most likely sequence $x_0^T = (x_0, \ldots, x_T)$ under the made observations, which is given by

$$\hat{x}_0^T = \text{argmax}_{x_0^T}\ P(X_0^T = x_0^T|y_1^T).$$

The most likely sequence can be computed with dynamic programming in a forward and backwards recursion.

Taking the log and using the previous decomposition (21),

$$\begin{aligned}\log P(X_0^T = x_0^T | Y_1^T = y_1^T) &\propto \log P(X_0^T = x_0^T, Y_1^T = y_1^T) \\ &= \log P(X_0 = x_0, X_1 = x_1, \ldots, X_T = x_T, Y_1 = y_1, Y_2 = y_2, \ldots Y_T = y_T) \\ &= \log P(X_0 = x_0) + \sum_{t=1}^{T} \Big[\log P(X_t = x_t | X_{t-1} = x_{t-1}) + \log P(Y_t = y_t | X_t = x_t)\Big] \\ &= \log \pi^0(x_0) + \sum_{t=1}^{T} \Big[\log \mathbf{A}_{x_t, x_{t-1}} + \log \mathbf{B}_{y_t, x_t}\Big]\end{aligned}$$

The optimization problem can be seen as one of minimizing the cost of traversing time from 0 to T and passing through $x_0, x_1, \ldots, x_T$ along the way where we incur

(i) Cost for passing through the initial state x_0 which depends on the prior distribution π^0 for X_0:

$$-\log \pi^0(x_0)$$

(ii) Cost for passing from state x_{t-1} to state x_t at every time $t = 1, \ldots, T$:

$$-\log(\mathbf{A}_{x_t, x_{t-1}})$$

(iii) Cost for state x_t at every time $t = 1, \ldots, T$ (depending on the observation y_t at this point).

$$-\log(\mathbf{B}_{y_t, x_t})$$

We can first make a forward recursion going thorugh $t = 1, \ldots, T$, and record in $\psi_t(x)$ the lowest cost (negative log-likelihood) achievable up to this point t in time if we end up in position x at time t, that is

$$\psi_t(x) = \min_{(x_0, \ldots, x_{t-1}), x_t = x} \Big(- \log \pi^0(x_0) + \sum_{t'=1}^{t} \big[- \log \mathbf{A}_{x_{t'}, x_{t'-1}} + - \log \mathbf{B}_{y_{t'}, x_{t'}} \big] \Big).$$

Note that

$$\hat{x} = \text{argmin}_{(x_0, \ldots, x_T)} \Big(- \log \pi^0(x_0) + \sum_{t'=1}^{T} \big[- \log \mathbf{A}_{x_{t'}, x_{t'-1}} + - \log \mathbf{B}_{y_{t'}, x_{t'}} \big] \Big)$$

and hence

$$\hat{x}_T = \text{argmax}_x \; \psi_T(x)$$

The function ψ_t can be calculated now for $t = 1, 2, \ldots$ in a forward recursion as

$$\begin{aligned}
\psi_0(x) &= -\log\pi^0(x) \\
\psi_t(x) &= \min_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}}) - \log(\mathbf{B}_{y_t,x})\big) \qquad \text{for } t = 1, \ldots, T.
\end{aligned}$$

We also record the value of x_{t-1} (the back-pointer) for which the minimum was achieved at time $t-1$ if we pass through x at time t as

$$\begin{aligned}
\xi_{t-1}(x) &= \text{argmin}_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}}) - \log(\mathbf{B}_{y_t,x})\big) \\
&= \text{argmin}_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}})\big).
\end{aligned}$$

The optimal path $(\hat{x}_0, \ldots, \hat{x}_T)$ is then calculated in a backwards recursion as

$$\begin{aligned}
\hat{x}_T &= \text{argmin}_x \psi_T(x) \\
\hat{x}_{t-1} &= \xi_{t-1}(\hat{x}_t) \qquad \text{for } t = T-1, T-2, \ldots, 0.
\end{aligned}$$

This is sometimes called the Viterbi algorithm.

4.5 Parameter estimation via the EM-algorithm

Assume we have a distribution with discrete observed variables Y and latent X and unknown parameter θ (same ideas work for continuous variables). We would like to get the Maximum-likelihood estimate of θ as

$$\hat{\theta} = \text{argmax}_\theta \; \ell(\theta),$$

where the log-likelihood $\ell(\theta)$ is given by $\log P_\theta(Y = y)$ if the observations of Y are y.

The problem is that $P_\theta(Y)$ is not easily available in tractable form. What is available is the likelihood $P_\theta(Y, X)$ if we could observe the latents X as well as. The EM (Expectation-Maximization; also called Baum-Welch for HMMs) algorithm greedily optimizes the likelihood by alternating between

(i) estimating the latent variables X, given the observed variables Y and the current parameter estimate, and then

(ii) updating the parameter estimates in a second step.

Starting from some initial estimate $\theta^{(1)}$, the steps are for an iteration $t = 1, \ldots,$

E-step (Expectation): Compute the conditional distribution of the latent variables X, given the observed variables and the current parameter estimates $\theta^{(t)}$:

$$P_{\hat{\theta}^{(t)}}(X|Y = y).$$

This is similar to type of inference we discussed. Define the expected log-likelihood under the distribution of X implied by the current parameter estimate as

$$Q_t(\theta) := E_{\hat{\theta}^{(t)}}\Big[\log P_\theta(Y = y, X)|Y = y\Big],$$

where the expectation is with respect to the random X, conditional on $Y = y$ and the current parameter estimate $\hat{\theta}^{(t)}$.

M-step (Maximization): update the parameters as

$$\hat{\theta}^{(t+1)} = \text{argmax}_\theta\ Q_t(\theta).$$

We get monotonically increasing likelihood

$$\ell(\hat{\theta}^{(t)}) \geq \ell(\hat{\theta}^{(t-1)}),$$

that is the parameter estimates $\hat{\theta}^{(t)}$ will converge to a **local** maximum of the likelihood for $t \to \infty$. Depending on the starting value we might reach the global optimum but this is not guaranteed.

The EM algorithm appears very often in practice. Compare also the applications in eg clustering we discuss. Sometimes the expectation is replaced with just computing the most-likely state of x', given $Y = y$ and $\hat{\theta}^{(t)}$ and setting $Q_t(\theta) := \log P_\theta(Y = y, x')$. This is what happens in the K-means clustering algorithm, for example. For HMMs it corresponds to computing the most likely sequence with the Viterbi algorithm, as discussed. The approach is sometimes referred to as hard-EM. The result will again depend on the chosen starting values in general.

But why do we get monotonically increasing likelihood? The proof sheds some more light on EM. To start with, it holds for all functions f over the space $\mathcal{X}$ of the hidden variables with $\sum_{x\in\mathcal{X}} f(x) = 1$ and $f(x) \geq 0$ for all $x \in \mathcal{X}$,

$$\begin{aligned}\ell(\theta) &= \log P_\theta(Y = y)\\ &= \log\sum_{x\in\mathcal{X}} P_\theta(Y = y, X = x)\\ &= \log\sum_{x\in\mathcal{X}} f(x)\frac{P_\theta(Y = y, X = x)}{f(x)}\\ &\geq \sum_{x\in\mathcal{X}} f(x)\log\frac{P_\theta(Y = y, X = x)}{f(x)},\end{aligned}$$

where the last inequality uses Jensens inequality and the fact that log is a concave function. Note that equality holds iff

$$\frac{P_\theta(Y = y, X = x)}{f(x)}$$

does not depend on x, for example if $f(x) = P_\theta(X = x|Y = y)$.

Hence there exists a constant $c > 0$ at each time-step[3] such that

$$\ell(\theta') \geq Q_t(\theta') + c \;\text{ for all } \theta' \qquad \text{and } \ell(\hat{\theta}^{(t)}) = Q_t(\hat{\theta}^{(t)}) + c.$$

Hence

$$\ell(\hat{\theta}^{(t+1)}) \geq Q_t(\hat{\theta}^{(t+1)}) + c \geq Q_t(\hat{\theta}^{(t)}) + c = \ell(\hat{\theta}^{(t)}),$$

where the first inequality is due to the argument just above and the second is true as $\hat{\theta}^{(t+1)}$ is, by definition, maximizing Q_t.

4.6 Kalman filter

Suppose we have a linear dynamical system

$$\begin{aligned} X_t &= AX_{t-1} + V_t \\ Y_t &= BX_t + W_t, \end{aligned}$$

where $X_t \in \mathbb{R}^p$ is the latent state, $Y_t \in \mathbb{R}^q$ the made observations, W_t the so-called process noise and V_t the measurement noise.

Under a Gaussian noise assumption

$$W_t \sim \mathcal{N}(0, W), \qquad V_t \sim \mathcal{N}(0, V),$$

the joint vector of latent and observations over $t = 1, \ldots, T$ will have a joint Gaussian distribution. In principle it is thus easy to derive the conditional distribution of, say, $X_t|Y_1^t$. Let $Z \in \mathbb{R}^p$ be a random vector with a Gaussian distribution,

$$Z \sim \mathcal{N}(\mu, \Sigma).$$

Then the conditional distribution of Z_k, conditional on $Z_S = z_s$ for some $S \subseteq \{1, \ldots, p\}$, will again be Gaussian

$$Z_k|Z_s = z_s \quad \sim \mathcal{N}(\mu_{k|S}, \Sigma_{k,S}),$$

with

$$\begin{aligned} \mu_{k|S} &= \mu_k + \Sigma_{k,S}\Sigma_{S,S}^{-1}(z_S - \mu_S) \\ \Sigma_{k|S} &= \Sigma_{k,k} - \Sigma_{k,S}^t\Sigma_{S,S}^{-1}\Sigma_{S,k} \end{aligned}$$

The problem with the direct approach is that the dimensionality of S grows like pT if we condition on the observations $Y_1^T = (Y_1, \ldots, Y_T)$.

[3] namely entropy of $f(x) = P_{\hat{\theta}^{(t)}}(X = x|Y = y)$ as entropy is $-\sum_x f(x)\text{log}f(x)$

Using the structure of the HMM again and the same message-passing as in the discrete case, we can define

$$\begin{aligned}\hat{X}_{t|t} &= E(X_t|Y_1^t)\\ \hat{X}_{t+1|t} &= E(X_{t+1}|Y_1^t)\\ \Sigma_{t|t} &= Cov(X_t|Y_1^t)\\ \Sigma_{t+1|t} &= Cov(X_{t+1}|Y_1^t)\end{aligned}$$

The updates are usually split again into two part,s the time- and the measurement update. The time-update concerns

$$\begin{aligned}\hat{X}_{t|t} &\to \hat{X}_{t+1|t}\\ \Sigma_{t|t} &\to \Sigma_{t+1|t},\end{aligned}$$

while the measurement update concerns

$$\begin{aligned}\hat{X}_{t+1|t} &\to \hat{X}_{t+1|t+1}\\ \Sigma_{t+1|t} &\to \Sigma_{t+1|t+1},\end{aligned}$$

taking into account the new observation made at time t.

Conditioning on Y_1^t, we get

$$\begin{aligned}X_{t+1}|Y_1^t &= (AX_t + V_t)|Y_1^t = AX_t|Y_1^t + V_t\\ Y_{t+1}|Y_1^t &= (BX_{t+1} + W_t)|Y_1^t = BX_{t+1}|Y_1^t + W_t.\end{aligned}$$

The first equation yields the so-called **time-update** (updating $t \to t+1$ without using the new observation at time $t+1$)

$$\begin{aligned}\hat{X}_{t+1|t} &= E(X_{t+1}|Y_1^t) = A\hat{X}_{t|t}\\ \Sigma_{t+1|t} &= A\Sigma_{t|t}A^t + V\end{aligned}$$

The second equation yields the measurement update (where the new observation Y_{t+1} is used to update the conditional distribution). Note that the distribution of $(X_{t+1}, Y_{t+1})|Y_1^t$ has a multivariate Gaussian distribution

$$(X_{t+1}, Y_{t+1})|Y_1^t \quad \sim \quad \mathcal{N}(\mu, S),$$

with

$$\mu = \begin{pmatrix}\hat{X}_{t+1|t}\\ B\hat{X}_{t+1|t}\end{pmatrix}, \qquad S = \begin{pmatrix}\Sigma_{t+1|t} & \Sigma_{t+1|t}^t B^t\\ B\Sigma_{t+1|t} & B^t\Sigma_{t+1|t}B + W\end{pmatrix}.$$

Hence we get the **measurement update** as

$$\begin{aligned}\hat{X}_{t+1|t+1} &= \hat{X}_{t+1|t} + \Sigma_{t+1|t}^t B^t (B^t\Sigma_{t+1|t}B + W)^{-1}(Y_t - B\hat{X}_{t+1|t})\\ \Sigma_{t+1|t+1} &= \Sigma_{t+1|t} - \Sigma_{t+1|t}^t B^t (B^t\Sigma_{t+1|t}B + W)^{-1}B\Sigma_{t+1|t}\end{aligned}$$

Perhaps surprisingly, the error covariance can be computed ahead of time (without seeing any observations).

Steady-state Kalman filter If $\Sigma_{t+1|t}$ converges to a Σ^* (which it will in general), then Σ^* is the solution of a Ricatti-type equation

$$\Sigma^* = A\Sigma^* A^t + V - A(\Sigma^*)^t B^t (B^t \Sigma^* B + W)^{-1} B \Sigma^* A^t.$$

The estimated means follow then the recursion

$$\hat{X}_{t+1|t} = A\hat{X}_{t|t-1} + L(Y_t - B\hat{X}_{t|t-1}),$$

where $L = A(\Sigma^*)^t B^t (B^t \Sigma^* B + W)^{-1}$ is the so-called Kalman gain. The first term updates the guess according to the dynamics of the system while the second corrects it by using the newly available information via the new observation.