

A times series is  $X_t, t \in 1, \dots, T$  where

$X_t \sim F_t$  and  $F_t$  is a univariate distribution function.

An observed times series is seen as a realization of the random vector  $X_t = \{x_1, \dots, x_T\}$ .

Notice therefore that for each random variable  $X_t \in F_t$  you observe a single observations  $x_t$ .

Without imposing some restriction on the series it would be impossible to make any conjecture on  $F_t$  from a single observation.

We therefore introduce the concept of **stationarity**.

Under this idea the probabilistic nature of the series does not change over time, such that the joint distribution of  $X_t, \dots, X_{tk}$  does coincide with  $X_s, \dots, X_{sk}$  if  $t, s, k$ .

As this is impossible to verify in practice as you would need an exquisitely long time series, the concept of weak stationarity is introduced under which the condition on the probabilistic character of the series is restricted to the first moments.

$$E(X_t) = \mu$$

$$\text{Var}(X_t) = \sigma^2$$

$$\text{Cov}(X_t, X_{t+h}) = \gamma(h)$$

Notice that times series might be stationary even when

$$\text{Cov}(X_t, X_{t+h}) = \gamma(h) \neq 0, E(X_t | X_{t+1}, \dots, X_{t+k}) = \mu_t \text{ and}$$

$\text{Var}(X_t | X_{t+1}, \dots, X_{t+k}) = \sigma_t^2$ ; i.e. there is some minor short run behaviour in the times-series and deviation from the long run moments.

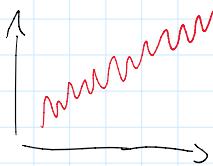
### Times-Series Decomposition

When analyzing a series you might find a suitable stochastic or deterministic model describing it in its entirety. A second approach would be to decompose it into single parts/components and model them individually. When using such models for forecasting you might need to put them back together by reverse engineering and putting the components together.

Two major approaches in this sense are:

1. Additive decomposition

$$X_t = m_t + s_t + r_t$$



2. Multiplicative Decomposition

$$X_t = m_t \cdot s_t \cdot r_t$$

$m$  = trend component  
 $s$  = seasonal component  
 $r$  = residual



### Modeling Trend Component:

Options:

1. Removing the trend by differencing the series

$$X_t = m_t + r_t \quad \{ \text{true model}$$

$$\tilde{X}_t = X_t - X_{t-1} = R_t - R_{t-1} \quad \{ \text{no trend.}$$

Notice that here the trend is fully eliminated just if it is exactly linear:  $m_t = \alpha + \beta t$

$\Rightarrow$  in such case

$$\tilde{X}_t = X_t - X_{t-1} = \beta(t-t-1) + R_t - R_{t-1}$$

but notice that if  $m_t = \alpha + \beta t^2$

$$\begin{aligned} \tilde{X}_t &= X_t - X_{t-1} = \beta(t^2) - \beta(t-1)^2 + R_t - R_{t-1} \\ &= \cancel{\beta t^2} - (\cancel{\beta t^2} + \cancel{\beta t - \beta}) + R_t - R_{t-1} \end{aligned}$$

not fully removed.

It would need more complex differencing structures

Important moreover is to see that the dependence structures of  $R_t$  ≠ to the one of  $R_t - R_{t-1}$  so that new artificial dependencies might result.

$\hookrightarrow$  to see that

$$\begin{aligned} \text{Cov}(\tilde{X}_t, \tilde{X}_{t-1}) &= \text{Cov}(R_t - R_{t-1}, R_{t-1} - R_{t-2}) \\ &= -\text{Cov}(R_{t-1}, R_{t-1}) \neq 0 \quad \text{even when } R_t \sim N(0, \sigma^2) \forall t. \end{aligned}$$

Notice that you might use the same method for removing the seasonal component.

$$\tilde{X}_t = (1 - \beta^p) X_t = X_t - X_{t-p}$$

This will allow you to remove the seasonal component at lag  $p$ .

Important: Notice that when removing the seasonal component in such a way you are automatically removing a trend component at yearly lags.

↳ For instance in the case of a linear component

$$X_t = m_t + R_t \quad m_t = \beta t + \alpha$$

$$X_t - X_{t-p} = \beta p + R_t - R_{t-p}$$

not dependent on time!

so linear trend  $\downarrow$

## Trend Removal - Smoothers

A second option to remove the trends is to use smoothers. These can be

- moving averages
- linear regressions
- loess smoothers

For moving averages you would model the trend via

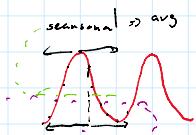
$$m_i = \sum_{j=-p}^q a_j X_{t+j}$$

so that you can apply different weights  $a_j$  for different lags.

↳ In the plain vanilla running mean  $p=q$  and  $a_i = \frac{1}{(2p+1)}$

## Smoothing in the case of Seasonal Data.

↳ Here you have to make sure that you eliminate the seasonal trend by taking the average over the entire seasonal period and removing such component.



## Parametric Models

One of the issues of removing trend and seasonality by simple smoothers is that such an option in terms of degrees of freedom.

↳ Due to the averaging process you loose many degrees of freedom and therefore datapoints.

A more parsimonious way to remove trend and seasonality components is in this sense to use models of the form:

$$X_t = \beta_0 + \underbrace{\beta_1 t}_{\substack{\text{parametric} \\ \text{linear} \\ \text{trend}}} + \underbrace{\beta_2 \sin(2\pi t) + \beta_3 \cos(2\pi t)}_{\substack{\text{seasonal} \\ \text{parametric} \\ \text{trend}}} + R_t$$

} just 3 d.f.

↳ Notice: In the case of correlated residuals - very likely in the series, you might not get BLUE estimators with OLS.

Should correlation exists then you would well yet unbiased estimators however their SE would not be reliable.

A different possibility to model trend and seasonal component of a series is to use a GAM formulation at the form:

$$X_t = f(t) + a_i s_i + R_t$$

$\curvearrowleft$  flexible trend component  
 $\curvearrowleft$  dummy for seasonal component

⇒ this estimated via cross-validation. See lecture on regression for more.

↳ This allows to well model cases when a linear-cyclic (trend-seasonal) decomposition is not sensible. GAM models are also parsimonious in the choice of degrees of freedom.

## Autocorrelation

An important concept when analyzing a series in order to understand whether the stationary condition holds and whether there is significant presence of seasonality is the concept of autocorrelation.

In the times series field it is a natural extension to the cross-data covariance

$$\text{Corr}(X_{t+k}, X_t) = \frac{\text{Cov}(X_{t+k}, X_t)}{\sqrt{\text{Var}(X_{t+k}) \cdot \text{Var}(X_t)}}$$

Notice that in the case of a truly stationary series; this simplifies to:

$$\text{Corr}(X_{t+k}, X_t) = \frac{\text{Cov}(X_{t+k}, X_t)}{\text{Var}(X)} = \rho(k) \Rightarrow \text{not depend on the time dimension anymore.}$$

Notice hence that in the most general case you compute the correlation at lag  $k$  as:

$$\hat{\rho}(k) = \frac{\sum_{s=1}^{n-k} (X_{s+k} - \bar{X}_{(k)}) (X_s - \bar{X}_{(1)})}{\sqrt{\sum_{s=k+1}^n (X_s - \bar{X}_{(k)})^2 \cdot \sum_{t=1}^{n-k} (X_t - \bar{X}_{(1)})}},$$

$$\text{where } \bar{X}_{(k)} = \frac{1}{n-k} \sum_{i=k+1}^n x_i, \quad \bar{X}_{(1)} = \frac{1}{n-k} \sum_{i=1}^{n-k} x_i$$

Important: From the very definition of the covariance and the estimation of it, it is clear that, for

$$\hat{\rho}(1), \text{ # } n-1 \text{ observations}$$

$$\hat{\rho}(2), \text{ # } n-2 \text{ observation}$$

:

$$\hat{\rho}(m); \text{ # } n-m \text{ observations}$$

So that with increasing lag our empirical estimators  $\hat{\rho}(k)$  become more variable and less trustworthy as their confidence intervals explode.

To mitigate the above effect, - that by the way you would obtain looking at the Pearson correlation coefficient of lagged scatterplots - it is recommended to work with plug-in estimators, i.e.

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}, \text{ where}$$

$$\hat{\gamma}(k) = \frac{1}{n} \sum_{s=1}^{n-k} (X_{s+k} - \bar{X})(X_s - \bar{X})$$

$$\hat{\gamma}(0) = \frac{1}{n} \sum_{s=1}^n (X_s - \bar{X})^2$$

$$\bar{X} = \frac{1}{n} \sum_{t=1}^n x_t$$

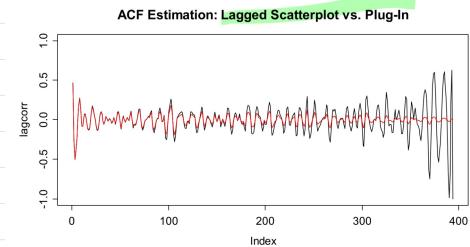
in such a way you would not get biased estimators but less volatile ones. You would hence get a more conservative estimation. This is what is used in the `acf()` function in R.

Notice finally that in the formula above:

Notice finally that in the formula above:

$$\hat{p}(k) = \frac{\cancel{\frac{1}{n}} \sum_{s=1}^{n-k} (x_{s+k} - \bar{x})(x_s - \bar{x})}{\cancel{\frac{1}{n}} \sum_{i=1}^n (x_i - \bar{x})^2}$$

So that there is a shrinkage towards 0 as the lag increases due to the denominator being  $\gg$  nominator.



Finally notice that plug-in estimators are consistent, meaning that  $\lim_{n \rightarrow \infty} \text{bias} \rightarrow 0$ ; and that for i.i.d. series:

$$\hat{p}(k) \sim N(0, \frac{1}{n})$$

→ from here you could infer CI for i.i.d. series. Notice however that for dependent series the CI properties are much harder to define.

**Important!**: notice finally that by mathematical constraint  $\sum_{k=1}^{n-1} \hat{p}(k) = -\frac{1}{2}$ . This generates quite severe artifacts it is worth to consider when looking at ACF plots.

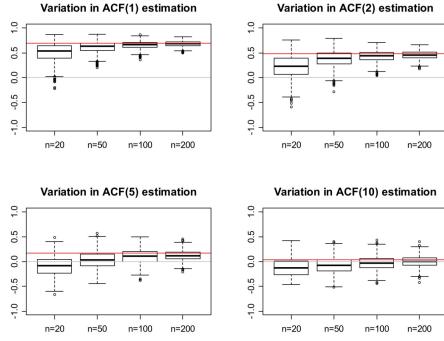
On outliers in the autocorrelation context:

These have an especially strong effect as when computing the correlations, each outlier enters multiple times in the covariance calculation.

In order to treat outliers in times series that have been identified as such by the expert it is advised to:

- replace the value with the global mean
- using a local mean, i.e. +/- 3 observations
- model based forecasting computation

On the simulation studies results



We observe that for "short" series with less than 100 observations, estimating the ACF is difficult: the  $\hat{p}(k)$  are strongly biased, and there is huge variability. Only for longer series, the consistency of the estimator "kicks in", and yields estimates which are reasonably precise. For lag  $k=10$ , on the other hand, we observe less bias, but the variability in the estimate remains large, even for "long" series.

## On the Confidence Interval for TS-Mean

This might be important when checking for a statistical significant mean in the series.

The idea is that for i.i.d. series the mean is easy to compute

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$$

and

$$\text{Var}(\hat{\mu}) = \frac{\sigma_x^2}{n}, \text{ so that you can plug-in the } \hat{\sigma}_x^2 \text{ to obtain your mean value.}$$

On the other hand in the case of dependence, it is true that:

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}\left(\frac{1}{n} \sum X_t\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum X_t\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \gamma(|i-j|) \\ &= \frac{\gamma(0)}{n^2} \sum_{i=1}^n \sum_{j=1}^n p(|i-j|) \\ &= \frac{\gamma(0)}{n^2} \cdot \left(n+2 \sum_{k=1}^{n-1} (n-k)p(k)\right) \end{aligned}$$

So that if the Variance of the mean is higher or lower than its i.i.d. counterpart depends on all of the autocorrelations  $p(k)$ .

Due to the difficult reliability on higher estimation lags it is in practice the compromise looks as follows

$$\hat{\mu} \pm 1.96 \cdot \sqrt{\frac{\gamma(0)}{n^2} \left( n + 2 \cdot \sum_{k=1}^{10 \log_2(n)} (n-k) \rho(k) \right)}$$

## Some Property of PACF

- theory omitted -

act at lag 1

$$\pi(1) = \rho(1)$$

↳ this natural as  $\text{Cov}(X_{t+1}, X_t | X_t) = \text{Cov}(X_{t+1}, X_t)$

Moreover for an AR(p) process:

$$\alpha_p = \pi(p)$$

$$\pi(j) = 0 \quad \forall j > p.$$

## Univariate Linear Models

We will briefly introduce AR, MA and ARMA models. All of them are based on the assumption that the series can be formulated as

$$X_t = \mu_t + E_t, \quad \text{with}$$

$E_t$  = white noise

$\mu_t$  = short-term mean of the series,  
this will be modeled by the above models. Notice that  $\mu_t \neq 0$  but  $\mu = c$ , i.e. the long run mean must be constant such that  $\mu_t$  can just be short term deviations.

As we will see shortly in the models,

$$\mu_t = f(X_{t-1}, \dots, X_{t-p}, E_{t-1}, \dots, E_{t-p})$$

where  $f(\cdot)$  is a linear function in its arguments.

## Auto regressive Models

An AR(p) model is of the form:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + E_t.$$

! Important: As mentioned above AR() models must be only fitted to stationary series. Treat the series before if that does not hold.

Notice that for an AR( $p$ ) model

$$\begin{aligned}\mu &= E(X_t) = E(a_1 X_{t-1} + \dots + a_p X_{t-p} + \epsilon_t) \\ &= (a_1 + \dots + a_p) \mu \quad \text{due to stationarity}\end{aligned}$$

$$\mu (1 - (a_1 + \dots + a_p)) = 0$$

$$\mu = 0$$

However,

$$\mu_t = E(X_t | X_{t-1}, \dots, X_{t-p}) = a_1 X_{t-1} + \dots + a_p X_{t-p} \neq 0$$

Finally, for the stationarity of the AR process a further requirement on the coefficients is necessary:

The (potentially complex) roots of the characteristic polynomial  $\Phi(B)$  must all exceed 1 in absolute value for an AR( $p$ ) process to be stationary.

This is easy to understand in the AR(1) case and for higher lags it is just an algebraic exercise.

Finally notice the following pattern in AR(1) models

$$p(k) = a^k$$

To see this:

$$\begin{aligned}p(1) &= \frac{\gamma(1)}{\gamma(0)} = \frac{\text{Cov}(a_1 X_{t-1} + \epsilon_t, X_{t-1})}{\gamma(0)} = \frac{a_1 \gamma(0) + 0}{\gamma(0)} \\ p(2) &= \frac{\gamma(2)}{\gamma(0)} = \frac{\text{Cov}(a_1^2 X_{t-2} + a_1 \epsilon_t, X_{t-1})}{\gamma(0)} = a_1^2 \\ &\vdots\end{aligned}$$

We thus observe an exponential decay for the ACF in case of AR(1).

On the Yule-Walker equations - The way to obtain ACF coef for AR( $p$ ) processes.

Consider an AR(3) process; then:

$$\begin{aligned}p(3) &= \gamma(0)^{-1} \text{Cov}(X_{t+3}, X_t) \\ &= \gamma(0)^{-1} \text{Cov}(a_1 X_{t+2} + a_2 X_{t+1} + a_3 X_t + \epsilon_t, X_t) \\ &= \gamma(0)^{-1} (a_1 \text{Cov}(X_{t+2}, X_t) + a_2 \text{Cov}(X_{t+1}, X_t) \\ &\quad + a_3 \text{Cov}(X_t, X_t))\end{aligned}$$

$$\begin{aligned}
 & + \alpha_3 \operatorname{Cov}(X_t, X_{t-1}) \\
 & = \alpha_1 p(2) + \alpha_2 p(1) + \overbrace{\alpha_3 p(0)}^{=1} \\
 p(1) & = \alpha_1 p(0) + \alpha_2 p(-1) + \alpha_3 p(-2) \\
 p(2) & = \alpha_1 p(1) + \alpha_2 p(0) + \alpha_3 p(-1)
 \end{aligned}$$

so that with the fact that  $\operatorname{Cov}(X_{t+k}, X_t) = \operatorname{Cov}(X_{t-k}, X_t)$   
due to stationarity you get a system of  $p \times p$  with  
 $p$  unknowns you might easily solve with standard methods.

Notice finally that AR( $p$ ) models rely just on the previous  $p$  observations in the series and are fully explained by them. It is clear that PACF  $\pi(j)$  with  $j > p = 0$ .

### On AR( $p$ ) model fit

This subchapter addresses the topic and AR coefficients estimation and the different methodologies

First of all the AR model must be selected. In order to do that consider the following heuristics of AR models:

- exponential decay of ACF
- clear-cut in PACF  $p < 5-10$ .
- stationary series

Notice that through the mathematical construct of  $\sum_{k=1}^{p-1} p(k) = -1/\epsilon$  and the empirical estimation issue; the PACF will hardly drop to exactly 0. It is then an expert judgement to tell when a drop can be considered as such.

Given the decision on the AR-order the estimation can be done via:

- OLS
- Burg's Algorithm
- Yule-Walker Approach
- MLE

### OLS

Here the idea is the one of estimating the coefficients of the model,

Here the idea is the one of estimating the coefficients of the model,

$$Y_t = m + \mu_t + E_t$$

In the following way  $m = \frac{1}{n} \sum_{t=1}^n Y_t$  and then to estimate

$$Y_t = 2 \bar{X} + \varepsilon_t \quad \text{via OLS without intercept}$$

Important is however to understand here that the first  $p$  observations, are used just as independent variables and the model is therefore fitted on  $n-p$  observations.

### Brug's Algorithm

This overcomes the asymmetry of the OLS above where the first  $p$  observations are lost.

It is based on the idea that an  $AR(p)$  process is also an  $AR(p)$  process

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$$

$$-\varepsilon_t + X_t - \alpha_1 X_{t-1} = \alpha_2 X_{t-2}$$

$$X_{t-2} = -\frac{\alpha_1}{\alpha_2} X_{t-1} + \frac{1}{\alpha_2} X_t - \frac{1}{\alpha_2} \varepsilon_t$$

$$\Leftrightarrow X_{t-2} = \hat{\alpha}_1 X_{t-1} + \hat{\alpha}_2^* X_t + \hat{\varepsilon}_t^*$$

In this sense the Brug's Algorithm minimizes:

$$\sum_{t=p+1}^n \left\{ (X_t - \sum_{k=1}^p \alpha_k X_{t-k})^2 + (X_{t-p} - \sum_{k=1}^p \alpha_k X_{t-p-k})^2 \right\}$$

Notice that the coeff is the same for the inverted AR model.

Nowadays generally accepted that the Brug's solution is better for finite samples.

### Yule-Walker Method

This involves in leveraging the  $p \times p$  system derived above and instead of computing the theoretical covariances given an  $AR(p)$  model with known coefficients you use the system to ~~infer~~ the  $AR(p)$  coefficients given the plug in estimator for the covariances.

This generally performs worse than the Brug's Algorithm when looking at the (Gaussian) Likelihood. It is therefore generally accepted to rather rely on the first.

## Maximum Likelihood estimator

This requires you to specify a distribution for the AR( $p$ ) process.

Then you would simultaneously estimate  $\{\mu, \sigma_E^2, \alpha_1, \dots, \alpha_p\}$ .

You might even easily test different parametrizations through the well known distribution of MLE estimators.

Finally notice that due to the simultaneous estimation of  $\sigma_E^2$  and the coefficients  $\{\alpha_1, \dots, \alpha_p\}$  a recursive algorithm is required.

## Moving Average Models

Here

$$\mu_t = f(\varepsilon_{t-1}, \dots, \varepsilon_{t-p}), \text{ in the specific } .$$

$$y_t = a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + \dots + a_p \varepsilon_{t-p} + \varepsilon_t$$

Notice that this process is always stationary independently from the coefficients, however it is not always i.i.d.

On the moments of MA( $p$ ).

$$E(X_t) = 0 \text{ by definition as } E(\varepsilon_t) = 0 \forall t.$$

and with  $\varepsilon_t \sim N(0, \sigma_E^2) \quad \forall t$ , i.e. gaussian white noise

$$\begin{aligned} \text{Var}(X_t) &= \text{Var}\left(\sum_{i=0}^p a_i \varepsilon_{t-i}\right) \\ &= \sum_{i=0}^p \sum_{j=0}^p \text{Cov}(a_j \varepsilon_{t-j}, a_i \varepsilon_{t-i}) \quad \left. \begin{array}{l} \text{As} \\ \text{Cov}(\varepsilon_{t-j}, \varepsilon_{t-i}) = 0 \\ \forall j \neq i \end{array} \right. \\ &= \sum_{j=0}^p a_j^2 \cdot \sigma_E^2 = \text{constant} \end{aligned}$$

with  $a_0 = 1$ .

Now given on MA( $p$ ) notice that

$$\begin{aligned} \gamma(1) &= \text{Cov}(X_t, X_{t-1}) \\ &= \text{Cov}(a_1 \varepsilon_{t-1} + \dots + a_p \varepsilon_{t-p}, a_1 \varepsilon_{t-2} + \dots + a_p \varepsilon_{t-p-1}) \end{aligned}$$

$$\begin{aligned}
 &= \text{Cov}(a_2 E_{t-2}, a_1 E_{t-2}) + \text{Cov}(a_3 E_{t-3}, a_2 E_{t-3}) + \dots \\
 &= a_2 \cdot a_1 \sigma_E^2 + a_3 \cdot a_2 \sigma_E^2 + \dots \\
 &= \text{constant}
 \end{aligned}$$

} this also proves  
 a relation among  
 the model's parameters  
 and the correlation.

Notice moreover that in the case of MA( $p$ ) processes:

$$\begin{aligned}
 \gamma(p+1) &= \text{Cov}(X_{t-p}, X_t) \\
 &= \text{Cov}(E_{t-p} + a_1 E_{t-p-1} + \dots + a_p E_{t-1}, a_1 E_{t-1} + \dots + a_p E_{t-p} + E_t) \\
 &= 0
 \end{aligned}$$

as no common  $E_{t-i}$ .

From the above we see that for MA( $p$ ) processes:

- Always stationary
- clear cut-off in the ACF
- finally notice for the exam that  $\rho(1)$  for MA( $N$ )  $\leq 0,5$ . Prove is trivial.

### Invertibility of MA processes:

Notice that the coefficients of MA processes are not identifiable as,

$$\rho(1) = \frac{\beta_1}{1 + \beta_1^2} = \frac{1/\beta_1}{1 + (1/\beta_1)^2} = \frac{1/\beta_1}{\beta_1^2 + 1} = \frac{1}{\beta_1^2 + 1}$$

so that for instance

$$X_t = \frac{1}{2} E_{t-1} + E_t$$

has the very same dependency structure as

$$X_t = 2 E_{t-1} + E_t$$

Notice now that in the case of an MA(1) it is possible to see:

$$\begin{aligned}
 E_t &= \beta_1 E_{t-1} - X_t \\
 &= \beta_1 (X_{t-1} - \beta_1 E_{t-2}) - X_t \\
 &= \beta_1 X_{t-1} - X_t - \beta_1^2 (X_{t-2} - E_{t-3})
 \end{aligned}$$

$$\begin{aligned}
 &= \beta_1 X_{t-1} - X_t - \beta_1^2 (X_{t-2} - E_{t-3}) \\
 &= \beta_1 X_{t-1} - X_t - \beta_1^2 X_{t-2} + E_{t-3} \beta_1^2 \quad \Rightarrow \text{replace } E_{t-3} \text{ etc.}
 \end{aligned}$$

It follows therefore that such AR( $\infty$ ) representation of the MA(1) will converge just when  $|\beta| < 1$ .

Hence just the MA(.) respecting this AR( $\infty$ ) and representing the dependency structure is deemed to be **invertible**.

The generalization to MA( $p$ ) is straightforward and the model is invertible if all the roots of the characteristic polynomial

$$\phi(B) = (1 - \beta_1 B - \beta_2 B^2 - \dots + \beta_p B^p)$$

lies outside of the unit root.

### Fitting the MA model

Fitting the MA model is tricky as there is no efficient explicit estimator and one has to rely on numerical optimization.

The idea here is to leverage the AR( $\infty$ ) representation of MA processes.

$$E_t = X_t - \beta_1 X_{t-1} + \beta_1^2 X_{t-2} + \dots + \beta_1^t E_0$$

Now given an initial estimate for  $E_0$  we have a formula for  $E_t$  which depend just on  $X_{t-i}$  &  $\beta_1$   $i=1, \dots, t$ .

You can then determine the coefficients minimizing  $\sum E_t^2$ . Due to the higher order polynomial there is no simple solution to it and you must rely on numerical optimization.  $\Rightarrow$  Conditional sum of squares.

### MLE

Usually it is preferred to numerically obtain a first estimator for the coefficients via CSS and then to use MLE to estimate  $\sigma^2_e$ ,  $\beta_j$ ,  $m$  via MLE.

## ARMA Models

They are a hybrid between MA(.) and AR(.) models. It is useful as it sometimes allows to model complex dependency structures via a few autoregressive and moving average coefficients. It is therefore parsimonious.

An ARMA model is of the form:

$$X_t = a_1 X_{t-1} + \dots + a_p X_{t-p} + \beta_1 E_{t-1} + \dots + \beta_q E_{t-q} + E_t$$

$$\phi(1 - a_1 B - a_2 B^2 - \dots - a_p B^p) X_t = \Theta(1 + \beta_1 B + \dots + \beta_q B^q) E_t$$

$$\phi(B) X_t = \Theta(B) E_t$$

it is now clear that

If the roots of  $\Phi(\cdot)$  are outside of the unit circle, the process will be stationary and have mean zero. On the other hand, if the roots of  $\Theta(\cdot)$  are outside of the unit circle, the process is invertible. Both properties are important for practical application. If they are met, we can rewrite any  $ARMA(p,q)$  in the form of a  $AR(\infty)$  or an  $MA(\infty)$ . This explains why fitting an  $ARMA(p,q)$  can in practice often be replaced by fitting  $AR$ - or  $MA$ -models with high orders

Model	ACF	PACF
$AR(p)$	infinite / exp. decay	cut-off at lag $p$
$MA(q)$	cut-off at lag $q$	infinite / exp. decay
$ARMA(p,q)$	infinite / mix of decay & cut-off	infinite / mix of decay & cut-off

A natural extension is the one of introducing a differentiation component for the series. This can handle the trend component and unit roots in the series.

This is especially useful for forecasting as you might not have to recompute the series individually.

The model will look as follows

$$\phi(B)(1-B)^d X_t = \Theta(B) E_t$$

differentiated  
series to which  
the ARMA is  
fitted

Practically this is done in the following way:

- 1) Choose the appropriate order of differencing, usually  $d=1$  or (in rare cases)  $d=2$ , such that the result is a stationary series.
- 2) Analyze ACF and PACF of the differenced series. If the stylized facts of an  $ARMA$  process are present, decide for the orders  $p$  and  $q$ .
- 3) Fit the model using the `arima()` procedure. This can be done on the original series by setting  $d$  accordingly, or on the differences, by setting  $d=0$  and argument `include.mean=FALSE`.
- 4) Analyze the residuals; these must look like White Noise. If several competing models are appropriate, use AIC to decide for the winner.

A further extension involves in also handling the seasonality on top of the trend in one single shot through the following model

$$\phi(B) \phi_s(B^s) Z_t = \Theta(B) \Theta_s(B^s) E_t$$

$$Z_t = (1 - B)^d (1 - B^s)^D X_t$$

$B^s$  = backshift operator at lag  $s$ .

$\phi(B)$  = characteristic polynomial for AR component

1) Perform seasonal differencing on the data. The lag  $s$  is determined by the periodicity of the data, for the order, in most cases  $D=1$  is sufficient.

2) Do a time series plot of the output of the above step. Decide whether it is stationary, or whether additional differencing at lag 1 is required to remove a potential trend. If not, then  $d=0$ , and proceed. If yes,  $d=1$  is enough for most series.

3) From the output of step 2, i.e. series  $Z_t$ , generate ACF and PACF plots to study the dependency structure. Look for coefficients/cut-offs at low lags that indicate the direct, short-term dependency and determine orders  $p$  and  $q$ . Then, inspect coefficients/cut-offs at multiples of the period  $s$ , which imply seasonal dependency and determine  $P$  and  $Q$ . !

4) Fit the model using procedure `arima()`. In contrast to ARIMA fitting, this is now exclusively done on the original series, with setting the two arguments `order=c(p,d,q)` and `seasonal=c(P,D,Q)` accordingly.

5) Check the accuracy of the fitted model by residual analysis. These must look like White Noise. If thus far, there is ambiguity in the model order, AIC analysis can serve to come to a final decision.

Idea: it is normal to observe some movement due to the differentiation at the seasonal lag.

## ARCH-GARCH Models

This is typical in the case of volatility modeling.

Idea: To model the conditional heteroskedasticity of the returns

$$X_t = \mu_t + \varepsilon_t \quad [J]$$

with

$$\begin{aligned} E_t &= V_t \cdot \sqrt{AR(p)} \\ &= V_t \cdot \sigma_t \end{aligned}$$

with

$\varepsilon_t \leftrightarrow$  innovation; for instance gaussian white noise  $N(0, 1)$

It is then possible to see that for detrended returns you have

$$E(E_t) = 0$$

$$\text{So that } E_{t-1}^2 = (E_{t-1} - \mu)^2 = \text{Var}(E_{t-1}) = \text{Var}(\sigma_{t-1})$$

as  $\text{Var}(E_{t-1})$

$$= E(E_{t-1}^2) = \underbrace{E(\varepsilon_{t-1}^2)}_{=1} \cdot E(\sigma_{t-1}^2) = \sigma_{t-1}^2$$

What you actually have is therefore

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 \sigma_{t-2}^2 + \dots + \alpha_p \sigma_{t-p}^2.$$

Notice finally that you might generalise the above and model volatility through an ARMA model. This is called a GARCH model and has the property:

$$E_t = W_t \cdot \sqrt{H_t}$$

$$H_t = \underbrace{\alpha_0 + \alpha_1 E_{t-1}^2 + \dots + \alpha_p E_{t-p}^2}_{\text{MA}} + \underbrace{\beta_1 H_{t-1} + \dots + \beta_p H_{t-p}}_{\text{AR}}$$

## Time Series Regression

The idea here is to model a series by means of another one:

$$Y_t = \beta_0 + \beta_1 X_{t+1} + \dots + \beta_p X_{t+p} + E_t$$

The issue is then that if the independent and/or dependent series display autocorrelation you might not use OLS as that would not be efficient.

In fact, there are not many restrictions for the time series regression model. As we have seen, it is perfectly valid to have non-stationary series as either the response or as predictors. However, it is crucial that there is no feedback from  $Y_t$  to the  $X_i$ . Additionally, the error  $E_t$  must be independent of the explanatory variables, but it may exhibit serial correlation.

In order to deal with correlated errors two methods are proposed:

### Cochrane - Orcutt Method

Consider the time series regression

$$Y_t = \beta_0 + \beta_1 X_{t+1} + \beta_2 X_{t+2} + E_t$$

Then after inspecting the model you would observe some autocorrelation in the error term  $E_t$  that can be modeled via a AR(1) process.

$$E_t = \alpha_{t-1} E_{t-1} + U_t \quad U_t \sim N(0, \sigma_u^2) \text{ i.i.d.}$$

If it is a well known fact that due to the correlation the OLS estimates are inefficient despite being unbiased.

The solution proposed by the method is the following

Compute

$$Y'_t = Y_t - \alpha Y_{t-1}$$

Then due to the AR(1) representation of  $E_t$  you get

$$E_t = \alpha E_{t-1} + U_t$$

$$U_t = E_t - \alpha E_{t-1}, \text{ so that}$$

$$\begin{aligned} Y'_t &= \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + E_t - \alpha(\beta_0 + \beta_1 x_{t-1,1} + \beta_2 x_{t-1,2} + E_{t-1}) \\ &= \beta_0(1-\alpha) + \beta_1(x_{t1} - x_{t-1,1}) + \beta_2(x_{t2} - x_{t-1,2}) + E_t - \alpha E_{t-1} \\ &= \beta'_0 + \beta'_1 x'_{t1} + \beta'_2 x'_{t2} + U_t \end{aligned}$$

Here the coefficients are the same. Notice however that for fitting the regression by OLS you need the modified  $Y'_t$ . To get these you need to know the coefficient  $\alpha$ .

Hence, in practice this method works iteratively by:

- 1) Run OLS regression to obtain estimates  $\hat{\beta}_0, \dots, \hat{\beta}_p$
- 2) Estimate an AR(1) on the OLS residuals to obtain  $\hat{\alpha}$
- 3) Determine the prime variables  $Y', x'$  and derive  $\hat{\beta}'_0, \hat{\beta}'_1, \dots, \hat{\beta}'_p$  by OLS

For higher lags of the AR(1) representation you should experiment with similar algebraic manipulations.

An alternative method that does not require the iterative procedure is the GLS method. You have also packages for it in standard stat-software.

### Generalized Least Squares

The idea here is that in the case of OLS you assume for the errors

$$\text{Var}(\vec{E}) = \sigma^2 I, \text{ where } \vec{E} = \begin{pmatrix} E_0 \\ \vdots \\ E_T \end{pmatrix}$$

However in the case of correlated errors you would have

$$\text{Var}(\vec{E}) = \sigma^2 \cdot \underset{\text{correlation structure}}{\mathcal{E}}$$

A solution to deal with that is to decompose  $\mathcal{E}$  by Cholesky:

$$\varepsilon = S S^T$$

So that

$$\begin{aligned}\vec{y} &= \mathbb{X} \vec{\beta} + \vec{E} \\ S^{-1} \vec{y} &= S^{-1} \mathbb{X} \vec{\beta} + S^{-1} \vec{E} \\ \vec{y}' &= \mathbb{X}' \vec{\beta} + \vec{E}'\end{aligned}$$

So you simply have to determine  $\varepsilon$  and from that you have an uncorrelated structure which you can apply a standard OLS.

$$\begin{aligned}\text{Var}(\varepsilon) &= \text{Var}(S^{-1} \vec{E}) = S^{-1} \text{Var}(\vec{E}) S^{-T} = \cancel{S^{-1}} \sigma^2 S S^T \cancel{S^{-T}} \\ &= \sigma^2 I\end{aligned}$$

The question is now on how to estimate  $\varepsilon$  from the data.

#### 4 Approach I:

Use Cochrane-Orcutt; i.e.

- Fit OLS
- Get ARMA representation of  $\vec{E}'$
- Infer the correlation from the fitted ARMA model; for instance using the Yule-Walker equation.

#### Approach II (recommended)

Use MLE to estimate regression coeff and ARMA parameters for the error term simultaneously.

## Forecasting

### AR Model

AR(1)

$$X_t = \alpha_1 X_{t-1} + E_t$$

$$E(X_{t+1}) = \alpha_1 \cdot X_t + E(E_t)$$

$$E(X_{t+2}) = \alpha_1 (X_{t+1})$$

$$= a_1^2 X_t$$

$$E(X_{t+k}) = a_1^k x_t \Rightarrow \text{for}$$

For higher lags a similar exercise..

### Moving Average

Consider  $M(1)$ ; Then it is obvious that for  $k \geq 2$

$$E(X_{t+2}) = a_1 E(E_{t+1}) + E(E_{t+2})$$

$$= a_1 \cdot 0 + 0 = 0$$

while for  $k=1$

$$E(X_{t+1}) = a_1 E(E_t) + \cancel{E(E_{t+1})} \neq 0$$

The generalization for higher order processes is straight for an  $MA(q)$  process you would have for  $k > q$

$$E(X_{t+k}) = 0$$

### ARMA

For the ARMA you would proceed as above. Getting first the estimates of the MA component  $E(E_{t+k-j} | X_{-\infty}^t)$  by reformulating it as an  $AR(\infty)$  process.

Then you would obtain the estimates  $E(X_{t+k-j})$  and you would combine the two.

Notice therefore that you would get an exponential decay to the global mean for  $k > q$ .

### ARIMA

Here you have modeled an ARMA on the differenced series. It is clear that to get the forecast on the original scale:

$$X_{t+k} = X_t + \sum_{j=t}^k Y_j$$

Then it is clear due to the exponential decay to the mean of the forecasted  $Y_j$ , the forecast of  $X_{t+k}$  will tend to a constant despite the unit root in the series.

It follows that forecasting via ARIMA is not compatible with a series displaying a deterministic trend.

In order to fit a more accurate forecast in such case you must add a non-zero global mean to the differenced series when fitting your ARMA. This can be done with the `xreg` argument in the `arima` function in R.

## Exponential Smoothing

Here the idea is to model

$$X_t = \mu_t + \bar{E}_t$$

by

$$\mu_t = a_t = \alpha \cdot X_t + (1-\alpha) a_{t-1} \Leftrightarrow \alpha(X_t - a_{t-1}) + a_{t-1}$$

Then it is possible to see by repetitive substitution that:

$$\begin{aligned} a_t &= \alpha \cdot X_t + (1-\alpha) \cdot (\alpha a_{t-1} + (1-\alpha) a_{t-2}) \\ &= \alpha X_t + \alpha a_{t-1} - \alpha^2 a_{t-1} + (1-\alpha)^2 a_{t-2} \\ &= \alpha X_t + \alpha (1-\alpha) a_{t-1} + (1-\alpha)^2 a_{t-2} \\ &= \dots + (1-\alpha)^2 (\alpha a_{t-2} + (1-\alpha) a_{t-3}) \\ &= \dots " \dots + \alpha (1-\alpha)^2 a_{t-2} + (1-\alpha)^3 a_{t-3} \\ &= \alpha a_t + \alpha (1-\alpha) a_{t-1} + \alpha (1-\alpha)^2 a_{t-2} + \dots \end{aligned}$$

So that you see that exponential smoothing is a geom series.

The parameter  $\alpha$  decides how much the series will be influenced by the innovation coming.

The parameter can be estimated by minimizing the sum of squared 1-step prediction errors.

Notice moreover that for the estimation:

$$E(X_{t+1}) = \alpha E(\mu_{t+1}) + \cancel{E(E_{t+1})}$$

$$= \alpha E(X_{t+1}) + (1-\alpha) a_t$$

$$(1-\alpha) E(X_{t+1}) = (1-\alpha) E(a_t)$$

$$E(X_{t+1}) = a_t$$

For

$$E(X_{t+2}) = \alpha \cdot E(X_{t+1}) + (1-\alpha) E(\mu_{t+1})$$

$$\begin{aligned}
 E(X_{t+2}) &= \alpha \cdot E(X_{t+1}) + (1-\alpha) E(\mu_{t+1}) \\
 (\cancel{\alpha}) E(X_{t+2}) &= (1-\alpha) (\alpha a_t + (1-\alpha) a_t) \\
 &= (1-\alpha) (a_t) \\
 &= a_t
 \end{aligned}$$

So that

$$E(X_{t+k}) = a_t \quad \forall k > 0$$

### Holt-Winters Model

This leverages the idea of exponential smoothing and fits an exponential smoother to the three components of the decomposed series: trend + season + level.

$$\begin{aligned}
 a_t &= \alpha(x_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \\
 b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\
 s_t &= \gamma(x_t - a_t) + (1-\gamma)s_{t-p}
 \end{aligned}$$

obs - season previous level + previous trend

s<sub>t-p</sub> a<sub>t-1</sub> + b<sub>t-1</sub>

interesting  
not corrected for trend.

← previous seasonal observation

with:

- $a_t$  = level at time  $t$
- $b_t$  = trend at time  $t$
- $s_t$  = seasonal at time  $t$ .

### Multivariate Time Series

An important concept when analyzing multiple times series is to look at the correlation among the series.

One way to do that is to look at:

$$\begin{aligned}
 \gamma_{11}(k) &= \text{Cov}(X_{1,t+k}, X_{1,t}) \\
 \gamma_{22}(k) &= \text{Cov}(X_{2,t+k}, X_{2,t}) \\
 \gamma_{12}(k) &= \text{Cov}(X_{1,t+k}, X_{2,t}) \\
 \gamma_{21}(k) &= \text{Cov}(X_{2,t+k}, X_{1,t})
 \end{aligned}$$

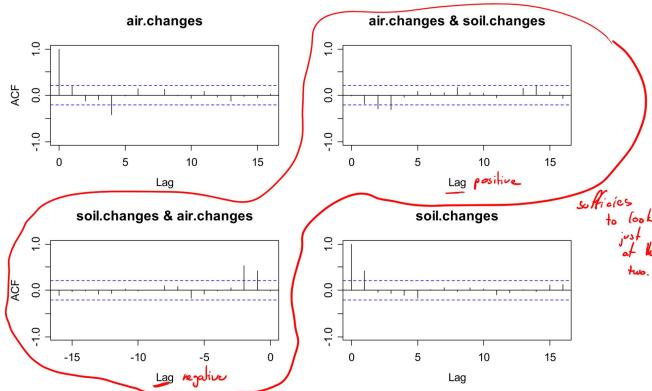
Then by inspecting such you can get a first glimpse on the correlation among the series.

Notice moreover that due to the assumed stationarity of the series you would have that the correlation among the series is dependent on the lag( $k$ ) but not the time. But then:

$$\gamma_{12}(-k) = \text{Cov}(X_1, -k, X_2, +) \xleftarrow{\text{shifting time by } k} \text{Cov}(X_1, +, X_2, +k) = \gamma_{12}(k)$$

Hence in practice, you would inspect correlation for:

$$\gamma_{11}(k), \gamma_{22}(k), \gamma_{12}(k), \gamma_{21}(-k)$$



Important: Notice that the interpretation of cross-correlation is difficult as you would actually have a mix of the within-series dependencies and the plain correlations that results in the fact that the correlations displayed are a mix of the two correlations.

Moreover, the confidence intervals are completely off. It is possible to prove

#### Case 1: No correlation between the two series for large lags

In the case where the cross correlation  $\rho_{12}(j)=0$  for  $|j| \geq m$ , we have for  $|k| \geq m$ :

$$\text{Var}(\hat{\rho}_{12}(k)) \approx \frac{1}{n} \sum_{j=-\infty}^{\infty} \{\rho_{11}(j)\rho_{22}(j) + \rho_{12}(j+k)\rho_{21}(j-k)\}.$$

as  $n \rightarrow \infty$

#### Case 2: No correlation between the series for all lags

If the two processes  $X_1$  and  $X_2$  are independent, i.e.  $\rho_{ij}(j)=0$  for all  $j$ , then the variance of the cross correlation estimator simplifies to:

$$\text{Var}(\hat{\rho}_{12}(k)) \approx \frac{1}{n} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j).$$

#### Case 3: No cross correlations for all lags and one series uncorrelated

Only now, in this special case, the variance of the cross correlation estimator is significantly simplified. In particular, if  $X_1$  is a White Noise process which is independent of  $X_2$ , we have, for large  $n$  and small  $k$ :

$$\text{Var}(\hat{\rho}_{12}(k)) \approx \frac{1}{n}.$$

Hence in order to have reliable estimates for the confidence intervals, we would like to be in case 3.

To do that you apply prewhitening.

### Prewhitening

Idea here is to transform one series such that it is uncorrelated from the other.

The idea here is to find an AR( $\infty$ ) representation for the series. We know from the above theory that this is possible for invertible ARMA(p, q) processes.

Then you would have:

$$\begin{aligned} U_t &= \sum_{i=0}^{\infty} a_i X_{1,t-i} \\ V_t &= \sum_{i=0}^{\infty} b_i X_{2,t-i} \end{aligned}$$

and

$$\rho_{UV}(k) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i b_j \rho_{X_1 X_2}(k+i-j)$$

It is then possible to see that when  $\rho_{X_1 X_2}(k) = 0 \quad \forall k$  then  $\rho_{UV}(k) = 0 \quad \forall k$ .

Conversely when  $\rho_{UV}(k) = 0 \quad \forall k \rightarrow \rho_{X_1 X_2}(k) = 0 \quad \forall k$ .

It is hence possible to see that we are in case 3 above and you can easily test whether  $\rho_{X_1 X_2}(k) = 0$  with standard confidence intervals applied on  $\rho_{UV}(k)$ .

→ Notice however that if a lag is significantly different from 0 for  $\rho_{UV}(k)$ , you can infer that also  $\rho_{X_1 X_2}(k)$  is, however you cannot infer the level from the cross-correlation.

### Transfer Function Models

! Important: You can apply such model just if one series influences another but the other has no influence on the other.

Furthermore the influence does have to go just in the future or simultaneously but not on past values

A good example

prewhitening {

```
> fit.air <- arima(diff(air.na), order=c(5,0,0))
> fit.soil <- arima(diff(soil.na), order=c(0,0,1))
> u.air <- resid(fit.air); v.soil <- resid(fit.soil)
> acf(ts.union(u.air, v.soil), na.action=na.pass)
```

u.air

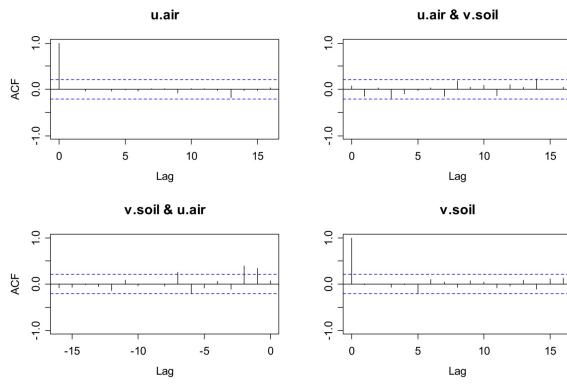
u.air & v.soil

prewhitening }  

```

> fit.air <- arima(diff(air.na), order=c(5,0,0))
> fit.soil <- arima(diff(soil.na), order=c(0,0,1))
> u.air <- resid(fit.air); v.soil <- resid(fit.soil)
> acf(ts.union(u.air, v.soil), na.action=na.pass)

```



Then in such case you might use the following modeling technique

$$X_{2,t} - \mu_2 = \sum_{j=0}^{\infty} \nu_j (X_{1,t-j} - \mu_1) + E_t$$

where  $E_t = 0$  and  $\text{Cov}(E_t, X_{1,s}) = 0 \forall t, s$ .

Assume w.l.o.g  $\mu_1 = \mu_2 = 0$ . Then

$$\begin{aligned} \gamma_{22}(k) &= \text{Cov}(X_{2,t+k}, X_{1,t}) = \text{Cov}\left(\sum_{j=0}^{\infty} \nu_j X_{1,t+j-k}, X_{1,t}\right) \\ &= \sum_{j=0}^{\infty} \nu_j \cdot \text{Cov}(X_{1,t+j-k}, X_{1,t}) \\ &= \sum_{j=0}^{\infty} \nu_j \cdot \gamma_{11}(|k-j|) \end{aligned}$$

And in the case of uncorrelated  $X_{1,t}$

$$\gamma_{22}(k) = \nu_k \cdot \gamma_{11}(0)$$

So that you might estimate all of the  $\nu_k$  by

$$\nu_k = \frac{\hat{\gamma}_{22}(k)}{\hat{\gamma}_{11}(0)}$$

Again, in order to get the desired uncorrelated structure for the  $X_{1,t}$  we can leverage prewhitening.

i.e. transform

$$X_{1,t} = AR(p)$$

$\Rightarrow$

$$\text{error of } AR(p) := D_p = (\underbrace{I - a_1 B + a_2 B^2 + \dots + a_p B^p}_\infty) X_{1,t}$$

$$\text{error at AR}(p) := D_t = \underbrace{(1 - a_1 B + a_2 B^2 + \dots + a_p B^p)}_{\phi(B)} X_{1,t}$$

Then applying the same characteristic polynomial to  $X_{2,t}$ , you would get:

$$Z_t = \phi(B) \cdot X_{2,t}$$

and as well to the Transfer Function Model Error:

$$U_t = \phi(B) \cdot E_t$$

You can rewrite the Transfer Function model as:

$$\begin{aligned} Z_t &= \sum_{j=0}^{\infty} v_j \cdot D_t + U_t && \left\{ \begin{array}{l} \text{Notice that } U_t \text{ and } Z_t \text{ are usually correlated} \\ \text{so that OLS is not efficient.} \end{array} \right. \\ \cancel{\phi(B)} \cdot X_{2,t} &= \sum_{j=0}^{\infty} v_j \cdot \cancel{\phi(B)} \cdot X_{1,t} + \cancel{\phi(B)} \cdot E_t \end{aligned}$$

$$X_{2,t} = \sum v_j X_{1,t} + E_t$$

You see that the coef are the same. You can therefore estimate:

$$v_{1c} = \frac{\hat{\gamma}_{Z,D}(k)}{\hat{\gamma}_{D,D}(k)}$$

## Spectral Analysis

The idea here is to leverage harmonic series (waves) to decompose a signal.

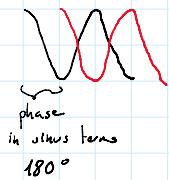
Recall that a wave is defined by:

- Frequency: # wave peaks in 1 sec.  
Unit = Hz. So 3 wave peaks  $\Leftrightarrow$  3 Hz.

- Amplitude: wave height:



- phase: where the wave starts; i.e.



It is then possible to represent any wave through:

$a$  = amplitude

$$y_t = a \cdot \cos(2\pi \cdot v \cdot t - \phi) \quad \text{, where}$$

$v$  = frequency

$t$  = time.

notice that by the very nature of waves  
 $y_t$  is periodic with periods

$\phi$  = phase

$$T = \frac{1}{v}$$

Then by standard trigonometric manipulation it is possible to see that

$$y_t = \alpha \cdot \cos(2\pi \cdot v \cdot t) + \beta \cdot \sin(2\pi \cdot v \cdot t)$$

$$\text{with } \alpha = a \cdot \cos(\phi) \quad \beta = a \cdot \sin(\phi)$$

The above has the merit of being a linear equation in the time domain  $y(t)$ , instead of a non-linear system as above, for fixed frequencies  $v$ .

Important is moreover to notice two key ingredients:

(I) the sample interval and the data influence the possible identifiable frequencies. It is impossible to infer  $v > \frac{1}{T}$  i.e. periodicity  $T > 2$ ; wave that oscillates more than 2 times every observed time frame  $T$ .

(II) There is an identification problem. If frequency  $v$  perfectly fit the data and manages to reconstruct the signal, so does  $v+1, v+2, \dots, v+k$ ,

$$\text{where } k \in \mathbb{N}$$

In practice now it is possible to represent every signal - times series - through an infinite series of waves:

$$y_t = a_0 + \sum_{k=1}^{\infty} (\alpha_k \cos(2\pi v_k \cdot t) + \beta_k \sin(2\pi v_k \cdot t))$$

In practice given that you cannot determine such infinite series as the problem cannot be

In practice given that you cannot determine such infinite series as the problem cannot be treated analytically you would just restrict the problem to the set of identifiable frequencies  $\nu > 1/2$ .

You would then have

$$y_t = a_0 + \sum_{k=1}^m (\alpha_k \cos(2\pi\nu_k t) + \beta_k \sin(2\pi\nu_k t)) + \epsilon_t$$

With

$$\nu_k = \frac{1}{k} \quad \text{and} \quad k = 1, 2, \dots, m \quad m = \frac{k}{2}$$

→ Notice that with this formulation  $\nu_k \leq \frac{1}{2}$   $\forall k$ .

→ and this system has full rank  $\rightarrow 1, \dots, N$  observations and  $N$  unknowns, i.e.  $\hat{\epsilon}_t = 0 \forall t$ .

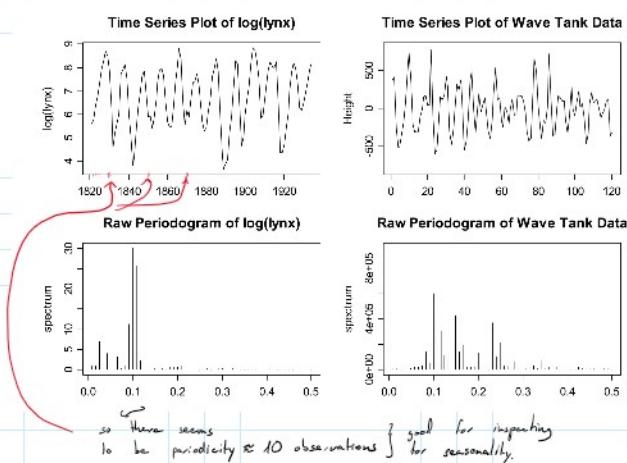
The frequencies above are Fourier Frequencies. It is moreover possible to prove that the regression above has an orthogonal design.

Finally it is possible to prove that by omitting  $\nu_k$  your sum of squared residuals would increase by

$$\sum_{t=1}^n \epsilon_t^2 \uparrow \text{by } \frac{n}{2} (\hat{\alpha}_k^2 + \hat{\beta}_k^2)$$

You can use this property for inspecting the importance of each individual frequency for reconstructing the series.

This is in fact what the periodogram does by sequentially removing a frequency  $\nu_k$  and inspecting the increase in SSE.



### Time Series Spectrum

The spectrum is a continuous function telling the presence of particular frequencies in the variation of the series.

This corresponds in fact to the continuous representation of the periodogram, telling the contribution of a particular frequency in the variation of the series (recall

This corresponds in fact to the continuous representation of the periodogram, telling the contribution of a particular frequency in the variation of the series (recall the derivation).

↳ Notice however that in practice there is often a big gap between periodogram and spectrum. This because

While the estimates for different Fourier frequencies  $\nu_j$  and  $\nu_k$  are asymptotically independent and unbiased, they are inconsistent. This means that they do not gain in precision with increasing sample size, which is owed to the fact that the periodogram always uses a saturated fit to the observed series. Hence, the asymptotics do not work in this case. They key in obtaining a better spectral estimate lies in the asymptotic unbiasedness and independence of adjacent periodogram values.

This is in fact what it is done when estimating the spectrum that leverage the above properties to get consistent estimates by smoothing the individual periodogram  $\frac{1}{q}(\hat{\alpha}_k^2 + \hat{\beta}_k^2) := I_n$  as

$$f(\hat{\nu}_j) = \frac{1}{2L+1} \sum_{k=-L}^{+L} w_k I_n(\nu_{j+k}),$$

i.e. through a running mean. This is called **Daniel Smoother**.

A further improvement is to taper the series. I like the following description reframing the above and explaining the point of tapering

In windowed Fourier analysis, we divide or **window** our time series of length  $N$  into shorter segments of length  $N_w$ , each covering a time interval  $T = N_w \Delta t$ . For a stationary time series, the periodogram of each window of data gives an independent unbiased estimate of the power spectrum (if the time series were not stationary, the power spectrum from each window would not be expected to have the same statistical character so this averaging would be invalid). If there are  $n = N/N_w$  windows, we can average these estimates. The resulting estimator (which is proportional to a  $\chi_{2n}^2$  distribution), has a relative standard deviation of  $n^{-1/2}$  instead of 1 for the unwindowed periodogram, giving a more robust estimate of the true power spectrum. It takes 100 windows to bring down the relative standard deviation of each spectral estimate down to 10%!

The tradeoff is that since each window is a factor  $n$  shorter, the frequency separation between power spectral estimates  $\Delta\omega$  is  $n$  times as coarse. In this sense, there is a close analogy between windowed Fourier analysis and averaging the periodogram across blocks of  $n$  adjacent frequencies.

If the data has substantial low-frequency variability, a problem with windowing is that there may be substantial endpoint discontinuities that contaminate the spectral estimates. The usual strategy to remove possible endpoint discontinuities is to multiply each window of the time series by a weight function or **taper** that smoothly goes to zero at the end points. This can be shown to affect the estimated power spectrum at the lowest frequencies (where it is fairly unreliable anyway) but not at frequencies corresponding to periods much less than the window period  $T$ .

## State Space Models

State space Models are useful when the observations are blurred with noise, or in situations where parameters vary over time.

State space models consists of two equations:

- (i) the state equation.
- (ii) the observation equation

### State Equation:

The linear transformation of the state at time  $t-1$  is denoted with a matrix  $G_t$ , and the covariance matrix of the multivariate normal is denoted with  $w_t$ .

$$X_t = G_t X_{t-1} + W_t, \text{ where } W_t \sim N(0, w_t)$$

### Observation Equation:

The observation at time  $t$  is denoted by a column vector  $Y_t$  that is a linear combination of the states, determined by a matrix  $F_t$ , and random variation (measurement noise) from a normal distribution with covariance matrix  $v_t$ .

$$Y_t = F_t X_t + V_t, \text{ where } V_t \sim N(0, v_t)$$

Note that in this general formulation all of the transition matrices are allowed to vary over time.