Forecasting & Predictive Analytics

Guillaume Chevillon (chevillon@essec.edu)
and Pierre Jacob (jacob@essec.edu)

October-December 2021
2nd set of slides
Forecasting Principles

Overview

Aims: forecast univariate or multivariate series using available data to guide decision-making.

Time series data present unique features, such as non-exchangeability, trends, seasonality.

Forecasting with uncertainty can be done within a probabilistic framework, using stochastic processes as models.

Today we review classical, non-stochastic techniques and then lay the foundations for a probabilistic approach.

Transformation and decomposition

First processing steps

Adjustements (calendar, population, inflation).

Transformations, such as logarithms, Box–Cox transform:

$$\text{BoxCox}(y) = \begin{cases} \log(y) & \text{if } \lambda = 0, \\ (\text{sign}(y)|y|^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$

See the effect of λ here:

`https://otexts.com/fpp3/transformations.html`

Differencing

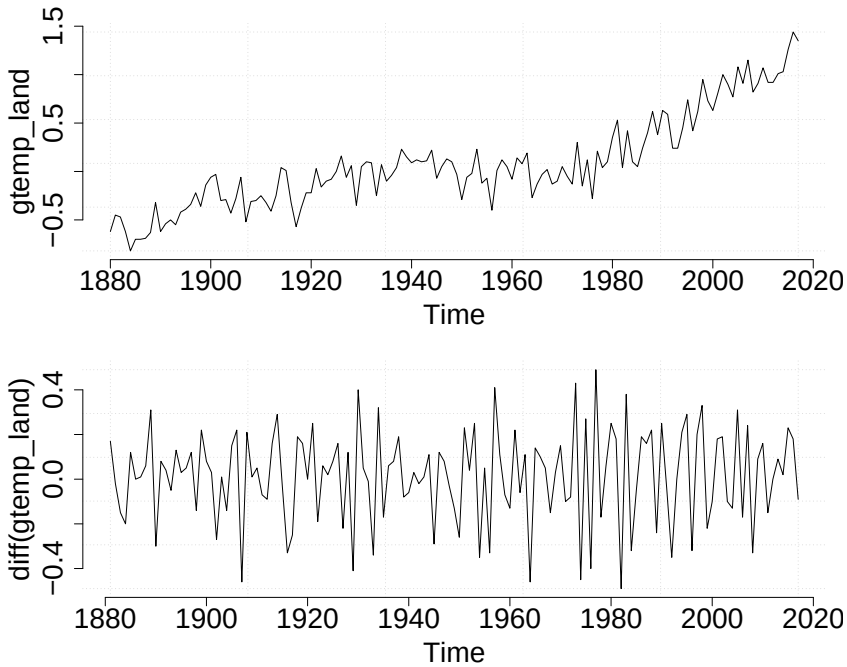


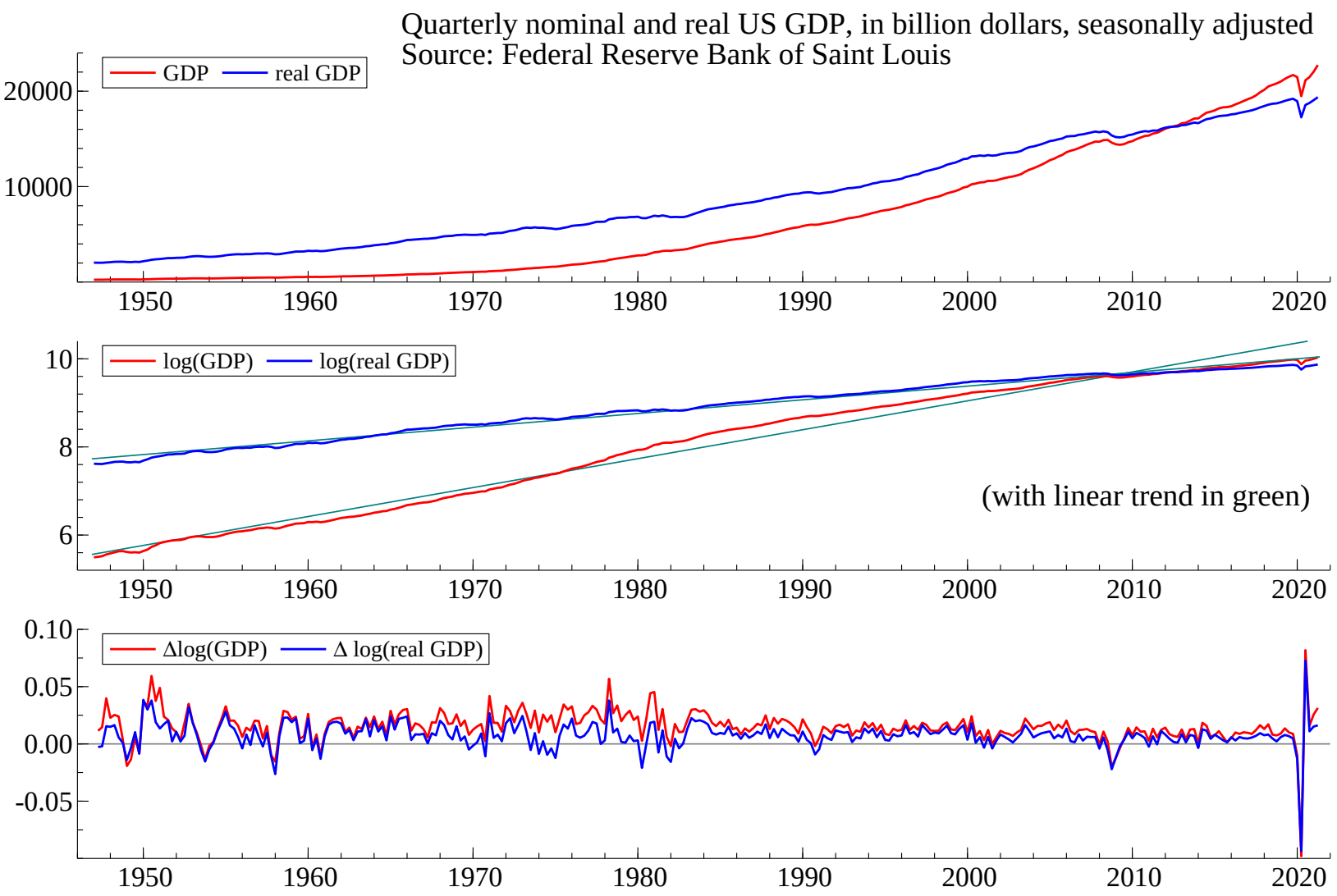
Figure: Annual temperature anomalies (in degress centigrade) averaged over the Earth's land area from 1880 to 2017. From `astsa` package.

First difference:

$$\nabla y_t = y_t - y_{t-1}.$$

Also seasonal differences of period m, differences order k.

Example: Quarterly US GDP



Moving averages

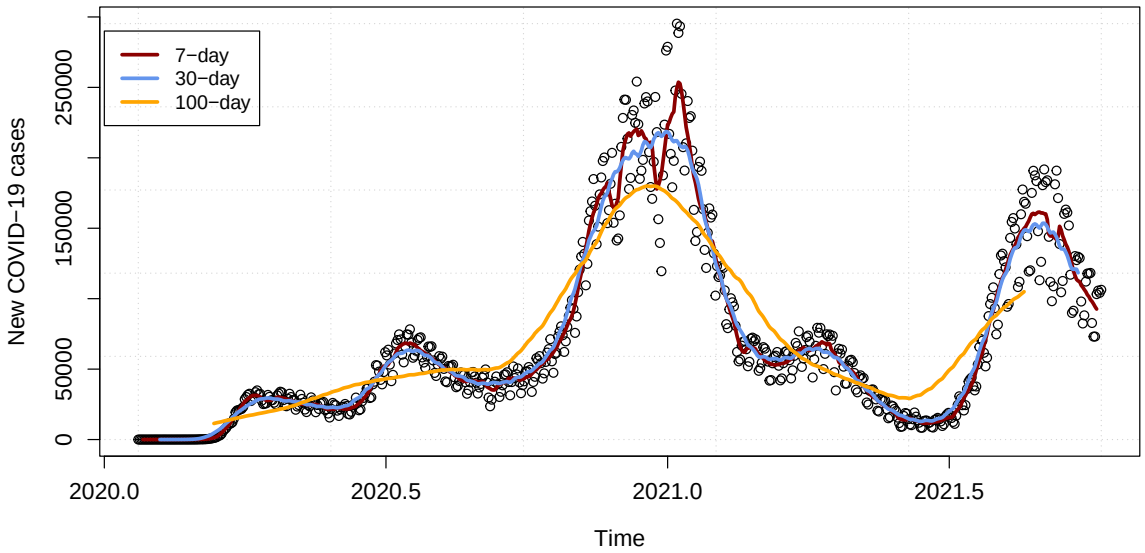


Figure: Daily COVID-19 cases in the US and moving averages. Source: CDC.

Moving average of order $m = 2k + 1$:

$$T_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}.$$

Smoothes out fluctuations in (y_t), extracts a "trend".

Decomposition

Represent series (y_t) as

$$y_t = T_t + S_t + R_t,$$

where T_t is a trend, S_t a seasonal component, and R_t is the residual, random or irregular element.

Seasonal component could be constant or time-varying.

Many variants exist under the names of X-11, SEATS, STL.

Additive decomposition versus multiplicative decomposition.

Decomposition

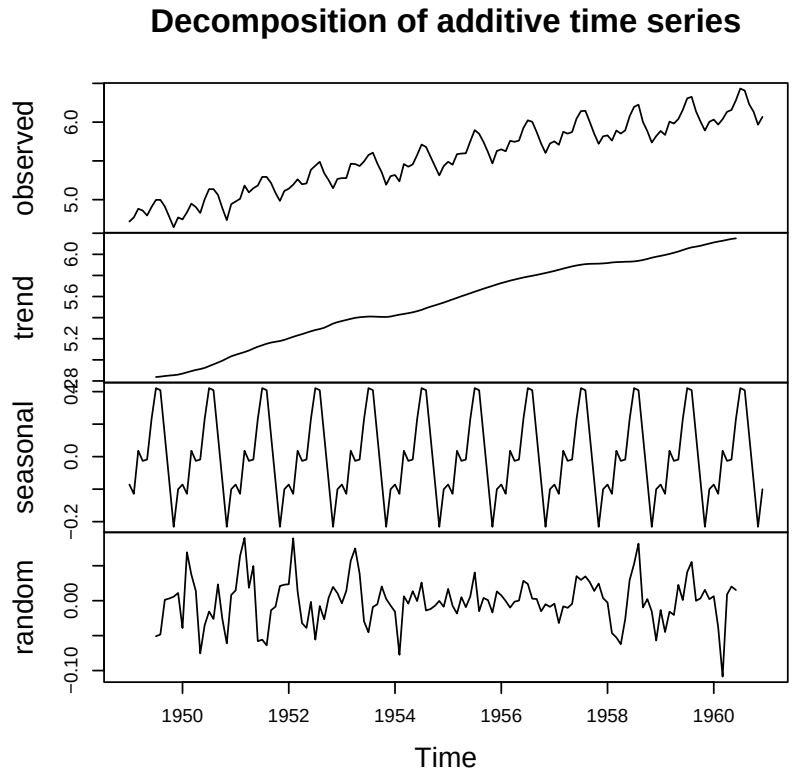


Figure: Additive decomposition of monthly totals of international airline passengers, 1949 to 1960, on the log scale.

Hodrick–Prescott filter

Extract a trend by solving

$$\min_{T_1,\ldots,T_n} \sum_{t=1}^{n} (y_t - T_t)^2 + \lambda \sum_{t=2}^{n-1} \left(\nabla^2 T_t\right)^2,$$

where $\lambda > 0$ and $\nabla^2 T_t = (T_t - T_{t-1}) - (T_{t-1} - T_{t-2})$.

The left-hand side encourages proximity to the data, the right-hand side encourages "smoothness".

Hodrick–Prescott filter

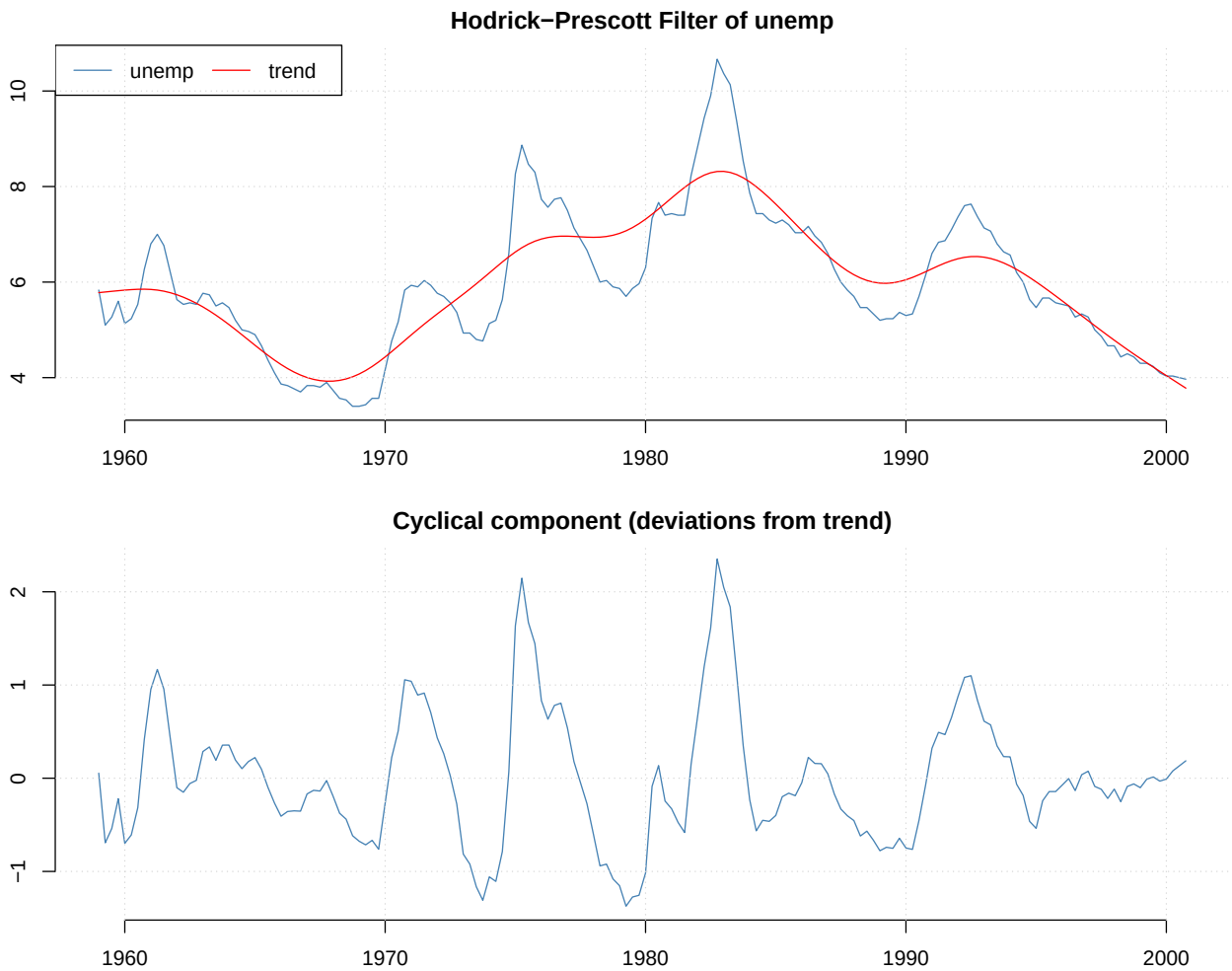


Figure: Quarterly US unemployment series for 1959.1 to 2000.4.

Change points

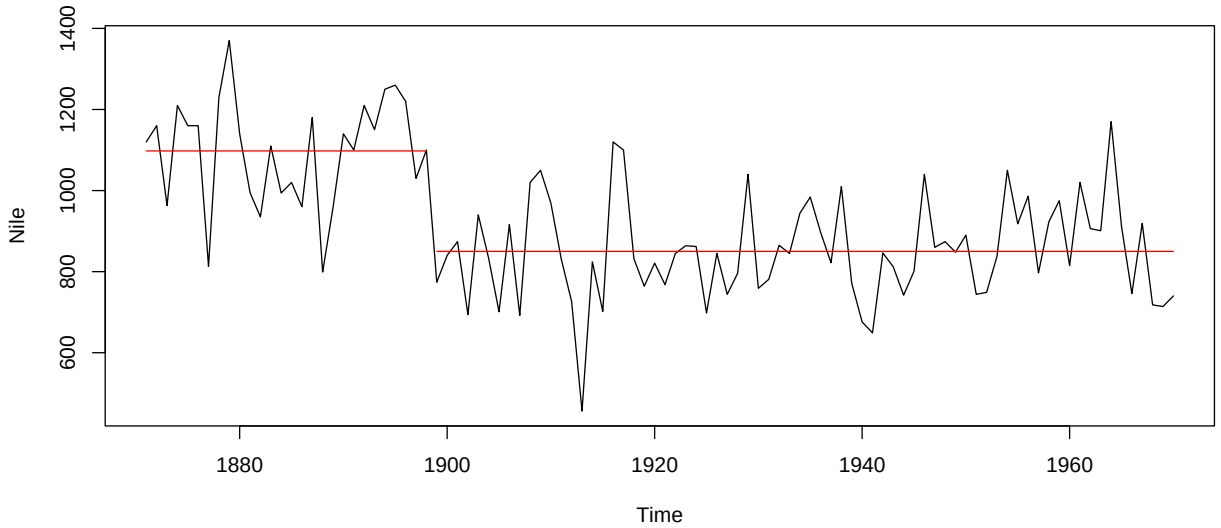


Figure: Measurements of the annual flow of the river Nile at Aswan 1871–1970, in $10^8\ m^3$.

Also called structural breaks. Offline versus online methods. Often based on statistical hypothesis testing.

Deterministic methods for point prediction

Two baselines

We have $(y_1, \ldots, y_n)$ and want to predict y_{n+1}.

- Use $n^{-1} \sum_{t=1}^{n} y_t$ to predict y_{n+1}.
- Use y_n to predict y_{n+1}.

The two strategies are instances of weighted averages.

Simple exponential smoothing

For some $\alpha \in [0, 1]$,

$$\hat{y}_{n+1} = \alpha y_n + \alpha(1-\alpha) y_{n-1} + \alpha(1-\alpha)^2 y_{n-2} + \dots.$$

Recursive form,

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t.$$

Recursively plugging $\hat{y}_t$ instead of y_t for $t > n$, we obtain $\hat{y}_{n+h} = \hat{y}_{n+1}$ for all $h \geq 1$.

Exponential smoothing

Simple version:

$$\hat{y}_{t+h} = \ell_t$$
$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1},$$

where ℓ_t is called the "level" at time t.

With trend,

$$\hat{y}_{t+h} = \ell_t + hb_t$$
$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \gamma(\ell_t - \ell_{t-1}) + (1-\gamma)b_{t-1}.$$

Exponential smoothing

With trend and seasonality,

$$\begin{aligned}
\hat{y}_{t+h} &= \ell_t + hb_t + s_{t+h-m(\lfloor (h-1)/m \rfloor + 1)} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \gamma(\ell_t - \ell_{t-1}) + (1-\gamma)b_{t-1} \\
s_t &= \delta(y_t - \ell_{t-1} - b_{t-1}) + (1-\delta)s_{t-m}.
\end{aligned}$$

Called Holt–Winters additive method. There's a multiplicative method, where the seasonal component multiplies the trend.

Exponential smoothing

```
m <- HoltWinters(co2)
p <- predict(m, 50, prediction.interval = TRUE)
plot(m, p)
```

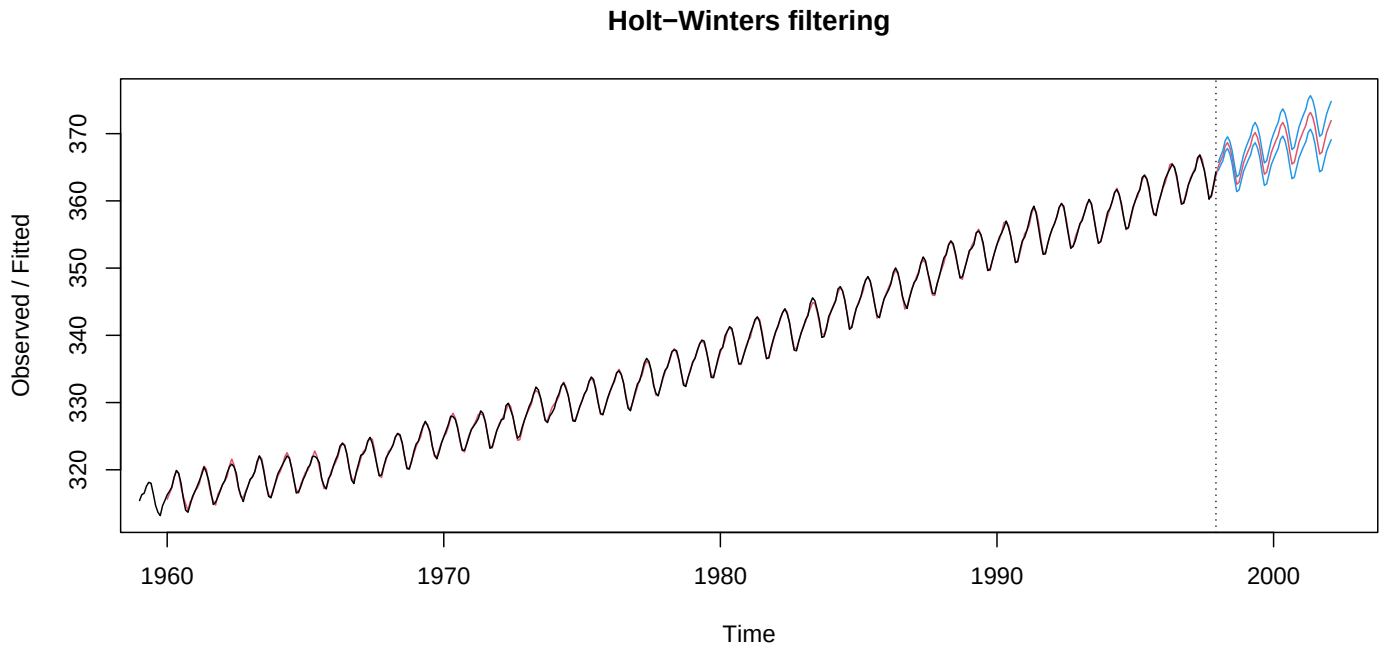


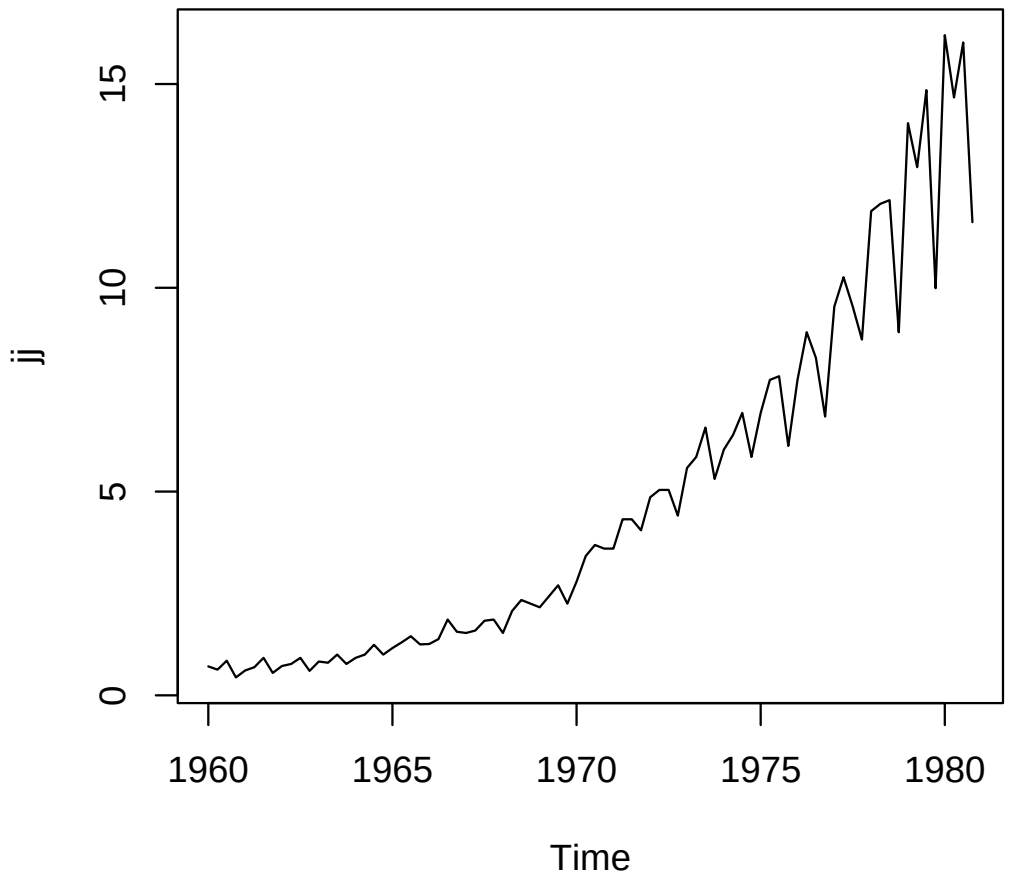
Figure: Atmospheric concentrations (monthly) of CO2 in Mauna Loa, expressed in parts per million (ppm), with predictions and intervals.

But how do we obtain these prediction intervals?

Forecasting interlude

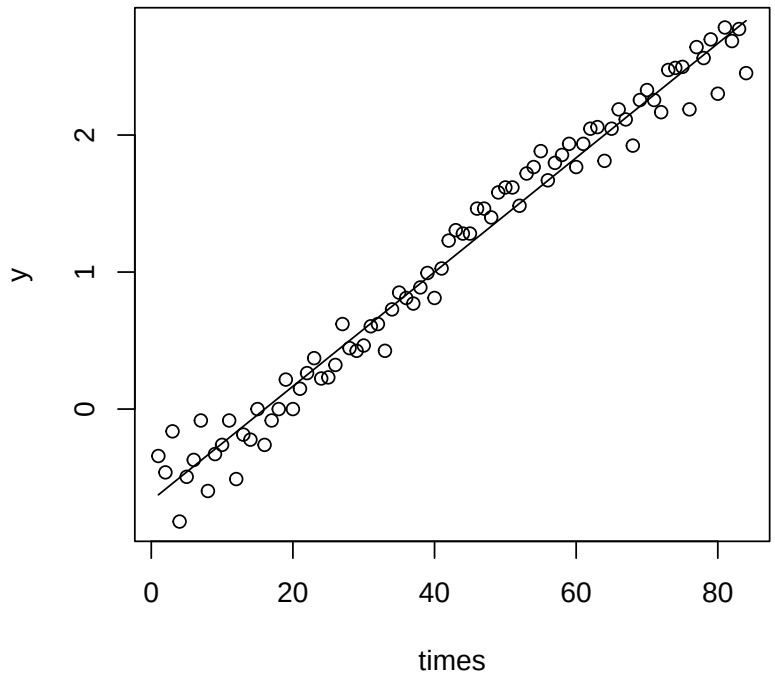
An example with Johnson & Johnson quarterly earnings

```
library(astsa)
plot(jj)
```



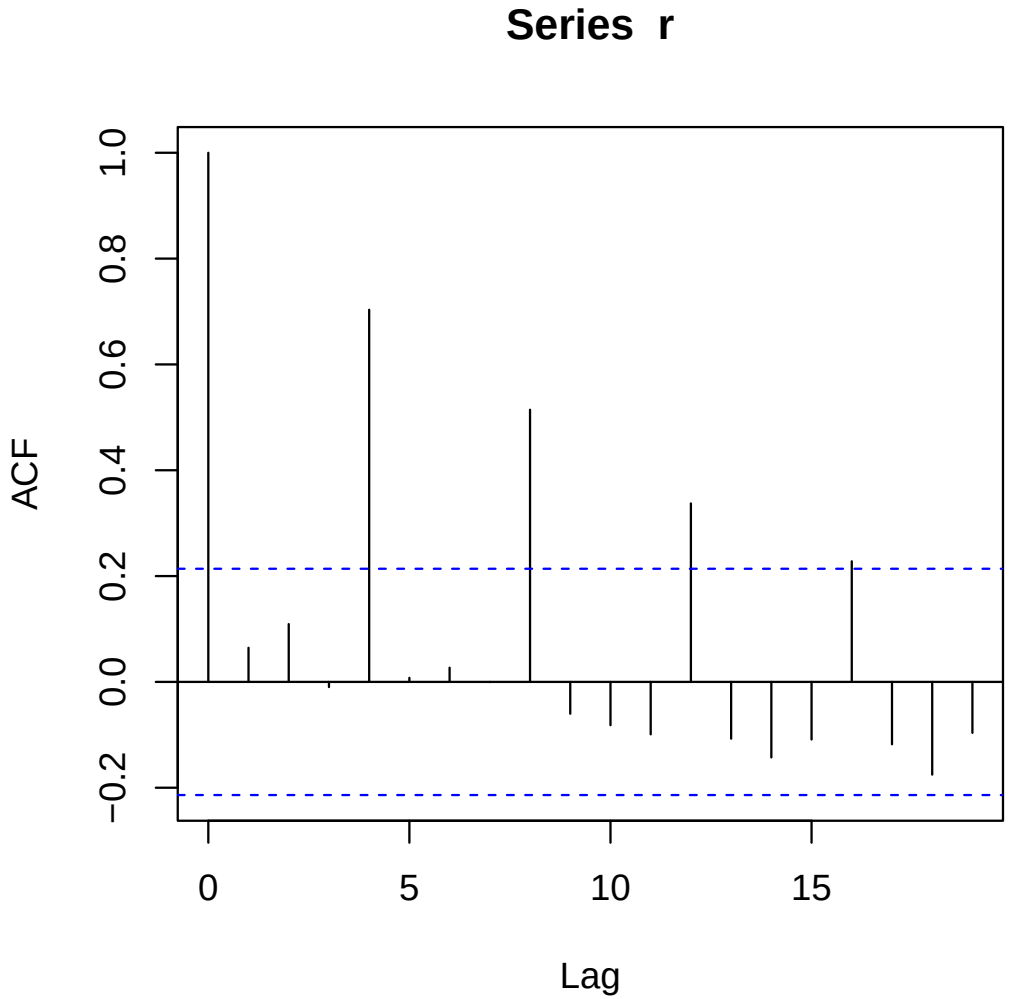
An example with Johnson & Johnson quarterly earnings

```
y <- as.numeric(log(jj)); times <- 1:length(y)
regression <- lm(y ~ times)
plot(x = times, y = y)
lines(x = times, y = predict(regression))
```



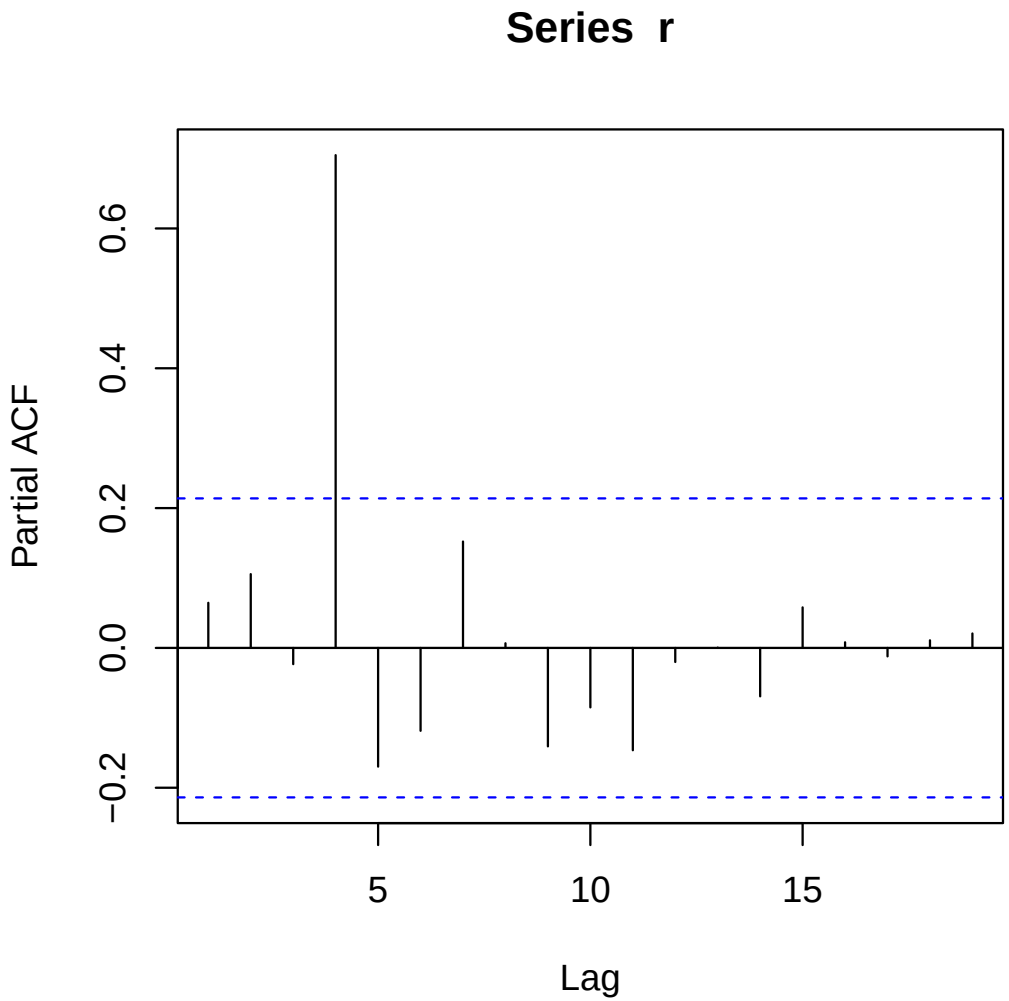
An example with Johnson & Johnson quarterly earnings

```
r <- residuals(regression)
acf(r)
```



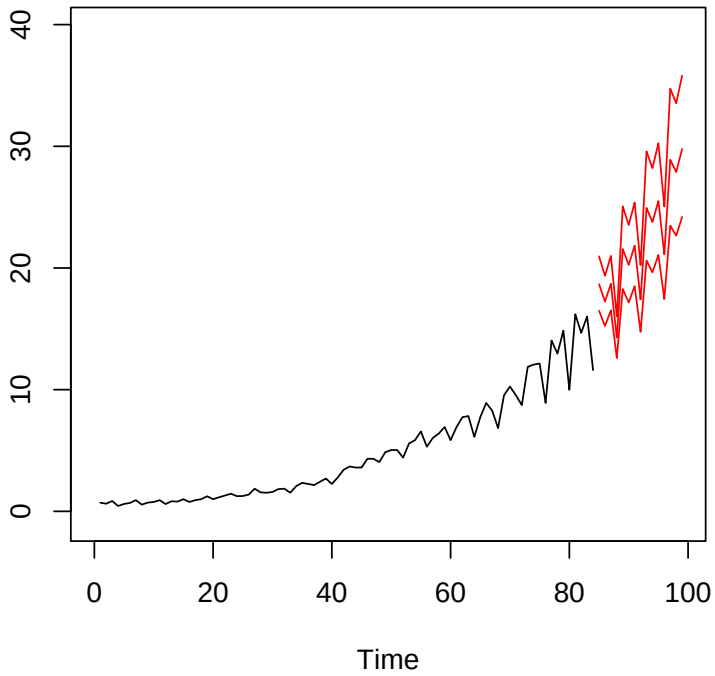
An example with Johnson & Johnson quarterly earnings

```
r <- residuals(regression)
pacf(r)
```

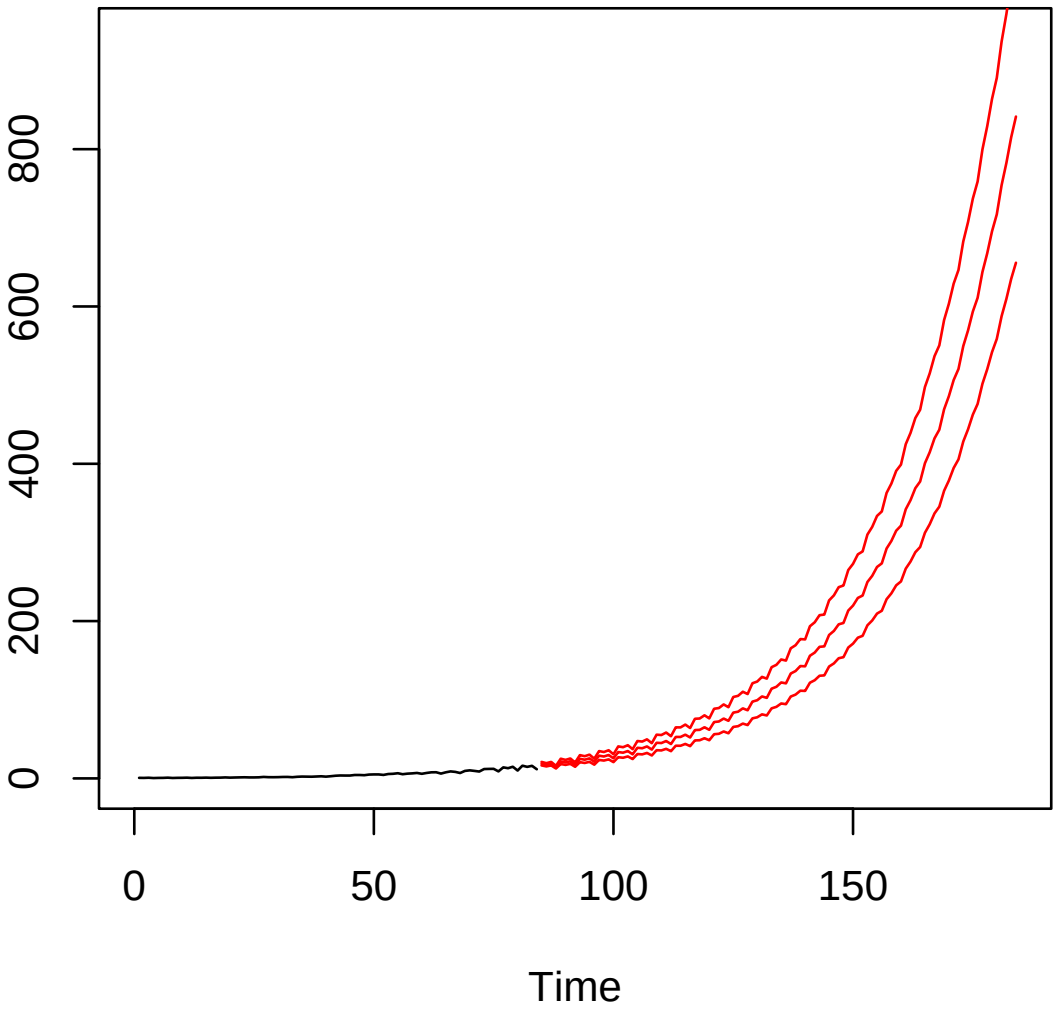


An example with Johnson & Johnson quarterly earnings

```
 fit <- Arima(jj, order = c(0,0,0), seasonal = list(order
= c(1,0,0), period = 4), biasadj = TRUE, lambda = 0,
include.drift = TRUE)
 forecast <- forecast(fit, h = n.ahead, biasadj = TRUE,
lambda = 0)
```



Going too far



Probabilistic reasoning in forecasting

Predict Y using X

Suppose that we observe X, and we want to predict Y. Any function of X, denoted by $c(X)$, is a potential predictor of Y.

We want to minimize $\mathbb{E}[(Y - c(X))^2]$.

After some calculations,

$$\mathbb{E}[(Y - c(X))^2] = \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]] + \mathbb{E}[(c(X) - \mathbb{E}[Y|X])^2].$$

Minimized by choosing $c(X) = \mathbb{E}[Y|X]$.

Predict Y using X

If we restrict $c(X)$ to be of the form $\alpha + \beta X$, we find the solution to be $\hat{\alpha} = \mathbb{E}[Y] - \hat{\beta}\mathbb{E}[X]$, $\hat{\beta} = \mathbb{C}\text{ov}(X, Y)/\mathbb{V}[X]$.

This is essentially linear regression.

If we aim at probabilistic prediction, and use a proper scoring rule as a loss, such as $S(p, y) = -\log p(y|X)$, then the expected loss

$$\mathbb{E}\left[S(p, y)\right]$$

is minimized over p by the predictive distribution $\hat{p}(y) = \text{dgp}(y|X)$, where the latter is the conditional distribution of Y given X under the data-generating process.

In practice we look for a parametric approximation of this conditional distribution, in a class of models.

Stationary processes

Empirical averages

- The law of large numbers says that if $\mathbb{E}\left[|Y|\right] < \infty$ and if $Y_{1:n}$ are i.i.d. copies of Y, then

$$\frac{1}{n}\sum_{t=1}^{n} Y_t \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\left[Y\right].$$

- This justifies the approximation of unknown expectations, e.g. $\mathbb{E}[Y]$, by sample averages, e.g. $\bar{y} = n^{-1}\sum_{t=1}^{n} y_t$.

- The law of large numbers also holds under different assumptions: in particular under *stationarity*.

Empirical correlation

- Empirical correlation:

$$\hat{\mathbb{C}}\mathrm{or}\left(x_{1:n}, y_{1:n}\right) = \frac{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\left(y_t - \bar{y}\right)}{\sqrt{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2 \sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2}},$$

  which might converge to $\mathbb{C}\mathrm{or}\left(X, Y\right)$ when $n \to \infty$.

- Related to linear regression of $y_{1:n}$ on $x_{1:n}$,
  and to linear regression of $x_{1:n}$ on $y_{1:n}$.

- Exactly the regression slope of standardized $x_{1:n}$ on standardized $y_{1:n}$.

Autocovariance

- Covariance between Y_t and Y_{t-1} within $Y_{1:n}$.
- It will be useful to consider $\mathbb{C}\text{ov}(Y_t, Y_{t-k})$ for all k.
- If for each t, $\mathbb{V}[Y_t] < \infty$, then we can introduce

$$\gamma(s,t) = \mathbb{C}\text{ov}(Y_s, Y_t),$$

  for all s and all t, called *autocovariance* function of $Y_{1:n}$.

- Measure linear relationships between different times.

Autocorrelation

- As before, we can normalize the covariances:

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

  for all s and all t, called *autocorrelation* function (ACF).

- Between -1 and 1 by Cauchy-Schwarz.

- Could be defined across time series as

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}},$$

  where

$$\gamma_{xy}(s,t) = \mathbb{C}\text{ov}(X_s, Y_t),$$

  called *cross-correlation* and *cross-covariance*.

The Challenge

- We can approximate $\mathbb{E}[X]$ by $n^{-1}\sum_{t=1}^{n} x_t$, assuming that $x_{1:n}$ constitute a representative sample distributed as X.

- If we want to approximate

$$\mathbb{C}\mathrm{ov}(Y_t, Y_{t+h}) = \mathbb{E}[(Y_t - \mathbb{E}[Y_t])(Y_{t+h} - \mathbb{E}[Y_{t+h}])],$$

  we need a representative sample of Y_t, Y_{t+h}, and $Y_t Y_{t+h}$.

- But we only observe one value y_t at time t and one value y_{t+h} at time $t+h$.

- If each Y_t had nothing to do with any other Y_s, it would be hopeless.

Weak stationarity

- Stationarity means that mean and autocovariances are stable over time.

- That is, a process is (weak) stationary when
  1. $\mathbb{V}[Y_t] < \infty$ for all t,
  2. $\mathbb{E}[Y_t] = \mu$ for all t,
  3. the autocovariance $\gamma(s,t)$ depends only on $|t-s|$.

- In particular $\gamma(t-1,t) = \gamma(t,t+1) = \gamma(1,2)$ for all t.

A stochastic process is "white noise"
if $\gamma(h) = \sigma^2$ for $h = 0$ and $\gamma(h) = 0$ for $h \neq 0$.

Examples

- Normal white noise process: $W_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
- Random walk process:

$$Y_t = Y_{t-1} + W_t.$$

- MA(1) process:

$$Y_t = W_t + \theta_1 W_{t-1}.$$

- AR(1) process:

$$Y_t = \varphi_1 Y_{t-1} + W_t.$$

Which ones are stationary? Under what assumptions?

Autocovariance of stationary processes

- The autocovariance $\gamma(s,t)$ depends only on $|t-s|$. We can write it as a function of one argument, $h = |t-s|$:

$$\gamma(h) = \mathbb{C}\text{ov}(Y_t, Y_{t+h}) = \mathbb{C}\text{ov}(Y_1, Y_{1+h}).$$

- The autocorrelation function is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\mathbb{C}\text{ov}(Y_1, Y_{1+h})}{\mathbb{V}(Y_1)}.$$

The following properties can be verified:

- $\gamma(0) = \mathbb{V}[Y_1] = \mathbb{E}[(Y_1-\mu)^2]$, thus $\rho(0) = 1$.
- $|\gamma(h)| \leq \gamma(0)$ for all h, thus $|\rho(h)| \leq 1$.
- $\gamma(h) = \gamma(-h)$ for all h, and $\rho(-h) = \rho(h)$.

Resolving the Challenge

- For stationary processes, the mean is constant $\mu = \mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \ldots = \mathbb{E}[Y_n]$.
- We observe y_t, a realization of each Y_t.
- Since all means are the same, it could be that $n^{-1} \sum_{t=1}^{n} Y_t$ converges to μ.
- Conceptually amazing: we can average over time, instead of averaging over repeated experiments.
- Convergence guaranteed if $\gamma(h) \to 0$ as $h \to +\infty$. Brockwell & Davis 1991 (Theorem 7.1.1), see also the lecture notes.

Empirical autocorrelations

- Sample autocovariance:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y}).$$

  Makes more sense if h is small compared to n.

- Sample autocorrelation:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}.$$

These might converge to their theoretical counterparts for stationary processes.

Correlograms

Lag-plots and correlograms to look into dependencies of a series.

- Autocorrelogram (ACF) shows an approximation of $\rho(h)$ against h, for $h = 0, \ldots, h_{\max}$.
- Partial autocorrelogram (PACF) shows the impact of Y_{t-h} on Y_t, taking $Y_{t-1}, \ldots, Y_{t-h+1}$ into account: it is the coefficient c_h in the linear regression

$$Y_t = c_1 Y_{t-1} + \ldots + c_h Y_{t-h} + \epsilon_t.$$

- Shiny apps for AR(2), MA(2)processes.

```
library(shiny);
runGitHub(repo="shinyapps",ref="main",
username="pierrejacob",subdir="acfautoregressive/")
runGitHub(repo="shinyapps",ref="main",
username="pierrejacob",subdir="acfma/")
```

Take Aways This Week

1. Forecasting starts with a number of adjustments, transformations and decompositions on the original data.
2. During that process we gather insights on aspects of the data, which can guide the choice of forecasting tools.
3. A number of deterministic forecasting tool offer useful baselines, such as exponential smoothing.
4. Probabilistic approaches enable uncertainty estimates and a coherent treatment of calibration and model comparison.
5. Stationarity is a key property that enables properties of stochastic processes to be learned from samples.
6. Dependency features of time series can be explored with autocorrelograms, partial autocorrelograms.