Contents

2 Basics of forecasting and modeling time series data

Time series present distinctive features, such as trends, seasonalities, a constant or a varying amplitude of the fluctuations around the trend, to name a few. As mentioned earlier, we intend to use stochastic processes as a way of modeling time series data. Here we continue this dual exploration of time series features and stochastic modeling. In Sections 2.1-2.2 -2.3 we review classic "deterministic" (non-random) techniques to extract information from time series data and to perform point predictions. Then, we move on to a statistical/stochastic framework from Section 2.4 onward. We start by describing some motivation for this statistical viewpoint, introduce some useful definitions such as "stationarity", and introduce two basic models: autoregressive (AR) and moving average (MA). Throughout we denote our time series of interest by $(y_t)_{t=1}^n = (y_1, \dots, y_n)$, with length n.

2.1 Basic transformations

The recommended starting point of all analysis and forecasting entreprise is a thorough visual inspection of the available data. The term "raw data" refers to data before any processing; although the data collection itself typically involves some form of processing. We have already seen "trace plots" (or "line plots") of various series; that is, displaying y_t against t for $t = 1, \ldots, n$. In the presence of multiple series, one can plot them against each other, i.e. y_t against x_t for $t = 1, \ldots, n$. When seasonality is important, we can cut a series into multiple series corresponding to different seasons, and overlay them on the same graph. That is, for a periodicity equal to m (e.g. 12 for monthly data), display in separate lines for $k = 0, 1, \ldots,$ the series y_{km+t} against t for $t = 1, \ldots, m$. This is sometimes called a seasonal plot; see "seasonal subseries plots" for an alternative visualization of seasonal series.

A first benefit of visualizing time series is that it might suggest simple transformations of the data. The term "adjustement" is used to designate common-sense modifications of raw series, such as calendar adjustments (months can have 28, 29, 30 or 31 days, which affects data that represent monthly counts), population adjustments, inflation adjustments, etc.

Beyond adjustments, it is common to apply a function to each element in the series. For example, taking the logarithm of a series of positive values might help identify whether a trend is polynomial or exponential, and might help to make the amplitudes of variations around a trend stable over time. A logarithmic transformation also turns a positive series into real values, which are then easier to model with probability distributions supported on real values such as Normals. The Box–Cox transformation is a generalization of the logarithm, with a parameter $\lambda \in \mathbb{R}$, defined as

$$\mathrm{BoxCox}(y) = \begin{cases} \log(y) & \text{if } \lambda = 0, \\ (\mathrm{sign}(y)|y|^\lambda - 1)/\lambda & \text{otherwise.} \end{cases} \tag{2.1}$$

As long as $\lambda \neq 0$ the argument y could be negative. For more on this and the selection of λ see Guerrero, Time-series analysis supported by Power Transformations, 1993. See https://otexts.com/fpp3/transformations.html for an interactive plot where one can play with λ and see the effect on Australian quarterly gas production.

Another simple transformation is "differencing", which refers to computing $\nabla y_t = y_t - y_{t-1}$ for all $t \geq 2$. For positive series it is common to combine logarithm and differences, i.e. to consider $\log y_t - \log y_{t-1}$ instead of y_t. We can iterate the differencing operation, for example $\nabla^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$, for all $t \geq 3$. Seasonal differencing refers to the computation of $y_t - y_{t-m}$, for m possibly larger than one.

Differencing can be used to get rid of trends in the original data, without having to fit a regression model; see an example in Figure 2.1. We will encounter differencing again, for example

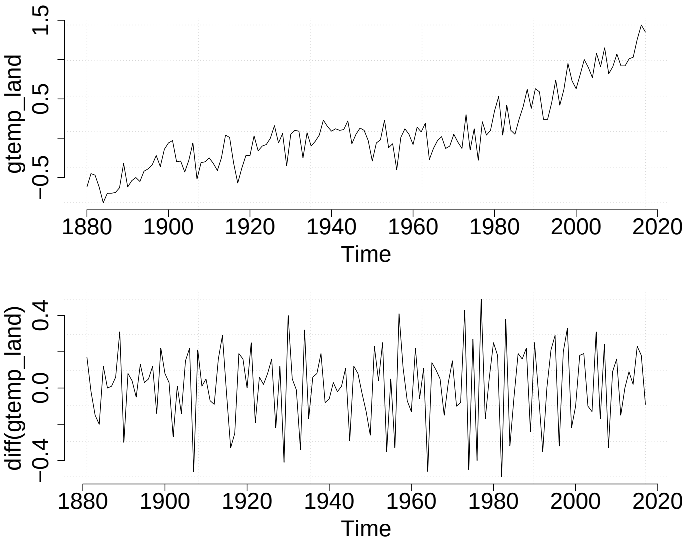


Figure 2.1: Annual temperature anomalies (in degress centigrade) averaged over the Earth's land area from 1880 to 2017. From `astsa` package. Top panel presents (y_t), bottom panel shows the first differences $(y_t - y_{t-1})$, which appear more "stable".

in the chapter on ARMA models.

If interest is ultimately in the original series, then the analyst needs to "back-transform" the forecast into the original scale. This might need a bit of care, especially with probabilistic predictions: we might need to make sure our transformations are one-to-one and apply "change of variable" formulas to obtain predictive distributions on the scale of interest.

Figure 2.2 shows series of quarterly US gross domestic products. The figure illustrates the effect of inflation adjustments (top panel), logarithmic transformations to help identify trends (middle panel), and first differences (bottom panel). Note that the 1970s were a period of high inflation in the US; this is particularly apparent in the bottom panel.

2.2 Decompositions

After having performed some transformations as above, we can consider a variety of manipulations to gain some insight on the features of the series under consideration.

Moving averages. Also called "rolling window" or "sliding window" averages, they refer to calculations such as

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

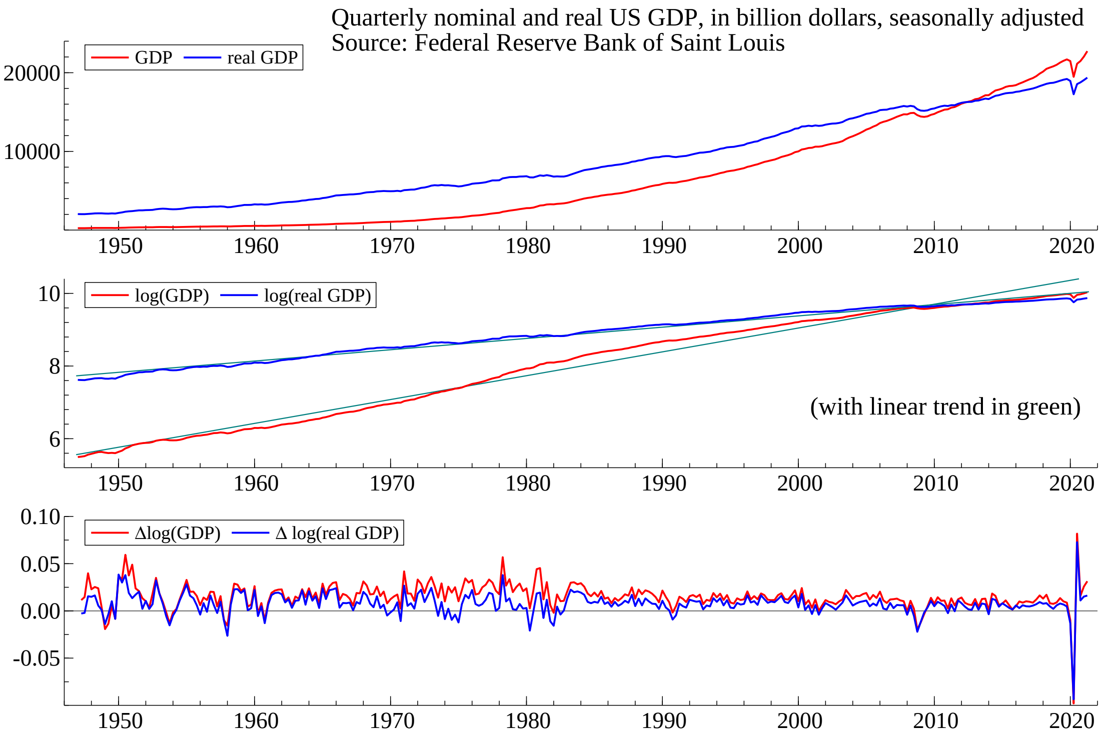


Figure 2.2: Quarterly US GDP. The top panel shows the nominal values, and those adjusted for inflation ("real"). The middle panel shows the log transforms, with fitted linear trends. The bottom panel shows the first differences.

where $m = 2k+1$ is the "order", $k \in \mathbb{N}$, and $t = 1+k, \dots, n-k$. This smoothes out some variations in the original series. If m is an even number, the sum is not symmetric around t; for example a moving average of order 4 can be defined as $T_t = \frac{1}{4}\sum_{j=-1}^{2} y_{t+j}$.

We can iterate and compute moving averages of moving averages. For example a moving average of order 2 applied to a moving average of order 4 results in

$$T_t = \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1} + \frac{1}{8}y_{t+2},$$

which has the appeal, in the case of quarterly data, of assigning equal weight $(1/4)$ to each quarter of the series. More generally we can compute "weighted moving averages" with arbitrary weights set as non-negative values summing to one. By playing with the order m and the weights we can extract a trend T_t that is more or less smooth. For example, the CDC website proposes visualizations of COVID-19 daily cases in the US, in the form of raw data and an overlaid curve computed as a

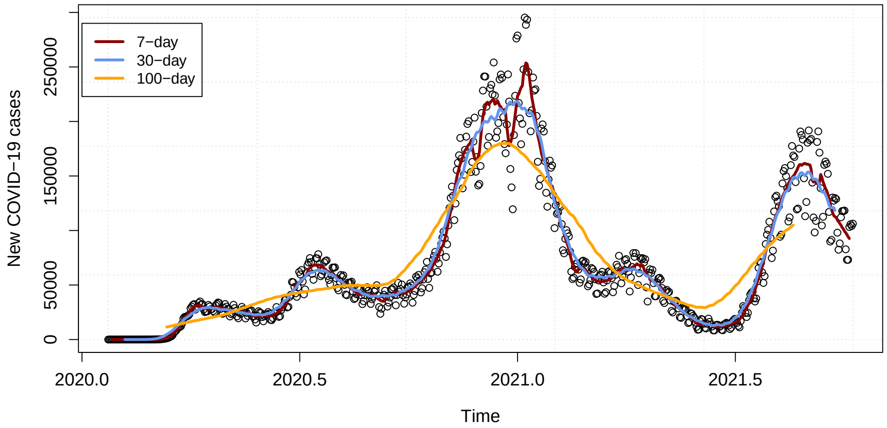


Figure 2.3: Daily COVID-19 cases in the US and moving averages. Data obtained from the CDC website.

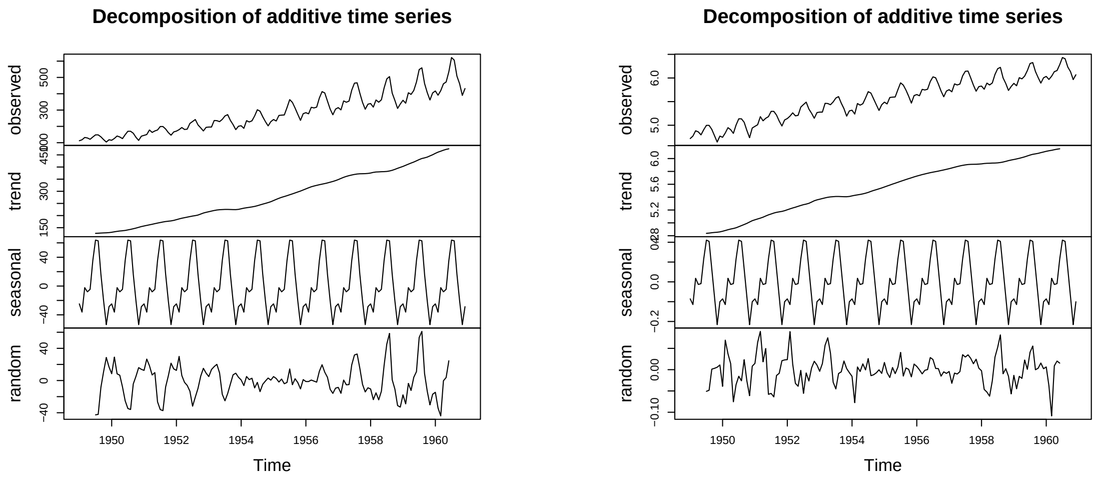
7-day moving average. Figure 2.3 shows the raw data along with 7-day, 30-day and 100-day moving averages.

Remark 2.1. We use uppercase for T_t but here it is not a random variable: it is computed deterministically from the observed series (y_t). It would be problematic to write t_t, but we could have used another letter, such as m_t.

Trend, seasonal and residual components. A most classical decomposition of time series is of the form $y_t = T_t + S_t + R_t$, where T_t refers to a trend, S_t as a seasonal component and R_t as the residuals. The series $T_t + R_t$, or equivalently $y_t - S_t$, is referred to as the "seasonally adjusted" series. In the most basic forms, the seasonal component is constant across periods, so that, with a periodicity denoted by m, $S_t = S_{t+km}$ for all $k \geq 1$, and the seasonal component is centered: $\sum_{j=1}^{m} S_{t+j} = 0$. In other words the seasonal component is made of m numbers summing to 0, repeated over successive periods.

There are many ways of obtaining such decomposition; we describe a basic one. We first obtain the trend using a moving average described above. To obtain the seasonal component, we can average for each season all the available values of $y_t - T_t$, the "detrended" series. That is, we average for $k = 1, \ldots, m$ all the detrended values for that season, i.e. $(y_{k+jm} - T_{k+jm})$ for $j = 1, \ldots, \lfloor (n-m)/m \rfloor$. We then adjust $S_1, \ldots, S_m$ so that they sum to zero, and repeat the values to obtain the entire seasonal component $S_1, \ldots, S_n$. Finally we can define the residuals as $R_t = y_t - T_t - S_t$.

The decomposition can also be multiplicative: $y_t = T_t \times S_t \times R_t$. If the series (y_t) is positive,



(a) Original series

(b) Logarithm of the series

Figure 2.4: Decomposition of the classic Box & Jenkins airline data: monthly totals of international airline passengers, 1949 to 1960. On the left, the additive decomposition is done on the original series, while it is applied to the log on the right.

this can be seen as an additive decomposition on the logarithmic scale. Otherwise, we can obtain the multiplicative decomposition by obtaining the trend T_t from a moving average, then defining the detrended series as y_t/T_t, obtaining the seasonal component S_t essentially as above, and finally the residuals as $y_t/(T_t \times S_t)$. Figure 2.4 shows an additive decomposition of the "airline data"; on the left, the residuals exhibit clear seasonal patterns. On the right, the additive decomposition has been performed on the logarithm of the series; the residuals look a bit less "seasonal".

Many variants of the above decomposition are available, under the names of X-11, SEATS, and STL (`stl` function in `R`). These methods allow the seasonal component to vary over time, provide explicit control on the smoothness of the trend, and include mechanisms to handle outliers.

For the purpose of forecasting, the seasonal component is the easiest to predict: we can simply replicate the previous values, assuming a constant or slowly-varying seasonality. The main challenge lies in forecasting the seasonally adjusted series, i.e. $T_t + R_t$ or $T_t \times R_t$. Any of the forecasting methods described below can be used for this purpose. Similar decompositions are the basis of Facebook's forecasting model humbly called "Prophet". There, the trend component is modeled as a piecewise-linear curve.

Most recent versions of seasonal adjustment techniques (e.g. X-12 or X-13 designed by the U.S. Census Bureau, or Facebook Prophet) take seriously the question of adjustments for bank holidays

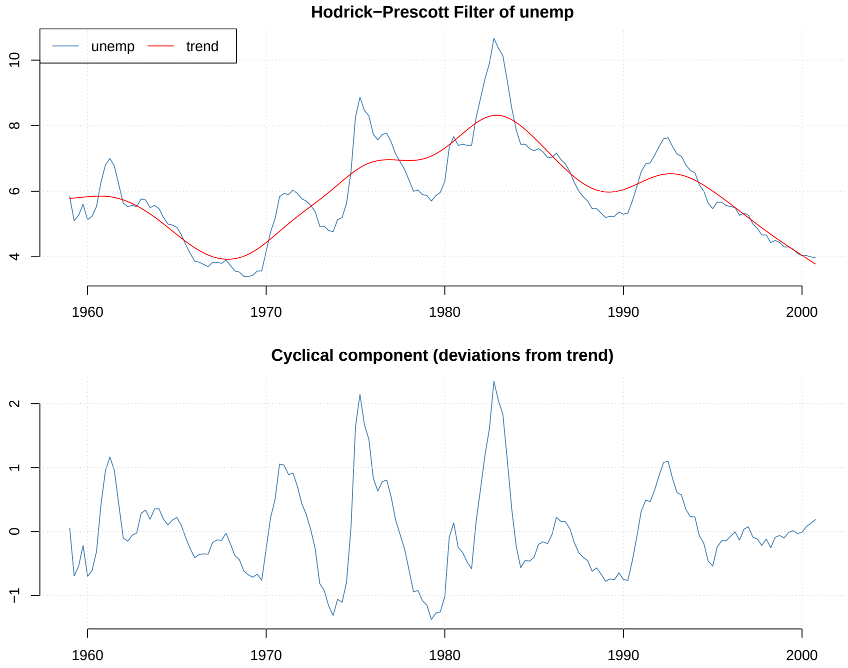


Figure 2.5: Quarterly US unemployment series for 1959.1 to 2000.4.

and the number of working days. Indeed, remember that there are about 200 working days per year. Hence one extra bank holiday counts for $1/200 = 0.5\%$ of total number of days per year. When we think that GDP growth in most advanced economies is about 1 to 3% per annum, and that in a country like France, the number of bank holidays oscillates from one year to the next in the [7,13] range, we see that this is a significant question.

We mention briefly the Hodrick–Prescott filter, which is a trend extraction method employed primarily in macroeconomics. The idea is to obtain the trend by solving an optimization program:

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

where $\lambda > 0$ is a tuning parameter that is calibrated according to the frequency of the data, and recall that $\nabla^2 T_t = (T_t - T_{t-1}) - (T_{t-1} - T_{t-2})$. The first sum encourages the trend to be close to the data, while the second sum penalizes changes in the trend (we can think of $\nabla^2 T_t$ as a sort-of second derivative of the curve T_t) and thus makes the solution smooth, and more so for large λ. See Hamilton's "Why You Should Never Use the Hodrick–Prescott Filter" for a critical assessment of this technique. Figure 2.5 shows the trend extracted using Hodrick–Prescott for a series of quarterly

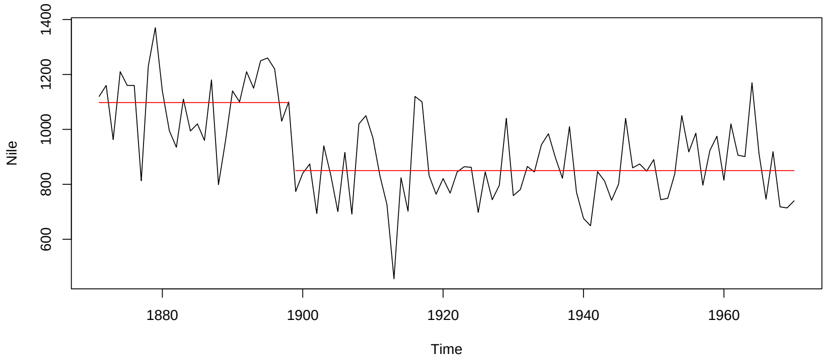


Figure 2.6: Measurements of the annual flow of the river Nile at Aswan 1871-1970, in 10^8 m^3.

US unemployment counts.

Change points. Another common task in the realm of time series decomposition is to identify times called "change points" or "structural breaks" that represents abrupt changes in the series. Upon identifying these change points, one can choose to filter out part of the series when performing predictions, or to include the possibility of future change points in the forecast. There are many ways of defining and identifying such change points, and many variants of the change point detection problem: offline or online, with a known or unknown number of change points, etc. Figure 2.6 shows a series of annual flows of the rive Nile, with an abrupt change around 1900, as identified by a change point algorithm. See Cobb, "The problem of the Nile: conditional to a change-point problem", 1978 for more on this data, and see "An Evaluation of Change Point Detection Algorithms", by van den Burg & Williams, 2020 for a recent overview of change point detection.

2.3 Deterministic point prediction and exponential smoothing

We now denote by (y_t) the series that we want to predict; perhaps obtained after transformations and/or decompositions such as described above. We start with the description of some basic or "common-sense" techniques that do not require much statistical formalism, at least if we temporarily ignore important issues of uncertainty quantification.

A first strategy to forecast future values of (y_t) is to ask an expert. Experts can be people with domain expertise, and we can gather multiple experts, hoping that a group would be more effective than individuals. The "Delphi method" is a method to elicit a consensus from various experts. In recent times, prediction markets, where many anonymous individuals bet on future outcomes, have been sometimes interpreted as a way of leveraging the "wisdom of crowds" in forecasting. Experts also include oracular animals, such as Paul the Octopus and Mani the Parakeet who predicted the results of the 2010 Football World Cup with some success.

Assuming that we have to produce the forecast ourselves, how might we go about it? Given the available series, two basic strategies are as follows. First, we might compute the average of the available values $(y_1, \ldots, y_n)$, and report it as the prediction for y_{n+1}. Or we might use the latest observation y_n as a prediction for y_{n+1}. Those two simple strategies correspond to two choices of weighted averages: the first assigns the same weight n^{-1} to all observations, while the second puts all the weight on y_n. They correspond to different assessments about the relevance of past data for the prediction of future values.

A popular forecasting method, which can be seen as an interpolation between these two strategies, is known as exponential smoothing. The simplest version predicts y_{n+1} with

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

where $\alpha \in [0,1]$ is a parameter. Since $\alpha \sum_{j \geq 0} (1-\alpha)^j = 1$, the forecast is a weighted average of past values; a little approximation is necessary in practice as our available data is of finite length. We can re-express the forecast in a recursive form,

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

for $t = 1, 2, \ldots, n$, where $\hat{y}_1$ has to be set somehow. If we want to forecast y_{n+2} and plug $\hat{y}_{n+1}$ in place of y_{n+1} in the formula, we obtain $\hat{y}_{n+2} = \hat{y}_{n+1}$: the forecast for all future times forms a flat line.

Owing to this limitation, exponential smoothing is of limited use beyond horizon $h = 1$. Instead, we can think of extrapolating the latest "local" trend. Thus, we can extend the above "simple exponential smoothing" technique to include trend and seasonal components. We first write the above model as

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

where ℓ_t is called the "level" at time t, and h is the horizon we want to predict over. We can then include a trend with the following modification,

$$\begin{aligned}\hat{y}_{t+h} &= \ell_t + hb_t\\ \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})\\ b_t &= \gamma(\ell_t - \ell_{t-1}) + (1-\gamma)b_{t-1},\end{aligned}$$

with two parameters, α and γ, ℓ_t remains the "level" and b_t is an estimate of the slope of a linear trend at time t. Similarly we can add equations and parameters to represent seasonality, and we can further include "damping" parameters so that the linear extrapolation flattens over time. This type of strategy is generally known as either exponential smoothing or as "Holt–Winters", due to these two early articles: Holt, Forecasting seasonals and trends by exponentially weighted averages, 1957 and Winters, Forecasting sales by exponentially weighted moving averages, 1960. Such strategies remain widely used today. An comprehensive reference on exponential smoothing is the book "Forecasting with Exponential Smoothing" by Hyndman, Koehler, Ord and Snyder, 2008.

Figure 2.7 shows a prediction made on a series of concentrations of CO2 in Hawaii, using a Holt–Winters method with seasonality. The figure shows prediction intervals around the prediction points. At this point it should be surprising: we have described Holt–Winters as a deterministic technique, so how can we associate these predictions with uncertainty estimates? It will turn out that there is a statistical interpretation of Holt–Winters that enables the construction of prediction intervals, as we will cover later on, when we discuss state space models.

2.4 A probabilistic approach to prediction

We now move towards a probabilistic framework for time series forecasting. The approach delivers numerical recipes comparable to the above techniques, and

- a way of obtaining prediction intervals or entire predictive distributions, and not only point predictions,
- a way of relating methods to clear objectives, defined in the language of loss functions or "decision theory", and thus to define a coherent workflow to forecasting,
- a quantitative way of comparing forecasting methods to one another.

We view time series data $(y_1, \ldots, y_n)$ as realizations of random variables $(Y_1, \ldots, Y_n)$. We allow successive variables Y_t and Y_{t+1} to be *dependent*.

As a starter consider the task of predicting a real-valued random variable Y, with a guess $c \in \mathbb{R}$. To evaluate our guess, we use a loss function $(c, y) \mapsto L(c, y)$ e.g. $L(c, y) = (c - y)^2$, the squared

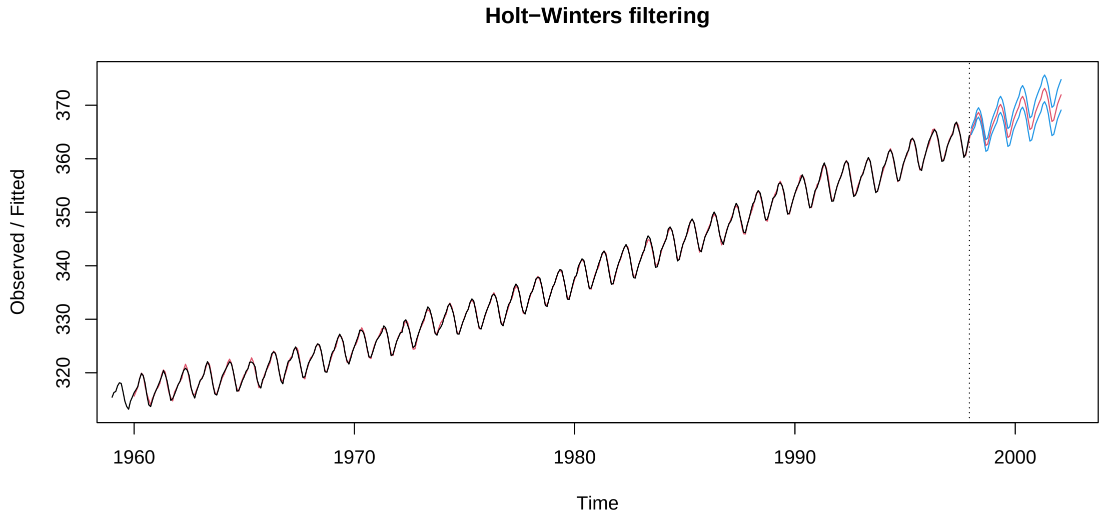


Figure 2.7: Atmospheric concentrations (monthly) of CO2 in Mauna Loa, expressed in parts per million (ppm), with predictions and intervals, and prediction using Holt–Winters.

loss. We define the objective of formulating a guess c that makes $L(c, y)$ small on average: we aim at minimizing $\mathbb{E}[(Y - c)^2]$, called the *mean squared error* (MSE). Some calculations show that $c = \mathbb{E}[Y]$ minimizes the MSE. This is useful guidance, at least insofar as we can estimate the solution.

Next, suppose that we observe X and we want to predict Y given X. A function of X, denoted by $c(X)$, is a potential predictor of Y. By the tower property, we can write $\mathbb{E}[(Y - c(X))^2] = \mathbb{E}[\mathbb{E}[(Y - c(X))^2|X]]$. We can always write $Y - c(X)$ as $Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - c(X)$, and then we can expand the square in the above expression to obtain, after some calculations,

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

The first term is constant in c. We can minimize the second term by choosing $c(X) = \mathbb{E}[Y|X]$, which makes it zero; and it cannot be less than zero. Thus, the optimal prediction of Y given X is the conditional expectation $\mathbb{E}[Y|X]$. Again, this is useful if we can estimate such quantity.

It might be difficult to estimate the conditional expectation $\mathbb{E}[Y|X]$. We can look at a simpler task: the *best linear prediction*. That is, we restrict the function $x \mapsto c(x)$ to be a linear function of x, and we find coefficients $\alpha, \beta \in \mathbb{R}$ that minimize $\mathbb{E}[(Y - (\alpha + \beta X))^2]$. By differentiating with respect to α and β, we obtain two equations:

$$\mathbb{E}\left[(Y - (\alpha + \beta X))\right] = 0 \text{ and } \mathbb{E}\left[X\left(Y - (\alpha + \beta X)\right)\right] = 0.$$

Two unknowns (α, β) and two equations: we solve and find $\hat{\beta} = \mathbb{C}\text{ov}(X,Y)/\mathbb{V}[X]$, $\hat{\alpha} = \mathbb{E}[Y] - \hat{\beta}\mathbb{E}[X]$. Note that the notion of covariance between X and Y appears in the solution of a prediction problem.

If we are interested in probabilistic prediction instead of point prediction, and if we use a "proper scoring rule" as a loss, such as $S(p,y) = -\log p(y|X)$, where p is the proposed predictive distribution and y is the realization of Y to predict, then the expected loss $\mathbb{E}\left[S(p,y)\right]$ is minimized over all p by the predictive distribution $\text{dgp}(y|X)$, that is, the conditional distribution of Y given X under the data-generating process. In practice we would look for a parametric approximation of this conditional distribution, in a class of working models. See Gneiting and Raftery, Proper Scoring Rules, Prediction and Estimation, 2007.

2.5 Reminders on random variables

Expectation and independence. We start with a quick review of some properties of a generic real-valued random variable, denoted by X.

The expectation of X, $\mathbb{E}[X]$, is defined by $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$ where $x \mapsto f_X(x)$ is the density of X. The integral is not always finite. Similarly we define $\mathbb{E}[h(X)] = \int_{-\infty}^{+\infty} h(x) f_X(x) dx$ for a measurable function h, and again the integral might be infinite. The expectation $\mathbb{E}$ satisfies some fundamental properties. It is linear: if X is equal to a constant real number c, then $\mathbb{E}[X] = c$; for any pair of random variables X and Y, and any two real numbers a and b, we have $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. Furthermore, we can always write $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$, where $\mathbb{E}[X|Y]$ is the conditional expectation of X given Y. This is the "tower property".

The linearity of expectation can be used to establish properties of the variance $\mathbb{V}[X]$, defined as $\mathbb{E}[(X - \mathbb{E}[X])^2]$. For example we can develop the square and obtain $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

If X and Y are real-valued random variables with marginal density functions f_X and f_Y, and joint density $f_{X,Y}$, they are said to be *independent* if the joint density $f_{X,Y}$ factorizes into the product of marginals:

$$\forall x, y \in \mathbb{R} \quad f_{X,Y}\left(x,y\right) = f_X\left(x\right) f_Y\left(y\right). \tag{2.2}$$

Note that for any joint distribution $f_{X,Y}$, we can always write $f_{X,Y}\left(x,y\right) = f_X\left(x\right) f_{Y|X}\left(y|x\right) = f_Y\left(y\right) f_{X|Y}\left(x|y\right)$, where $f_{Y|X}$ (respectively $f_{X|Y}$) denotes the density of Y given X (respectively of X given Y). Thus, independence between X and Y is equivalent to saying that

$$\forall x, y \in \mathbb{R} \quad f_{Y|X}\left(y|x\right) = f_Y\left(y\right) \text{ and } f_{X|Y}\left(x|y\right) = f_X\left(x\right). \tag{2.3}$$

This follows the informal idea of "independence": conditioning Y on X does not "change" its distribution. In other words, knowing the value of X does not inform about the values taken by Y. And vice versa; the notion of independence is symmetric in X and Y.

A useful property about independent variables is that the expectation $\mathbb{E}[XY]$ is equal to $\mathbb{E}[X]\,\mathbb{E}[Y]$. This can be derived either from first principles, using the definition of expectation and Eq. (2.2), or using the tower property and the fact that $\mathbb{E}[X|Y] = \mathbb{E}[X]$ by independence. Another useful property is that, for any two functions g and h, if X and Y are independent then $g(X)$ and $h(Y)$ are independent.

We say that X and Y are dependent if they are not independent; so two variables are either independent or dependent.

Covariance and correlation. The notion of independence is considered uncontroversial: most people agree on its mathematical definition and on its use in data analysis. On the other hand, there have many attempts at quantifying the amount of dependence between variables. The correlation coefficient is one of them. It is somewhat deceptive, and yet incredibly useful.

First, the covariance between X and Y is defined as

$$\mathbb{C}\text{ov}\,(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\,. \tag{2.4}$$

The second line can be checked by developing the product of the first line, and using the linearity of expectations. The covariance is not a very intuitive notion but note that if X and Y are independent, then $\mathbb{C}\text{ov}(X, Y) = 0$ because then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. If $\mathbb{C}\text{ov}(X, Y) = 0$ we say that X and Y are uncorrelated; otherwise they are correlated.

Therefore, *independent variables are always uncorrelated.* The other direction is not true: there are plenty of pairs of variables X and Y such that $\mathbb{C}\text{ov}\,(X, Y) = 0$ and yet X and Y are dependent.

Example 2.1. Consider X following a symmetric distribution around 0, such as a centered Normal distribution or a Uniform distribution on $[-1, 1]$. Define Y as $Y = X^2$. Then X brings a lot of information on Y (in fact, X determines Y completely), so intuitively the two variables are dependent (and we could check it formally). On the other hand, we can compute $\mathbb{C}\text{ov}\,(X, Y) = \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right]$. Since X is symmetric around 0, we have $\mathbb{E}\left[X^3\right] = 0$ and $\mathbb{E}[X] = 0$, and thus $\mathbb{C}\text{ov}\,(X, Y) = 0$.

Example 2.2. If two variables are not only uncorrelated but also jointly normally distributed, then they are independent.

We state some properties of the covariance that can be checked using the properties of expectation.

- With $X = Y$, we obtain $\mathbb{C}\text{ov}\,(X, X) = \mathbb{V}[X]$.
- The covariance is symmetric: $\mathbb{C}\text{ov}\,(X, Y) = \mathbb{C}\text{ov}\,(Y, X)$.

- The covariance is *bilinear*: for numbers a, b, c and random variables X, Y, W, $\mathbb{C}\text{ov}\left(aX + bY, cW\right) = ac\,\mathbb{C}\text{ov}\left(X, W\right) + bc\,\mathbb{C}\text{ov}\left(Y, W\right)$.
- The covariance is invariant by shifts: for all $a \in \mathbb{R}$, $\mathbb{C}\text{ov}\left(X + a, Y\right) = \mathbb{C}\text{ov}\left(X, Y\right)$.

The correlation coefficient is a standardized version of the covariance, originally introduced by Francis Galton and Karl Pearson in 1885, as a measure of association between samples. The correlation between X and Y is defined as

$$\mathbb{C}\text{or}\left(X, Y\right) = \frac{\mathbb{C}\text{ov}\left(X, Y\right)}{\sqrt{\mathbb{V}\left[X\right]\mathbb{V}\left[Y\right]}}. \tag{2.5}$$

Some properties of the correlation are derived from those of the covariance (symmetry, invariance by shifts). Some properties are specific to the correlation.

- The correlation is invariant by scalings: $\mathbb{C}\text{or}\left(aX, Y\right) = \mathbb{C}\text{or}\left(X, Y\right)$ for all $a \in \mathbb{R}$. Invariance by shifts and scalings means that the correlation is insensitive to the units used for X and Y.
- The correlation is always between -1 and $+1$. Indeed, for any pair of random variables X and Y with finite first two moments (i.e. $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[Y^2\right]$ are finite), the Cauchy-Schwarz inequality states that $\mathbb{E}\left[XY\right]^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$. If we apply this inequality to the variables $X - \mathbb{E}\left[X\right]$ and $Y - \mathbb{E}\left[Y\right]$, we obtain $\mathbb{C}\text{ov}\left(X, Y\right)^2 \leq \mathbb{V}\left[X\right]\mathbb{V}[Y]$ and thus $\mathbb{C}\text{or}\left(X, Y\right) \in [-1, 1]$. Furthermore, the equality holds only if $Y = aX + b$ for some real numbers a and b. Therefore, we have $\mathbb{C}\text{or}\left(X, Y\right) = 1$ (resp. $= -1$) if and only if $Y = aX + b$ with $a > 0$ (resp. with $a < 0$). Maximally correlated variables are perfectly aligned.

The latter property hints at the limitations of the correlation coefficient: it really only captures linear associations between variables; Example 2.1 showed that a strong nonlinear association can be missed completely by the correlation coefficient.

Empirical covariance and correlation We can define *empirical versions* of expectations, covariances and correlations, computed from data. Consider two samples, $x = (x_1, \ldots, x_n)$, and $y = (y_1, \ldots, y_n)$, considered to be realizations of random variables $X_{1:n}$ and $Y_{1:n}$, all distributed identically as X and Y. We replace expectations such as $\mathbb{E}\left[X\right]$ by empirical averages such as $n^{-1}\sum_{t=1}^{n} x_t$, denoted by $\bar{x}$. The reasoning is that the empirical and the theoretical quantities should be similar if the sample size is large enough, *under some assumptions.* Indeed, if the variables were i.i.d. and $\mathbb{E}\left[|X|\right] < \infty$, then the law of large numbers would state that

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow[n\to\infty]{a.s.} \mathbb{E}\left[X\right],$$

which means that $\mathbb{P}\left(\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} X_t = \mathbb{E}\left[X\right]\right) = 1$; in words: for every single experiment there is a n large enough so that $\bar{x}$ is close to $\mathbb{E}\left[X\right]$. Of course with time series data we will not be directly able to use the law of large numbers for i.i.d. variables. From Eq. (2.4), replacing expectations by averages we get the empirical covariance

$$\hat{\mathbb{C}}\text{ov}\left(x_{1:n}, y_{1:n}\right) = \frac{1}{n}\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\left(y_t - \bar{y}\right) = \frac{1}{n}\sum_{t=1}^{n} x_t y_t - \bar{x}\bar{y}. \tag{2.6}$$

With the same reasoning, if we replace $\mathbb{V}\left[X\right]$ by the empirical variance $\hat{\sigma}_x^2$ defined as $\hat{\mathbb{C}}\text{ov}\left(x_{1:n}, x_{1:n}\right)$, and if we replace $\mathbb{V}\left[Y\right]$ by $\hat{\sigma}_y^2 = \hat{\mathbb{C}}\text{ov}\left(y_{1:n}, y_{1:n}\right)$, then we can consider the empirical correlation defined as

$$\hat{\mathbb{C}}\text{or}\left(x_{1:n}, y_{1:n}\right) = \frac{n^{-1}\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\left(y_t - \bar{y}\right)}{\sqrt{\hat{\sigma}_x^2\hat{\sigma}_y^2}} = \frac{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)\left(y_t - \bar{y}\right)}{\sqrt{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2}}. \tag{2.7}$$

As mentioned earlier there are strong connections with linear regressions. Denote by $\hat{\beta}_{y|x}$ the slope of the regression of $y = y_{1:n}$ on $x = x_{1:n}$, and by $\hat{\beta}_{x|y}$ the regression slope for x on y. We can write these slopes in terms of the samples $x_{1:n}$ and $y_{1:n}$ explicitly:

$$\hat{\beta}_{y|x} = \frac{n^{-1}\sum_{t=1}^{n} x_t y_t - \bar{x}\bar{y}}{n^{-1}\sum_{t=1}^{n} x_t^2 - \left(\bar{x}\right)^2} \quad , \text{ and } \quad \hat{\beta}_{x|y} = \frac{n^{-1}\sum_{t=1}^{n} x_t y_t - \bar{x}\bar{y}}{n^{-1}\sum_{t=1}^{n} y_t^2 - \left(\bar{y}\right)^2}.$$

Then, the correlation coefficient in Eq. (2.7) satisfies the relationship $\hat{\mathbb{C}}\text{or}\left(x_{1:n}, y_{1:n}\right) = \hat{\beta}_{y|x}\hat{\sigma}_x/\hat{\sigma}_y$. By symmetry, we also have $\hat{\mathbb{C}}\text{or}\left(x_{1:n}, y_{1:n}\right) = \hat{\beta}_{x|y}\hat{\sigma}_y/\hat{\sigma}_x$ where we have simply swapped x and y. The correlation coefficient is therefore always somewhere between the two regression coefficients $\hat{\beta}_{y|x}$ and $\hat{\beta}_{x|y}$ (depending on $\hat{\sigma}_y/\hat{\sigma}_x$ being greater or less than one). More precisely, the correlation coefficient is the geometric mean of $\hat{\beta}_{y|x}$ and $\hat{\beta}_{x|y}$: $\hat{\mathbb{C}}\text{or}\left(x_{1:n}, y_{1:n}\right)^2 = \hat{\beta}_{y|x}\hat{\beta}_{x|y}$.

Consider standardized samples, that is, samples $x_{1:n}$ and $y_{1:n}$ with empirical mean of zero and empirical variance of one. Then $\hat{\mathbb{C}}\text{or}\left(x_{1:n}, y_{1:n}\right) = \hat{\beta}_{y|x} = \hat{\beta}_{x|y}$. Therefore, the correlation coefficient between two samples is the slope of the linear regression of the samples, one onto the other, up to standardization. If you picture a scatterplot of $x_{1:n}$ and $y_{1:n}$, then the correlation coefficient is one number that summarizes the whole graph. There is strictly more information in the plot than in the correlation coefficient that summarizes it.

2.6 Stationarity, autocorrelations and convergence

Back to stochastic processes and time series data, we are interested in the association between successive measurements in time. We define the autocovariance function γ and the autocorrelation

ρ as

$$\forall s,t \in \{1,\ldots,n\} \quad \gamma\left(s,t\right) = \mathbb{C}\mathrm{ov}\left(Y_s,Y_t\right) \qquad \text{and} \qquad \rho\left(s,t\right) = \frac{\mathbb{C}\mathrm{ov}\left(Y_s,Y_t\right)}{\sqrt{\mathbb{V}\left(Y_s\right)\mathbb{V}\left(Y_t\right)}}, \tag{2.8}$$

provided that these quantities exist (i.e. assuming that $\mathbb{V}\left[Y_t\right] < \infty$ for all t). Autocovariance and autocorrelation satisfy some properties such as symmetry, invariance by shifts, etc, inherited from the properties of covariance and correlation.

Now, let's try to define empirical versions of these quantities. Similarly to the previous section, we can replace expectations by empirical averages if we have representative samples from the underlying distributions. Recalling that $\mathbb{C}\mathrm{ov}\left(Y_s,Y_t\right) = \mathbb{E}\left[Y_sY_t\right] - \mathbb{E}\left[Y_s\right]\mathbb{E}\left[Y_t\right]$, we could replace these expectations by empirical averages if we had representative samples of the variables Y_sY_t, Y_s and Y_t. But here we face the fundamental difficulty with time series, alluded to in the first chapter: with a time series $(y_1,\ldots,y_n)$ we only have one realization of Y_t at each time t. One value of y_sy_t, one value of y_s and one value of y_t. That does not seem quite enough to approximate $\mathbb{C}\mathrm{ov}\left(Y_s,Y_t\right)$.

Stationarity. To resolve that conundrum we introduce *stationary processes*. We say that a process (Y_t) with $t \in \mathbb{Z}$ (that is, t is any integer between $-\infty$ and $+\infty$) is stationary if the following properties hold:

- The expectation of Y_t is constant for all $t \in \mathbb{Z}$: for some $\mu \in \mathbb{R}$, $\mathbb{E}\left[Y_t\right] = \mu$.
- The variance $\mathbb{V}\left[Y_t\right]$ is finite for each time t, and the autocovariance function as defined in Eq. (2.8) is such that $\gamma\left(s,t\right) = \gamma\left(s',t'\right)$ whenever $s,t,s',t' \in \mathbb{Z}$ are such that $|t-s| = |t'-s'|$.

This is sometimes called *weak stationarity.* There is a notion of *strict* or *strong* stationarity: a process (Y_t) is strictly stationary, if for any $k \in \mathbb{N}$, the distribution of any vector $(Y_{t_1},\ldots,Y_{t_k})$ is invariant by time shifts, i.e. is the same distribution as the distribution of $(Y_{h+t_1},\ldots,Y_{h+t_k})$. Strict stationarity implies weak stationarity, but the converse is false. For example, consider a sequence of independent random variables (Y_t) where $Y_t \sim \mathcal{N}\left(1,1\right)$ if t is an odd number, and $Y_t \sim \mathrm{Exponential}\left(1\right)$ if t is an even number. One can check that this process is weakly stationary but not strictly stationary.

A stationary process is a collection of variables that might be dependent and evolving over time. By definition, for stationary processes, $\gamma\left(s,t\right)$ is equal to $\gamma\left(s-1,t-1\right)$, and to $\gamma\left(s-2,t-2\right)$, and to $\gamma\left(0,t-s\right)$: it only depends on $t-s$, not on the particular values of t and s. This seems useful: we have only one value of y_s and y_t, but we have many pairs (y_s,y_t) with a fixed time shift $|t-s|$. Since $\gamma(s,t)$ depends only on $|t-s|$, we re-define autocovariance and autocorrelation for stationary processes:

$$\forall h \in \mathbb{Z} \quad \gamma(h) = \mathbb{C}\mathrm{ov}\left(Y_1,Y_{1+h}\right) \qquad \text{and} \qquad \rho(h) = \frac{\gamma\left(h\right)}{\gamma\left(0\right)} = \frac{\mathbb{C}\mathrm{ov}\left(Y_1,Y_{1+h}\right)}{\mathbb{V}\left[Y_1\right]}. \tag{2.9}$$

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

Empirical approximations. Focusing on stationary processes, we go back to the problem of devising an approximation of expectations and autocovariances from samples.

Since the mean $\mathbb{E}[Y_t] = \mu$ is constant over t, it sounds reasonable to approximate it by the empirical average $\bar{y} = n^{-1}\sum_{t=1}^{n} y_t$ Indeed the expectation of $\bar{Y}$ is exactly μ: $\mathbb{E}\left[\bar{Y}\right] = \mu$. This does not necessarily mean that $\bar{Y}$ is “close” to μ, though. Via Chebyshev’s inequality, we have, for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\bar{Y} - \mu\right| > \epsilon\right) \leq \frac{\mathbb{V}[\bar{Y}]}{\epsilon^2}.$$

Thus, to deal with the probability that $\left|\bar{Y} - \mu\right|$ is large, we can ensure that $\mathbb{V}[\bar{Y}]$ is small. This will be the case, for large enough n, under some assumptions on $(Y_t)_{t\in\mathbb{Z}}$. So, let’s compute the variance of $\bar{Y}$; first, we can always write

$$\mathbb{V}[\bar{Y}] = \frac{1}{n^2}\sum_{s,t=1}^{n} \mathbb{C}\mathrm{ov}\left(Y_t, Y_s\right) = \frac{1}{n^2}\sum_{s,t=1}^{n} \gamma(t,s) = \frac{1}{n^2}\sum_{s,t=1}^{n} \gamma(|t-s|).$$

We can write $\sum_{s,t=1}^{n}\gamma(|t-s|)$ as $n\gamma\left(0\right)+2\left(n-1\right)\gamma(1)+2\left(n-2\right)\gamma(2)+\ldots+2\left(n-\left(n-1\right)\right)\gamma(n-1)$; to see this, it might be useful to see $\gamma(t,s)$ for $t,s \in \{1,\ldots,n\}$ in a big $n \times n$ matrix, with $\gamma(0)$ on the diagonal, $\gamma(1)$ just below and above the diagonal, etc., up to $\gamma(n-1)$ in the corners. This leads to

$$\mathbb{V}\left[\bar{Y}\right] = \frac{1}{n}\left(\gamma\left(0\right) + 2\left(1-\frac{1}{n}\right)\gamma(1) + 2\left(1-\frac{2}{n}\right)\gamma(2) + \ldots + 2\left(1-\frac{n-1}{n}\right)\gamma(n-1)\right).$$

The above formula holds for the average of any stationary process. We argue that if $\left|\gamma\left(h\right)\right|$ goes to zero when h goes to infinity, then $\mathbb{V}\left[\bar{Y}\right]$ will go to zero when n goes to infinity. First, note that $1 - t/n \leq 1$ for all $t \geq 1$, and therefore

$$\mathbb{V}[\bar{Y}] \leq \frac{1}{n}\sum_{h=-(n-1)}^{n-1} |\gamma(h)|.$$

The right hand side is going to zero if $\left|\gamma\left(h\right)\right| \to 0$. Indeed, it is essentially the average of a convergent sequence going to zero. More formally, $\left|\gamma\left(h\right)\right| \to 0$ means that for all $\epsilon > 0$, there exists $N \in \mathbb{N}$

such that for all $h \geq N$, $|\gamma(h)| \leq \epsilon$. Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{h=-(n-1)}^{n-1}|\gamma(h)| &= \frac{1}{n}\left(\gamma(0) + 2\sum_{h=1}^{N-1}|\gamma(h)|\right) + 2\frac{1}{n}\sum_{h=N}^{n-1}|\gamma(h)| \\ &\leq \frac{1}{n}\left(\gamma(0) + 2\sum_{h=1}^{N-1}|\gamma(h)|\right) + 2\epsilon,\end{aligned}$$

where the last inequality comes from each $|\gamma(h)|$ being less than ϵ in $\sum_{h=N}^{n-1}|\gamma(h)|$, and there being less than n terms in that sum. When n is large enough, $n^{-1}(\gamma(0) + 2\sum_{h=1}^{N-1}|\gamma(h)|)$ is arbitrarily small since the term in parenthesis is constant in n; therefore it eventually becomes less than ϵ. We conclude that the right hand side of the last inequality is less than 3ϵ. Since this is true for any ϵ, we conclude that $\mathbb{V}[\bar{Y}]$ goes to zero when $n \to \infty$. Therefore, we have proved the following statement, which is a "law of large numbers" for stationary processes:

Proposition. *Let $(Y_t)_{t\in\mathbb{Z}}$ be a stationary process with mean μ and autocovariance function γ that satisfies $\gamma(h) \to 0$ when $h \to \infty$. Then for all $\epsilon > 0$, $\mathbb{P}\left(\left|n^{-1}\sum_{t=1}^{n}Y_t - \mu\right| > \epsilon\right)$ goes to zero when $n \to \infty$. In other words $n^{-1}\sum_{t=1}^{n}Y_t$ converges to μ in probability.*

Similarly we can approximate the variance $\mathbb{V}[Y_t]$ by the empirical variance $\hat{\sigma}_y^2 = n^{-1}\sum_{t=1}^{n}y_t^2 - (\bar{y})^2$. Note how amazing this is: by definition, $\mathbb{E}[Y_t]$ refers to the expectation of the random variable Y_t over *repeated independent draws* at time t; but we can in fact approximate it by averaging over *time*, i.e. "ergodic averages" converge to "population averages". To approximate $\gamma(h) = \mathbb{C}\mathrm{ov}(Y_1, Y_{1+h}) = \mathbb{C}\mathrm{ov}(Y_2, Y_{2+h}) = \ldots = \mathbb{C}\mathrm{ov}(Y_t, Y_{t+h})$ for all t, we use all the pairs (y_t, y_{t+h}), for $t \in \{1, \ldots, n-h\}$:

$$\forall h \in \mathbb{Z} \quad \hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(y_t - \bar{y})(y_{t+h} - \bar{y}). \tag{2.10}$$

There are only $n - h$ terms in the sum, but we normalize by n, which is simpler, and also preserves non-negative definiteness of the autocovariance function. We also define the sample autocorrelation

$$\forall h \in \mathbb{Z} \quad \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}(y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}. \tag{2.11}$$

Note that the sample autocovariance (resp. autocorrelation) is very similar to the sample covariance (resp. correlation) between $y_1, \ldots, y_{n-h}$ and $y_{1+h}, \ldots, y_n$, i.e. obtained with Eq. (2.6). The slight advantage of Eq. (2.10) lies in the approximation of the mean $\mathbb{E}[Y_t]$ by all of the sample $y_{1:n}$, instead of the smaller samples $y_{1:n-h}$ and $y_{1+h:n}$. Likewise, we can think of $\hat{\rho}(h)$ essentially as the empirical correlation between $y_{1:n-h}$ and $y_{1+h:n}$, and thus the intuition developed for the correlation coefficient (e.g. the relationship with linear regression) carries over to the empirical autocorrelation.

Correlograms. As a visual diagnostic, it is common to represent the empirical autocorrelation $\hat{\rho}\left(h\right)$ as a function of h; h is called the *lag*. Since $\hat{\rho}\left(h\right) = \hat{\rho}\left(-h\right)$ for all h, it is enough to plot $\hat{\rho}\left(h\right)$ for $h \geq 0$; note that $\hat{\rho}\left(0\right) = 1$ always. This kind of plot is called an *autocorrelogram* or *autocorrelation plot*, and can be produced by the function `acf` of `R`.

Since the empirical correlation $\hat{\rho}\left(h\right)$ is essentially the slope in the regression of $y_{1+h:n}$ on $y_{1:n-h}$, it can be informative to plot $y_{1+h:n}$ on the y-axis and $y_{1:n-h}$ on the x-axis. This is called a *lag-plot* with lag h, and can be produced by the function `lag.plot` of `R`. The advantage of the autocorrelogram lies in the possibility of looking at many lags h quickly. On the other hand, it reduces each lag-plot to a single number and is thus necessarily restrictive.

Another common tool is the *partial autocorrelation*, useful in part because it allows the identification of the order of autoregressive processes that we will introduce later. The partial autocorrelation can be understood informally as removing the impact of Y_{t-1} on both Y_t and Y_{t-2}, and looking at the residual correlations. To remove the impact, consider a regression of Y_t on Y_{t-1} and of Y_{t-2} on Y_{t-1}. That is, find $\hat{a}_1$ and $\hat{b}_1$ that minimize

$$\mathbb{E}\left[\left(Y_t - a_1 Y_{t-1}\right)^2\right] \text{ and } \mathbb{E}\left[\left(Y_{t-2} - b_1 Y_{t-1}\right)^2\right]$$

over all a_1 and b_1, respectively. Define the residuals as $\varepsilon_t = Y_t - \hat{a}_1 Y_{t-1}$, and , $\delta_{t-2} = Y_{t-2} - \hat{b}_1 Y_{t-1}$. Then the partial autocorrelation of lag 2 is defined as $\alpha\left(2\right) = \mathbb{C}\text{or}\left(\varepsilon_t, \delta_{t-2}\right)$. By convention $\alpha\left(1\right) = \rho\left(1\right) = \mathbb{C}\text{or}\left(Y_t, Y_{t+1}\right)$. For $h > 2$, we define similarly the regressions onto the $h-1$ intermediate variables, and the residuals

$$\begin{aligned} &\varepsilon_t = Y_t - \left(\hat{a}_1 Y_{t-1} + \hat{a}_1 Y_{t-2} + \ldots + \hat{a}_{h-1} Y_{t-h+1}\right) \text{ and} \\ &\delta_{t-h} = Y_{t-h} - \left(\hat{b}_1 Y_{t-h+1} + \hat{b}_2 Y_{t-h+2} + \ldots + \hat{b}_{h-1} Y_{t-1}\right). \end{aligned}$$

Then the partial autocorrelation of lag h is defined as $\alpha\left(h\right) = \mathbb{C}\text{or}\left(\varepsilon_t, \delta_{t-h}\right)$.

Remark 2.2. In the above equation, $\hat{a}_1$ and $\hat{b}_1$ are different for each value of $h \geq 2$. To be more rigorous, we would need to index them with h, e.g. writing $\hat{a}_{h1}$ and $\hat{b}_{h1}$.

Let's consider another way of assessing the relationship between Y_t and Y_{t-h} given the intermediate variables, by performing the regression:

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

Then we could interpret c_h as the impact of Y_{t-h} on Y_t, taking $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-h+1}$ into account. It turns out that c_h is exactly equal to $\alpha\left(h\right)$ defined before. In `R` the `pacf` function produces a plot of partial autocorrelations as a function of the lag h.

2.7 AR and MA models

We conclude these notes by introducing two classic families of models: autoregressive and moving average models. Both build upon a white noise process, (W_t), with zero mean and variance σ_W^2, but combine successive terms differently so as to create non-trivial correlations between successive variables.

Autoregressive process. We define an autoregressive process (Y_t) of order p as a process satisfying

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + W_t, \tag{2.12}$$

where $p \in \mathbb{N}$ and $\varphi_1, \ldots, \varphi_p$ are the autoregressive coefficients in $\mathbb{R}$, with $\varphi_p \neq 0$. Either we define (Y_t) for $t \in \mathbb{Z}$, and then the process is infinitely long. Or we start the process at some time, e.g. $t = 1$, and then we need to specify the initial distribution of $(Y_1, Y_2, \ldots, Y_p)$. Depending on the coefficients and the initial distribution, the process can be stationary, and such that $\gamma(h) \to 0$ as $h \to \infty$. The computation is particularly simple when $p = 1$, in which case we find that $\rho(h) = \varphi_1^{|h|}$ for all $h \in \mathbb{Z}$. Remarkably, for an autoregressive process of order p, the partial autocorrelations at lag $h > p$ are all equal to zero.

The name "autoregression" comes from an analogy with linear regressions: Y_t is modeled (or "predicted") as a linear combination of the past p observations, plus some noise. Autoregressive processes are typically fit by maximum likelihood estimation, but we could also use the method of moments as follows.

From the definition $Y_t = \sum_{i=1}^p \varphi_i Y_{t-i} + W_t$, multiply both sides by Y_{t-h} for $h \geq 0$ and take the expectation. This gives the following "Yule-Walker" equations:

$$\begin{aligned} \forall h \in \{1, \ldots, p\} \quad \gamma(h) &= \sum_{i=1}^p \varphi_i \gamma(h-i) \\ \text{and} \quad \sigma^2 &= \gamma(0) - \sum_{i=1}^p \varphi_i \gamma(i). \end{aligned}$$

Replacing each $\gamma(h)$ by an estimate $\hat{\gamma}(h)$, we have $p+1$ equations and $p+1$ unknowns, we can these equations to obtain parameter estimates.

Moving averages. The second fundamental class of time series models is called MA(q) models (for *moving average* of order q). We say that (Y_t) is a MA(q) process if

$$Y_t = W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q}, \tag{2.13}$$

where $\theta_q \neq 0$. There is an implicit coefficient $\theta_0 = 1$ in the above equation; if θ_0 was different than 1, we could always rescale the variance σ_W^2 of the noise terms (W_t) to obtain an equivalent process with $\theta_0 = 1$. By direct calculation we can obtain the autocovariance function of MA(q) processes as

$$\forall h \text{ such that } |h| \leq q \quad \gamma(h) = \sigma_W^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|},$$

where $\theta_0 = 1$. When $|h| > q$, we have $\gamma(h) = 0$. In other words, an MA(q) process has an autocovariance of zero beyond lag q. For any choice of coefficients $\theta_1, \ldots, \theta_q$, the MA(q) process is stationary.

We could try to use the method of moments for MA models. For instance, for an MA(1) model we have $\gamma(0) = \sigma^2(1 + \theta_1^2)$ and $\gamma(1) = \sigma^2\theta_1$. These two equations might be enough to identify the two parameters (θ_1, σ^2). We can replace the autocovariances by empirical counterparts, and divide the second equation by the first to obtain $\hat{\theta}_1/(1 + \hat{\theta}_1^2) = \hat{\gamma}\,(1)\,/\hat{\gamma}\,(0) = \hat{\rho}\,(1)$. This is a quadratic equation in $\hat{\theta}_1$, that can be written $\hat{\rho}\,(1)\,\hat{\theta}_1^2 - \hat{\theta}_1 + \hat{\rho}\,(1) = 0$. First problem: if $|\hat{\rho}\,(1)| > \frac{1}{2}$, the equation does not have any real-valued solution. Second problem: if $|\hat{\rho}\,(1)| < \frac{1}{2}$, the equation has two solutions. Indeed, one can check that if $\hat{\theta}_1$ is solution, then $1/\hat{\theta}_1$ is also solution. It will be simpler to calibrate both AR and MA models by maximum likelihood estimation, as done in the `arima` function of `R`.

The calculation of the likelihood for both AR and MA models will be covered in a later chapter, as part of a description of the encompassing class of ARMA models, which are arguably the most important statistical models for time series analysis.