Machine Learning of Dynamic Processes with Applications to Time Series Forecasting

Lyudmila Grigoryeva

University of St. Gallen, Switzerland

Emergent Algorithmic Intelligence Winter School 2023
JGU Research Center for Algorithmic Emergent Intelligence
Mainz (Nierstein), 2023

Outline

1. Forecasting time series

2. Classical time series models

3. Wold decomposition theorem

Outline for section 1

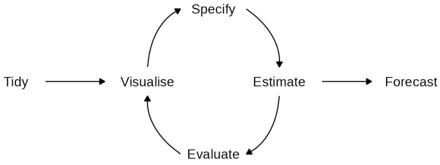
1. Forecasting time series

2. Classical time series models

3. Wold decomposition theorem

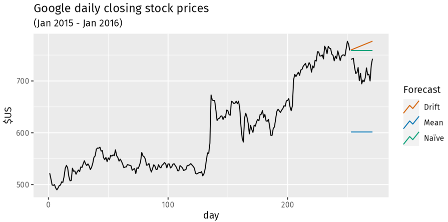
Forecasting time series

- data from past may contain information on the future development of a variable
- forecasting future developments requires certain regularities or structures in the data
- time series analysis helps to detect such characteristics and helps to understand the 'data generating mechanism'



Forecasting time series: simple

- Mean: $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)\,/\,T$.
- Naïve: $\hat{y}_{T+h|T} = y_T$.
- Seasonal naïve: $\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$, where $m =$ the seasonal period, and k is the integer part of $(h-1)/m$ (i.e., the number of complete years in the forecast period prior to time $T+h$).
- Drift: $\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right)$.



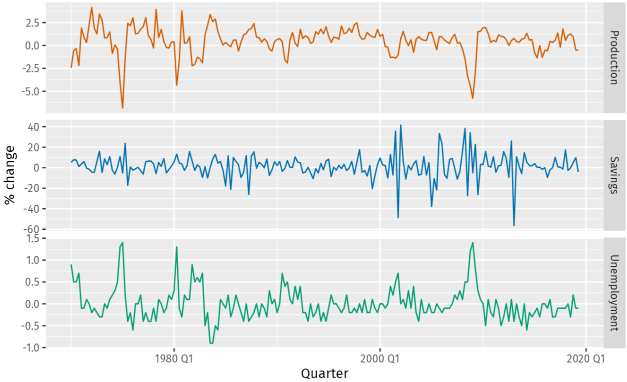
Forecasting time series: linear regression (revisited)

In the simplest case, the regression model allows for a linear relationship between the forecast variable y and k predictor variable series x:

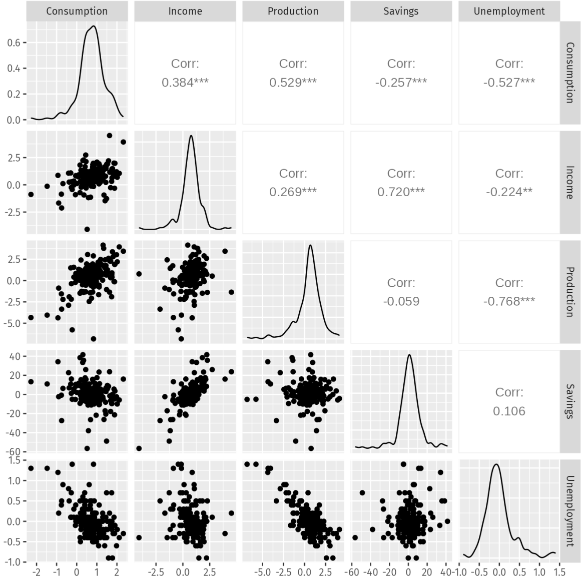
$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t.$$

1. Estimate coefficients using the data up to T
2. To form a forecast h steps into the future we use as predictors their lagged values.:
$$y_{t+h} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \cdots + \hat{\beta}_k x_{k,t}$$
for $h = 1, 2 \ldots$. The predictor set is formed by values of the x s that are observed h time periods prior to observing y. Therefore when the estimated model is projected into the future, i.e., beyond the end of the sample T, all predictor values are available.

Forecasting time series: linear regression (revisited)



Forecasting time series: linear regression (revisited)



Outline for section 2

1. Forecasting time series

2. Classical time series models

3. Wold decomposition theorem

Autoregressive model

An *autoregressive model* of order p or (AR(p)) can be written as

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t = \sum_{i=1}^{p} \varphi_i L^i X_t + \varepsilon_t, \quad \text{or}$$

$$\varepsilon_t = \left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) X_t = \Phi(L) X_t,$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, ε_t is white noise, L is the lag operator, and $\Phi(L)$ is the lag polynomial of order p.

For an AR(p) model to be weak-sense stationary, the roots of the polynomial $1 - \sum_{i=1}^{p} \varphi_i z^i$ must lie outside the unit circle, that is $|z_i| > 1$ should hold for all $i = 1, \ldots, p$.

Moving average model

An *moving average model* of order q or (MA(q)) can be written as

$$X_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t = \Theta(L)\varepsilon_t,$$

where $\theta_1, \ldots, \theta_q$ are the parameters of the model, ε_t is white noise, L is the lag operator, and $\Theta(L)$ is the lag polynomial of order q.

For a MA(q) model to be invertible, the roots of the polynomial $\Theta(z) := 1 - \sum_{i=1}^{q} \theta_i z^i$ must lie outside the unit circle, that is $|z_i| > 1$ should hold for all $i = 1, \ldots, q$.
For invertible MA(q) one has

$$\varepsilon_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)^{-1} X_t = \theta(L)^{-1} X_t,$$

Autoregressive moving average model

An *autoregressive moving average model* ARMA(p, q)) can be written as

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t,$$

or

$$\Phi(L) X_t = \Theta(L) \varepsilon_t.$$

For invertible $\Theta(L)$ part one has

$$\frac{\Phi(L)}{\Theta(L)} X_t = \varepsilon_t$$

or for stationary autoregressive component

$$X_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t = \Psi(L) \varepsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

Autoregressive integrated moving average model

Consider *autoregressive moving average model* ARMA$(\tilde{p}, q)$)

$$\left(1-\sum_{i=1}^{\tilde{p}} \tilde{\varphi}_i L^i\right) X_t = \left(1+\sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

or

$$\tilde{\Phi}(L) X_t = \Theta(L) \varepsilon_t.$$

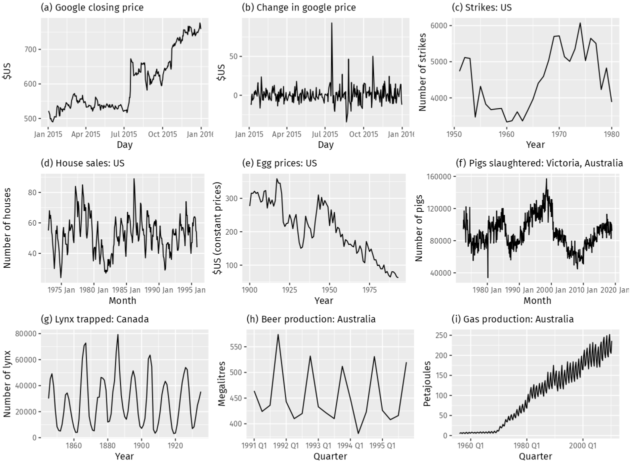
Let $\tilde{\Phi}(L)$ have a unit root of multiplicity d. Then:

$$\left(1-\sum_{i=1}^{\tilde{p}} \alpha_i L^i\right) = \left(1-\sum_{i=1}^{\tilde{p}-d} \varphi_i L^i\right) (1-L)^d.$$

An ARIMA (p, d, q) process expresses this polynomial factorisation property with $p = \tilde{p} - d$, and is given by:

$$\left(1-\sum_{i=1}^{p} \varphi_i L^i\right) (1-L)^d X_t = \left(1+\sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

Stationarity of time series



(a) Google closing price

(b) Change in google price

(c) Strikes: US

(d) House sales: US

(e) Egg prices: US

(f) Pigs slaughtered: Victoria, Australia

(g) Lynx trapped: Canada

(h) Beer production: Australia

(i) Gas production: Australia

Remedies

- Differencing:

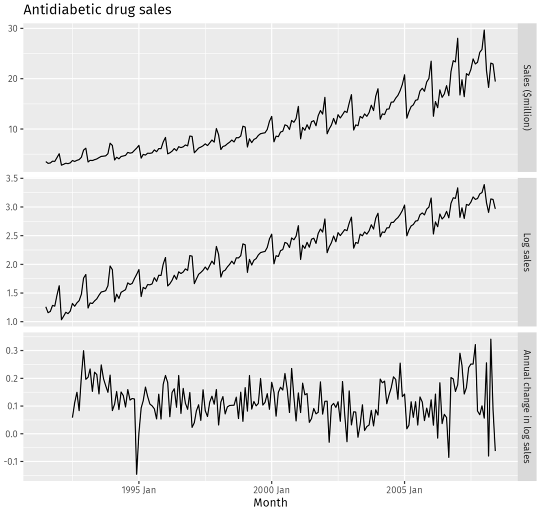
$$\Delta y_t = y'_t = y_t - y_{t-1}.$$

- Second-order differencing:

$$\begin{aligned}\Delta(\Delta y_t) = y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}\end{aligned}$$

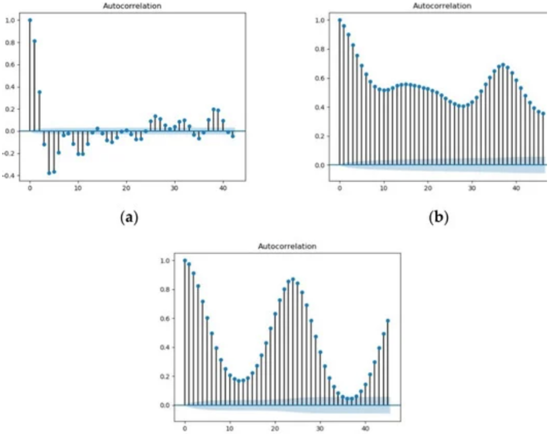
- Seasonal differencing: $y'_t = y_t - y_{t-m}$.

Remedies


Antidiabetic drug sales
30
20
10
Sales ($million)
3.5
3.0
2.5
2.0
1.5
1.0
Log sales
0.3
0.2
0.1
0.0
-0.1
Annual change in log sales
1995 Jan
2000 Jan
2005 Jan
Month

How to detect whether differencing is needed?

We plot autocorrelations and study their decay!



(a)

(b)

Outline for section 3

1. Forecasting time series
2. Classical time series models
3. Wold decomposition theorem

Wold decomposition theorem (1936)

Any zero-mean covariance-stationary process $\{z_t, t \in \mathbb{Z}\}$ can be represented in the form $z_t = u_t + d_t$, where $\{u_t\}$ and $\{d_t\}$ are the decorrelated MA(∞) and a deterministic process, respectively. Let $\mathcal{M}_t = \overline{\mathrm{span}}\{z_s, s \in \mathbb{Z}, s \leq t\}$, the one-step mean squared error $\sigma^2 := \mathbb{E}[|z_{t+1} - P_{\mathcal{M}_t} z_{t+1}|^2]$ and the closed linear subspace $\mathcal{M}_{-\infty}$

$$\mathcal{M}_{-\infty} = \cap_{t=-\infty}^{\infty} \mathcal{M}_t$$

of the Hilbert space $\mathcal{M} = \overline{\mathrm{span}}\{z_t, t \in \mathbb{Z}\}$. All subspaces and orthogonal complements should be interpreted as relative to $\mathcal{M}$.

Remark

The process $\{d_t\}_{t\in\mathbb{Z}}$ is said to be deterministic if and only if $\sigma^2 = 0$, or equivalently if and only if $d_t \in \mathcal{M}_{-\infty}$, for each t.

Wold decomposition theorem (1936)

Theorem

Any zero-mean covariance-stationary process $\{z_t\}_{t\in\mathbb{Z}}$ *with* $\sigma^2 > 0$ *can be represented as*

$$z_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} + d_t,$$

where

(i) $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,

(ii) $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$

(iii) $z_t \in \mathcal{M}_t$, *for each* $t \in \mathbb{Z}$

(iv) $\mathbb{E}[\epsilon_t d_s] = 0$, *for all* $t, s \in \mathbb{Z}$

(v) $d_t \in \mathcal{M}_{-\infty}$, *for each* $t \in \mathbb{Z}$

(vi) $\{d_t\}$ *is deterministic.*

Wold decomposition theorem (1936)

In this theorem the sequences defined as

(i) $\epsilon_t = z_t - P_{\mathcal{M}_{t-1}} z_t = z_t - P_{z_{t-1}} z_t$ with $\mathbb{P}_{z_{t-1}} : \mathcal{L}^2(\Omega, \mathbb{R}) \to \overline{\mathrm{span}} \{z_{t-1}, z_{t-2,1}, \ldots\}$ in the linear projector

(ii) $\psi_j = \langle z_t, \epsilon_{t-j} \rangle / \sigma^2$, with $\sigma^2 = \mathrm{Var}(\epsilon_t)$

(iii) $d_t = z_t - \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$

satisfy conditions (i)-(vi) above and can be shown to be unique. Additionally,

$$\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = P_{\varepsilon_t} z_t$$

and, as d_t is deterministic then

$$d_t = P_{d_{t-1}} d_t.$$

Fit an ARMA model to the time series. Extract the fitted values as the deterministic component. Calculate the residuals, representing the stochastic component. Plot the original time series, fitted values, and residuals.