Data Science
Autoregessive models

Estimating linear temporal latent factors

Stéphane Marchand-Maillet

Department of Computer Science

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Master en Sciences Informatiques - Autumn semester

Table of contents

What is the lecture about?

- ⋆ To understand temporal data modeling
- ⋆ To understand temporal data analysis
- ⋆ To understand the concept of auto-regression

$\rightarrow$ linear temporal latent model

- ⋆ To present the basic of autoregressive models (AR(p))

Reading: [3] and [1] (chap 13)

Definition

- A temporal series is a sequence of observations depending on time

$$y_1, y_2, \ldots, y_t, \ldots, y_T \qquad y_t \in \Omega \subseteq \mathbb{R}^D$$

- In general, we assume a constant time interval
- For statistical temporal data analysis, we study causality

$\Rightarrow$ y_t depends on the p preceding values $y_{t-1}, \ldots, y_{t-p}$

- Causality is the information we want to model (give/measures parameters)

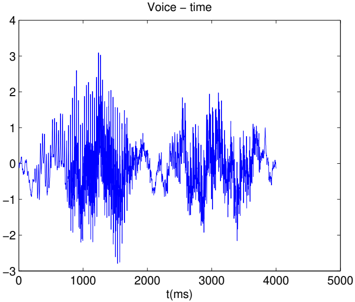
Applications

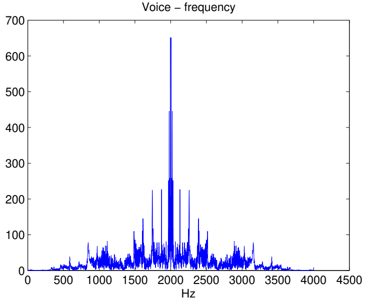
- ⋆ Économetry (trend prediction, risk analysis...)
- ⋆ Social (demography, migrations ...)
- ⋆ Physical measures (explanation of physical constants/values)
- ⋆ Communication and information $\rightarrow$ telecom, network, coding, speech, video ...

Examples

- ⋆ Temporal data analysis (mean, variance, correlation)
- ⋆ Frequency analysis (periodicity)
- ⋆ Combination time/frequency, time/scale

→ Non-stationary series





Temporal series

- Definition

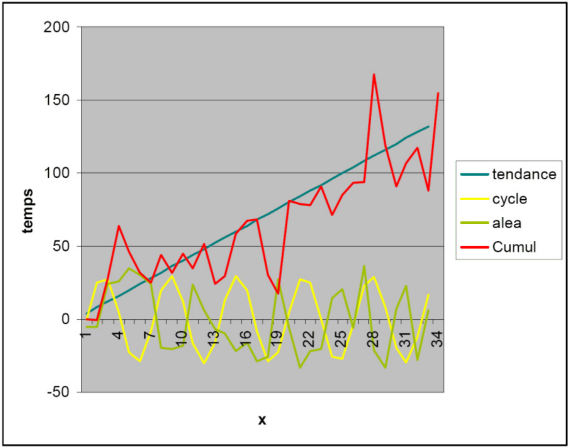
$$y_t = g(t) + \epsilon_t, \quad t \in [\![T]\!]$$

- y_t scalar or vector
- $g(t)$ structure of the series
- ϵ_t noise centered, uncorrelated (white noise)

$\Rightarrow$ ϵ_t random component of y_t

Components

* ⋆ Long-term: the trend
* ⋆ Periodic: the seasonality
* ⋆ Random: the residual



Components

* ⋆ Additive scheme $y_t = f_t + s_t + \epsilon_t$
* ⋆ Multiplicative scheme $y_t = f_t \odot s_t \odot (1 + \epsilon_t)$
  (and variations)

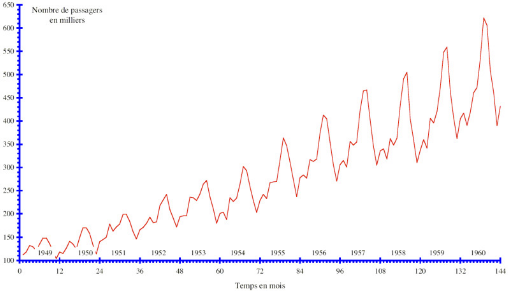


FIG. 4 – Nombre mensuel de passagers internationaux aux États Unis de 1949 à 1960

Note: Multiplicative is equivalent to the additive scheme when considering the log of the series

Component estimation

A rigorous analysis requires the estimation of the trend, independent from seasonality and noise

Smoothing

By definition, s_t and ϵ_t are 0-mean:

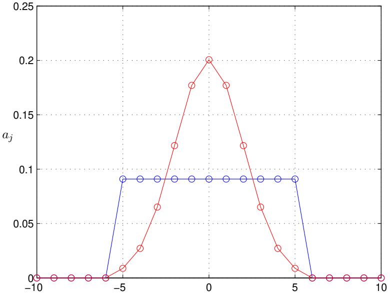
$$\mathbb{E}S = \mathbb{E}E = 0$$

A moving average of y_t preserves the trend only

Moving average Ψ_p

$$\Psi_p(y_z) = \bar{y}_t = \sum_{j=-p}^{p} a_j y_{t-j}$$

with $a_{-j} = a_j$ (symmetric filter) with $\sum_j a_j = 1$.



Note: This is equivalent to a convolution

Correcting the seasonal variations

- ⋆ Apply moving average to y_t

$$\mathbf{\Psi}_p(\mathbf{y}_t) = \mathbf{\Psi}_p(\mathbf{f}_t) + \mathbf{\Psi}_p(\mathbf{s}_t) + \mathbf{\Psi}_p(\boldsymbol{\epsilon}_t)$$

- ⋆ if $\mathbf{s}_t$ and $\boldsymbol{\epsilon}_t$ centered, $\mathbf{\Psi}_p(\mathbf{s}_t)$, $\mathbf{\Psi}_p(\boldsymbol{\epsilon}_t) \approx 0$

$$\hat{\mathbf{f}}_t = \mathbf{\Psi}_p(\mathbf{y}_t)$$

$\Rightarrow$ To "erase" a seasonality with period p, apply moving average $\mathbf{\Psi}_p$

Prediction

- ⋆ Assume we know $\mathbf{y}_1, \dots, \mathbf{y}_T$ stationary
- $\Rightarrow$ We wish to infer step $t + p$ ($\mathbf{y}_{t+p}$) from steps until t ($\mathbf{y}_t$) $\rightarrow \hat{\mathbf{y}}_t(p)$
- ⋆ k is the horizon
- ⋆ In general $\hat{\mathbf{y}}_t(p) \neq \mathbf{y}_{t+p}$ (prediction error)$\rightarrow$ we seek minimum error

No prediction is possible if there is no dependence/causality between values $\mathbf{y}_t$, $t \in [\![T]\!]$

For the rest of the lecture: $D = 1$ (scalar series):

- ⋆ Can be generalized per component
- ⋆ i.e we consider independent components
- $\Rightarrow$ simplify notation ($y_t \in \mathbb{R}$)

Linear prediction

We assume that y_t depends linearly on the previous values

$$y_t = \sum_{k=1}^{n} \alpha_k y_{t-k}$$

$$y_t = \boldsymbol{\alpha}^\mathsf{T} \mathbf{y}_{t-1}$$

with $\boldsymbol{\alpha} = [\alpha_1, \ldots \alpha_n]^\mathsf{T}$ and $\mathbf{y}_{t-1} = [y_{t-1}, \ldots, y_{t-n}]^\mathsf{T}$

$\Rightarrow$ Linear relationship $\rightarrow$ correlation between values y_t φ

Stationary process

- ⋆ The series is seen as a stochastic process
  - ∘ i.e y_t is a realization (draw/instance) of random variable Y_t
  - ∘ In general we know only one realization of $\{Y_t\}_{t \in [\![T]\!]}$

$\Rightarrow$ impossible to use correlation between $y_{t-1}, \ldots, y_{t-p}$, to predict y_t

Stationary hypothesis (constant mean):

Covariance $\mathrm{cov}(y_t, y_{t-p}) = \gamma_p$ does not depend on t

$$\gamma_p = \mathrm{cov}(y_t, y_{t-p}) = \frac{1}{T-p} \sum_{t=p}^{T} (y_t - \mu)(y_{t-p} - \mu)$$

Prediction by Exponential smoothing

$$\hat{y}_t(p) = (1 - \tau) \sum_{j=0}^{p} \tau^j y_{t-1-j}$$

with $\tau \in [0, 1]$

- ⋆ Most basic method
- ⋆ The predicted value is the mean of past values
- ⋆ forget effect (exponential decay) making most recent values important

Exponential smoothing

φ Show:

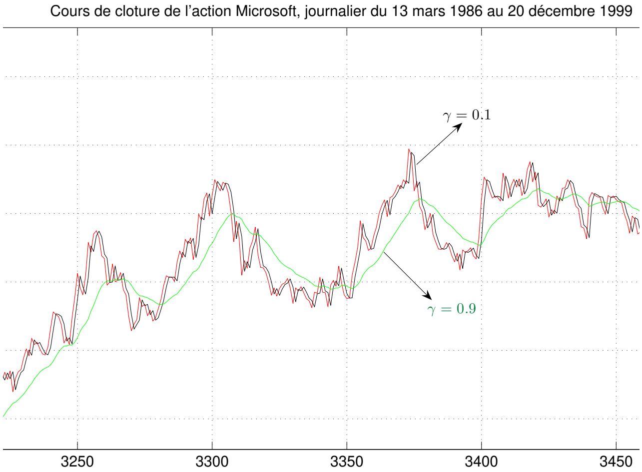
$$\hat{y}_{t+1}(p) = (1-\tau)y_t + \tau\hat{y}_t(p-1)$$

or:

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

- ⋆ $\tau \to 1$ prediction accounts for far past (smooth)
- ⋆ $\tau \to 0$ prediction only depends on immediate past (less smooth)

Exponential smoothing



Auto-regressive models (AR)

- ⋆ order-p AR model: AR(p):

$$y_t = \sum_{j=1}^{p} a_j y_{t-j} + \epsilon_t$$

- ⋆ Values at t linearly depends the on the p preceding values
- ⋆ ϵ_t is a white noise with variance σ_ϵ^2 ($\epsilon_t \sim \mathcal{N}(0, \sigma)$)

Parameter estimation

- ⋆ Mean-squared error regression:

$$\mathbf{a}^* = \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}[e_t^2] = \operatorname*{argmin}_{\mathbf{a}} \sum_{t=0}^{T} \left(y_t - \sum_{j=1}^{p} a_j y_{t-j} \right)^2$$

- ⋆ Solution

$$\frac{\partial \mathbb{E}[e_t^2]}{\partial a_j} = 0 \qquad j \in [\![p]\!]$$

Correlation and regression coefficient

- ⋆ We get for j:

$$\frac{\partial E[e_t^2]}{\partial a_j} = 2\mathbb{E}_t\left[\left(y_t - \sum_{i=1}^{p} a_i y_{t-i}\right) y_{t-j}\right]$$
$$= \gamma_j - \sum_{i=1}^{p} a_i \gamma_{i-j} = 0$$

- ⋆ Recall $\gamma_j = \mathrm{cov}(y_t, y_{t-j}) = \gamma_{-j}$ and y_t stationary

Yule-Walker Equations

Let $\rho_k = \frac{\gamma_k}{\gamma_0}$ be the correlation function
$\Rightarrow$ the minimization of mean squared error leads to

$$\rho_k = \sum_{i=1}^{p} a_i \rho_{k-i} \qquad k > 0$$

$$\begin{cases} \rho_1 = a_1\rho_0 + a_2\rho_1 + a_3\rho_2 + \cdots + a_p\rho_{p-1} & k = 1 \\ \rho_2 = a_1\rho_1 + a_2\rho_0 + a_3\rho_1 + \cdots + a_p\rho_{p-2} & k = 2 \\ \vdots & \vdots \\ \rho_p = a_1\rho_{p-1} + a_2\rho_{p-2} + a_3\rho_{p-3} \cdots + a_p\rho_0 & k = p \end{cases}$$

Auto correlation matrix

In matrix form, Yule-Walker equations are:

$$\mathbf{R}_p \mathbf{a} = \boldsymbol{\rho}$$

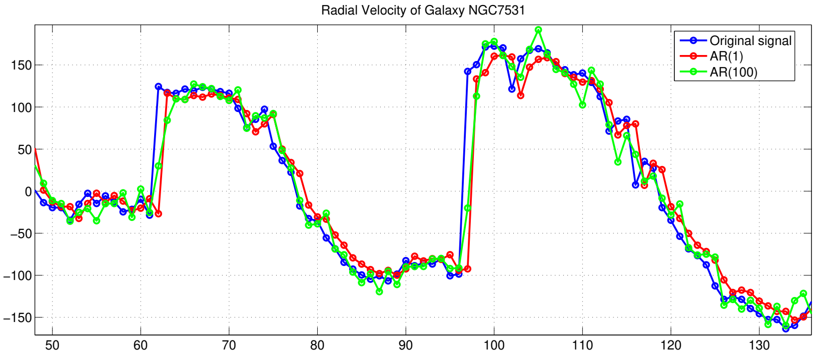
with $\mathbf{a} = [a_1, \ldots, a_p]^T$, $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_p]^T$ and

$$\mathbf{R}_p = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \ldots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \ldots & \rho_{p-2} \\ & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \ldots & \ldots & 1 \end{pmatrix}$$

If $\mathbf{R}_p$ is invertible (positive definite)

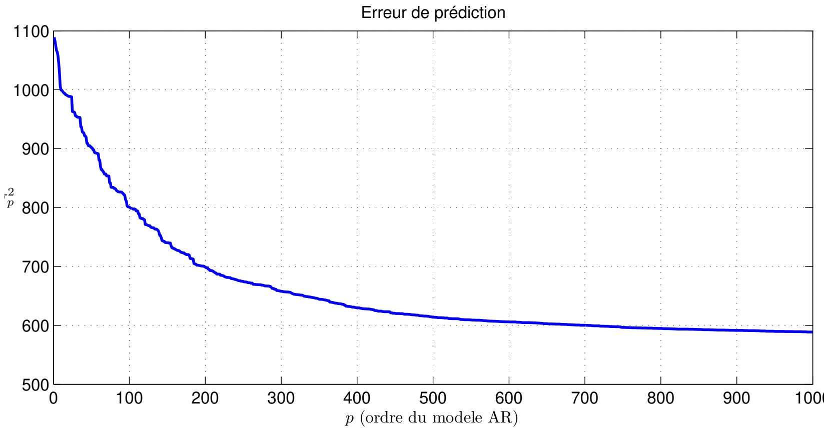
$$\mathbf{a} = \mathbf{R}_p^{-1} \boldsymbol{\rho}$$

Example



⇒ Choice of optimal order p ?

Example



$\Rightarrow$ Choice of optimal order p ?

Order selection

The higher p the better, but :

- ⋆ Requires the estimation of p correlation coefficients $\rightarrow$ Confidence over a finite series?
- $\Rightarrow$ Rule of thumb for length N: $p < N/3$
- ⋆ Model parsimony : in applications, it is generally better to keep a simple model
- $\Rightarrow$ Define parsimony (low p?)

Order selection

Whitening test (AR model)

$$\epsilon_t = y_t - \sum_{i=1}^{p} a_i y_{t-i} \qquad \text{white noise? } (\text{i.i.d} \sim \mathcal{N}(0, \sigma))$$

⇒ Tests: Durbin-Watson, Fisher,...

- ⋆ Check whether values ϵ_t ($t \in [\![T]\!]$) are significantly decorrelated

Final Prediction Error (FPE)

Definition of a global error, combining the variance of the prediction error σ_p and a variance over the imprecision of parameter estimation:

$$\mathrm{FPE}(p) = \sigma_p \, . \, \frac{T+p+1}{T-p-1}$$

decrease with $p \quad \longleftrightarrow \quad$ increase with p

Criteria for parsimony

Generally from Information Theory :

- ⋆ Minimum Description Length (MDL)

$$\mathrm{MDL}(p) = N \log(\sigma_p^2) + p \log(T)$$

- ⋆ Akaike Information Criterion (AIC)

$$\mathrm{AIC}(p) = N \log(\sigma_p^2) + 2(p + 1)$$

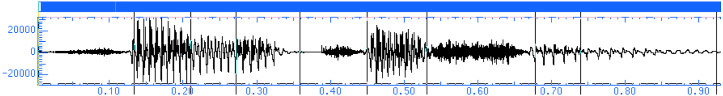
⇒ MDL is a good criterion with low number of samples

Limitations

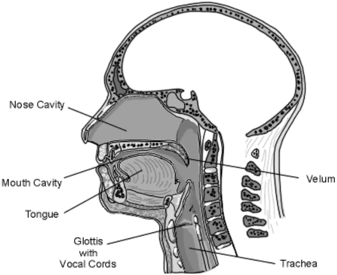
- ⋆ Stationarity is a strong assumption
- ⋆ Econometric data is rarely stationary (trend, multiplicative seasonality)
- → Trend removal, log model, ...
- → More complex models (ARMA, ARIMA, ARCH...) at the price of a more complex estimation
- ⋆ Example of stationary signal: speech (at least locally)

Speech analysis and compression using AR models

Hypothesis: local stationarity



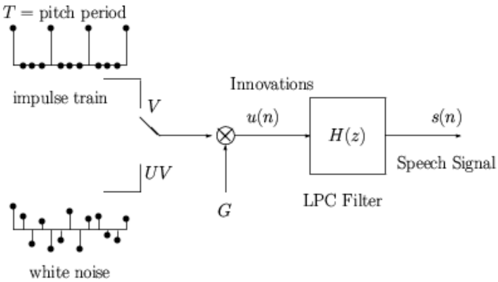
Linear model for speech production



Air flow :

- ⋆ Excited by vocal chords
- ⋆ Echo effect on vocal cavity (linear system)

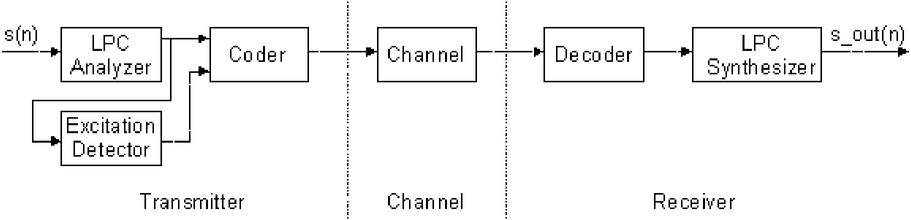
Modeling



- ⋆ Can show that coefficients $\mathrm{H}(z)$ are the parameters of the AR model
- ⋆ Characterize the parameters of speech production
- ⋆ Important for speech analysis and compression

Compression by Linear Predictive Coding (LPC)

Vocoder



$\Rightarrow$ used in GSM

Summary

- ⋆ As opposed to static data, temporal data includes causality
- ⋆ Temporal data present a unique sample
- ⋆ Correlation is found by considering stationarity

$\Rightarrow$ need to remove trend ans seasonality

- ⋆ AR(p) models are the base for temporal signal modeling
- ⋆ They are linear (latent) models (coefficient are used for interpretation, generation, ...)

- ⋆ Extension to Neural Nets for sequential data: e.g [2] (chap 15)

Example questions [mostly require formal – mathematical – answers]

- ⋆ What are trend, seasonality and how to remove them? What are the assumptions made?
- ⋆ What is stationarity? What does it imply and allow?
- ⋆ What is exponential smoothing (and exponential decay)?
- ⋆ Provide the base models for linear auto-regression
- ⋆ What is the horizon (and the lag)?

φ It is strongly advised to develop the algebra contained in this chapter
(inc relation to auto-correlation and convolution)

References I

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. (available online).

[2] Kevin P. Murphy. *Probabilistic Machine Learning: an Introduction*. MIT Press, 2022. (available online).

[3] Robert H. Shumway. *Time Series Analysis and its Applications, with R Examples*. Springer, 2017. (available online).

License

The text of this document and its illustrations are published under the Creative Commons BY-NC-SA 4.0 International License.