DS-GA 1018.001

Probabilistic time series analysis

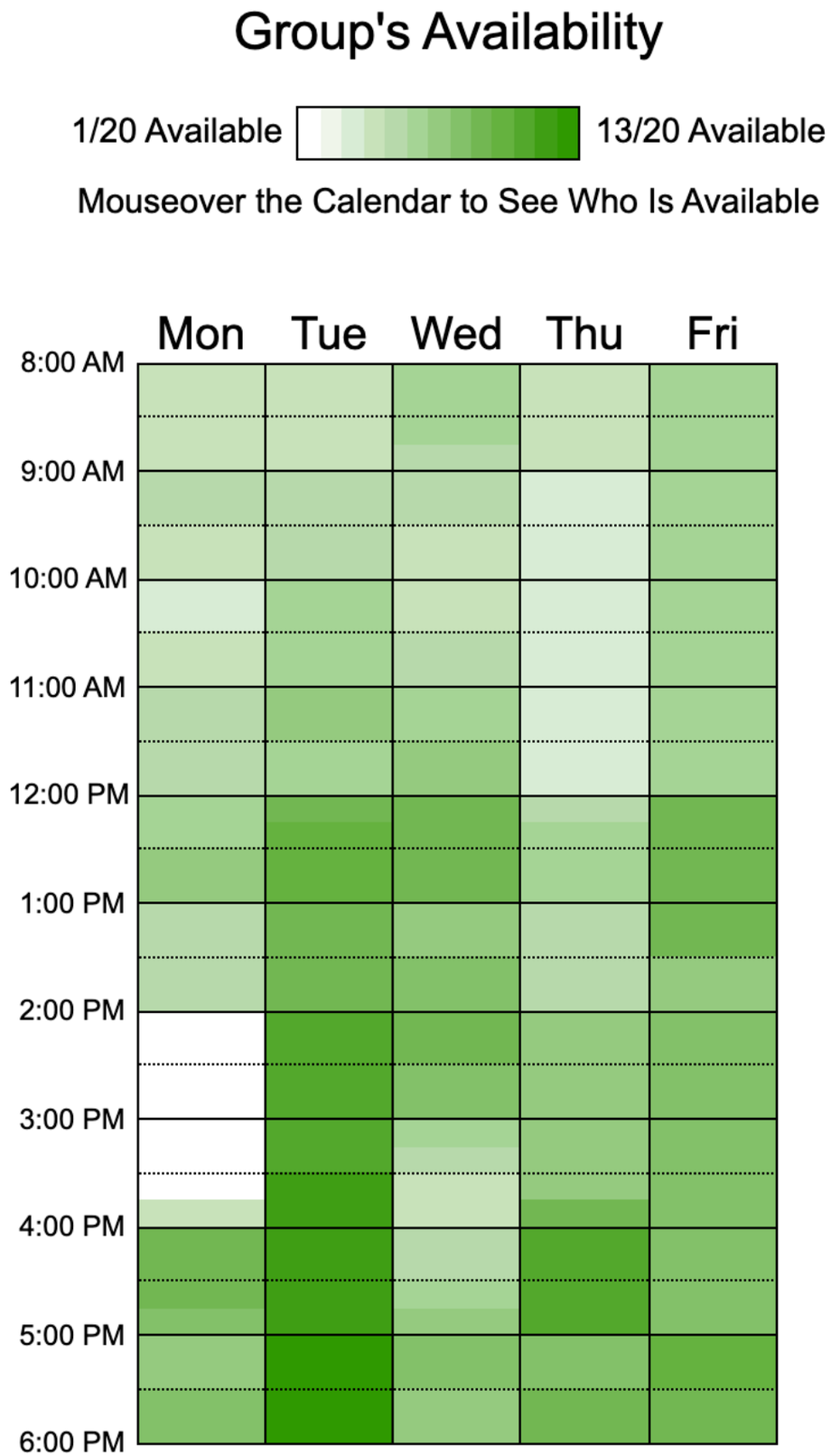
Lecture 1

Basic statistics of time series analysis. AR

Instructor: Cristina Savin

NYU, CNS & CDS

OH logistics



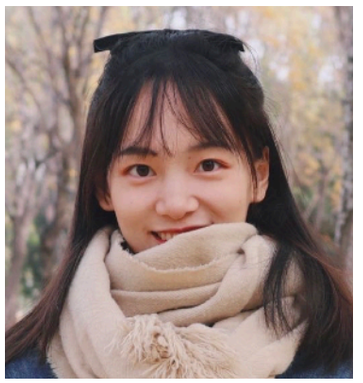
Course logistics

Instructor
Cristina Savin, csavin@nyu.edu
Office hours: Tue 5 pm, in CDS

TAs

- Section 002 - Haresh Rengaraj Rajamohan (hrr288@nyu.edu)
- Section 003 - Ying Wang (yw3076@nyu.edu)

Office hours: Thu 4pm via zoom



Couse page: https://github.com/savinteachingorg/pTSAfall2021.git
Piazza: piazza.com/nyu/fall2021/dsga3001001
!!! **access code** pTSA21

Definition. Probabilistic models

What is a time series?

Formally, a collection of random variables indexed by time, t*

**Usually discrete time (digital data collection), but continuous time can be convenient in some cases*

$$\{X_1, X_2, \ldots, X_t \ldots\}$$

"stochastic process"

data = "realization"

Unlike the traditional case, NOT I.I.D. !!!

These **dependencies** are the main point; it's what makes prediction possible.

Fully specified by joint*:

$$\mathrm{P}(X_1 \leq x_1, \ldots, X_t \leq x_t \ldots)$$

These **dependencies** are also the main problem: they make the math hard.

***Intractable in general, we limit ourselves to more structured classes of distributions*

What helps us: regularities/structure

Chain rule:

$$\mathrm{P}(X_{1:T}) = \mathrm{P}(\mathrm{X}_1)\mathrm{P}(\mathrm{X}_2|\mathrm{X}_1)...\mathrm{P}(\mathrm{X}_\mathrm{T}|\mathrm{X}_{1:\mathrm{T}-1})$$

Markov assumption (order K=1):

$$\mathrm{P}(X_{1:T}) = \mathrm{P}(\mathrm{X}_1)\mathrm{P}(\mathrm{X}_2|\mathrm{X}_1)\mathrm{P}(\mathrm{X}_3|\mathrm{X}_2)...\mathrm{P}(\mathrm{X}_\mathrm{T}|\mathrm{X}_{\mathrm{T}-1})$$

Basic statistics of a time series

Mean

$$\mu_X(t) = \mathbb{E}(X_t)$$

Covariance

$$R_X(t,u) = \operatorname{cov}(X_t, X_u)$$

Auto-Correlation Function (ACF)

$$\rho_X(t,u) = \frac{R_X(t,u)}{\sqrt{R_X(t,t), R_X(u,u)}}$$

measures linear predicability of X_t from X_s

$$-1 \leq \rho_X(t,u) \leq 1$$

Cross-Covariance

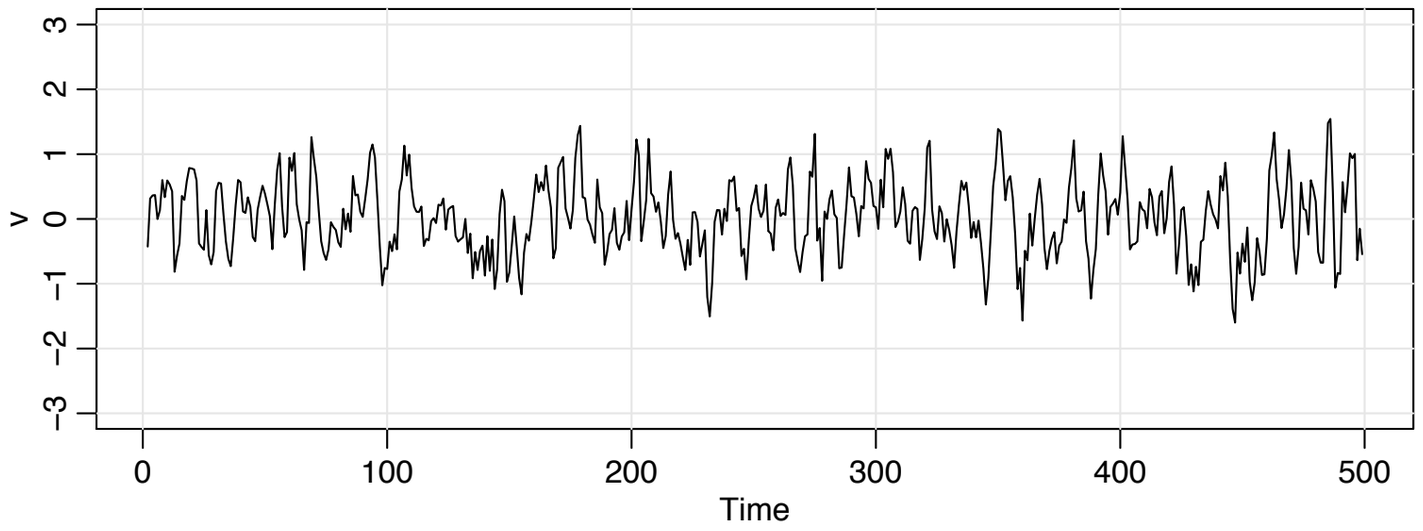
$$R_{X,Y}(t,u) = \operatorname{cov}\left(X_t, Y_u\right)$$

Cross-Correlation Function (ACF)

$$\rho_{X,Y}(t,u) = \operatorname{corr}(\mathrm{X_t}, \mathrm{Y_u}) = \frac{\mathrm{R_{X,Y}(t,u)}}{\sqrt{\mathrm{R_{X,X}(t,t) R_{Y,Y}(u,u)}}}$$

Example stochastic processes

White noise $X_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

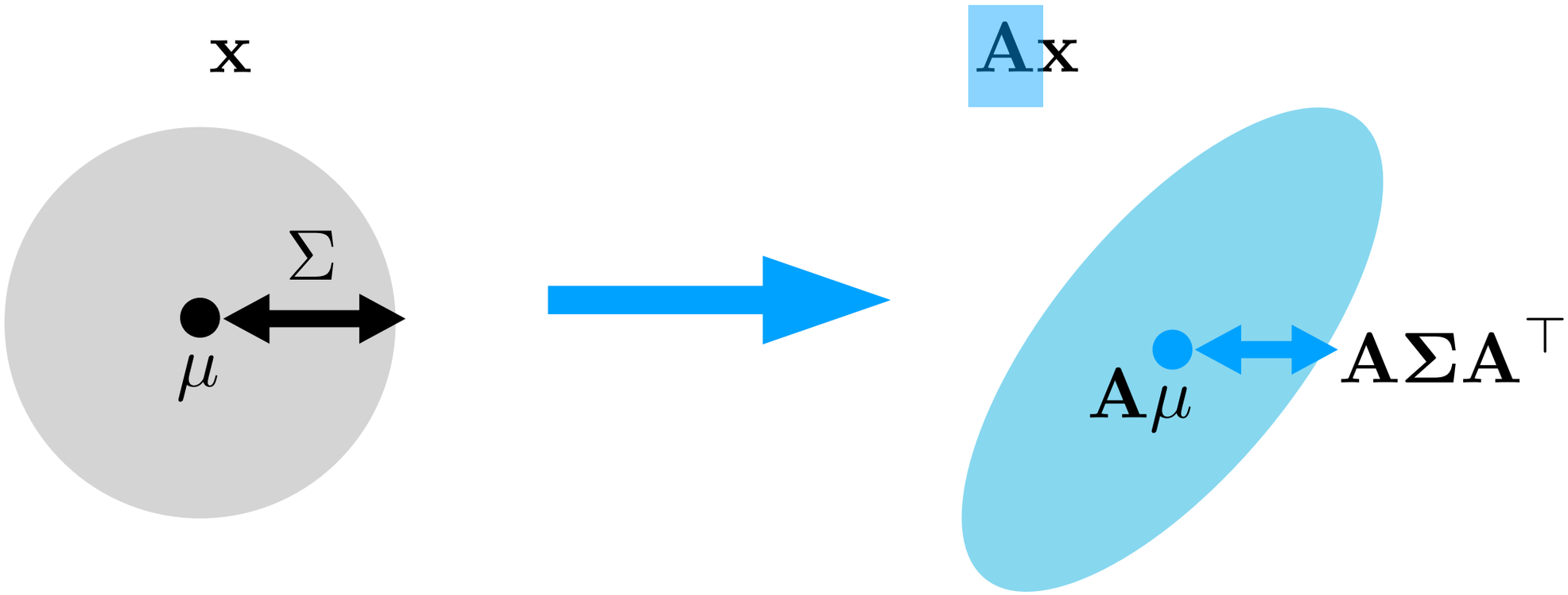


Trivially, white noise has

$$\mu_X(t) = 0$$

$$R_X(t, u) = \begin{cases} \sigma^2, t = u \\ 0, t \neq u \end{cases}$$

Linear transformations of gaussian variables

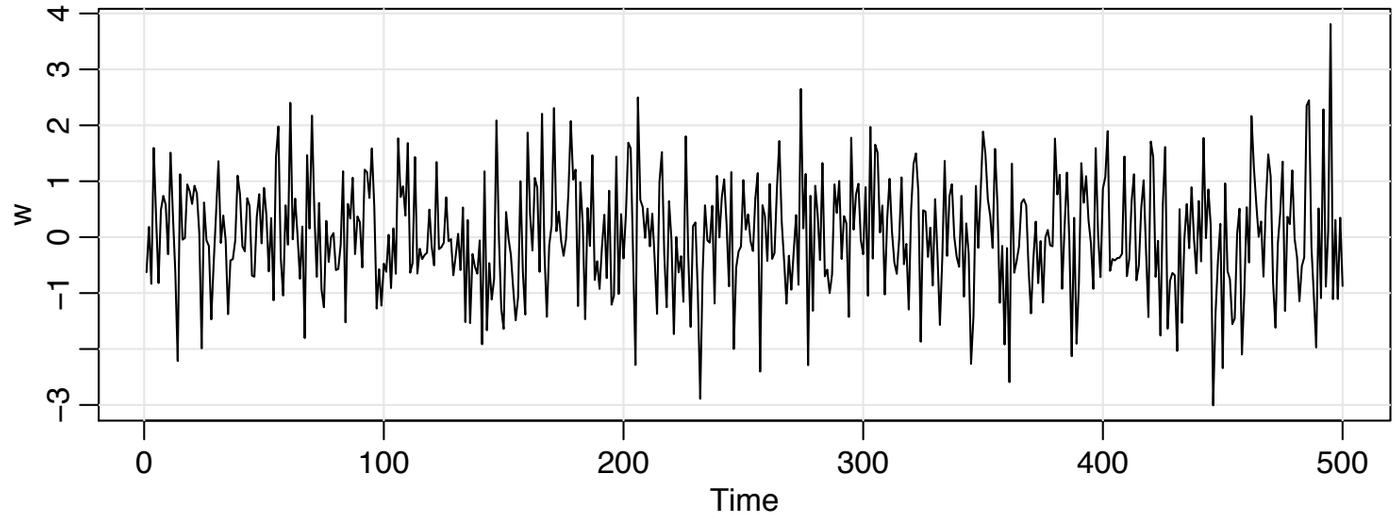


Cov. of linear combinations

$$U = \sum_i a_i X_i$$
$$V = \sum_i b_i Y_i$$
$$\mathrm{cov}(V, U) = \sum_{i,j} a_i b_j \mathrm{cov}(X_i, Y_j)$$

Example stochastic processes

Moving average (MA) $v_t = \frac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right)$



filtered white noise

$$\mu_V(t) = 0$$

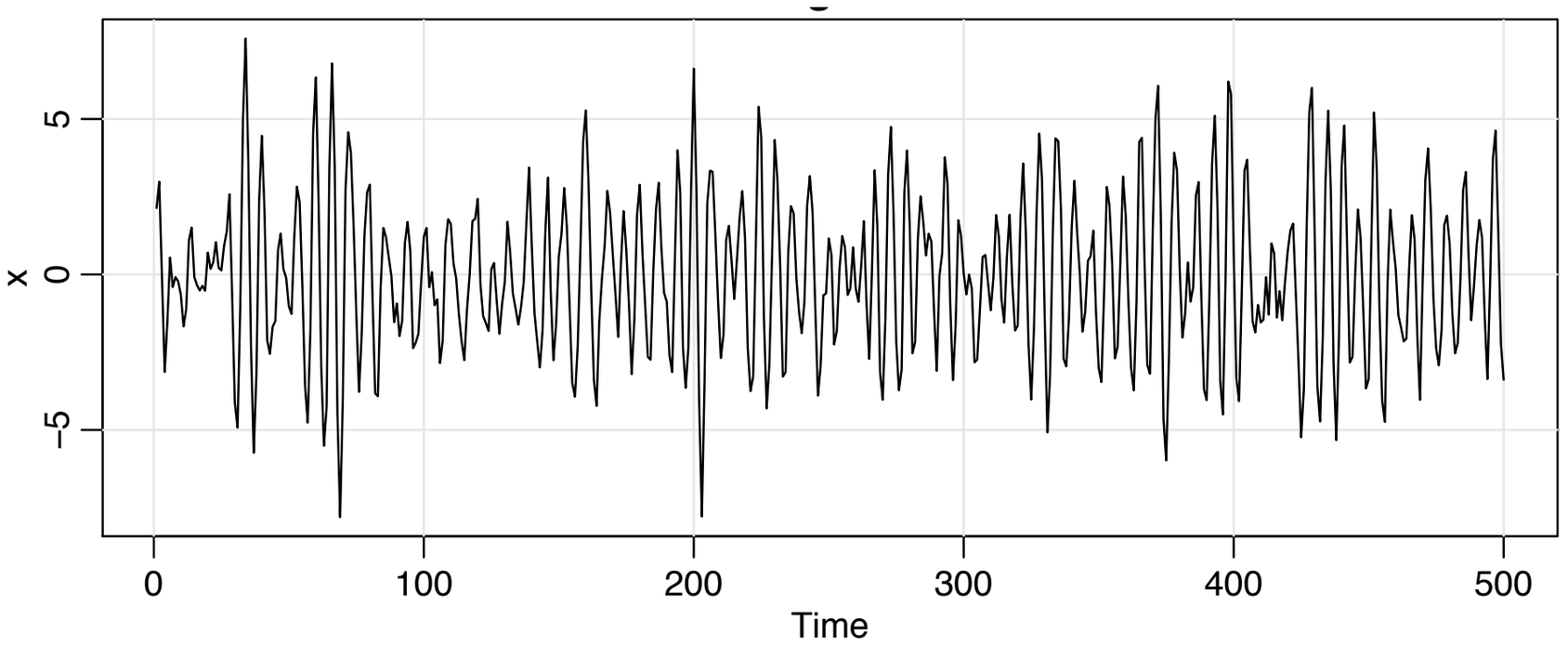
$$R_V(t,u) = \begin{cases} 1/3\ \sigma^2 & , t = u \\ 2/9\ \sigma^2 & , |t-u| = 1 \\ 1/9\ \sigma^2 & , |t-u| = 2 \\ 0, \quad & |t-u| > 2 \end{cases}$$

***Useful: Cov. of linear combinations**

$$U = \sum_i a_i X_i$$
$$V = \sum_i b_i Y_i$$
$$\mathrm{cov}(V,U) = \sum_{i,j} a_i b_j \mathrm{cov}(X_i, Y_j)$$

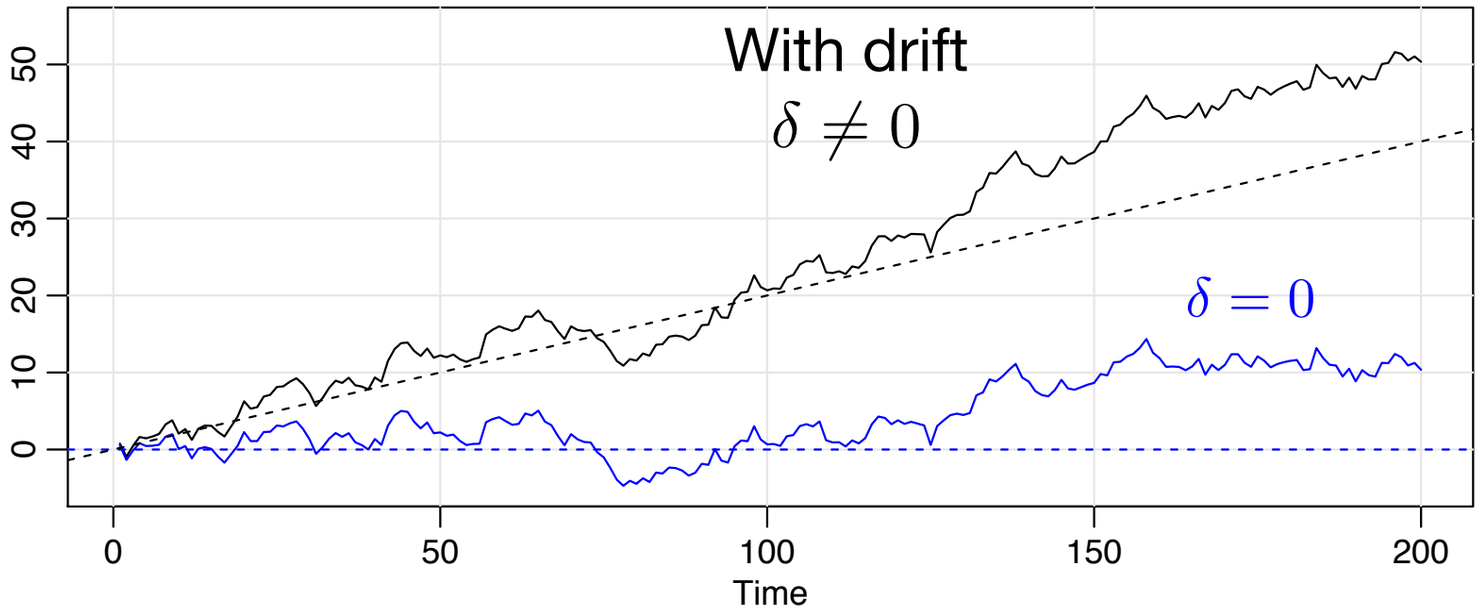
Autoregressive process (AR)

$$x_t = x_{t-1} - 0.9x_{t-2} + w_t$$



These are simple examples of ARIMA models —discussed in L3 .

Random walk $x_t = \delta + x_{t-1} + w_t$



If we unfold recursion:

$$x_t = t\delta + \sum_{i \leq t} w_i$$

$$\mu_X(t) = t\delta$$

$$R_X(t, u) = \min(t, u)\sigma^2$$

Cross-Covariance

$$R_{X,Y}(t,u) = \operatorname{cov}\left(X_t, Y_u\right)$$

Cross-Correlation Function (ACF)

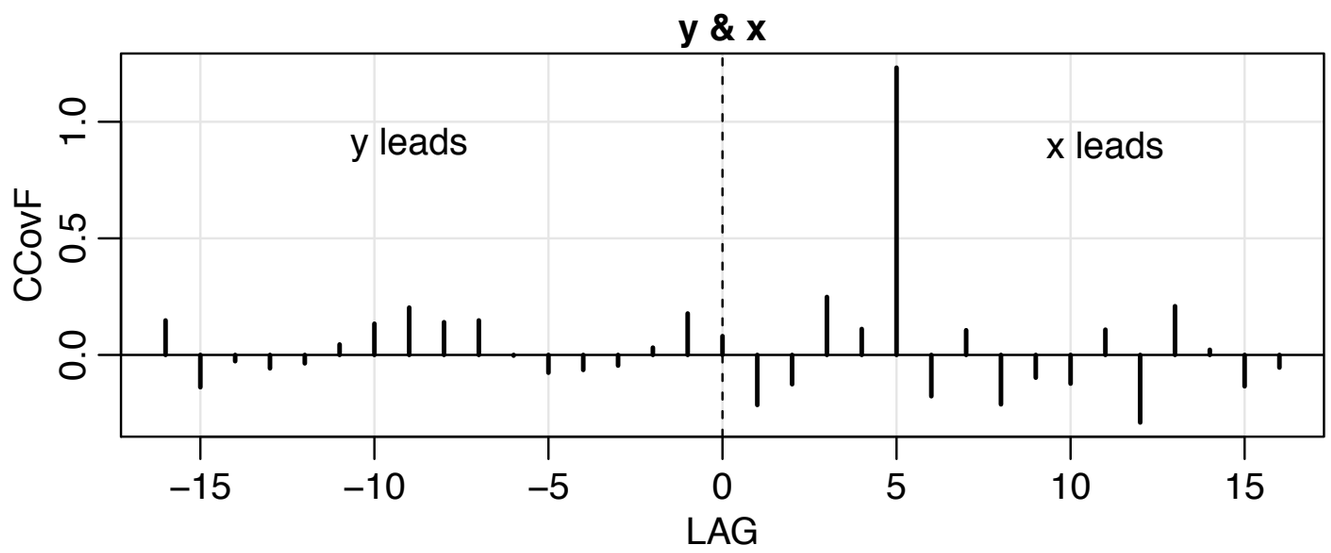
$$\rho_{X,Y}(t,u) = \frac{R_{X,Y}(t,u)}{\sqrt{R_{X,Y}(t,t), R_{X,Y}(u,u)}}$$

E.g. $x_t = w_t + w_{t-1}$ and $y_t = w_t - w_{t-1}$,

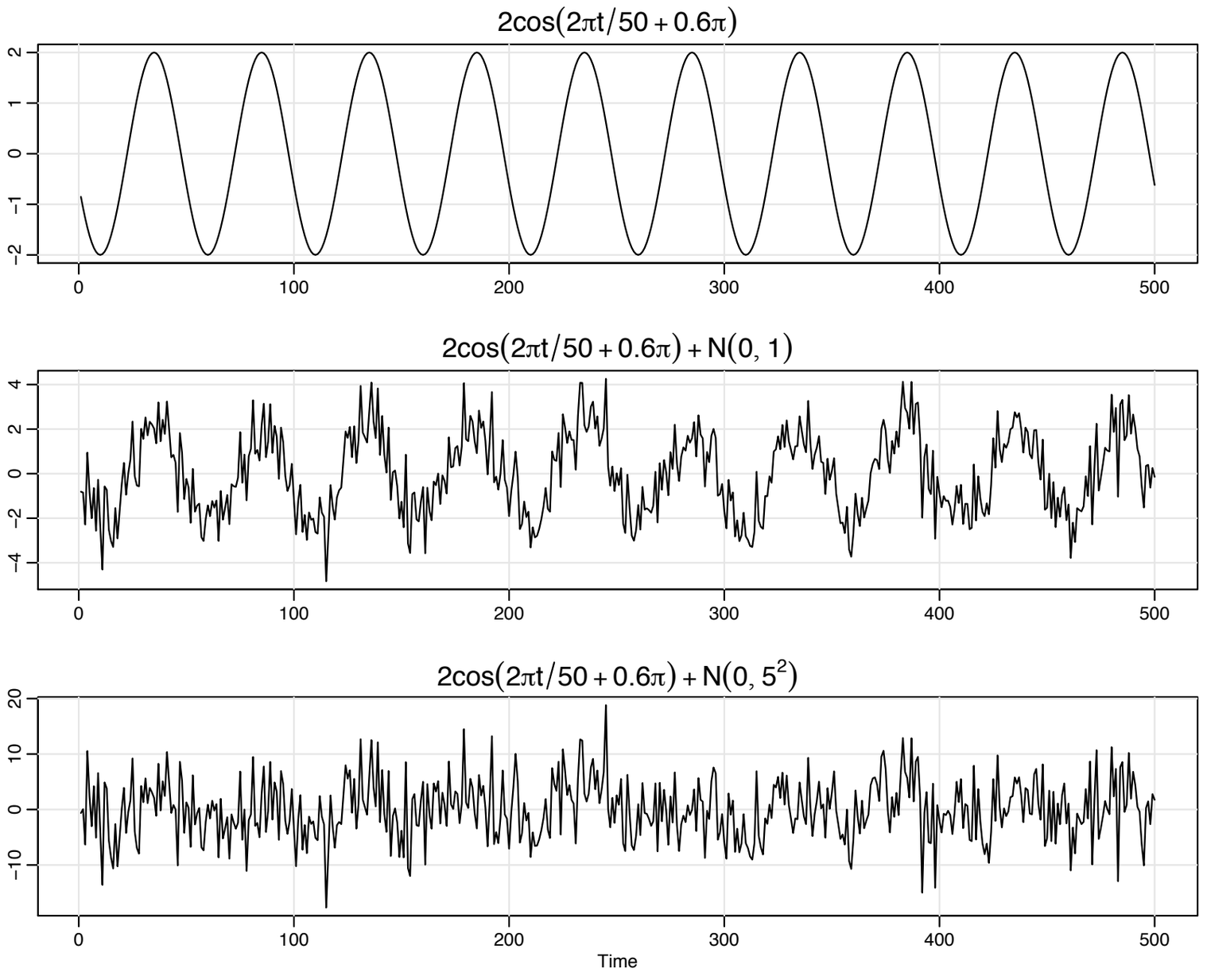
$$\rho_{xy}(h) = \begin{cases} 0 & h = 0, \\ 1/2 & h = 1, \\ -1/2 & h = -1, \\ 0 & |h| \geq 2. \end{cases}$$

Lead-lag $y_t = Ax_{t-\ell} + w_t$

$$\begin{aligned}\gamma_{yx}(h) &= \text{cov}(y_{t+h}, x_t) \\ &= \text{cov}(Ax_{t+h-\ell} + w_{t+h}, x_t) \\ &= \text{cov}(Ax_{t+h-\ell}, x_t) \\ &= A\gamma_x(h-\ell).\end{aligned}$$



Signal vs noise



denoising in frequency domain,
inferring latent structure in temporal domain

Basic statistics of a time series

“Strong stationarity”

$$\{X_t, ..., X_{t+K}\}$$
$$\{X_{t+h}, ..., X_{t+h+K}\}$$

Identically distributed subsets
for all t,h,K

Consequences:

For single variables (K=0) this implies same marginals everywhere

$\mathrm{P}(X_t < x) = \mathrm{P}(X_{t+h} < x)$ for all t,h, and so $\mu_X(t) = \mathrm{cte}$

For single variables (K=1) this implies same pairwise dependencies

Basic statistics of a time series

“(Weak) stationarity”

$$\mu_X(t) = \text{const.}$$
$$R_X(\Delta t) = \text{cov}(X_t, X_{t+\Delta t})$$

+finite variance

Example: moving averages

A strongly stationary process with finite variance is weakly stationary

Converse is more complicated:
a gaussian weakly stationary process is strictly stationary

*Note: change in notation, for stationary processes R_x has a single argument

Basic statistics of a time series

“Trend stationarity”

$$\mu_X(t) \neq \text{const.}$$

$$R_X(\Delta t) = \text{cov}(X_t, X_{t+\Delta t})$$

This means that data can be partitioned into a
time-dependent term + zero-mean stationary process

e.g. sigmoid + white noise

Final note: random walks are non-stationary

Basic statistics of a time series

Linear process

A general version of filtered white noise

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Causality

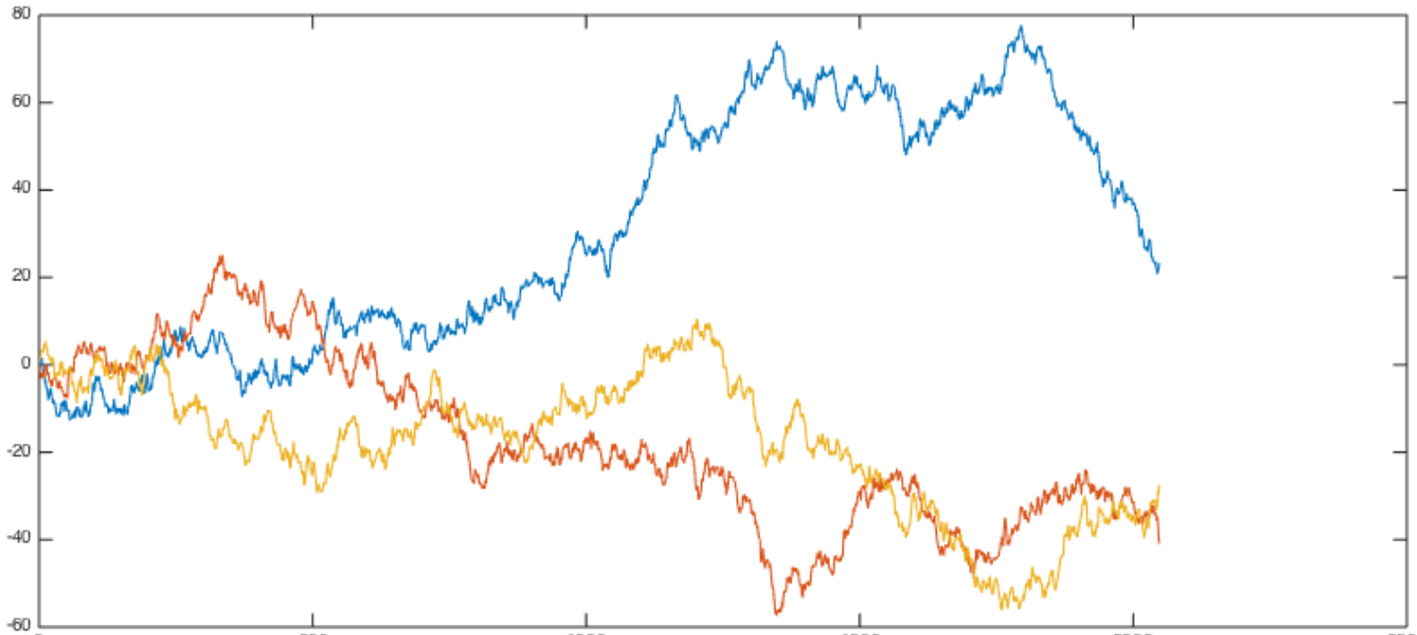
Present depends on past but not on future

It's often a natural assumption for real world data

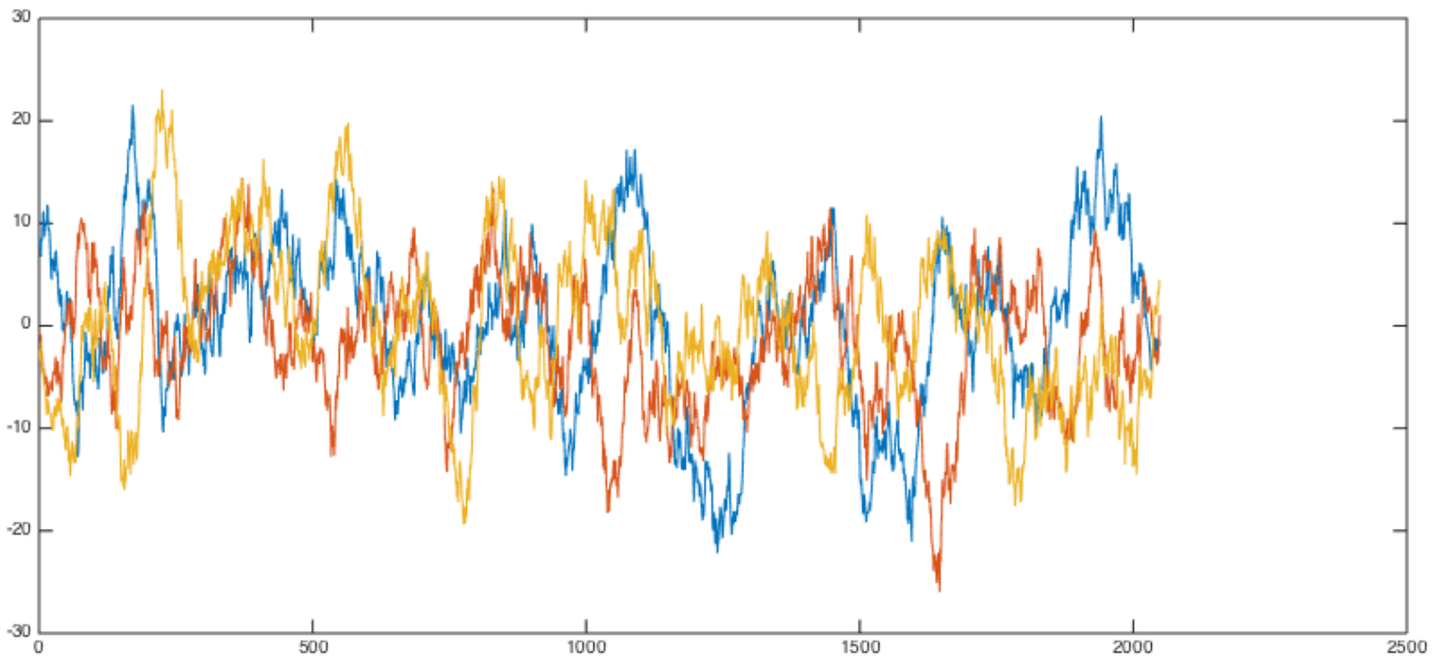
Not necessarily true for all sequential data (e.g. DNA)

Causal linear process:

$\psi_j = 0$ for future components (j<0)



Random walk
(non stationary)



stationary

Simpler structure,
Easier to estimate

Empirical measurements (stationary process)

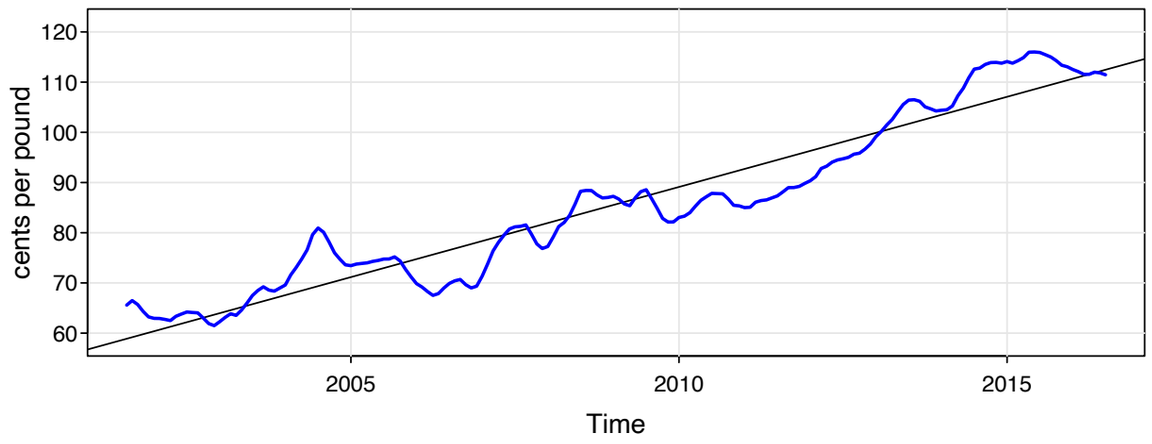
$$\hat{\mu}_x = \frac{1}{T}\sum_t x_t$$

$$\begin{aligned} R_x(\Delta t) &= \mathrm{cov}(x_t, x_{t+\Delta t}) \\ &= \frac{1}{T}\sum (x_{t+\Delta t} - \bar{\mu}_x)(x_t - \bar{\mu}_x) \end{aligned}$$

* both are biased estimates, in general, see yellow textbook for details

Exploratory data analysis, de-trending

Linear regression



$$x_t = \beta_0 + \beta_1 z_t + w_t$$

Fig. 2.1. *The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.*

Parameter learning: Maximum likelihood $\boldsymbol{\beta} = \mathrm{argmax} \mathrm{P}(\mathcal{D}|\boldsymbol{\beta})$

OLS:
(Ordinary Least Squares)

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

$$\frac{dQ}{d\beta} = 0$$

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^{n} (z_t - \bar{z})^2}$$

$$\hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \, \bar{z},$$

$$\bar{x} = \sum_t x_t / n$$

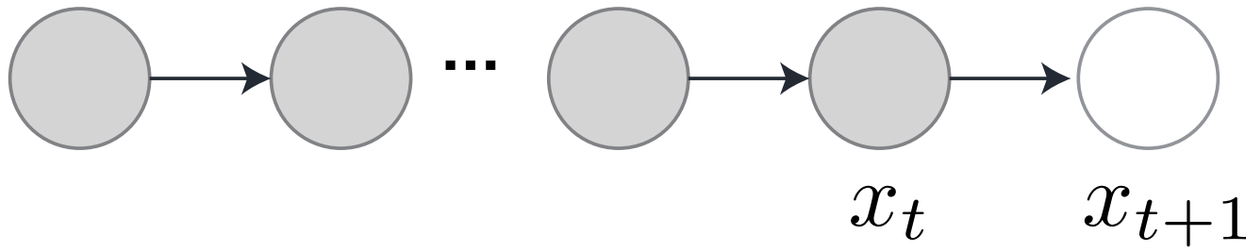
$$\bar{z} = \sum_t z_t / n$$

Multivariate:

$$\hat{\beta} = \left(\sum_{t=1}^{n} z_t z_t' \right)^{-1} \sum_{t=1}^{n} z_t x_t \, .$$

After detrending, check for independence,
if there is lingering temporal structure, go for more analysis

A simple autoregressive process AR(1)



Forecasting/prediction

$$\mathrm{P}(x_{t+1}|x_{1:t})$$

Autoregressive process AR(1)

$$X_t = \lambda X_{t-1} + W_t$$

Where $\{W_t\}$ is white noise and $|\lambda| < 1$

By expanding the recursion we get: $X_t = W_t + \lambda W_{t-1} + \lambda^2 W_{t-2} + \cdots$

$$\mu_X = \mathbb{E}\left[\sum_{h=0}^{\infty} \lambda^h W_{t-h}\right] = 0$$

$$\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[\sum_h \lambda^{2h} W_{t-h}^2\right] = \sigma^2 \sum \lambda^{2h} = \frac{\sigma^2}{1-\lambda^2}$$

For now, assume $h > 0$

$$\begin{aligned} R_x(h) = \operatorname{cov}(X_t, X_{t+h}) &= \operatorname{cov}(X_t, \lambda X_{t+h-1} + W_{t+h}) \\ &= \lambda \operatorname{cov}(X_t, X_{t+h-1}) \\ &= \lambda^h \operatorname{cov}(X_t, X_t) \\ &= \sigma^2 \frac{\lambda^{|h|}}{1-\lambda^2} \end{aligned}$$

stationary

*Check other direction at home

Least squares and ACF

Least squares estimation reminder

$$\hat{f} = \text{argmin}_f (Y - f)^2$$

$$\hat{f} = \mathbb{E}[Y|X]$$

With MSE $\quad \text{var}[Y|X]$

We can compute a least square estimate of X_{t+h} given X_t

Since everything is Gaussian, conditional expectations are easy!

If $\{X_t\}$ is jointly gaussian $\quad f_X(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

The pair (X_t, X_{t+h}) is also jointly gaussian, with covariance

$$\begin{pmatrix} \sigma_t^2 & \rho(t,t+h)\sigma_t\sigma_{t+h} \\ \rho(t,t+h)\sigma_t\sigma_{t+h} & \sigma_{t+h}^2 \end{pmatrix}$$

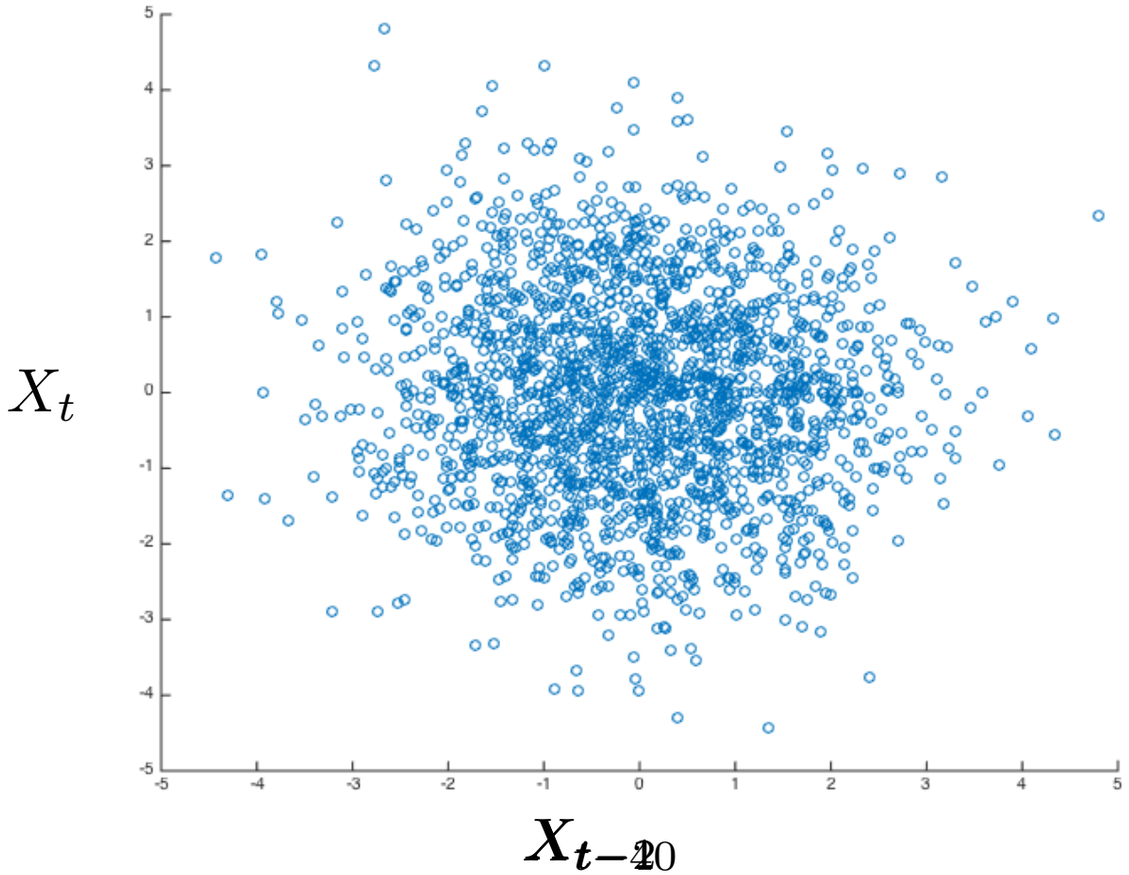
$$X_{t+h}|X_t = x_t \qquad \mathcal{N}\left(\mu_{t+h} + \frac{\sigma_{t+h}\rho(t,t+h)(x_t-\mu_t)}{\sigma_t}, \sigma^2(1-\rho(t,t+h)^2)\right)$$

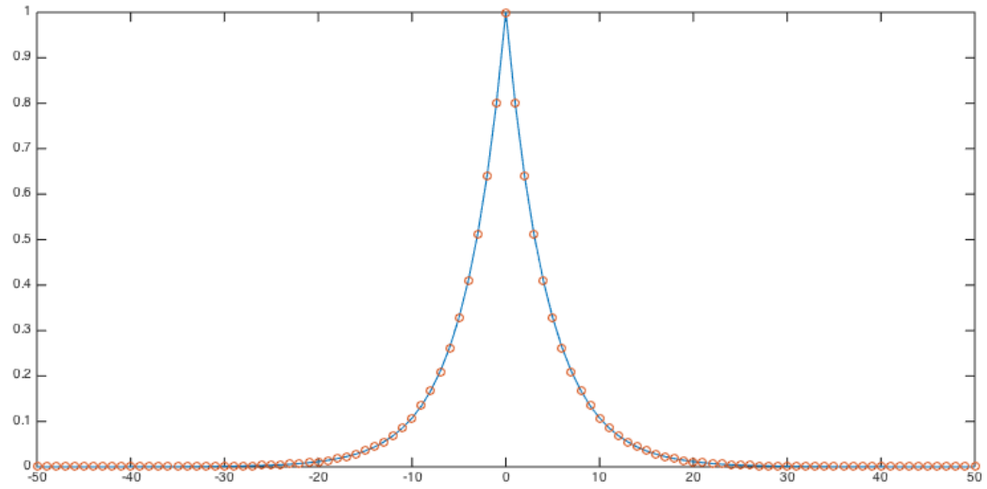
*Gaussian identities .pdf

Prediction

How does the prediction depend on ACF?

example: AR(1)

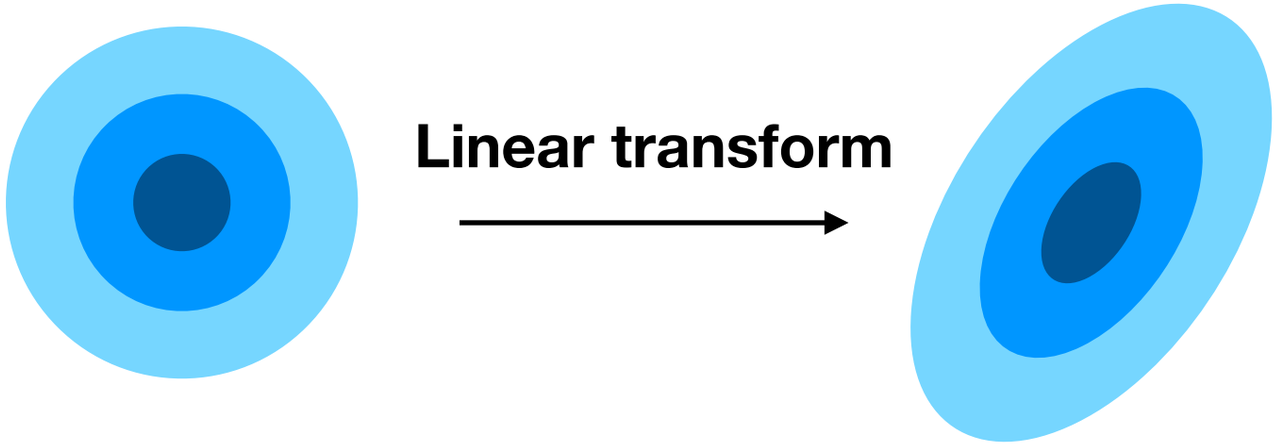




ACF determines
linear predictability

Philosophy or AR(I)MA models

White noise

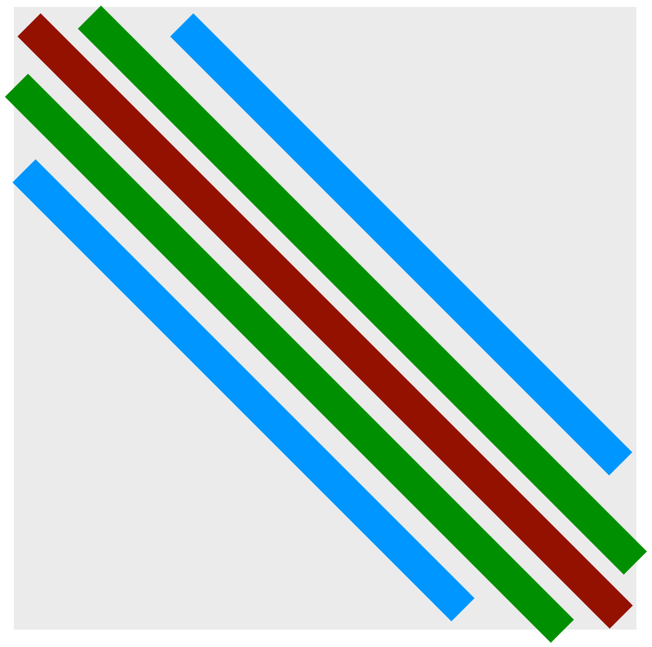


models require a variable no. of parameters

The goal is to capture the cov. structure of the data with **as few as possible** parameters

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow (\mathbf{A}\mathbf{x} + \mathbf{y}) \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu} + \mathbf{y}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right)$$

Most generally,
TxT covariance
Additional assumptions
on shared parameters



$\mathrm{Cov}(x_t, x_{t+2})$
$\mathrm{Cov}(x_t, x_{t+1})$
$\mathrm{Var}(x_t)$

Next week:
the ARIMA framework

A unified way of identifying
properties of linear gaussian processes