Inference and Representation

Lecture 8: Introduction to Time Series

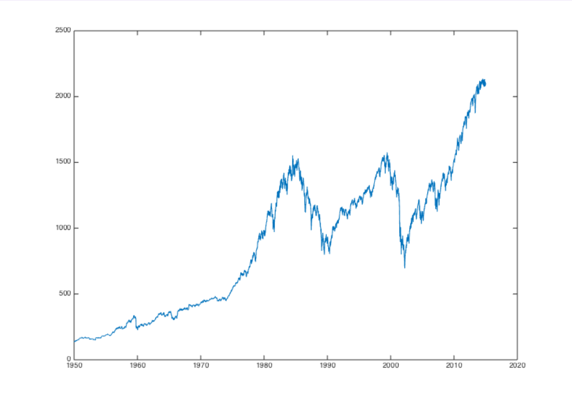
Joan Bruna

Courant Institute
NYU

November 14, 2016

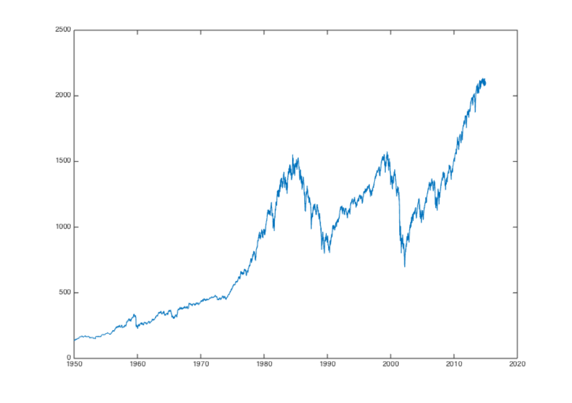
What is a Time Series?

What is a Time Series?



What is a Time Series?

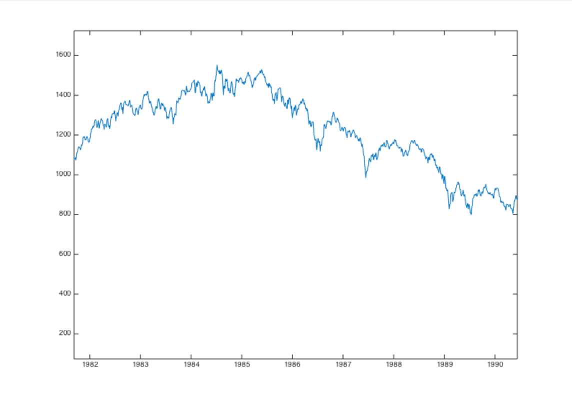
S&P 500 index



Statistical Model necessary.

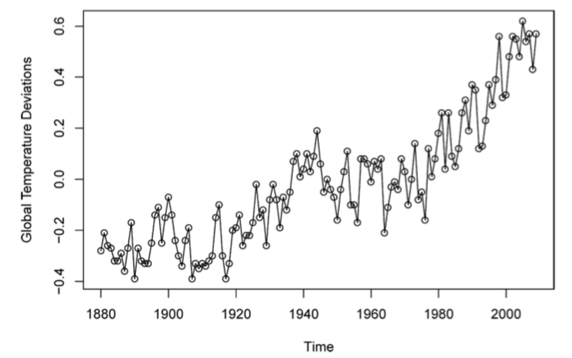
What is a Time Series?

S&P 500 index, zoomed

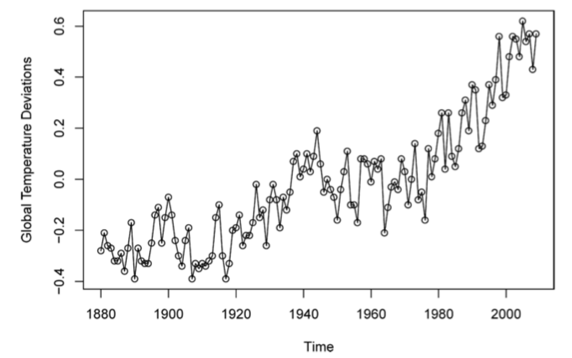


Interaction of *random* and *non-random* behavior.

What is a Time Series?

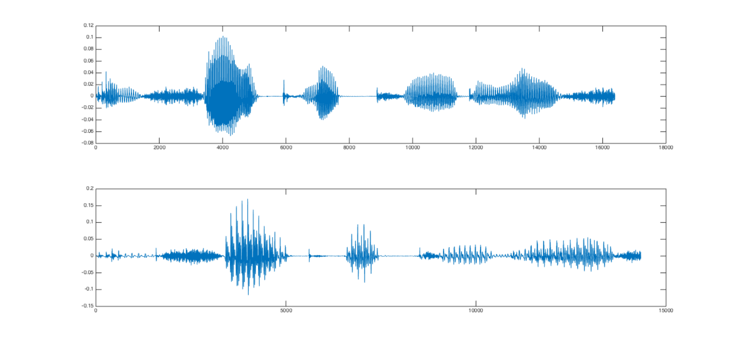


What is a Time Series?

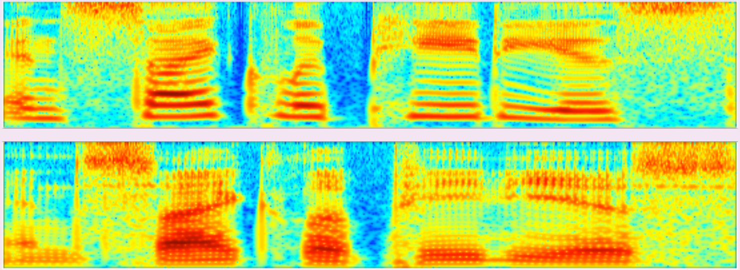


- trend
- seasonality (periodicity)

What is a Time Series?



What is a Time Series?



Periodic, stationary phenomena is studied with spectrograms.

What is a Time Series?

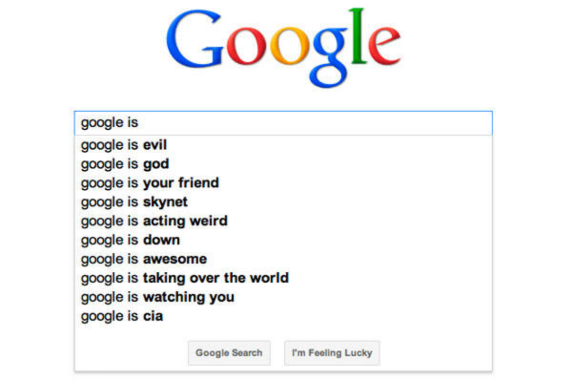
Fig. 1. An image of the traffic sequences.

Fig. 2. An example of trajectories involved in a traffic conflict.

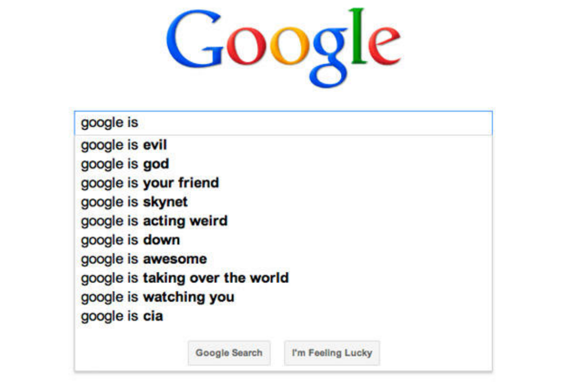
from Saunier and Sayed

- Robotics
- Control Theory (Kalman)
- Self-driving cars ...

What is a Time Series?

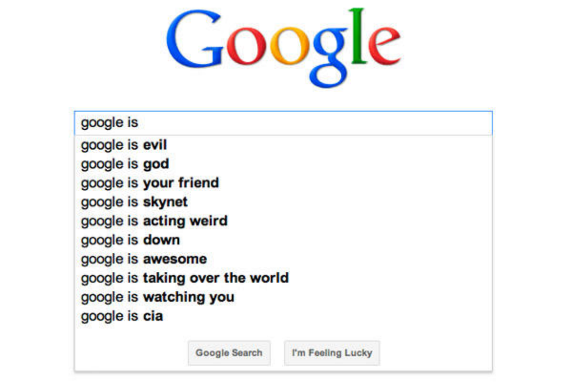


What is a Time Series?



- State space models (language).

What is a Time Series?



- State space models (language).
- Prediction, compression, ...

This is not an IID world

Fundamental characteristic of time-series: in general, samples are correlated (thus statistically dependent).

This is not an IID world

Fundamental characteristic of time-series: in general, samples are correlated (thus statistically dependent).

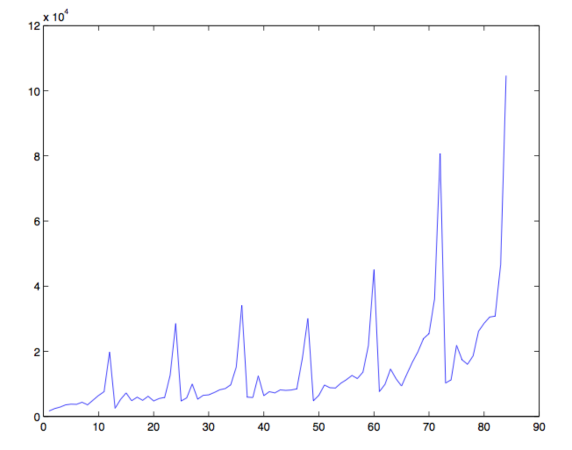
Estimating, modeling and analyzing this dependency/correlation is the main objective of Time Series Analysis.

Objectives of Time Series Analysis

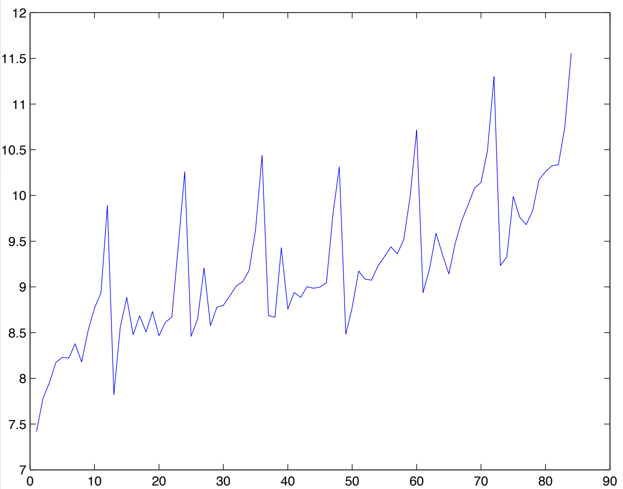
- Compact description of the data: statistical modeling.
- Interpretation
- Forecasting/Prediction
- Control
- Simulation
- Hypothesis testing

Example: Monthly Sales of a Souvenir Shop

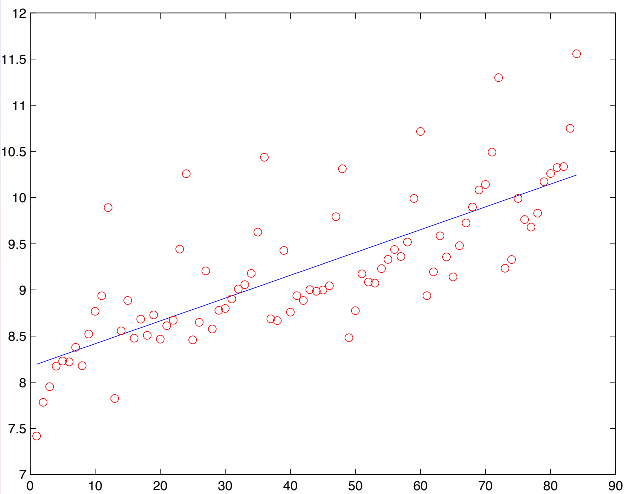
(from P.Bartlett's slides, Makridakis, Wheelwright and Hyndman, 1998)



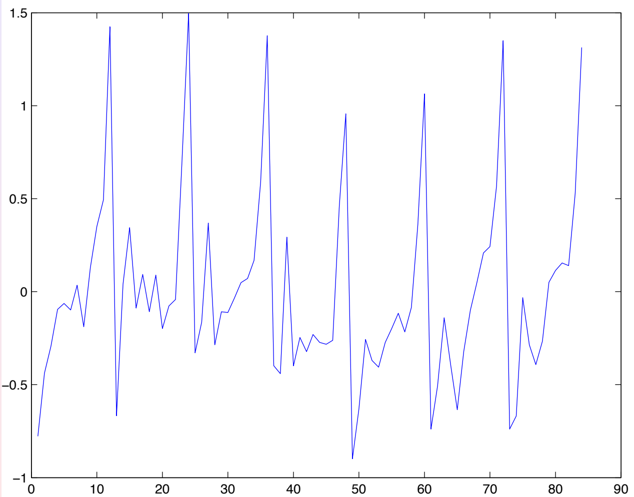
Example: Transformation of the data



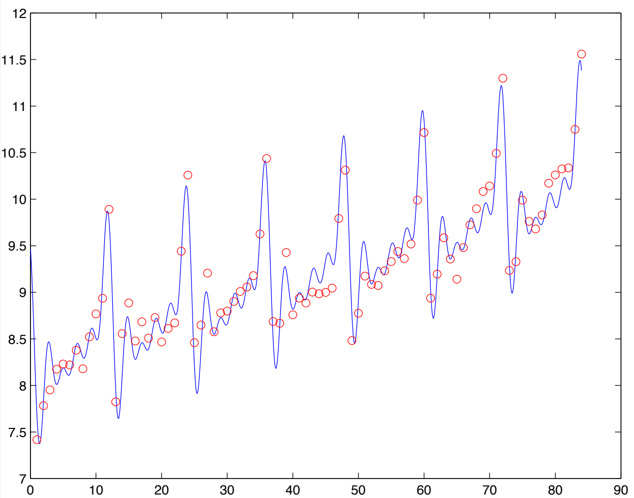
Example: Trend



Example: Look at the residuals



Example: Modeling Seasonality and Trend



Example

- Compact Description: Decomposition model

$$X_t = T_t + S_t + R_t \ .$$

Example

- Compact Description: Decomposition model

$$X_t = T_t + S_t + R_t \ .$$

- This model is *interpretable*. (eg seasonal adjustment due to holidays).

Example

- Compact Description: Decomposition model

$$X_t = T_t + S_t + R_t \ .$$

- This model is *interpretable*. (eg seasonal adjustment due to holidays).
- Forecasting/Prediction: Expected Sales next month?

Overview of the Lecture

1. Time series basics.
2. Time domain Models.
3. Spectral Analysis.
4. State Space and Discrete Models.

Time Series Modeling/Notation

How do we define a Time Series?

Time Series Modeling/Notation

How do we define a Time Series?

A (stochastic) Time Series is a collection $\{X_t\}$ of random variables indexed by a temporal index t.

- In this course, we will mostly consider discrete time series: $t = 1, 2, 3, \ldots$ is a discrete index.

Time Series Modeling/Notation

How do we define a Time Series?

A (stochastic) Time Series is a collection $\{X_t\}$ of random variables indexed by a temporal index t.

- In this course, we will mostly consider discrete time series: $t = 1, 2, 3, \ldots$ is a discrete index.
- $\{X_t\}$ will **always** denote a stochastic process.
- $\{x_t\}$ will **always** denote a particular realization.

Time Series Modeling

How do we specify a Time Series Model?

Time Series Modeling

How do we specify a Time Series Model?

A Time Series model is (fully) specified by giving the joint distribution of $\{X_t\}$*:*

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_t \leq x_t) \text{ for all } t, x_1, \ldots, x_t \ .$$

The Curse of Dimensionality

- As t increases, the complexity of the previous model grows *exponentially*.

The Curse of Dimensionality

- As t increases, the complexity of the previous model grows *exponentially*.
- Intractable in general.

The Curse of Dimensionality

- As t increases, the complexity of the previous model grows *exponentially*.
- Intractable in general.
- We will resort to low-order statistics only (mostly first and second order).

White Noise

$\{X_t\}$ is a white noise if for all t,

1. $\mathbf{E}(X_t) = 0$,
2. $\mathrm{var}(X_t) = \sigma^2$,
3. X_t and X_u are uncorrelated for $t \neq u$.

White Noise

$\{X_t\}$ is a white noise if for all t,

1. $\mathbf{E}\left(X_t\right) = 0$,
2. $\text{var}\left(X_t\right) = \sigma^2$,
3. X_t and X_u are uncorrelated for $t \neq u$.

In particular, if $\{X_t\}$ are i.i.d with zero mean, $\{X_t\}$ is a white noise. Also,

$$P(X_1 \leq x_1, \ldots, X_t \leq x_t) = \prod P(X_i \leq x_i) \ .$$

White Noise

$\{X_t\}$ is a white noise if for all t,

1. $\mathbf{E}(X_t) = 0$,
2. $\mathrm{var}(X_t) = \sigma^2$,
3. X_t and X_u are uncorrelated for $t \neq u$.

In particular, if $\{X_t\}$ are i.i.d with zero mean, $\{X_t\}$ is a white noise. Also,

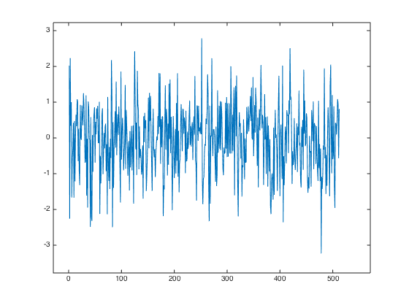
$$P(X_1 \leq x_1, \dots, X_t \leq x_t) = \prod P(X_i \leq x_i) \ .$$

Forecasting is not possible under iid noise:

$$P(X_t \leq x_t \,|\, X_1, \dots, X_{t-1}) = P(X_t \leq x_t) \ .$$

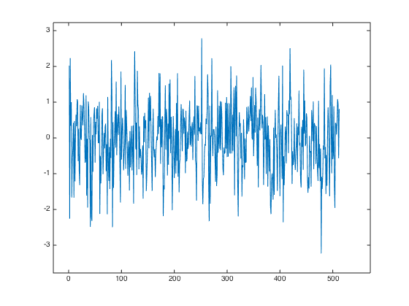
Most Famous Time Series

Gaussian White Noise: $X_t \sim \mathcal{N}(0, \sigma^2)$ iid.



Most Famous Time Series

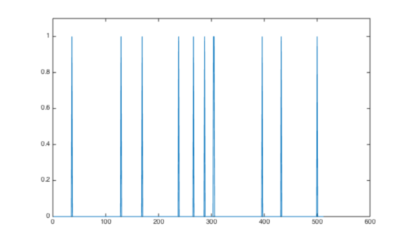
Gaussian White Noise: $X_t \sim \mathcal{N}(0, \sigma^2)$ iid.



It cannot model any time-dependencies.

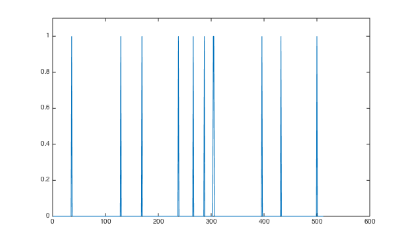
Bernouilli White Noise

$X_t \sim \mathrm{Bern}(p)$, with $p \in [0, 1]$.



Bernouilli White Noise

$X_t \sim \text{Bern}(p)$, with $p \in [0, 1]$.



Eg: models the success at a casino roulette.

Moving Averages

A simple way to model dependencies across samples is to average across time.

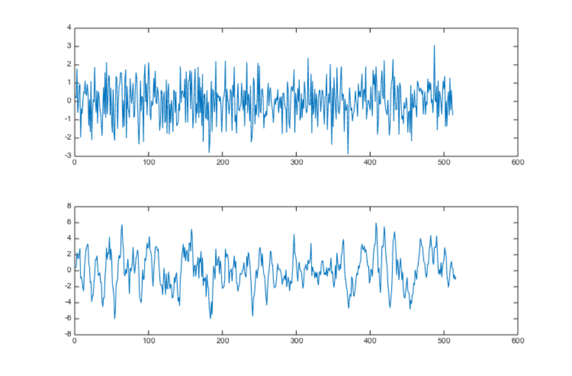
If $\{W_t\}$ is white noise, the series

$$X_t = \sum_{k=-\Delta}^{\Delta} \lambda_k W_{t+k}$$

is called a moving average ($MA(2\Delta)$).

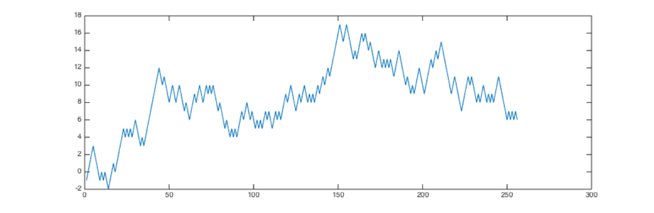
Moving Averages

$$X_t = \sum_{k=-\Delta}^{\Delta} \lambda_k W_{t+k} \ .$$



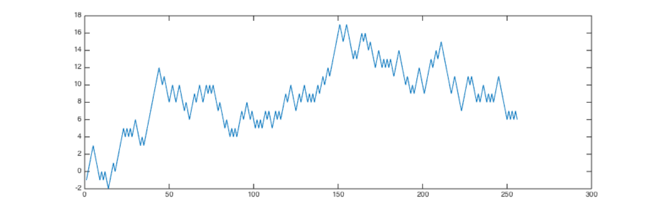
Random Walks

Consider $\{W_t\}$ white noise, and $X_t = \sum_{i \leq t} W_i$.



Random Walks

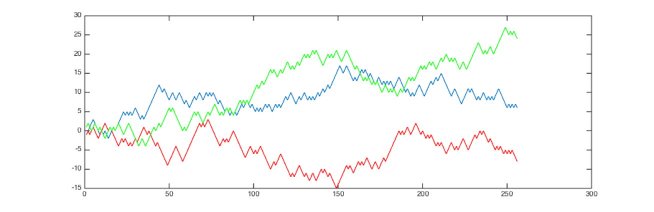
Consider $\{W_t\}$ white noise, and $X_t = \sum_{i \leq t} W_i$.



Differences $\nabla X_t = X_t - X_{t-1} = W_t$.

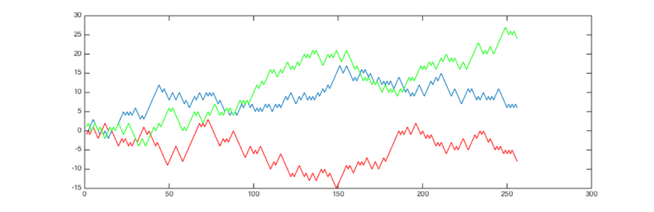
Random Walks

$$\mathbf{E}(X_t) = \quad , \quad var(X_t) =$$



Random Walks

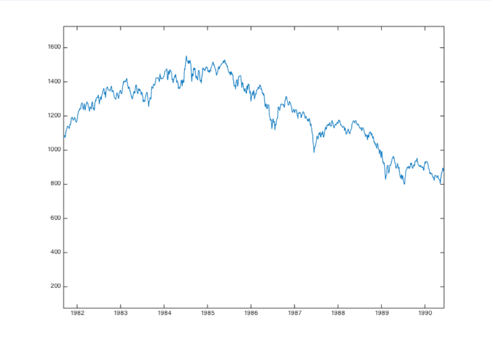
$$\mathbf{E}(X_t) = 0 \quad , \quad var(X_t) = t\sigma^2 \ .$$



Variance increases with time!

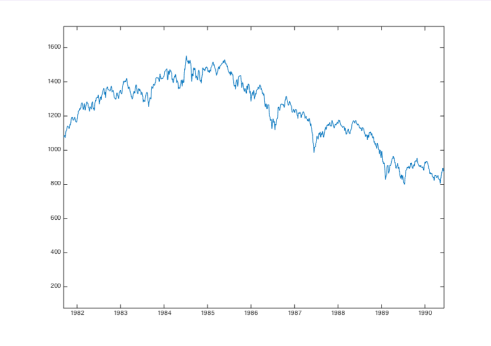
Random Walks

Recall S&P data.



Random Walks

Recall S&P data.



Random walks and their generalizations (Brownian Motions) are good financial models.

Signal with Noise

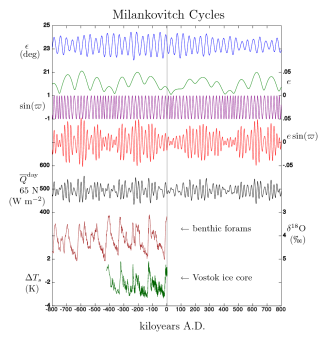
In many problems, good models combine deterministic with stochastic aspects:

$$X_t = F(t) + W_t \ , \ \text{with } F \text{ deterministic} \ .$$

Signal with Noise

In many problems, good models combine deterministic with stochastic aspects:

$$X_t = F(t) + W_t \ , \text{ with } F \text{ deterministic} \ .$$



Mean and Covariance of a Time Series

Definition

The mean function of a time series $\{X_t\}$ is

$$\mu_X(t) = \mathbf{E}\left(X_t\right) \; .$$

Mean and Covariance of a Time Series

Definition

The mean function of a time series $\{X_t\}$ is

$$\mu_X(t) = \mathbf{E}(X_t) \; .$$

Definition

The autocovariance function of a time series $\{X_t\}$ is

$$R_X(t,s) = \mathrm{cov}(X_t, X_s) = \mathbf{E}((X_t - \mathbf{E}(X_t))(X_s - \mathbf{E}(X_s))) \; .$$

Examples

$\{X_t\}$ Random Walk: $X_t = \sum_{i \leq t} W_i$, with $\{W_t\}$ iid white noise.

Examples

$\{X_t\}$ Random Walk: $X_t = \sum_{i \leq t} W_i$, with $\{W_t\}$ iid white noise.

$$\mu_X(t) = \mathbf{E}\left(X_t\right) = \mathbf{E}\left(\sum_{i \leq t} W_i\right) = \sum_{i \leq t} \mathbf{E}\left(W_i\right) = 0 \ .$$

Examples

$\{X_t\}$ Random Walk: $X_t = \sum_{i\leq t} W_i$, with $\{W_t\}$ iid white noise.

$$\mu_X(t) = \mathbf{E}\left(X_t\right) = \mathbf{E}\left(\sum_{i\leq t} W_i\right) = \sum_{i\leq t} \mathbf{E}\left(W_i\right) = 0 \ .$$

$$R_X(s,t) = \mathrm{cov}(\sum_{i\leq s} W_i, \sum_{i'\leq t} W_{i'}) = \sum_{i\leq s, i'\leq t} \mathrm{cov}(W_i, W_{i'}) = \min(s,t)\sigma^2 \ .$$

Examples

$\{X_t\} = F(t) + \{W_t\}$: signal with white noise.

Examples

$\{X_t\} = F(t) + \{W_t\}$: signal with white noise.

$$\mu_X(t) = \mathbf{E}(X_t) = \mathbf{E}(F(t) + W_t) = F(t) \ .$$

Examples

$\{X_t\} = F(t) + \{W_t\}$: signal with white noise.

$$\mu_X(t) = \mathbf{E}\left(X_t\right) = \mathbf{E}\left(F(t) + W_t\right) = F(t)~.$$

$$R_X(s,t) = \text{cov}(F(s)+W_s, F(t)+W_t) = \text{cov}(W_s, W_t) = \begin{cases} \sigma^2 & \text{if s=t}~, \\ 0 & \text{otherwise} \end{cases}~.$$

Autocorrelation function (ACF)

In some situations, it is better to normalize the autocovariance.

Autocorrelation function (ACF)

In some situations, it is better to normalize the autocovariance.

Definition

The Autocorrelation function (ACF) of a time series $\{X_t\}$ is defined as

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

Autocorrelation function (ACF)

In some situations, it is better to normalize the autocovariance.

Definition

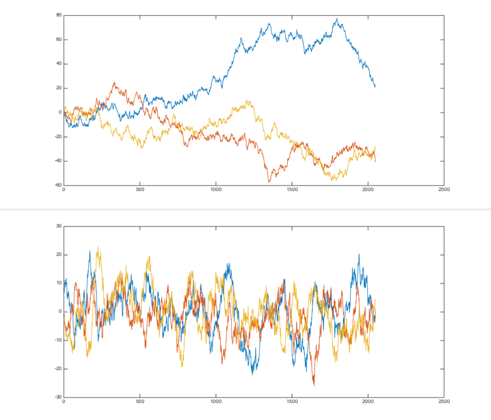
The Autocorrelation function (ACF) of a time series $\{X_t\}$ is defined as

$$\rho_X(s,t) = \frac{R_X(s,t)}{\sqrt{R_X(s,s)R_X(t,t)}} \; .$$

- Measures linear predicability of X_t from X_s.
- It satisfies

$$-1 \leq \rho_X(s,t) \leq 1 \; .$$

Stationarity: Motivation



Q: Main difference between figures?

Strict Stationarity

Definition

A Time Series $\{X_t\}$ is strictly Stationary if

$$P(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_k} \leq c_k) =$$
$$P(X_{t_1+h} \leq c_1, X_{t_2+h} \leq c_2, \ldots, X_{t_k+h} \leq c_k) \ , \ \forall \, k, t_i, c_i, h \ .$$

Strict Stationarity

Definition

A Time Series $\{X_t\}$ is strictly Stationary if

$$P(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_k} \leq c_k) =$$
$$P(X_{t_1+h} \leq c_1, X_{t_2+h} \leq c_2, \ldots, X_{t_k+h} \leq c_k) \ , \ \forall\, k, t_i, c_i, h \ .$$

- **Time is relative:** statistical properties do not depend upon time reference.

Strict Stationarity

Definition

A Time Series $\{X_t\}$ is strictly Stationary if

$$P(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_k} \leq c_k) =$$
$$P(X_{t_1+h} \leq c_1, X_{t_2+h} \leq c_2, \ldots, X_{t_k+h} \leq c_k) \ , \ \forall\, k, t_i, c_i, h \ .$$

- **Time is relative:** statistical properties do not depend upon time reference.
- In particular, using $k = 1$, $P(X_t \leq c)$ is independent of t.

Strict Stationarity

Definition

A Time Series $\{X_t\}$ is strictly Stationary if

$$P(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_k} \leq c_k) =$$
$$P(X_{t_1+h} \leq c_1, X_{t_2+h} \leq c_2, \ldots, X_{t_k+h} \leq c_k) \; , \; \forall \, k, t_i, c_i, h \; .$$

- **Time is relative:** statistical properties do not depend upon time reference.
- In particular, using $k = 1$, $P(X_t \leq c)$ is independent of t.
- Far LESS parameters to describe the process.

Strict Stationarity: Consequences

- Using previous property for $k = 1$, $P(X_t \leq c) = P(X_s \leq c)$ for all t, s implies

$$\mu_X(t) = cte \ .$$

Strict Stationarity: Consequences

- Using previous property for $k=1$, $P(X_t \leq c) = P(X_s \leq c)$ for all t, s implies

$$\mu_X(t) = cte \ .$$

- Using previous property for $k=2$, joint law (X_t, X_s) = joint law (X_{t+h}, X_{s+h}), hence

$$\forall\, h\ ,\ R_X(t,s) = R_X(t+h, s+h) \Longrightarrow R_X \textit{ depends only on } |t-s|\ .$$

Strict Stationarity: Consequences

- Using previous property for $k = 1$, $P(X_t \leq c) = P(X_s \leq c)$ for all t, s implies

$$\mu_X(t) = cte \ .$$

- Using previous property for $k = 2$, joint law (X_t, X_s) = joint law (X_{t+h}, X_{s+h}), hence

$$\forall\, h \ , \ R_X(t,s) = R_X(t+h, s+h) \Longrightarrow R_X \ depends\ only\ on\ |t-s| \ .$$

$$(\text{since } R(0, t-s) = R(s,t) = R(t,s) = R(0, s-t)) \ .$$

Weak Stationarity

In practice, strict Stationarity is hard to impose/estimate.

Weak Stationarity

In practice, strict Stationarity is hard to impose/estimate.

Definition

A Time series $\{X_t\}$ is weakly stationary if it has finite variance, $\mu_X(t) = cte$ and $R_X(t,s) = f(|t-s|)$.

Weak Stationarity

In practice, strict Stationarity is hard to impose/estimate.

Definition

A Time series $\{X_t\}$ is weakly stationary if it has finite variance, $\mu_X(t) = cte$ and $R_X(t,s) = f(|t-s|)$.

- If $\{X_t\}$ is strictly stationary (and has finite variance), then it is weakly stationary.

Weak Stationarity

In practice, strict Stationarity is hard to impose/estimate.

Definition

A Time series $\{X_t\}$ is weakly stationary if it has finite variance, $\mu_X(t) = cte$ and $R_X(t,s) = f(|t-s|)$.

- If $\{X_t\}$ is strictly stationary (and has finite variance), then it is weakly stationary.
- Converse not true, but ...

Weak Stationarity

In practice, strict Stationarity is hard to impose/estimate.

Definition

A Time series $\{X_t\}$ is weakly stationary if it has finite variance, $\mu_X(t) = cte$ and $R_X(t,s) = f(|t-s|)$.

- If $\{X_t\}$ is strictly stationary (and has finite variance), then it is weakly stationary.
- Converse not true, but ...
- ...If $\{X_t\}$ is Gaussian and weakly stationary, then it is strictly stationary.

Autocorrelation Function (ACF)

The Autocorrelation of a stationary process $\{X_t\}$ is

$$\rho_X(h) = \frac{R_X(h)}{R_X(0)} = \frac{cov(X_t, X_{t+h})}{var(X_t)} .$$

Examples: White Noise

$\{W_t\}$ white noise.

Examples: White Noise

$\{W_t\}$ white noise. It is weakly stationary since

$$\mu_w(t) = 0 \,\forall\, t \ , \ \text{and } R_w(s,t) = \sigma^2 \mathbf{1}(|s-t| = 0) \ .$$

Examples: Random Walk

$\{X_t\}$ Random Walk. We saw that

$$\mu_X(t) = 0 \ , \ R_X(s,t) = \sigma^2 \min(s,t) \ .$$

Examples: Random Walk

$\{X_t\}$ Random Walk. We saw that

$$\mu_X(t) = 0 \ , \ R_X(s,t) = \sigma^2 \min(s,t) \ .$$

Thus it is NOT stationary.

Examples: $MA(1)$ process

Moving average process $MA(1)$:

$$X_t = W_t + \lambda W_{t-1} \ , \ \{W_t\} \text{ white noise } .$$

Examples: $MA(1)$ process

Moving average process $MA(1)$:

$$X_t = W_t + \lambda W_{t-1} \ , \ \{W_t\} \text{ white noise} \ .$$

We have $\mu_X(t) = 0$, and

$$\begin{aligned} R_X(t, t+h) &= \mathbf{E}\left(X_t X_{t+h}\right) \\ &= \mathbf{E}\left((W_t + \lambda W_{t-1})(W_{t+h} + \lambda W_{t+h-1})\right) \\ &= \begin{cases} \sigma^2(1+\lambda^2) & \text{if } h = 0 \\ \sigma^2\lambda & \text{if } h = \pm 1 \\ 0 & \text{otherwise} \ . \end{cases} \end{aligned}$$

Examples: $MA(1)$ process

Moving average process $MA(1)$:

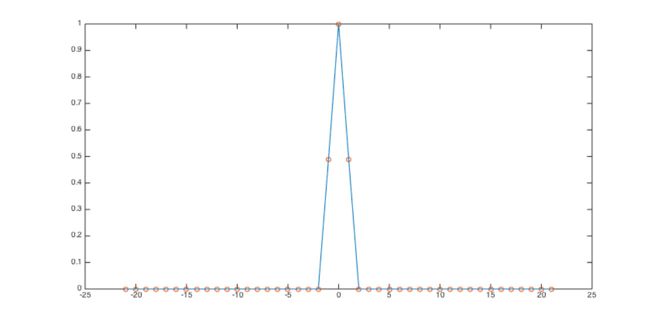
$$X_t = W_t + \lambda W_{t-1} \ , \ \{W_t\} \text{ white noise} \ .$$

We have $\mu_X(t) = 0$, and

$$\begin{aligned} R_X(t, t+h) &= \mathbf{E}\left(X_t X_{t+h}\right) \\ &= \mathbf{E}\left((W_t + \lambda W_{t-1})(W_{t+h} + \lambda W_{t+h-1})\right) \\ &= \begin{cases} \sigma^2(1+\lambda^2) & \text{if } h = 0 \\ \sigma^2\lambda & \text{if } h = \pm 1 \\ 0 & \text{otherwise} \ . \end{cases} \end{aligned}$$

Examples: $MA(1)$ Process.

ACF of $\{X_t\}$:



Last Important Example: Autoregressive Process

Autoregressive $AR(1)$ process:

$$X_t = \lambda X_{t-1} + W_t \ , \quad \text{with } \{W_t\} \text{ white noise and } |\lambda| < 1 \ .$$

Last Important Example: Autoregressive Process

Autoregressive $AR(1)$ process:

$$X_t = \lambda X_{t-1} + W_t \ , \quad \text{with } \{W_t\} \text{ white noise and } |\lambda| < 1 \ .$$

By substituting the recursion we obtain

$$X_t = W_t + \lambda W_{t-1} + \lambda^2 W_{t-2} + \lambda^3 W_{t-3} + \ldots$$

Last Important Example: Autoregressive Process

Autoregressive $AR(1)$ process:

$$X_t = \lambda X_{t-1} + W_t \ , \quad \text{with } \{W_t\} \text{ white noise and } |\lambda| < 1 \ .$$

By substituting the recursion we obtain

$$X_t = W_t + \lambda W_{t-1} + \lambda^2 W_{t-2} + \lambda^3 W_{t-3} + \dots$$

$$\mu_X(t) = \mathbf{E}\left(\sum_{k=0}^{\infty} \lambda^k W_{t-k}\right) = \sum_k \lambda^k \mathbf{E}\left(W_{t-k}\right) = 0 \ , \ \text{ and}$$

$$\mathbf{E}\left(X_t^2\right) = \sum_{k=0}^{\infty} \lambda^{2k} \sigma^2 = \frac{\sigma^2}{1-\lambda^2} \ .$$

Autoregressive Process

Suppose $h > 0$ first. Then

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

(check for $h < 0$ at home).

Autoregressive Process

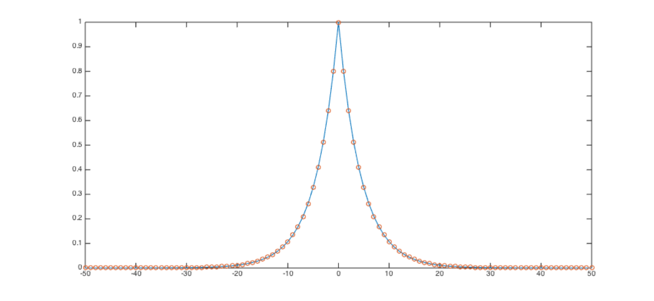
Suppose $h > 0$ first. Then

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

(check for $h < 0$ at home).
So $AR(1)$ is (weakly) stationary.

Example: $AR(1)$ Process.

ACF of $\{X_t\}$:



Linear Processes

So far, all stationary processes we have seen are of the form

$$X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k W_{t-k} \ ,$$

with

- $\{W_t\}$ white noise.
- $\sum_k |\psi_k| < \infty$.

Linear Processes

So far, all stationary processes we have seen are of the form

$$X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k W_{t-k} \ ,$$

with

- $\{W_t\}$ white noise.
- $\sum_k |\psi_k| < \infty$.

These are called **linear processes**.

Linear Processes

Proposition

Any linear Process $\{X_t\}$ is weakly stationary, with

$$\mu_X(t) = \mu \ ,$$

$$R_X(h) = \sigma^2 \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+h} \ .$$

Linear Processes

Proposition

Any linear Process $\{X_t\}$ *is weakly stationary, with*

$$\mu_X(t) = \mu \ ,$$

$$R_X(h) = \sigma^2 \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+h} \ .$$

Also, stationary processes are *essentially* linear.

Examples of Linear Processes

For white noise, choose $\mu = 0$ and

$$\psi_k = \begin{cases} 1 & \text{if } k = 0 \ , \\ 0 & \text{otherwise.} \end{cases}$$

Examples of Linear Processes

For a MA(1) process, choose $\mu = 0$ and

$$\psi_k = \begin{cases} 1 & \text{if } k = 0\ , \\ \lambda & \text{if } k = 1\ , \\ 0 & \text{otherwise.} \end{cases}$$

Examples of Linear Processes

For a AR(1) process, choose $\mu = 0$ and

$$\psi_k = \begin{cases} \lambda^k & \text{if } k \geq 0 \ , \\ 0 & \text{otherwise.} \end{cases}$$

Examples of Linear Processes

Q: What about a random walk?

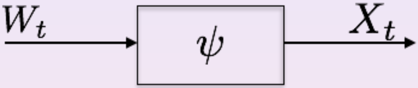
Examples of Linear Processes

Q: What about a random walk?

$$X_t = \sum_{0 \leq k \leq t} W_{t-k} \neq \sum_k \psi_k W_{t-k} \ .$$

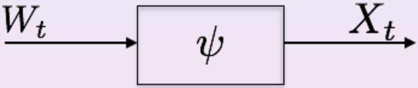
A slide for EE

We can view a linear process $\{X_t\}$ as



A slide for EE

We can view a linear process $\{X_t\}$ as



- $\{X_t\}$ is thus obtained by *filtering* a white noise with the filter ψ.
- The filter operation is known also as *convolution*:

$$(W * \psi)_t := \sum_k \psi_k W_{t-k} \ .$$

Key quantities to estimate

Suppose $\{X_t\}$ is a stationary process. Given a finite number of observations $x_1, \ldots, x_n$, we need to estimate

- Its mean μ.

Key quantities to estimate

Suppose $\{X_t\}$ is a stationary process. Given a finite number of observations $x_1, \ldots, x_n$, we need to estimate

- Its mean μ.
- Its autocovariance function

$$R_X(h) = \mathrm{cov}(X_t, X_{t+h}) \ ,$$

Key quantities to estimate

Suppose $\{X_t\}$ is a stationary process. Given a finite number of observations $x_1, \ldots, x_n$, we need to estimate

- Its mean μ.
- Its autocovariance function

$$R_X(h) = \mathrm{cov}(X_t, X_{t+h}) \ ,$$

- and its ACF:

$$\rho_X(h) = \frac{R_X(h)}{R_X(0)} \ .$$

Sample Mean and Autocorrelation

The sample mean is

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Sample Mean and Autocorrelation

The sample mean is

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

The sample autocorrelation is

$$\widehat{R_X}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(x_i - \widehat{\mu})(x_{i+h} - \widehat{\mu}) \ , \ (\text{for } h < n) \ .$$

Sample Mean and Autocorrelation

The sample mean is

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

The sample autocorrelation is

$$\widehat{R_X}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(x_i - \widehat{\mu})(x_{i+h} - \widehat{\mu}) \ , \ (\text{for } h < n) \ .$$

The sample ACF is

$$\widehat{\rho_X}(h) = \frac{\widehat{R_X}(h)}{\widehat{R_X}(0)} \ .$$

Sample Mean

Q: What is the variance of the sample mean estimator?

Sample Mean

Q: What is the variance of the sample mean estimator?

$$\begin{aligned}\text{var}\,(\widehat{\mu}) &= \frac{1}{n^2}\text{cov}(\sum_i x_i, \sum_{i'} x_{i'}) \\ &= \frac{1}{n^2}\left(nR_X(0) + (n-1)(R_X(1) + R_X(-1)) + \dots\right. \\ &\left. +(n-2)(R_X(2) + R_X(-2)) + \dots (R_X(n-1) + R_X(1-n))\right) \\ &= \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right) R_X(h)\ .\end{aligned}$$

Consequence: If $R_X(h)$ is smooth, variance of the sample mean increases.

Sample Autocovariance

$$\widehat{R_X}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(x_i - \widehat{\mu})(x_{i+h} - \widehat{\mu}) \ , \ (\text{for } h < n) \ .$$

Sample Autocovariance

$$\widehat{R_X}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(x_i - \widehat{\mu})(x_{i+h} - \widehat{\mu})\ ,\ (\text{for } h < n)\ .$$

Similar to the sample covariance of $(x_1, x_{1+h}), \ldots, (x_{n-h}, x_n)$, except

- divide by n instead of $n - h$,
- sample mean $\widehat{\mu}$ using all n samples.

Why?

Linear Prediction in Time Series

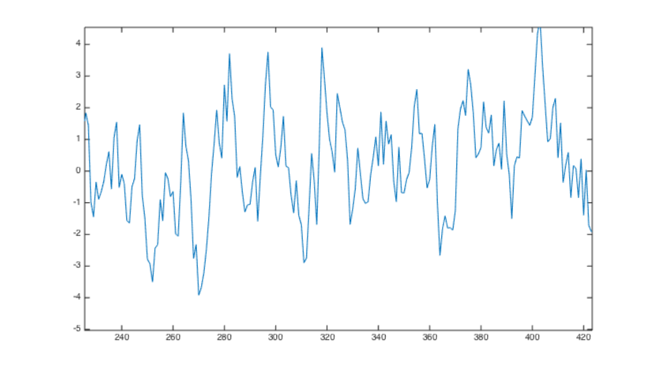
Q: For a given time series model, how well can X_t be predicted from the past?

Linear Prediction in Time Series

Q: For a given time series model, how well can X_t be predicted from the past?

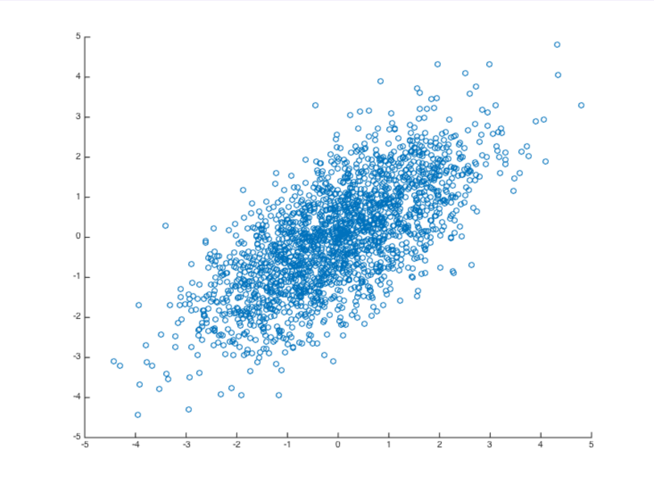
Q2: How is the prediction related to the Autocorrelation function?

Example: $AR(1)$ process



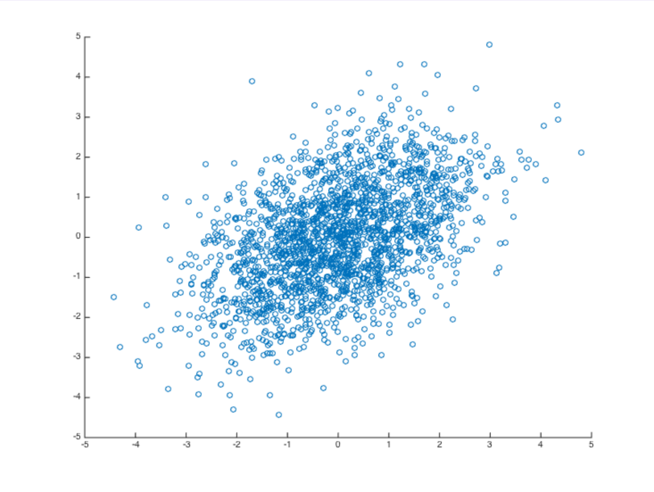
Example: $AR(1)$ process

Scatterplot between X_t and X_{t-1}:



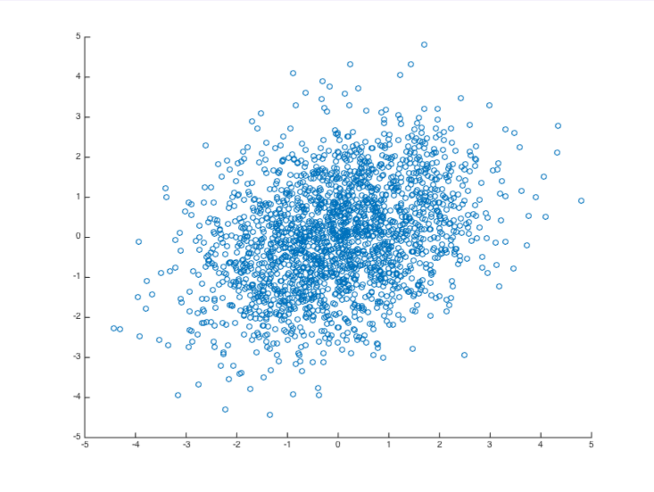
Example: $AR(1)$ process

Scatterplot between X_t and X_{t-2}:



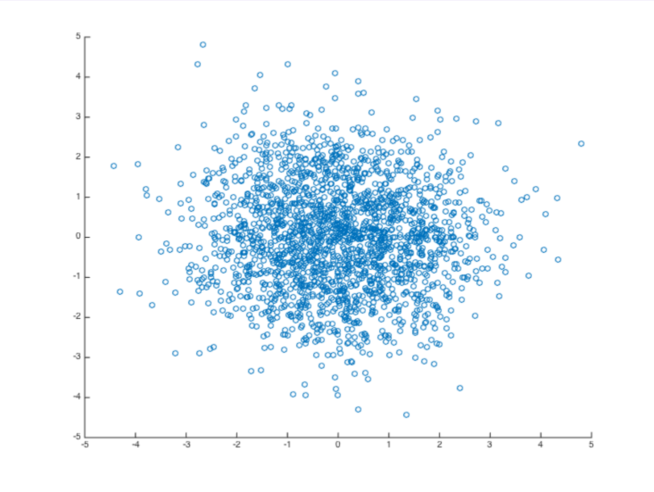
Example: $AR(1)$ process

Scatterplot between X_t and X_{t-3}:



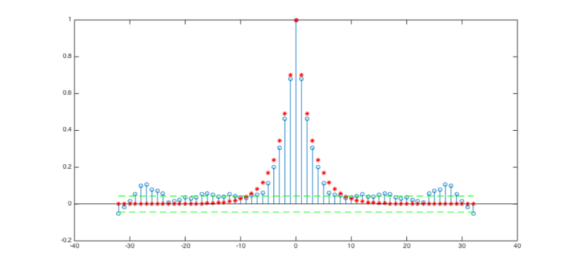
Example: $AR(1)$ process

Scatterplot between X_t and X_{t-40}:



Example: $AR(1)$ process

Recall the ACF function:



ACF controls quality of *linear* prediction.

Least Squares and ACF

- Least squares estimate of a random variable Y:

$$\min_f \mathbf{E}\left(|Y - f|^2\right) \Longrightarrow f =$$

Least Squares and ACF

- Least squares estimate of a random variable Y:

$$\min_f \mathbf{E}\left(|Y - f|^2\right) \Longrightarrow f = \mathbf{E}(Y) ,$$

with MSE $\mathrm{var}(Y)$.

Least Squares and ACF

- Least squares estimate of a random variable Y:

$$\min_f \mathbf{E}\left(|Y - f|^2\right) \implies f = \mathbf{E}(Y) \ ,$$

  with MSE $\mathrm{var}(Y)$.

- Least squares estimate of Y, given X:

$$\min_f \mathbf{E}\left(|Y - f(X)|^2 \mid X\right) \implies f(X) =$$

Least Squares and ACF

- Least squares estimate of a random variable Y:

$$\min_f \mathbf{E}\left(|Y - f|^2\right) \Longrightarrow \; f = \mathbf{E}\left(Y\right) \; ,$$

with MSE $\mathrm{var}\left(Y\right)$.

- Least squares estimate of Y, given X:

$$\min_f \mathbf{E}\left(|Y - f(X)|^2 \;|X\right) \Longrightarrow \; f(X) = \mathbf{E}\left(Y|X\right) \; ,$$

with MSE $\mathrm{var}\left(Y|X\right)$.

- Thus, least squares estimate of X_{t+h} given X_t: $\mathbf{E}\left(X_{t+h}|X_t\right)$.

Least Squares and ACF

Conditional expectations are easy under Gaussian distributions!

Least Squares and ACF

Conditional expectations are easy under Gaussian distributions!
Suppose $X_1, \ldots, X_n$ is jointly Gaussian, with density

$$f_X(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) .$$

Least Squares and ACF

Conditional expectations are easy under Gaussian distributions!
Suppose $X_1, \dots, X_n$ is jointly Gaussian, with density

$$f_X(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) .$$

In particular, joint law of (X_t, X_{t+h}) is also Gaussian, with mean (μ_t, μ_{t+h}) and covariance

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

Least Squares and ACF

Conditional expectations are easy under Gaussian distributions!
Suppose $X_1, \dots, X_n$ is jointly Gaussian, with density

$$f_X(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right) .$$

In particular, joint law of (X_t, X_{t+h}) is also Gaussian, with mean (μ_t, μ_{t+h}) and covariance

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

Conditional distribution X_{t+h} given $X_t = x_t$ is therefore

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

Least Squares and ACF

$\{X_t\}$ Gaussian and stationary. What is the optimal prediction of X_{t+h} given $X_t = x_t$?

Least Squares and ACF

$\{X_t\}$ Gaussian and stationary. What is the optimal prediction of X_{t+h} given $X_t = x_t$?

$$f(x_t) = \mathbf{E}\left(X_{t+h} | X_t = x_t\right) = \mu + \rho_X(h)(x_t - \mu)\ .$$

The resulting MSE is

$$\mathbf{E}\left(|X_{t+h} - f(x_t)|^2\ , |X_t = x_t\right) = \sigma^2(1 - \rho_X(h)^2)\ .$$

Least Squares and ACF

$\{X_t\}$ Gaussian and stationary. What is the optimal prediction of X_{t+h} given $X_t = x_t$?

$$f(x_t) = \mathbf{E}\left(X_{t+h} | X_t = x_t\right) = \mu + \rho_X(h)(x_t - \mu) \ .$$

The resulting MSE is

$$\mathbf{E}\left(|X_{t+h} - f(x_t)|^2 \ , |X_t = x_t\right) = \sigma^2(1 - \rho_X(h)^2) \ .$$

- Prediction gets better as $|\rho|$ increases.
- $f(x_t)$ is linear: $f(x_t) = \alpha + \beta t$.

Least Squares and ACF

For general stationary processes (not necessarily Gaussian), best predictor (in terms of MSE) has no closed form.

Least Squares and ACF

For general stationary processes (not necessarily Gaussian), best predictor (in terms of MSE) has no closed form.
However, we can consider optimal *linear* predictors.

$$\mathbf{E}\left(|X_{t+h} - \alpha - \beta X_t|^2\right) = E(\alpha, \beta) \ .$$

Setting $\partial_\alpha E(\alpha, \beta) = 0$ and $\partial_\beta E(\alpha, \beta) = 0$, we obtain

$$\alpha = \mu(1 - \rho_X(h)) \ , \ \beta = \rho_X(h) \ ,$$

with

$$MSE = \sigma^2(1 - \rho_X(h)^2) \ .$$

Least Squares and ACF

$$f(x_t) = \mu + \rho_X(h)(x_t - \mu)$$

- Optimal linear predictor for any stationary $\{X_t\}$.

Least Squares and ACF

$$f(x_t) = \mu + \rho_X(h)(x_t - \mu)$$

- Optimal linear predictor for any stationary $\{X_t\}$.
- Optimal predictor for stationary and gaussian $\{X_t\}$.

Least Squares and ACF

$$f(x_t) = \mu + \rho_X(h)(x_t - \mu)$$

- Optimal linear predictor for any stationary $\{X_t\}$.
- Optimal predictor for stationary and gaussian $\{X_t\}$.
- Corollary: For gaussian processes, linear prediction is optimal (in MSE).
- Extension to multiple time indices.

Wrap-up

- Two main quantities to estimate for stationary processes: mean and autocovariance (autocorrelation).
- Sample autocovariance/ACF is also positive semidefinite.
- Large sample distributions of sample mean, autocovariance, ACF available: *asymptotically Normal*.

AR(1) and the Backshift operator

We can write

$$\begin{aligned} X_t - \lambda X_{t-1} &= W_t \\ (1 - \lambda B) X_t &= W_t \\ P(B) X_t &= W_t \ , \end{aligned}$$

where $BX_t = X_{t-1}$ is the backshift or translation operator, and $P(B) = 1 - \lambda B$.

AR(1) and the Backshift operator

We can write

$$\begin{aligned} X_t - \lambda X_{t-1} &= W_t \\ (1 - \lambda B) X_t &= W_t \\ P(B) X_t &= W_t \ , \end{aligned}$$

where $BX_t = X_{t-1}$ is the backshift or translation operator, and $P(B) = 1 - \lambda B$.

- We can write the differentiation $\nabla X_t = X_t - X_{t-1} = (1 - B) X_t$.
- $B^2 X_t = BBX_t = BX_{t-1} = X_{t-2}$, and
- $B^k X_t = X_{t-k}$.

AR(1) and the Backshift operator

Also, the recurrence $X_t = \sum_{k=0}^{\infty} \lambda^k W_{t-k}$ can be written as

$$
\begin{aligned}
X_t &= \sum_{k=0}^{\infty} \lambda^k W_{t-k} \\
&= \sum_{k=0}^{\infty} \lambda^k B^k W_t \\
& Q(B) W_t \ ,
\end{aligned}
$$

where $Q(B) = \sum_{k \geq 0} \lambda^k B^k$.

AR(1) and the Backshift operator

Also, the recurrence $X_t = \sum_{k=0}^{\infty} \lambda^k W_{t-k}$ can be written as

$$X_t = \sum_{k=0}^{\infty} \lambda^k W_{t-k}$$
$$= \sum_{k=0}^{\infty} \lambda^k B^k W_t$$
$$Q(B) W_t \ ,$$

where $Q(B) = \sum_{k\geq 0} \lambda^k B^k$.
Why is this useful?

AR(1) and the Backshift operator

$P(B) = 1 - \lambda B$ *and* $Q(B) = \sum_{k \geq 0} \lambda^k B^k$ *are related by*

$$P(B)Q(B) = 1 \;, \quad \textit{or } Q(B) = P(B)^{-1} \;.$$

Since $P(B)X_t = W_t$, it results that

$$X_t = P(B)^{-1} W_t = Q(B) W_t \;.$$

AR(1) and the Backshift operator

$P(B) = 1 - \lambda B$ and $Q(B) = \sum_{k \geq 0} \lambda^k B^k$ are related by

$$P(B)Q(B) = 1 \ , \quad \textit{or } Q(B) = P(B)^{-1} \ .$$

Since $P(B)X_t = W_t$, it results that

$$X_t = P(B)^{-1} W_t = Q(B) W_t \ .$$

Remark: P and Q are operators that behave as polynomials:

$$\frac{1}{1 - \lambda z} = \sum_{k \geq 0} \lambda^k z^k \ , \ |\lambda| < 1, |z| \leq 1 \ .$$

AR(1) Process

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

Q: What happens when $|\lambda| > 1$?

AR(1) Process

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

Q: What happens when $|\lambda| > 1$?
$Q(B)W_t = \sum_{k\geq 0} \lambda^k B^k W_t$ does not converge.

AR(1) Process

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

Q: What happens when $|\lambda| > 1$?
$Q(B)W_t = \sum_{k\geq 0} \lambda^k B^k W_t$ does not converge.

$$\text{But} \quad \frac{1}{\lambda} X_t = \frac{\lambda}{\lambda} X_{t-1} + \frac{1}{\lambda} W_t$$

thus

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

By solving the recurrence, we obtain $X_t = -\sum_{k=1}^{\infty} \lambda^{-k} W_{t+k}$.
$\Rightarrow \quad X_t$ depends upon the future!

Review: AR(1) Process

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

- It has a unique, well-defined, stationary solution when $\lambda \neq \pm 1$.
- It has many non-stationary solutions, even for $|\lambda| = 1$.

Causality

Definition

A linear process $\{X_t\}$ is **causal** (with respect to $\{W_t\}$) if it can be written as

$$X_t = \psi(B) W_t \ ,$$

with $\psi(B) = \sum_{k \geq 0} \psi_k B^k$ and $\sum_{k \geq 0} |\psi_k| < \infty$.

Causality

Definition

A linear process $\{X_t\}$ is **causal** (with respect to $\{W_t\}$) if it can be written as

$$X_t = \psi(B)W_t \ ,$$

with $\psi(B) = \sum_{k\geq 0} \psi_k B^k$ and $\sum_{k\geq 0} |\psi_k| < \infty$.

- Thus, if $|\lambda| < 1$, AR model $P(B)X_t = W_t$ is causal wrt $\{W_t\}$.
- Conversely, if $AR(1)$ with parameter λ is casual, then $|\lambda| < 1$.
- Causality of a time series is relative.

MA(1) Process

Recall the MA(1) Process

$$X_t = W_t + \theta W_{t-1} = (1 + \theta B)W_t = P(B)W_t \ .$$

MA(1) Process

Recall the MA(1) Process

$$X_t = W_t + \theta W_{t-1} = (1+\theta B)W_t = P(B)W_t \ .$$

If $|\theta| < 1$, we can do

$$\begin{aligned} P(B)^{-1}X_t &= W_t \\ \frac{1}{1+\theta B}X_t &= W_t \\ \left(1-\theta B+\theta^2 B^2-\theta^3 B^3+\dots\right)X_t &= W_t \\ \sum_{k\geq 0}(-\theta)^k X_{t-k} &= W_t \ , \end{aligned}$$

MA(1) Process

Recall the MA(1) Process

$$X_t = W_t + \theta W_{t-1} = (1 + \theta B) W_t = P(B) W_t \ .$$

If $|\theta| < 1$, we can do

$$\begin{aligned} P(B)^{-1} X_t &= W_t \\ \frac{1}{1+\theta B} X_t &= W_t \\ \left(1 - \theta B + \theta^2 B^2 - \theta^3 B^3 + \dots\right) X_t &= W_t \\ \sum_{k \geq 0} (-\theta)^k X_{t-k} &= W_t \ , \end{aligned}$$

so $\{W_t\}$ is casual wrt $\{X_t\}$. We have *inverted* the roles of $\{X_t\}$ and $\{W_t\}$.

Invertibility

Definition

A linear process $\{X_t\}$ is **invertible** if there exist $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \dots$ with $\sum_k |\psi_k| < \infty$ and

$$\psi(B)X_t = W_t \ .$$

MA(1) and Invertibility

- Similarly as causality, invertibility involves both $\{X_t\}$ and $\{W_t\}$.
- In the MA(1) case, $X_t = (1 + \theta B)W_t$, $\{X_t\}$ is invertible if $|\theta| < 1$.
- Converse is also true.
- Equivalently, $\{X_t\}$ is invertible if and only if the root z_1 of the polynomial $1 + \theta z$ satisfies $|z_1| > 1$.

AR(1), MA(1), Invertibility, Causality

$$X_t - \lambda X_{t-1} = (1 - \lambda B) X_t = W_t \quad \textit{is}$$

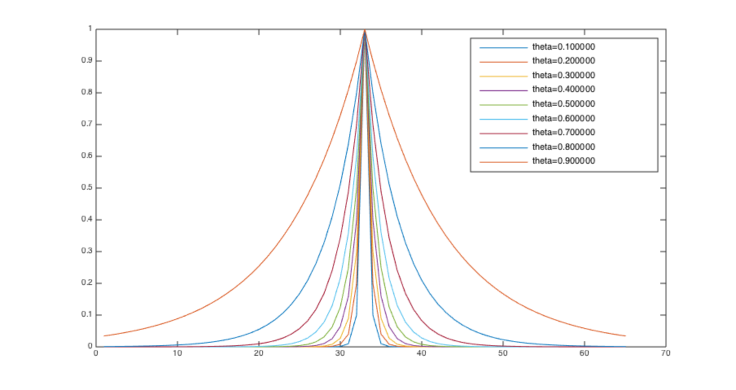
- *Causal (wrt $\{W_t\}$) iff $|\lambda| < 1$.*
- *Always invertible (wrt $\{W_t\}$).*

$$X_t = W_t + \theta W_{t-1} = (1 + \theta B) W_t \quad \textit{is}$$

- *Always causal (wrt $\{W_t\}$).*
- *Invertible (wrt $\{W_t\}$) iff $|\theta| < 1$.*

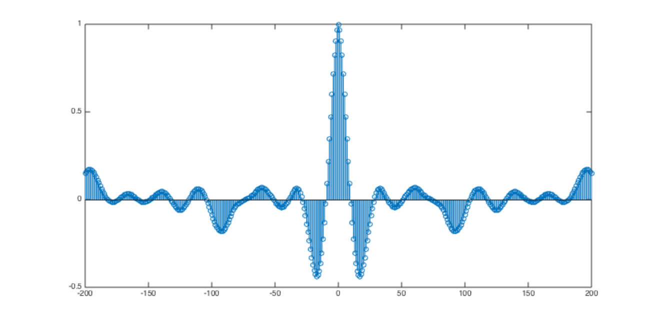
What can we model with AR(1) processes?

Typical ACF of a AR(1) process:



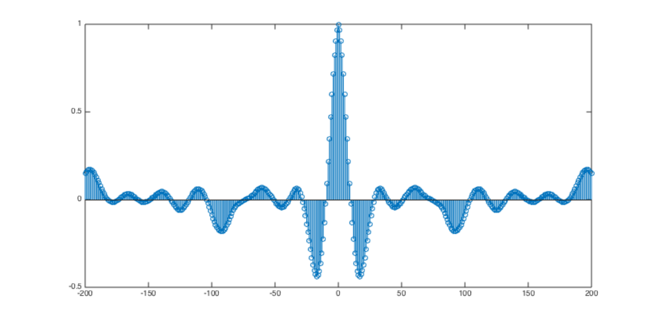
What can we model with AR(1) processes?

Pick a real-life example of fairly stationary time series: *here*
Its sample ACF:



What can we model with AR(1) processes?

Pick a real-life example of fairly stationary time series: *here*
Its sample ACF:



Need to make our models more *expressive*.

High-Order Autoregressive Models

High-Order Autoregressive Models

Definition

An AR(p) process $\{X_t\}$ is a stationary process satisfying

$$X_t - \lambda_1 X_{t-1} - \ldots \lambda_p X_{t-p} = W_t \ ,$$

where $\{W_t\}$ is a white noise.

It is equivalent to $P(B)X_t = W_t$, this time using

$$P(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \ldots \lambda_p B^p \ .$$

Constraints on the polynomial $P(B)$

Remember that for $p = 1$, we needed $|\lambda| \neq 1$ in order to have stationary solutions to $P(B)X_t = W_t$.

Constraints on the polynomial $P(B)$

Remember that for $p = 1$, we needed $|\lambda| \neq 1$ in order to have stationary solutions to $P(B)X_t = W_t$. Writing $P(z) = 1 - \lambda z$, this is equivalent to

$$\forall\ z \in \mathbb{R}\ , P(z) = 0 \Rightarrow z \neq \pm 1\ ,\ \ \mathit{or}$$
$$\forall\ z \in \mathbb{C}\ , P(z) = 0 \Rightarrow |z| \neq 1\ .$$

Constraints on the polynomial $P(B)$

Remember that for $p = 1$, we needed $|\lambda| \neq 1$ in order to have stationary solutions to $P(B)X_t = W_t$. Writing $P(z) = 1 - \lambda z$, this is equivalent to

$$\forall\ z \in \mathbb{R}\ , P(z) = 0 \Rightarrow z \neq \pm 1\ ,\ \ or$$
$$\forall\ z \in \mathbb{C}\ , P(z) = 0 \Rightarrow |z| \neq 1\ .$$

Q: What about the general case AR(p)?

Constraints on the polynomial $P(B)$

- In general, a polynomial $P(z) = 1 - \lambda_1 z - \lambda_2 z^2 - \ldots \lambda_p z^p$ will have *complex* roots (even if $\lambda_k \in \mathbb{R}$).

Constraints on the polynomial $P(B)$

- In general, a polynomial $P(z) = 1 - \lambda_1 z - \lambda_2 z^2 - \dots \lambda_p z^p$ will have *complex* roots (even if $\lambda_k \in \mathbb{R}$).
- In order to have a stationary solution, we want all the roots z_k^* of $P(z)$ to satisfy $|z_k^*| \neq 1$.

Stationarity and Causality

Theorem

1. *The equation* $P(B)X_t = W_t$ *has a unique stationary solution if and only if*

$$P(z) = 0 \Rightarrow |z| \neq 1 \ .$$

We call this unique solution an AR(p) process.

Stationarity and Causality

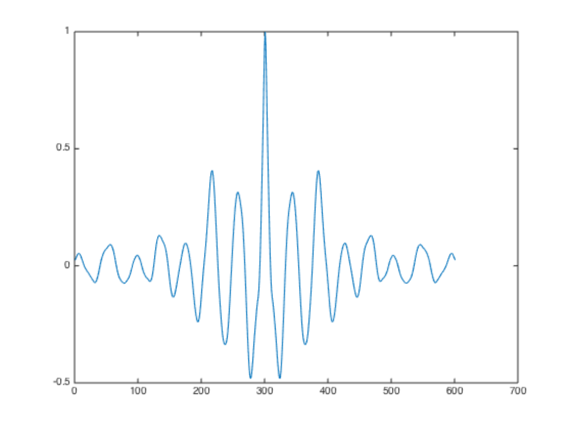
Theorem

1. *The equation $P(B)X_t = W_t$ has a unique stationary solution if and only if*
$$P(z) = 0 \Rightarrow |z| \neq 1 \ .$$
We call this unique solution an AR(p) process.
2. *Moreover, this process is causal if and only if*
$$P(z) = 0 \Rightarrow |z| > 1 \ .$$

The recurrence equations can be solved using linear differential equations methods.

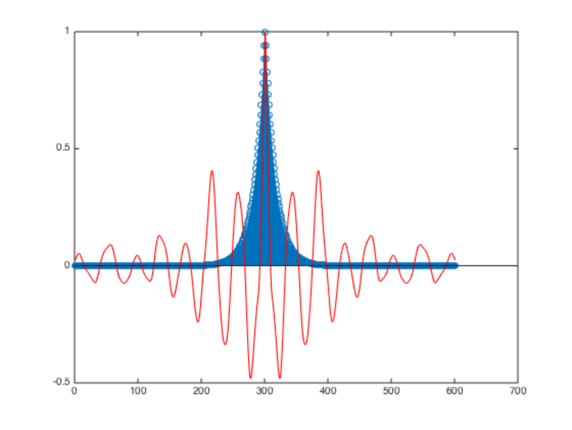
Example: AR(1) vs AR(r) processes

Recall the sound example. We estimated an ACF of the form



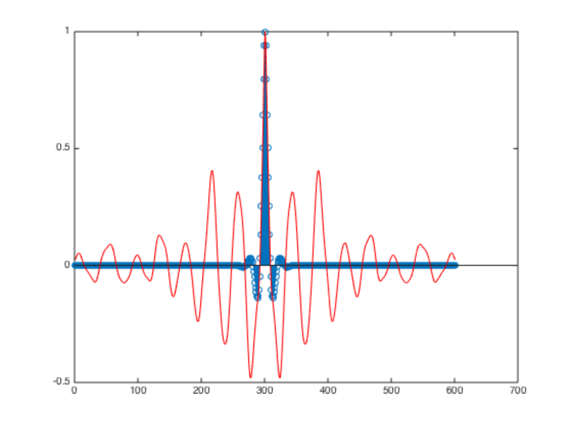
Example: AR(1) vs AR(r) processes

ACF when we fit an AR(1) model:



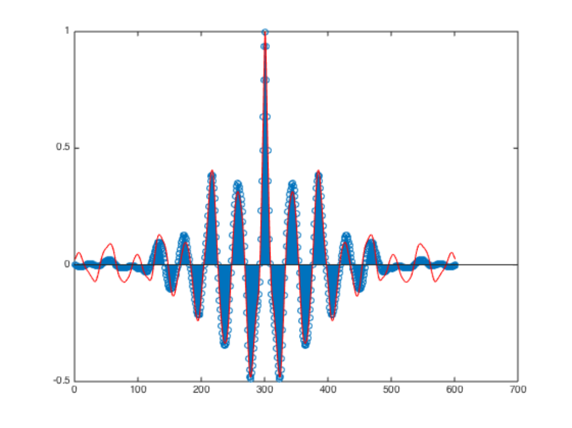
Example: AR(1) vs AR(r) processes

ACF when we fit an AR(4) model:



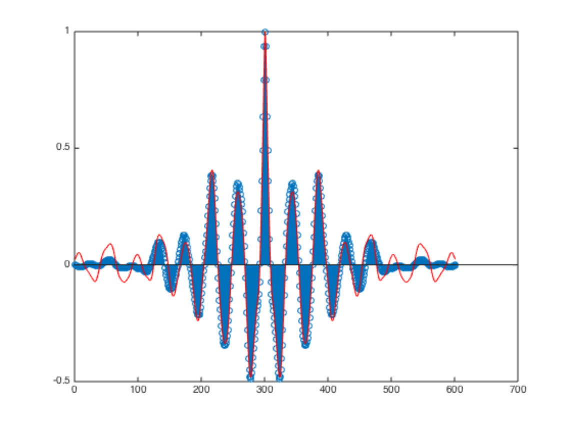
Example: AR(1) vs AR(r) processes

ACF when we fit an AR(16) model:



Example: AR(1) vs AR(r) processes

ACF when we fit an AR(16) model:



Other alternatives?

MA(q) process

Definition

The moving average model of order q, or MA(q), is defined as

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \dots \theta_q W_{t-q} \ ,$$

where $\{W_t\}$ is a white noise.

MA(q) process

Definition

The moving average model of order q, or MA(q), is defined as

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \dots \theta_q W_{t-q} ,$$

where $\{W_t\}$ is a white noise.

We can also write

$$X_t = \theta(B) W_t ,$$

with $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$.

ARMA Processes

Definition

An ARMA(p,q) process $\{X_t\}$ is a stationary process that satisfies

$$X_t - \lambda_1 X_{t-1} - \ldots \lambda_p X_{t-p} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q} \ ,$$

where $\{W_t\}$ is a white noise.

- AR(p) = ARMA(p, 0), ie $\theta(B) = 1$.
- MA(q) = ARMA(0,q), ie $P(B) = 1$.

ARMA Processes

Definition

An ARMA(p,q) process $\{X_t\}$ is a stationary process that satisfies

$$X_t - \lambda_1 X_{t-1} - \ldots \lambda_p X_{t-p} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q} ,$$

where $\{W_t\}$ is a white noise.

- AR(p) = ARMA(p, 0), ie $\theta(B) = 1$.
- MA(q) = ARMA(0,q), ie $P(B) = 1$.
- We ask that $\lambda_p \neq 0$, $\theta_q \neq 0$, and that $P(B)$, $\theta(B)$ have no common roots. (Why?)

ARMA Processes

We have a total of $p+q$ parameters. ARMA(p,q) can approximate many stationary processes:

For any stationary process with autocovariance R and any $k>0$, there is an ARMA process $\{X_t\}$ such that

$$R_X(h) = R(h)\ , h \leq k\ .$$

Complex Analysis Defrost: Polynomials of complex variable

A polynomial $Q(z)$ *can be factorized as*

$$Q(z) = a_0 + a_1 z + \ldots a_p z^p = a_p (z - z_1)(z - z_2) \ldots (z - z_p) \ ,$$

where $z_1, \ldots z_p \in \mathbb{C}$ *are the roots of* Q.

If $a_0, a_1, \ldots a_p \in \mathbb{R}$, *then the roots are either real or they come in conjugate pairs:* $z_i = \overline{z_j}$.

A complex number $z = a + ib$ *has real part* $\Re(z) = a$ *and imaginary part* $\Im(z) = b$. *Conjugate of* z *is* $\overline{z} = a - ib$, *modulus of* z *is* $|z| = \sqrt{a^2 + b^2}$ *and phase is* $\arg(z) = \tan^{-1}(b/a)$.

ARMA Processes

- AR(p) = ARMA(p, 0), ie $\theta(B) = 1$.
- MA(q) = ARMA(0,q), ie $P(B) = 1$.
- We ask that $\lambda_p \neq 0$, $\theta_q \neq 0$, and that $P(B)$, $\theta(B)$ have no common roots.
- The corresponding equation is written as

$$P(B)X_t = \theta(B)W_t \ ,$$

where $P(B)$ has degree p and $\theta(B)$ has degree q.

Causality and Invertibility

Definition

A linear process $\{X_t\}$ is **causal** (with respect to $\{W_t\}$) if it can be written as

$$X_t = \psi(B) W_t \ ,$$

with $\psi(B) = \sum_{k \geq 0} \psi_k B^k$ and $\sum_{k\geq 0} |\psi_k| < \infty$.

Definition

A linear process $\{X_t\}$ is **invertible** if there exist $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \dots$ with $\sum_k |\psi_k| < \infty$ and

$$\psi(B) X_t = W_t \ .$$

ARMA Stationarity, Causality and Invertibility

Theorem

- *If P and θ have no common factors, a stationary solution to $P(B)X_t = \theta(B)W_t$ exists iff the roots of $P(z)$ avoid the unit circle: $P(z) = 0 \Rightarrow |z| \neq 1$. This is called an ARMA(p,q) process.*
- *This process is* **causal** *iff the roots of $P(z)$ are* **outside** *the unit circle: $P(z) = 0 \Rightarrow |z| > 1$.*
- *This process is* **invertible** *iff the roots of $\theta(B)$ are* **outside** *the unit circle: $\theta(z) = 0 \Rightarrow |z| > 1$.*

ARMA vs AR vs MA

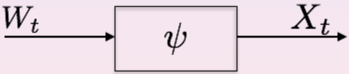
$$P(B)X_t = \theta(B)W_t \ , \ X_t = \psi(B)W_t \ .$$

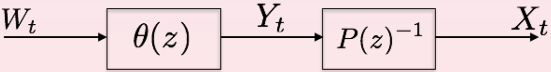
ARMA vs AR vs MA

$$P(B)X_t = \theta(B)W_t \ , \ X_t = \psi(B)W_t \ .$$

We can think an ARMA model as concatenating two models:

$$Y_t = \theta(B)W_t \ , \ \text{and} \ \ P(B)X_t = Y_t \ .$$





Autocovariance of an ARMA process

We know that

$$X_t - \lambda_1 X_{t-1} - \dots \lambda_p X_{t-p} = \theta_0 W_t + \cdots + \theta_q W_{t-q} \ .$$

$$\begin{aligned} \mathrm{cov}(X_t - \lambda_1 X_{t-1} - \cdots - \lambda_p X_{t-p}, X_{t-h}) = \\ \mathrm{cov}(X_t, X_{t-h}) - \lambda_1 \mathrm{cov}(X_{t-1}, X_{t-h}) - \cdots - \lambda_p \mathrm{cov}(X_{t-p}, X_{t-h}) = \\ \theta_0 \mathrm{cov}(W_t, X_{t-h}) + \cdots + \theta_q \mathrm{cov}(W_{t-q}, X_{t-h}) \ . \end{aligned}$$

So

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = \sigma^2 \left(\theta_h \psi_0 + \theta_{h+1}\psi_1 + \dots \theta_q \psi_{q-h}\right)$$

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = \sigma^2 \sum_{k=0}^{q-h} \theta_{k+h} \psi_k \ .$$

Autocovariance of an ARMA process

So the autocorrelation $R_X(h)$ also satisfies an homogeneous recurrence:

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = 0 \ , \ \text{for } h > q \ ,$$

with initial conditions given by

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = \sigma^2 \sum_{k=0}^{q-h} \theta_{k+h} \psi_k \ , \ (h = 0 \ldots, q-1$$

Autocovariance of an ARMA process

So the autocorrelation $R_X(h)$ also satisfies an homogeneous recurrence:

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = 0 \ , \ \text{for } h > q \ ,$$

with initial conditions given by

$$R_X(h) - \lambda_1 R_X(h-1) - \cdots - \lambda_p R_X(h-p) = \sigma^2 \sum_{k=0}^{q-h} \theta_{k+h} \psi_k \ , \ (h = 0 \ldots, q-1$$

How to solve these sort of equations?

Linear Homogeneous Equations

A linear homogeneous equation of order p is of the form

$$a_0 X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p} = 0 \ .$$

$$(a_0 + a_1 B + \cdots + a_p B^p) X_t = 0 \ .$$

$$a(B) X_t = 0 \ , \ \text{with } a(z) = a_0 + a_1 z + \cdots + a_p z^p \ .$$

Linear Homogeneous Equations

A linear homogeneous equation of order p is of the form

$$a_0 X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p} = 0 \ .$$

$$(a_0 + a_1 B + \cdots + a_p B^p) X_t = 0 \ .$$

$$a(B) X_t = 0 \ , \ \text{with } a(z) = a_0 + a_1 z + \cdots + a_p z^p \ .$$

$a(z)$ is the *characteristic polynomial*. Consider

$$a(z) = a_p (z - z_1)(z - z_2) \dots (z - z_p) \ .$$

Linear Homogeneous Equations

A linear homogeneous equation of order p is of the form

$$a_0 X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p} = 0 \ .$$

$$(a_0 + a_1 B + \cdots + a_p B^p) X_t = 0 \ .$$

$$a(B) X_t = 0 \ , \ \text{with } a(z) = a_0 + a_1 z + \cdots + a_p z^p \ .$$

$a(z)$ is the *characteristic polynomial.* Consider

$$a(z) = a_p (z - z_1)(z - z_2) \ldots (z - z_p) \ .$$

Thus

$$a(B) X_t = 0 \quad \Leftrightarrow \quad (B - z_1)(B - z_2) \ldots (B - z_p) X_t = 0 \ .$$

Linear Homogeneous Eqs: Summary

1. Goal: solve $a_0 X_t + \cdots + a_p X_{t-p} = 0$ with initial conditions $X_1, \ldots, X_p$.
2. Equivalent to factorizing the characteristic polynomial $a(z) = a_0 + \cdots + a_p z^p = 0$:

$$(z - z_1)^{m_1} \ldots (z - z_k)^{m_k} = 0 \ ,$$

   where z_l are the roots and m_l their corresponding multiplicity.
3. General solution:

$$\boxed{X_t = c_1(t) z_1^{-t} + \cdots + c_k(t) z_k^{-t}} \ ,$$

   where $c_l(t)$ are polynomials of degree $m_l - 1$.
4. The coefficients of these polynomials are adjusted using the initial conditions $X_1, \ldots, X_p$.

ARMA Parameter Estimation

We start by making the following two assumptions:

1. The model order is known, and
2. Data has zero mean (we can always remove the sample mean, fit, and then add the estimated sample mean otherwise).

ARMA Parameter Estimation

Two most famous parametric estimation methods:

ARMA Parameter Estimation

Two most famous parametric estimation methods:

- Maximum Likelihood.
- Method of Moments.

ARMA Maximum Likelihood Estimation

Only reasonable under Gaussian processes:

$$P(B)X_t = \theta(B)W_t \ ,$$

where W_t is an i.i.d Gaussian process with variance σ^2.

ARMA Maximum Likelihood Estimation

Only reasonable under Gaussian processes:

$$P(B)X_t = \theta(B)W_t \ ,$$

where W_t is an i.i.d Gaussian process with variance σ^2. The parameters of the model are λ_i, θ_j, $i = 1, \dots, p$, $j = 1, \dots q$. The data likelihood is

$$\mathcal{L}(\lambda, \theta, \sigma^2) = f_{\lambda,\theta,\sigma^2}(x_1, \dots, x_n) \ ,$$

where $f_{\lambda,\theta,\sigma^2}$ is the joint gaussian density of the ARMA model.

ARMA Maximum Likelihood Estimation

If $\mathbf{x} = (x_1, \dots, x_n)$, the likelihood becomes

$$\mathcal{L}(\lambda, \theta, \sigma^2) = (2\pi)^{-n/2} |\Gamma_n|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{x}^T \Gamma_n^{-1} \mathbf{x} \right) .$$

Notice that parameters are both inside and outside the exponential function.

ARMA Maximum Likelihood Estimation

Pros:

- Low variance estimates (efficient estimators).
- Gaussian Assumption is robust, ie even if $\{X_t\}$ is non-Gaussian, the asymptotic distribution of $(\hat{\lambda}, \hat{\theta})$ is the same as in the Gaussian case.

ARMA Maximum Likelihood Estimation

Pros:

- Low variance estimates (efficient estimators).
- Gaussian Assumption is robust, ie even if $\{X_t\}$ is non-Gaussian, the asymptotic distribution of $(\hat{\lambda}, \hat{\theta})$ is the same as in the Gaussian case.

Cons:

- Hard Optimization Problem: We require a good initial guess that then we define. How to obtain such cheap, initial guess?

Yule-Walker Equations

- In the AR(p) case, we saw that the forecasting coefficients

$$X_{t+1}^{t} = \phi_{t,1} X_t + \cdots + \phi_{t,t} X_1$$

  correspond exactly to the model parameters λ_i, $i = 1, \ldots, p$.
- So we can regress X_t onto $X_{t-1}, \ldots, X_{t-p}$ to estimate λ_i.

Yule-Walker Equations

- In the AR(p) case, we saw that the forecasting coefficients

$$X_{t+1}^{t} = \phi_{t,1} X_t + \cdots + \phi_{t,t} X_1$$

  correspond exactly to the model parameters λ_i, $i = 1, \ldots, p$.
- So we can regress X_t onto $X_{t-1}, \ldots, X_{t-p}$ to estimate λ_i.
- These are the so-called *Yule-Walker* equations.

Yule-Walker Equations

If $\{X_t\}$ is a casual AR(p) model $P(B)X_t = W_t$, it results that

$$\mathbf{E}\left(X_{t-i}\left(X_t - \sum_{j=1}^{p}\lambda_j X_{t-j}\right)\right) = \mathbf{E}\left(X_{t-i}W_t\right)\ ,\ (i = 0, \ldots, p) \Leftrightarrow$$

$$\boxed{R_X(0) - \lambda^T R_p = \sigma^2 \text{ , and } R_p = \Gamma_p \lambda}\ .$$

These are the same as the forecasting equations.

Yule-Walker Equations

Method of moments: Express moments in terms of parameters, and then substitute empirical moments.

Yule-Walker Equations

Method of moments: Express moments in terms of parameters, and then substitute empirical moments. In our setting, we use covariances:

Definition

The Yule-Walker equations for $\hat{\lambda}$ are

$$\hat{\lambda}^T \hat{R}_p = \hat{R_X}(0) - \hat{\sigma}^2 \text{ , and } \hat{\Gamma}_p \hat{\lambda} = \hat{R}_p \ .$$

- To solve this system efficiently , we can use the Durbin-Levinson algorithm.

State-Space Models: Motivation

Suppose we want to determine the precise location of a car over time. Two sources of measurement available:

- A GPS unit: provides estimates of the position within a few meters precision. Noisy estimate, but no global drift.
- Wheel revolutions and angle of the steering wheel. Smooth estimate, but has drift as errors accumulate.

Q: How to combine these measurements together? Can we improve our estimate by combining them?

Dynamic Linear Model

We can model a dynamic system using a vector of internal states $\mathbf{x}_t \in \mathbb{R}^p$ that is updated linearly:

$$\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \mathbf{w}_t \ , \ \text{where}$$

- $\Phi \in \mathbb{R}^{p \times p}$ is a state transition matrix,
- $\mathbf{w}_t$ is a Normal vector with zero mean and covariance Q.
- The initial condition $\mathbf{x}_0$ is modeled as Normal (μ_0, Σ_0) (thus $\mathbf{x}_t$ is Normal for all t).

Dynamic Linear Model

However, in general we do not observe $\mathbf{x}_t$ directly. Instead, we observe

$$\mathbf{y}_t = A_t \mathbf{x}_t + \mathbf{v}_t \in \mathbb{R}^q \ , \ \text{where}$$

- A_t is a $q \times p$ observation matrix,
- $\mathbf{v}_t$ is Normal with zero mean and covariance R.
- $\mathbf{v}_t$ and $\mathbf{w}_t$ are uncorrelated for simplicity.

Dynamic Linear Model

We can also consider fixed input variables $\mathbf{u}_t \in \mathbb{R}^r$ (eg, in control).

Dynamic Linear Model

We can also consider fixed input variables $\mathbf{u}_t \in \mathbb{R}^r$ (eg, in control). The model becomes

$$\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \Upsilon \mathbf{u}_t + \mathbf{w}_t \ ,$$
$$\mathbf{Y}_t = A_t \mathbf{x}_t + \Gamma \mathbf{u}_t + \mathbf{v}_t \ .$$

State-space models vs ARMA models

- There exists an equivalence between state-space models and ARMA models.
- In the case of missing data, multivariate systems, deterministic inputs, it is typically easier to use state-space models.

The Kalman Filter

Q: Given the observed data $y_1, \ldots, y_s$ and a state space model, how can we estimate the unobserved state of the system $X_1, \ldots, X_t$?

The Kalman Filter

Q: Given the observed data $y_1, \ldots, y_s$ and a state space model, how can we estimate the unobserved state of the system $X_1, \ldots, X_t$?

- When $t > s$, this is a *forecasting* problem.
- When $s = t$, this is a *filtering* problem.
- When $t < s$, this is a *smoothing* problem.

The Kalman Filter

Suppose we observe $Y_s = (y_1, \ldots, y_s)$. Let us define

$$\mathbf{X}_t^s = \mathbf{E}\left(\mathbf{X}_t \mid Y_s\right) \ , \ P_t^s = \mathbf{E}\left((\mathbf{X}_t - \mathbf{X}_t^s)(\mathbf{X}_t - \mathbf{X}_t^s)^T\right) \ .$$

The Kalman Filter

Suppose we observe $Y_s = (y_1, \ldots, y_s)$. Let us define

$$\mathbf{X}_t^s = \mathbf{E}\left(\mathbf{X}_t \mid Y_s\right) \ , \ P_t^s = \mathbf{E}\left((\mathbf{X}_t - \mathbf{X}_t^s)(\mathbf{X}_t - \mathbf{X}_t^s)^T\right) \ .$$

- Recall that, in the class of linear estimators, this estimator minimizes the mean-squared error.
- Thus we can use the multivariate projection operator:

$$\mathbf{X}_t^s = P(\mathbf{X}_t \mid y_1, \ldots, y_s) \ .$$

The Kalman Filter

The estimated states are given by

$$\mathbf{X}_t^{t-1} = \Phi \mathbf{X}_{t-1}^{t-1} + \Upsilon \mathbf{U}_t \ ,$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi^T + Q \ , \ \textit{with}$$

$$\mathbf{X}_t^t = \mathbf{X}_t^{t-1} + K_t (\mathbf{y}_t - A_t \mathbf{X}_t^{t-1} - \Gamma \mathbf{U}_t) \ ,$$

$$P_t^t = (I - K_t A_t) P_t^{t-1} \ , \ \textit{and}$$

$$K_t = P_t^{t-1} A_t^T (A_t P_t^{t-1} A_t^T + R)^{-1} \ .$$

The Kalman Filter: Interpretation

Assume for simplicity a state-space model with no inputs:

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{W}_t \ , \ \mathbf{Y}_t = A_t \mathbf{X}_t + \mathbf{V}_t \ ,$$

with

$$\mathbf{W}_t \sim N(0, Q) \ , \ \mathbf{V}_t \sim N(0, R) \ , \ \mathbf{X}_0 \sim N(\mu_0, \Sigma_0) \ .$$

The Kalman Filter: Interpretation

Assume for simplicity a state-space model with no inputs:

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{W}_t \ , \ \mathbf{Y}_t = A_t \mathbf{X}_t + \mathbf{V}_t \ ,$$

with

$$\mathbf{W}_t \sim N(0, Q) \ , \ \mathbf{V}_t \sim N(0, R) \ , \ \mathbf{X}_0 \sim N(\mu_0, \Sigma_0) \ .$$

- The state density can be described as
$$p(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \dots, \mathbf{X}_0) = p(\mathbf{X}_t \mid \mathbf{X}_{t-1}) = f_w(\mathbf{X}_t - \Phi \mathbf{X}_{t-1}) \ ,$$
where f_w is a p-dim Normal distribution with zero mean and covariance Q.
- Similarly, the relationship between observations and states is
$$p(\mathbf{y}_t \mid \mathbf{X}_t, Y_{t-1}) = f_v(\mathbf{y}_t - A_t \mathbf{X}_t) \ ,$$
where f_v is a q-dim Normal distribution with zero mean and covariance R.

The Kalman Filter: Interpretation

These conditional densities, together with $f_0(\mathbf{X}_0)$, completely specify the model:

$$\boxed{p(\mathbf{X}_0, \ldots, \mathbf{X}_t, \mathbf{y}_1, \ldots, \mathbf{y}_t) = f_0(\mathbf{X}_0) \prod_{j \le t} f_w(\mathbf{X}_j - \Phi \mathbf{X}_{j-1}) f_v(\mathbf{y}_j - A_j \mathbf{X}_j)} .$$

- Given the current filter density $p(\mathbf{X}_{t-1} \mid Y_{t-1})$, we generate a Gaussian forecast density $p(x_t \mid Y_{t-1})$.
- Given a new observation $\mathbf{y}_t$, we update the filter density $p(\mathbf{X}_t \mid Y_t)$.

Kalman Filter and Forecast

- The Kalman forecasting $\mathbf{X}_t^{t-1}$ has an associated error P_t^{t-1} larger than the Kalman filter $\mathbf{X}_t^t$:

$$Tr(P_t^{t-1}) \geq Tr(P_t^t) \ .$$

- We can also consider the Kalman *smoother*, which predicts $\mathbf{X}_t$ using past, present and future values of $\mathbf{y}_t$.

Non-linear state space models: Motivation

- So far, we have considered linear dynamic models consisting of Gaussian processes.
- They have the advantage of being simple to estimate and analyze, however they have limitations.

Non-linear state space models: Motivation

- So far, we have considered linear dynamic models consisting of Gaussian processes.
- They have the advantage of being simple to estimate and analyze, however they have limitations.
- Non-linear dynamics: systems with variable memory, hysteresis, etc.

Non-linear state space models: Motivation

- So far, we have considered linear dynamic models consisting of Gaussian processes.
- They have the advantage of being simple to estimate and analyze, however they have limitations.
- Non-linear dynamics: systems with variable memory, hysteresis, etc.
- Discrete states and/or discrete space of observations.

Hidden Markov Model (HMM)

Consider the following example: Victor goes to work every day. Let y_t his arrival time at day t. Victor has a car, but does not use it every day, in which case he comes by bike. We would like to find a good model for y_t.

Hidden Markov Model (HMM)

Consider the following example: Victor goes to work every day. Let y_t his arrival time at day t. Victor has a car, but does not use it every day, in which case he comes by bike. We would like to find a good model for y_t.

In that case, the state x_t is binary: `bike`, `car` . Moreover, x_t is not independent from the past (eg, if it is cold, the next day is likely to be cold as well).

Hidden Markov Model (HMM)

Consider the following example: Victor goes to work every day. Let y_t his arrival time at day t. Victor has a car, but does not use it every day, in which case he comes by bike. We would like to find a good model for y_t.

In that case, the state x_t is binary: `bike`, `car` . Moreover, x_t is not independent from the past (eg, if it is cold, the next day is likely to be cold as well).

Hidden Markov Model (HMM)

We can model x_t using a *Markov process*:

$$\begin{aligned} p(x_1, \dots, x_t) &= p(x_1)p(x_2|x_1)\dots p(x_t|x_1, \dots, x_{t-1}) \\ &= p(x_1) \prod_{1<i\leq t} p(x_i \mid x_{i-1}) \end{aligned}$$

If x_t is discrete, $x_t = m_k$, $k = 1 \dots L$. The transition probabilities are expressed with a probability matrix

$$p(x_i = m_k \mid x_{i-1} = m_l) = \pi_{k,l} \ .$$

Hidden Markov Models

The data likelihood can be written in terms of state transition probabilities:

$$
\begin{aligned}
p(y_1, \ldots, y_t) &= \prod_i p(y_i \mid Y_{i-1}) \\
&= \prod_i \sum_k p(y_i \mid Y_{i-1}, x_i = m_k) p(x_i = m_k \mid Y_{i-1}) \\
&= \prod_i \sum_{k,l} p(y_i \mid Y_{i-1}, x_i = m_k) \pi_{k,l} p(x_{i-1} = m_l \mid Y_{i-1}) \, .
\end{aligned}
$$

- The parameters of the model can be learnt using the EM algorithm (Expectation-Maximization).

HMMs

- We can incorporate a markov chain also in the parameters of a Dynamic Linear Model (example 6.17 from the book).
- HMM have been very successful in speech processing, handwritten digit recognition among others.
- However, optimization and inference are hard.
- Examples of HMM-based speech synthesis can be found for example here: `http://homepages.inf.ed.ac.uk/jyamagis/demos/page35/page35.html`

Recurrent Neural Networks

The Dynamic Linear model considered a state update of the form

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{V}_t \ .$$

Recurrent Neural Networks

The Dynamic Linear model considered a state update of the form

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{V}_t \ .$$

- Pros: Tractable model.
- Cons: Cannot account for non-linear phenomena such as hysteresis, variable memory, etc.
- Non-linear dynamical model?

Recurrent Neural Networks

Given a sequence of random vectors $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_t)$, the joint distribution can expressed conditionally as

$$p(\mathbf{Y}) = \prod_{i \leq t} p(\mathbf{y}_i \mid \mathbf{y}_1, \dots, \mathbf{y}_{i-1}) \ .$$

- This model is not *tractable*: as t increases, the complexity grows exponentially.
- Q: How can we use state variables to break the complexity explosion?

Recurrent Neural Networks

We can introduce a *state* variable $\mathbf{x}_i$ to decouple past from future observations.

Recurrent Neural Networks

We can introduce a *state* variable $\mathbf{x}_i$ to decouple past from future observations.
The joint distribution is thus modeled as

$$p(\mathbf{Y}) = \prod_{i \leq t} p(\mathbf{y}_i \mid g_i(\mathbf{y}_1, \dots, \mathbf{y}_{i-1})) ~,$$

where

$$\mathbf{x}_i = g_i(\mathbf{y}_1, \dots, \mathbf{y}_{i-1}) = f_\Theta(\mathbf{x}_{i-1}, \mathbf{y}_{i-1})$$

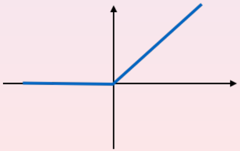
and $f_\Theta : \mathbb{R}^p \to \mathbb{R}^p$ is a generic non-linear function parametrized by Θ.

Recurrent Neural Networks

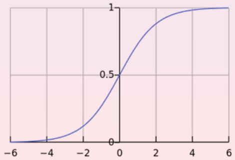
A flexible family of non-linear functions is given by Neural Networks:

$$f_\Theta(\mathbf{x}, \mathbf{y}) = \rho(A_x \mathbf{x} + A_y \mathbf{y} + b) \ , \ \text{with}$$

- $A_x \in \mathbb{R}^{p \times p}$, $A_y \in \mathbb{R}^{p \times q}$, $b \in \mathbb{R}^p$ representing an affine transformation.
- ρ is a *point-wise* non-linearity: $\rho(x_1, \dots, x_p) = (\rho(x_1), \dots, \rho(x_p))$. Some examples of non-linearities typically used in Neural networks:



Half-Rectifier



Sigmoid

Recurrent Neural Networks

Depending on the nature of the observations $\mathbf{y}_t$, we may want to choose different models for the conditional likelihood:

- On continuous $\mathbf{y}_t$, we can consider a Gaussian model:

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \mathcal{N}(\mu(\mathbf{x}_i), \Sigma(\mathbf{x}_i)) \ .$$

Recurrent Neural Networks

Depending on the nature of the observations $\mathbf{y}_t$, we may want to choose different models for the conditional likelihood:

- On continuous $\mathbf{y}_t$, we can consider a Gaussian model:

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \mathcal{N}(\mu(\mathbf{x}_i), \Sigma(\mathbf{x}_i)) \ .$$

- On discrete $\mathbf{y}_t$, we use parametrized multinomial distributions (softmax):

$$\mathbf{y_i} \sim \text{multinomial}(h(\mathbf{x_i})) \ .$$

Recurrent Neural Networks

Q: How to optimize the set of parameters?

Recurrent Neural Networks

Q: How to optimize the set of parameters?
We optimize the likelihood of the output sequence with respect to the parameters of the system.

Recurrent Neural Networks

Q: How to optimize the set of parameters?
We optimize the likelihood of the output sequence with respect to the parameters of the system.
Since there is no closed-form solution in general, we use stochastic gradient descent.

Recurrent Neural Networks

- In the previous slide, we defined an RNN to predict a given sequence from past observations (forecasting).

Recurrent Neural Networks

- In the previous slide, we defined an RNN to predict a given sequence from past observations (forecasting).
- It is easy to generalize to more general setup: Consider an input sequence $\mathbf{z}_t$ and an output sequence $\mathbf{y}_t$.
- In that case, we consider

$$p(\mathbf{Y} \mid \mathbf{Z}) = \prod_{i \leq t} p(\mathbf{y}_i \mid \mathbf{x}_i) \ ,$$

  with $\mathbf{x}_i = f_\Theta(\mathbf{x}_{i-1}, \mathbf{z}_{i-1})$.

Recurrent Neural Networks

- In the previous slide, we defined an RNN to predict a given sequence from past observations (forecasting).
- It is easy to generalize to more general setup: Consider an input sequence $\mathbf{z}_t$ and an output sequence $\mathbf{y}_t$.
- In that case, we consider

$$p(\mathbf{Y} \mid \mathbf{Z}) = \prod_{i \leq t} p(\mathbf{y}_i \mid \mathbf{x}_i) \ ,$$

  with $\mathbf{x}_i = f_\Theta(\mathbf{x}_{i-1}, \mathbf{z}_{i-1})$.
- We can also generalize RNNs by *stacking* layers of hidden state variables $\mathbf{x}_t^{(k)}$, by making the non-linear dynamic more complicated, by making it bi-directional, ...

Applications of RNNs

- One can train a large RNN to model piano music using midi files.
- Once the model is trained, we can use it to generate new music: we sample from the output distributions, and feedback the samples into the model to generate again.
- Example from Daniel Johnson: `http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/`

Applications of RNNs

- We consider input sequences in english and target sequences in french.
- An RNN is trained to maximize the output likelihood, conditioned on the input sequence.
- Several enhancements are made in the RNN architecture and the training procedure.

Applications of RNNs

Examples taken from "Neural Machine Translation by Jointly learning to align and Translate", Bahdanau et al, ICLR 2015.

Source	An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.
Reference	Le privilège d'admission est le droit d'un médecin, en vertu de son statut de membre soignant d'un hôpital, d'admettre un patient dans un hôpital ou un centre médical afin d'y délivrer un diagnostic ou un traitement.
RNNenc-50	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.
RNNsearch-50	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Google Translate	Un privilège admettre est le droit d'un médecin d'admettre un patient dans un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, fondée sur sa situation en tant que travailleur de soins de santé dans un hôpital.

Applications of RNNs

Examples taken from “Neural Machine Translation by Jointly learning to align and Translate”, Bahdanau et al, ICLR 2015.

Source	This kind of experience is part of Disney’s efforts to ”extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important,” he added.
Reference	Ce type d’expérience entre dans le cadre des efforts de Disney pour ”étendre la durée de vie de ses séries et construire de nouvelles relations avec son public grâce à des plateformes numériques qui sont de plus en plus importantes”, a-t-il ajouté.
RNNenc-50	Ce type d’expérience fait partie des initiatives du Disney pour ”prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.
RNNsearch-50	Ce genre d’expérience fait partie des efforts de Disney pour ”prolonger la durée de vie de ses séries et créer de nouvelles relations avec des publics via des plateformes numériques de plus en plus importantes”, a-t-il ajouté.
Google Translate	Ce genre d’expérience fait partie des efforts de Disney à “étendre la durée de vie de sa série et construire de nouvelles relations avec le public par le biais des plates-formes numériques qui deviennent de plus en plus important”, at-il ajouté.

Applications of RNNs

We can also use RNNs to model text as a character-based time series:

$$Y_t = \{\texttt{"I"}, \texttt{" "}, \texttt{"l"}, \texttt{"i"}, \texttt{"k"}, \texttt{"e"} \ldots\} \, .$$

Applications of RNNs

We can also use RNNs to model text as a character-based time series:

$$Y_t = \{\texttt{"I"}, \texttt{" "}, \texttt{"l"}, \texttt{"i"}, \texttt{"k"}, \texttt{"e"} \ldots\} \ .$$

State-space model that predicts an output distribution over the character set:

$$\mathbf{X}_t = f_\Theta(\mathbf{X}_{t-1}, Y_{t-1}) \ , \ \ Y_t \sim \text{multinomial}(h_\Lambda(\mathbf{X}_t)) \ .$$

Applications of RNNs

We can also use RNNs to model text as a character-based time series:

$$Y_t = \{\texttt{"I"}, \texttt{" "}, \texttt{"l"}, \texttt{"i"}, \texttt{"k"}, \texttt{"e"} \ldots\} \; .$$

State-space model that predicts an output distribution over the character set:

$$\mathbf{X}_t = f_\Theta(\mathbf{X}_{t-1}, Y_{t-1}) \; , \; Y_t \sim \text{multinomial}(h_\Lambda(\mathbf{X}_t)) \; .$$

We train the parameters of the model (Θ, Λ) by optimizing the data log-likelihood using stochastic gradient descent.

Applications of RNNs

- Examples taken from A. Karpathy' blog (`karpathy.github.io`).
- The model is trained on different text corpora.
- The model architecture includes several RNN layers.
- Text is generated by sampling from the multinomial distribution and feeding back the sample into the hidden-state equation.

Applications of RNNs

Example of generated text when training on Shakespeare texts:

```
PANDARUS: Alas, I think he shall be come approached
and the day When little srain would be attain'd
into being never fed, And who is but a chain and
subjects of his death, I should not sleep.
Second Senator:  They are away this miseries,
produced upon my soul, Breaking and strongly should
be buried, when I perish The earth and thoughts of
many states.
DUKE VINCENTIO: Well, your wit is in the care of
side and that.
Second Lord:  They would be ruled after this
chamber, and my fair nues begun out of the fact, to
be conveyed, Whose noble souls I'll have the heart
of the wars.
Clown:  Come, sir, I will make did behold your
worship.
```

Applications of RNNs

Examples when training on raw latex files containing algebraic geometry:

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. $\square$

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.