DATA ANALYTICS

Unit 3: Introduction Spectral Analysis of Time Series Analysis

Jyothi R.

Department of Computer Science
and
Engineering

Introduction

- Introduction
- Stationarity
- The Periodogram and the Spectral Density Periodogram and Regression
    - Spectral Density
- Smoothing and Tapering Smoothing
    - Tapering
- Example

Stationarity

- The assumption of stationarity imposes regularity on a time series model.
- We will need repeated observations with the same or similar relationship to one another in order to estimate the underlying relationships between observations.
- There are other ways to do this, but stationarity is the most common and perhaps most basic.

Strictly Stationary

A time series ..., x_{-1}, x_0, x_1, x_2, ... is *strictly stationary* if for a sequence of times t_1, t_2, ..., t_k

$$\{x_{t_1}, ..., x_{t_k}\}$$

has the same distributions as

$$\{x_{t_1+h}, ..., x_{t_k+h}\}$$

for every integer h. In other words,

$$P\{x_{t_1} \le c_1, ..., x_{t_k} \le c_k\} = P\{x_{t_1+h} \le c_1, ..., x_{t_k+h} \le c_k\}.$$

Weak Stationarity

- An important measure of dependency in time series is autocovariances.
- This is defined as
    - $\gamma(t, s) = E\,(x_t - \mu_t)(x_s - \mu_s)$
- where $\mu_t = Ex_t$.
- The time series x_t is *weakly stationary* if μ_t is constant and $\gamma(s, t)$ depends only on the distance $|s - t|$.
- In the case of Gaussian time series, these two concepts of stationarity overlap.

Autocovariances Notation

- For a weakly stationary time series, the notation used for auto covariance uses only lag:
- $\gamma(h) = E\,(x_t - \mu)(x_{t-h} - \mu)$ where μ is the constant variance.
- We also have a concept of the autocorrelation function which we saw in the first section in the ACF plot. The autocorrelation function is defined as

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}$$

Discrete Fourier Transform of the Time Series

- What if we are less interested in how our underlying process evolves in time and are more interested in the variance of the time series at certain frequencies?

- We may attempt to apply a Fourier transform to the data. For our time series, $x_1, ..., x_n$, the discrete Fourier transform would be

- where $\omega_j = 0, 1/n, ..., (n - 1)/n$.

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t exp(-2\pi i t\omega_j)$$

An Alternate Representation

- Note that we can break up $d\,(\omega_j)$ into two parts

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t cos(2\pi i\omega_j t) - in^{-1/2} \sum_{t=1}^{n} x_t sin(2\pi i\omega_j t)$$

- which we could write as a cosine component and a sine component

$$d\,(\omega_j) = d_c\,(\omega_j) - id_s\,(\omega_j)$$

An Alternate Representation

- We may use an inverse Fourier transform to rewrite the data as

$$
\begin{aligned}
x_t &= n^{-1/2} \sum_{j=1}^{n} d(\omega_j) e^{2\pi i \omega_j t} \\
&= n^{-1/2} \sum_{j=1}^{n} d(\omega_j) e^{2\pi i \omega_j t} \\
&= a_0 + n^{-1/2} \sum_{j=1}^{m} d(\omega_j) e^{2\pi i \omega_j t} + n^{-1/2} \sum_{j=m+1}^{n} d(\omega_j) e^{2\pi i \omega_j t} \\
&= a_0 + \sum_{j=1}^{m} \frac{2d_c(\omega_j)}{n^{-1/2}} \cos(2\pi i \omega_j t) + \sum_{j=1}^{m} \frac{2d_s(\omega_j)}{n^{-1/2}} \sin(2\pi i \omega_j t)
\end{aligned}
$$

$$\text{where } m = \left\lfloor \frac{n}{2} \right\rfloor$$

The Periodogram

- The Periodogram is defined as

$$I(\omega_j) = |d(\omega_j)|^2 = d_c^2(\omega_j) + d_c^2(\omega_j)$$

- If there is no periodic trend in the data, then $Ed(\omega_j) = 0$, and the Periodogram expresses the variance of x_t at frequency ω_j.
- If a periodic trend exists in the data, then $Ed(\omega_j)$ will be the contribution to the periodic trend at the frequency ω_j.

The Periodogram

- What are we trying to estimate with the Periodogram?
- We can use the Periodogram to find periodic trends in the data.
- Is there information left in the Periodogram after the trend is removed?
- Assuming that we have a stationary time series, what does the Periodogram estimate?

The Spectral Density

- The spectral density is the Fourier transform of the auto covariance function

$$f(\omega) = \sum_{h=-\infty}^{h=\infty} e^{-2\pi i \omega h} \gamma(h)$$

- for $\omega \in (-0.5, 0.5)$. Note that this is a population quantity. (i.e. This is a constant quantity defined by the model.)

The Spectral Density

- Why is the Periodogram an estimate for the spectral density?
- Let *m* be the sample mean of our data.

$$I(\omega_j) = |d(\omega_j)|^2 = n^{-1} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} \overline{\sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t}}$$

$$= |d(\omega_j)|^2 = n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} (x_t - m)(x_s - m) e^{-2\pi i \omega_j (t-s)}$$

$$= n^{-1} \sum_{h=-(n-1)}^{(n-1)} \sum_{t=1}^{n-|h|} (x_{t+|h|} - m)(x_t - m) e^{-2\pi i \omega_j (h)}$$

$$= \sum_{h=-(n-1)}^{(n-1)} \hat{\gamma}(h) e^{-2\pi i \omega_j (h)} \approx f(\omega_j)$$

The Spectral Density

-) Is the Periodogram a **good** estimator for the spectral density?
  - Not really!
-) The Periodogram, $I\ (\omega_1), ..., I\ (\omega_m)$, attempt to estimate parameters $f\ (\omega_1), ..., f\ (\omega_m)$. We have nearly the same number of parameters as we have data.
-) Moreover, the number of parameters grow as a constant proportion of the data. Therefore, the Periodogram is NOT a consistent estimator of the spectral density.

Moving Average

- A simple way to improve our estimates is to use a moving average smoothing technique

$$\hat{f}(\omega_j) = \frac{1}{2m+1} \sum_{k=-m}^{m} I(\omega_{j-k})$$

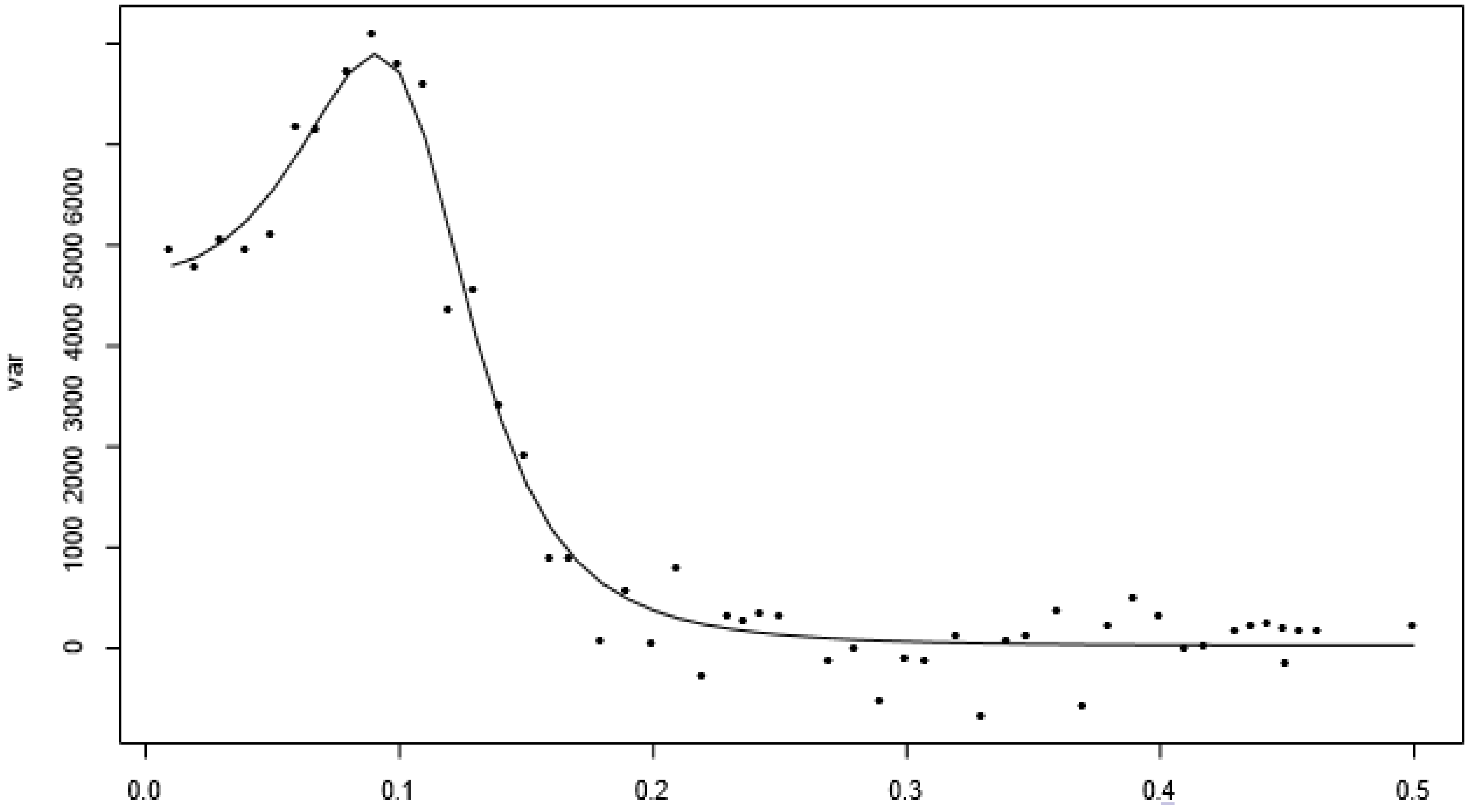
- We can also iterate this procedure of uniform weighting to be more weight on closer observations.

$$\hat{u}_t = \frac{1}{3} u_{t-1} + \frac{1}{3} u_t + \frac{1}{3} u_{t+1}$$

$$\hat{\hat{u}}_t = \frac{1}{3} \hat{u}_{t-1} + \frac{1}{3} \hat{u}_t + \frac{1}{3} \hat{u}_{t+1}$$

- Then, we iterate.
- Then, substitute to obtain better weights.

Moving Average



Smoothing Summary

- Smoothing decreases variance by averaging over the Periodogram of neighboring frequencies.
- Smoothing introduces bias because the expectation of neighboring Periodogram values are similar but not identical to the frequency of interest.
- Beware of over smoothing!

Tapering

- Tapering corrects bias introduced from the finiteness of the data.

- The expected value of the Periodogram at a certain frequency is not quite equal to the spectral density.

- It is affected by the spectral density at neighboring frequency points.

- For a spectral density which is more dynamic, more tapering is required.

Why do we need to taper?

- Our theoretical model ..., x_{-1}, x_0, x_1, ... consists of a doubly infinite time series
- We could think of our data, y_t as the following transformation of the model
- $y_t = h_t x_t$
- where h_t = 1 for t = 1, ..., n and zero otherwise. This has repercussions on the expectation of the Periodogram of our data.

$$E[I_y(\omega_j)] = \int_{-0.5}^{0.5} W_n(\omega_j - \omega) f_x(\omega) d\omega$$

- where $W_n(\omega) = |H_n(\omega)|^2$ and $H_n(\omega)$ is the Fourier transform of the sequence h_t.

The Taper

Specifically,

$$H_n(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} h_t e^{-2\pi i \omega t}$$

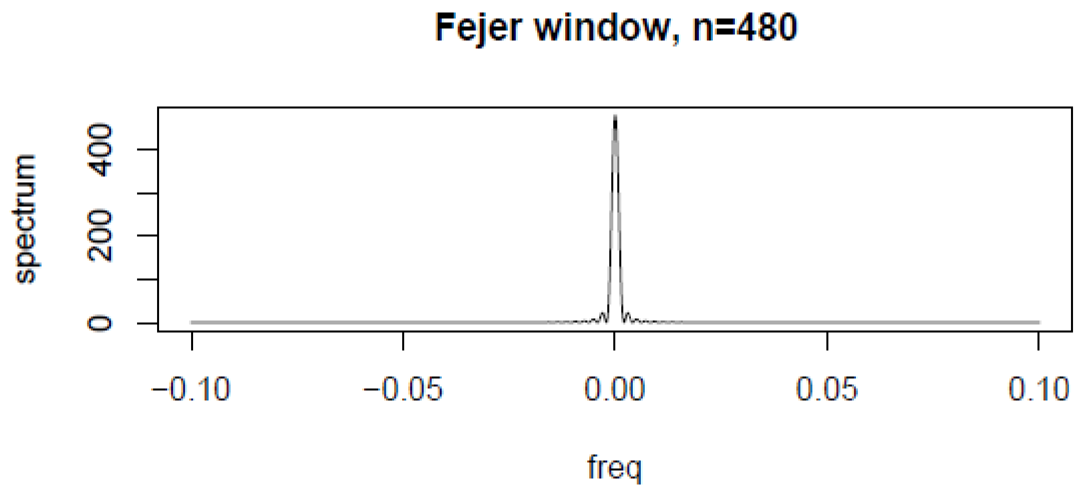
When we put in the h_t above, we obtain a spectral window of

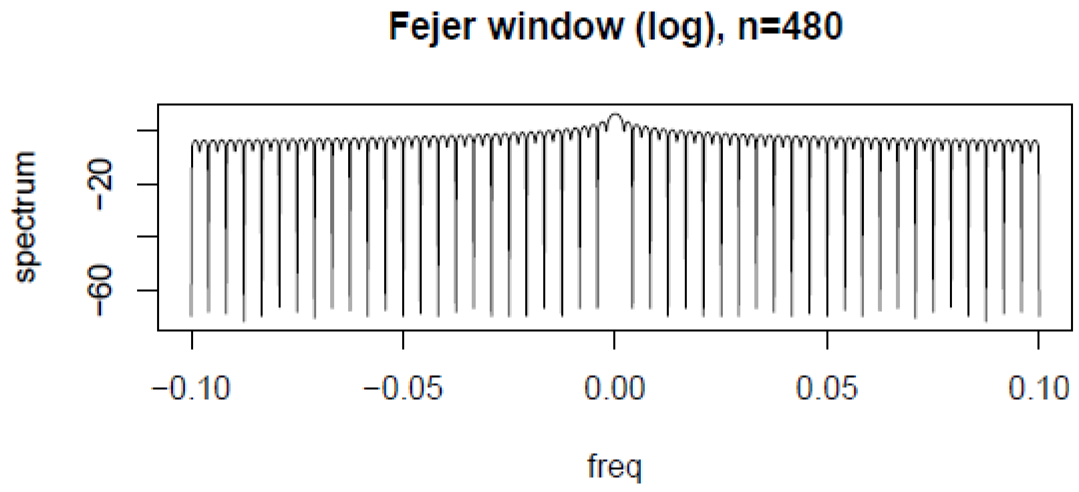
$$W_n(\omega) = \frac{\sin^2(n2\pi\omega)}{\sin^2(\pi\omega)}.$$

We set $W_n(0) = n$.

Smoothing and Tapering

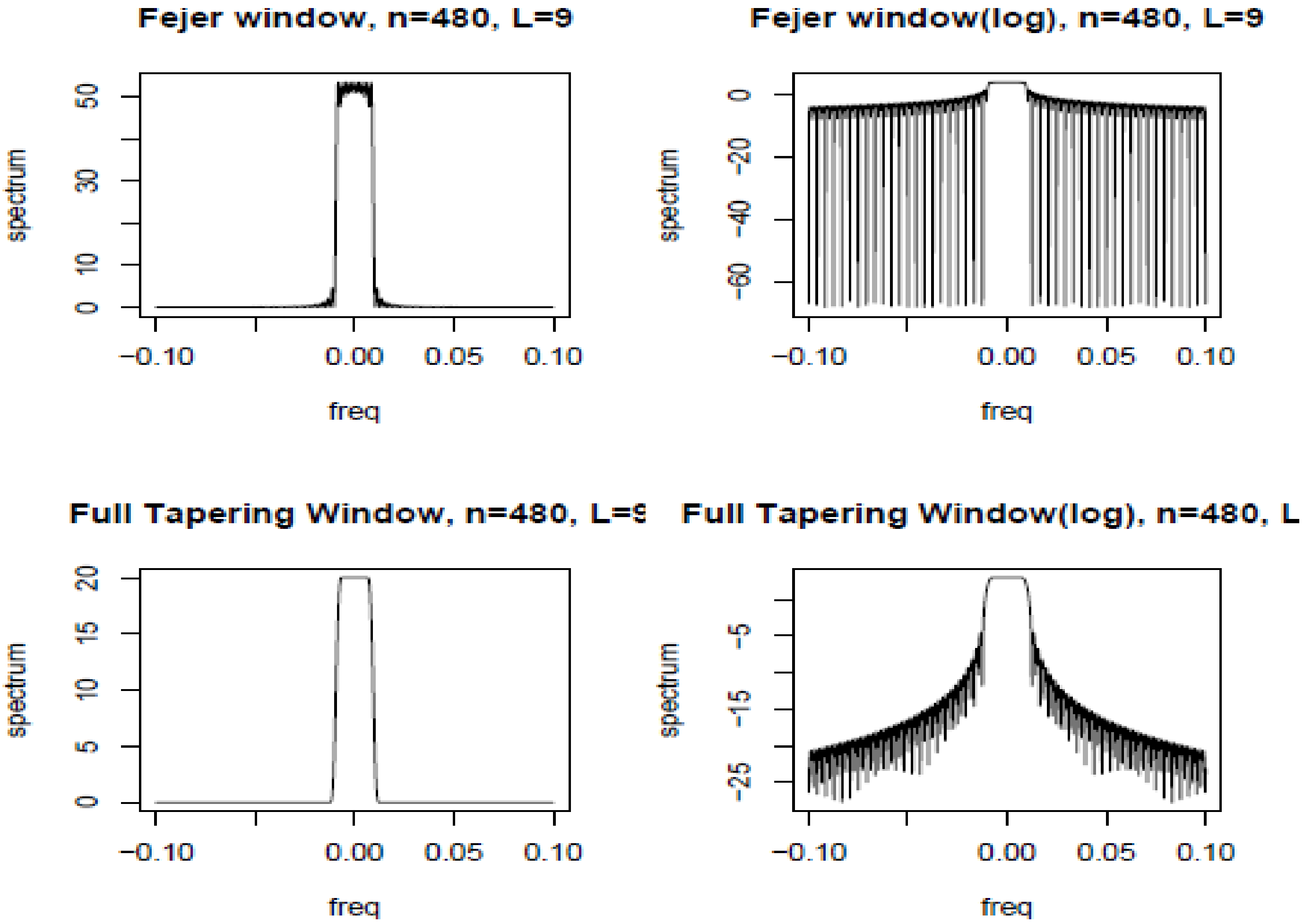
- There are problems with this spectral window, namely there is too much weight on neighboring frequencies (sidelobes).


Fejer window, n=480


Fejer window (log), n=480

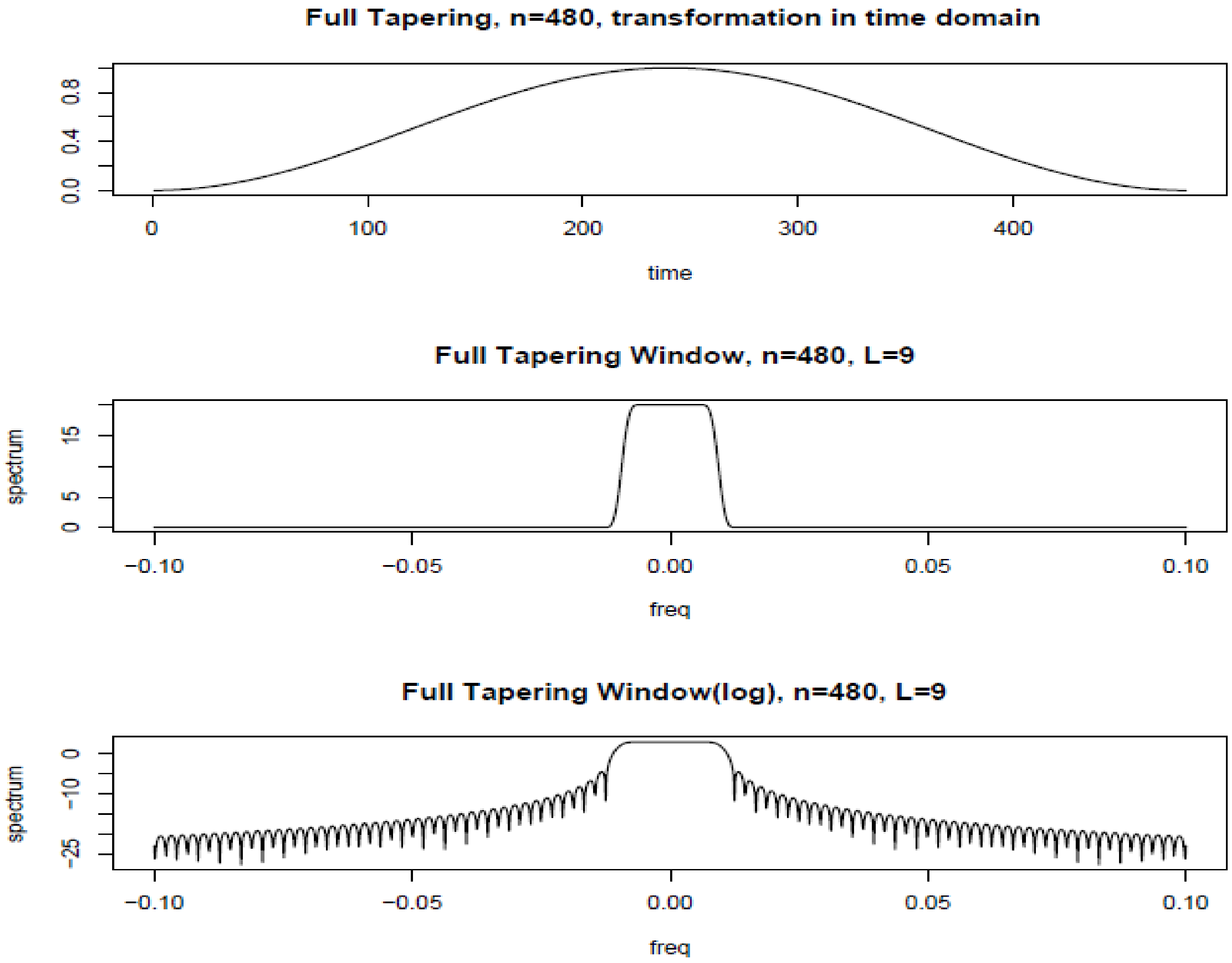
Smoothing and Tapering

One way to fix this is to use a Cosine taper. We select a transform h_t to be

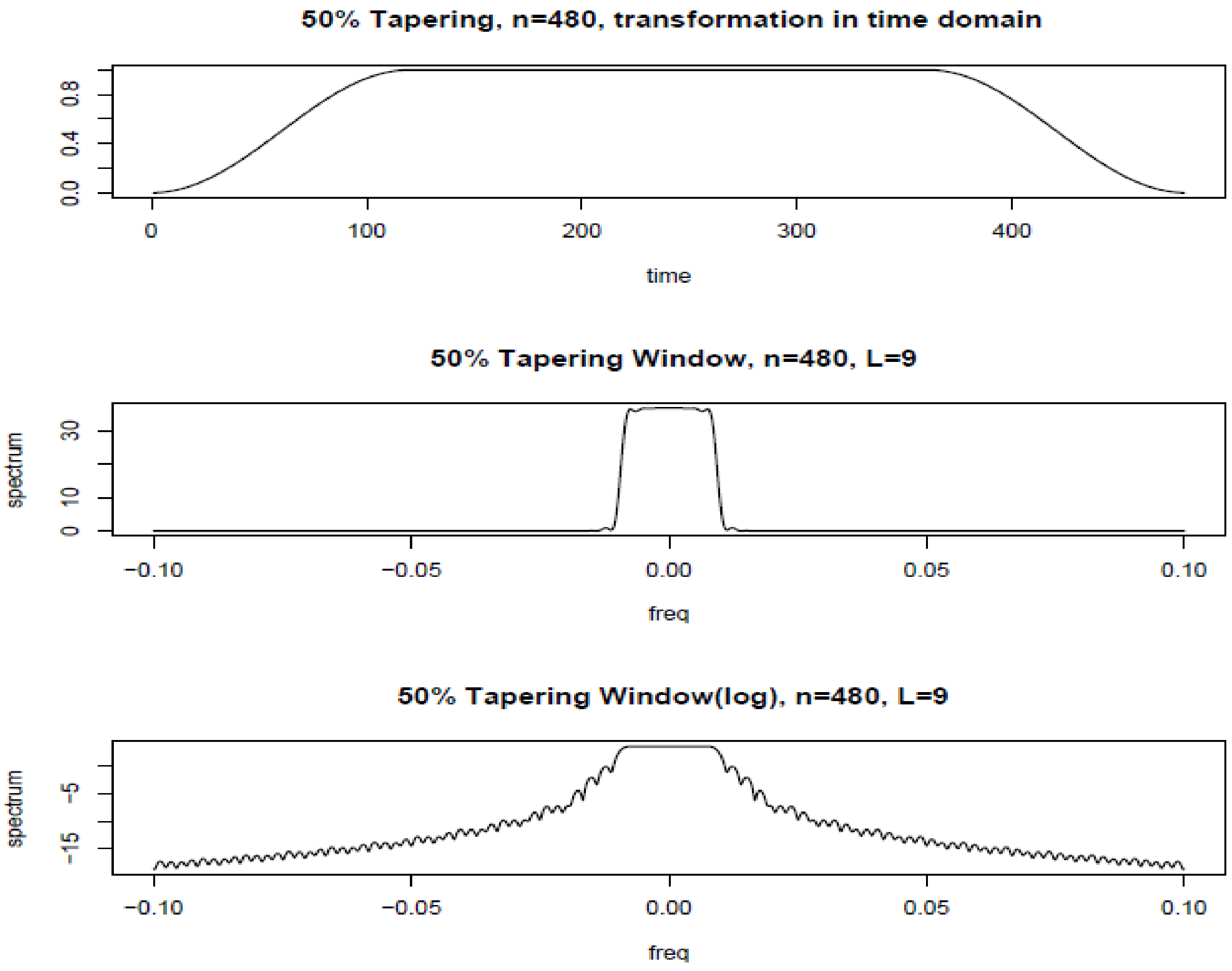


$$h_t = 0.5\left[1 + cos\left(\frac{2\pi(t - \bar{t})}{n}\right)\right]$$

Smoothing and Tapering



Smoothing and Tapering



50% Tapering, n=480, transformation in time domain

50% Tapering Window, n=480, L=9

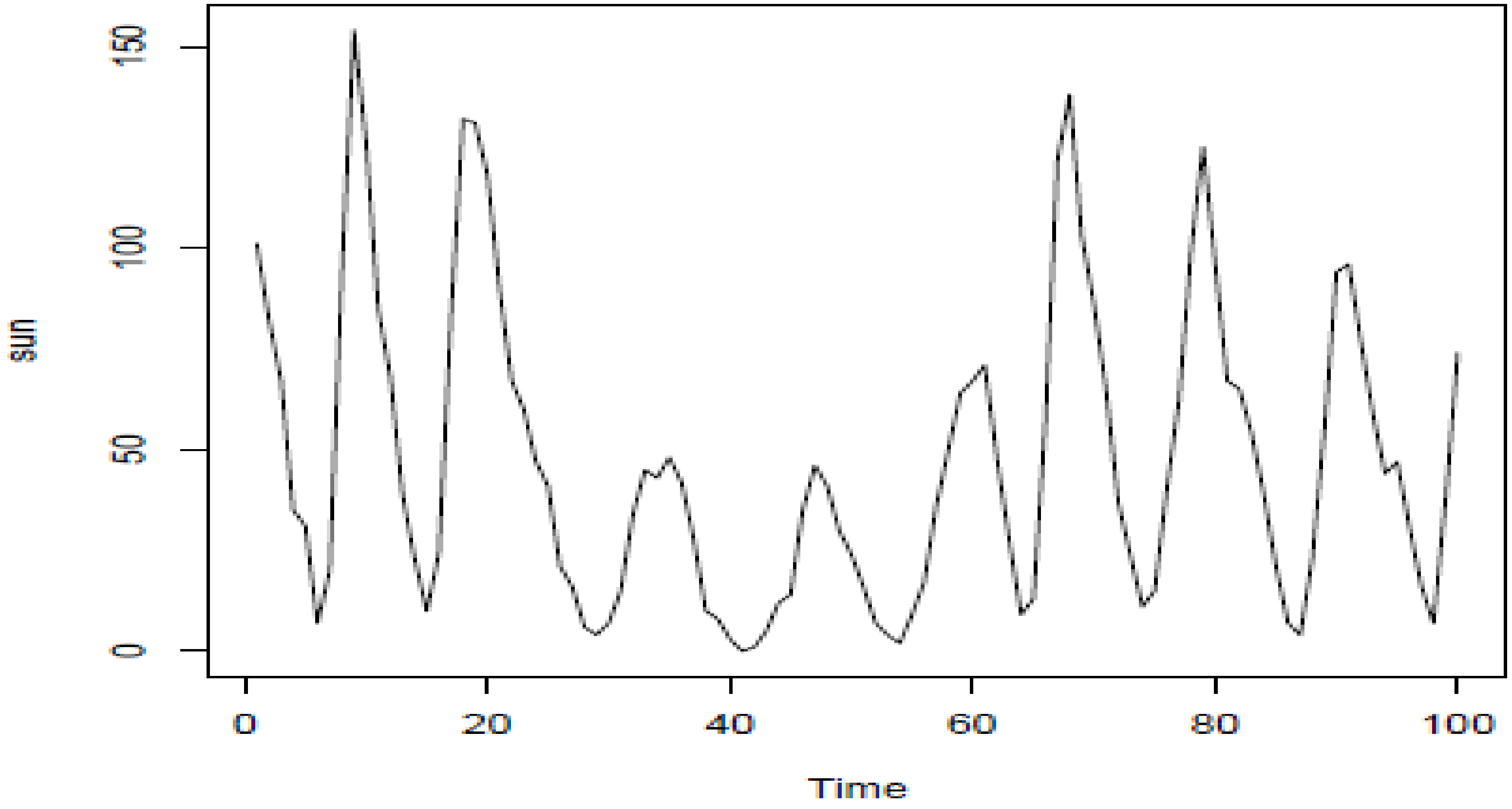
50% Tapering Window(log), n=480, L=9

Smoothing and Tapering

- Smoothing introduces bias, but reduces variance.
- Smoothing tries to solve the problem of too many “parameters”.
- Tapering decreases bias and introduces variance.
- Tapering attempts to diminish the influence of sidelobes that are introduced via the spectral window.

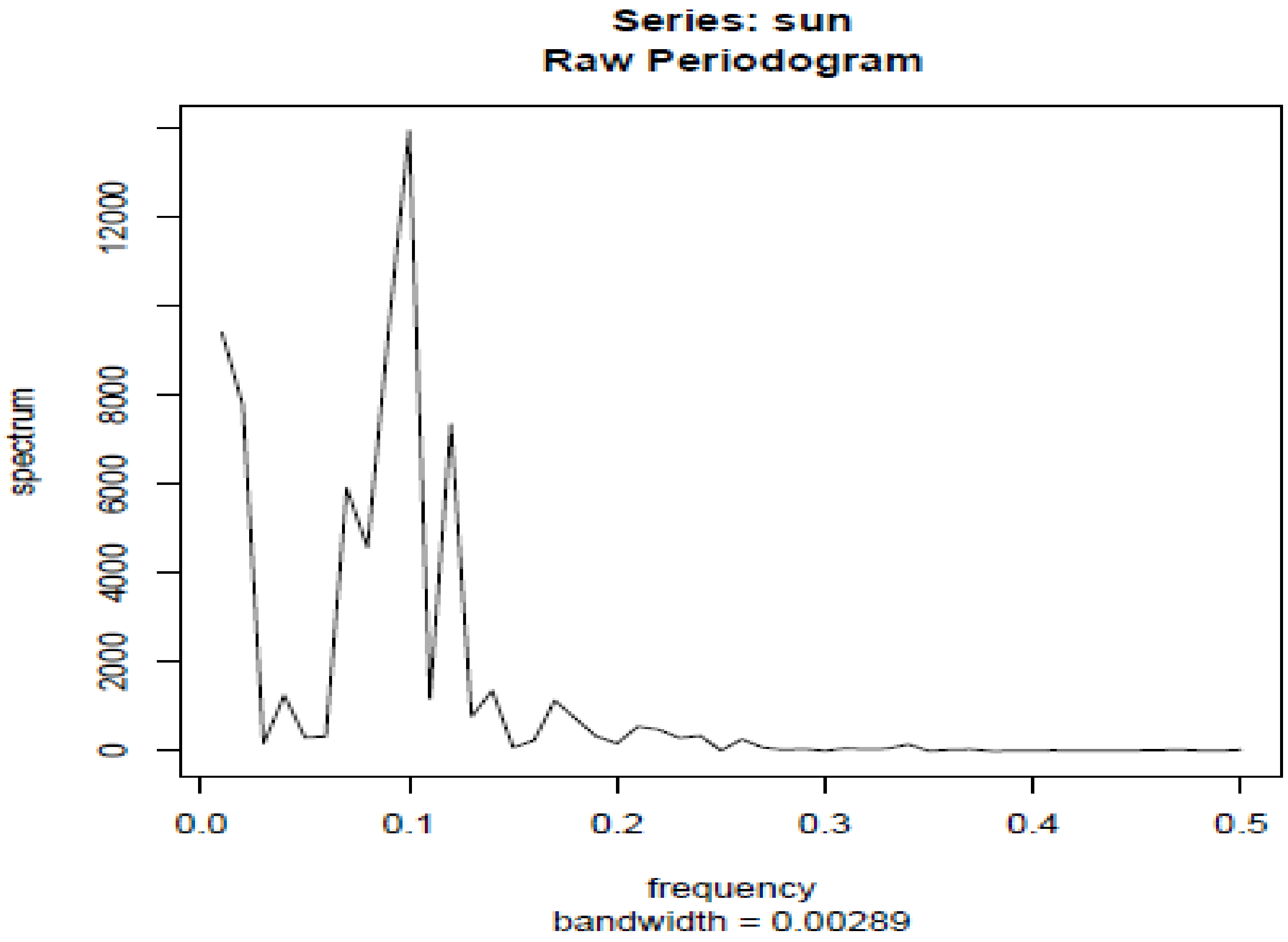
Examples

Wolfer sunspots 1770-1869



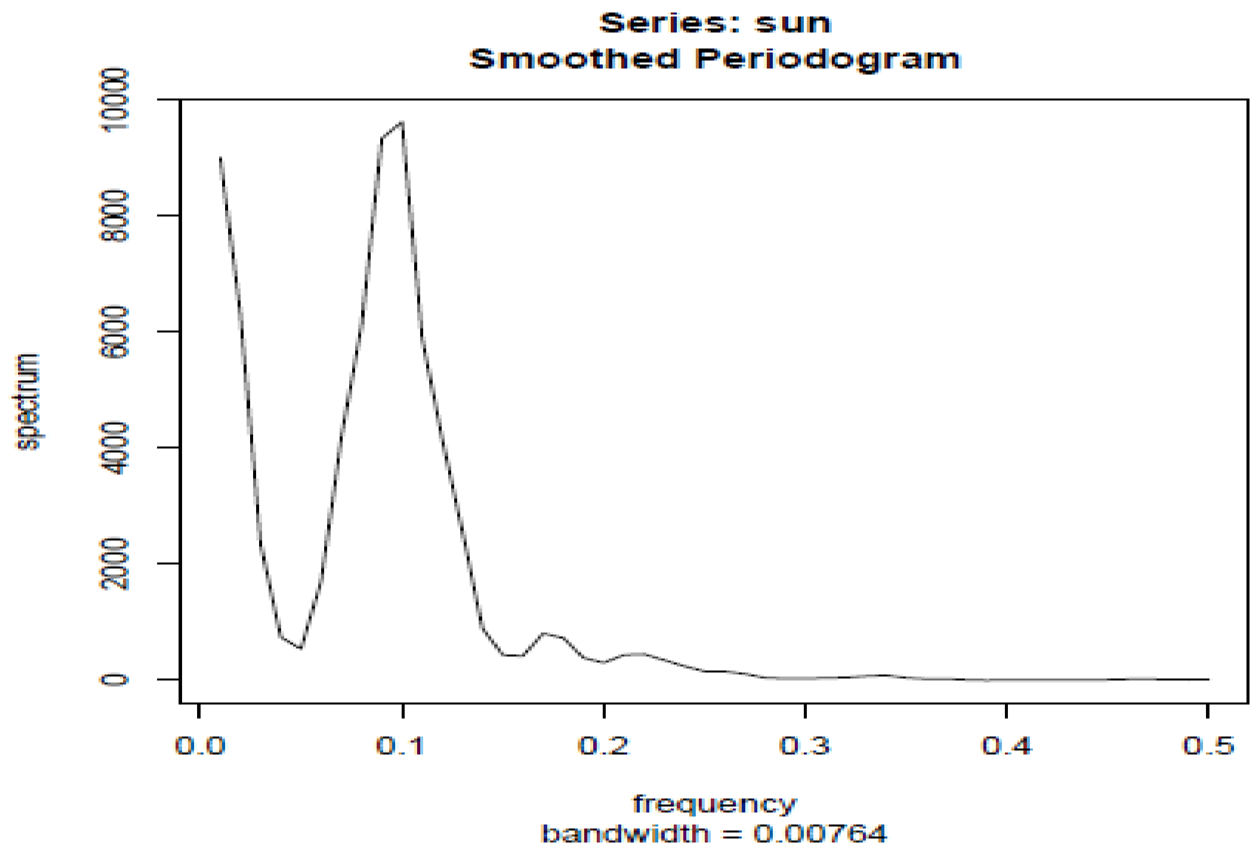
Examples

Raw Periodogram



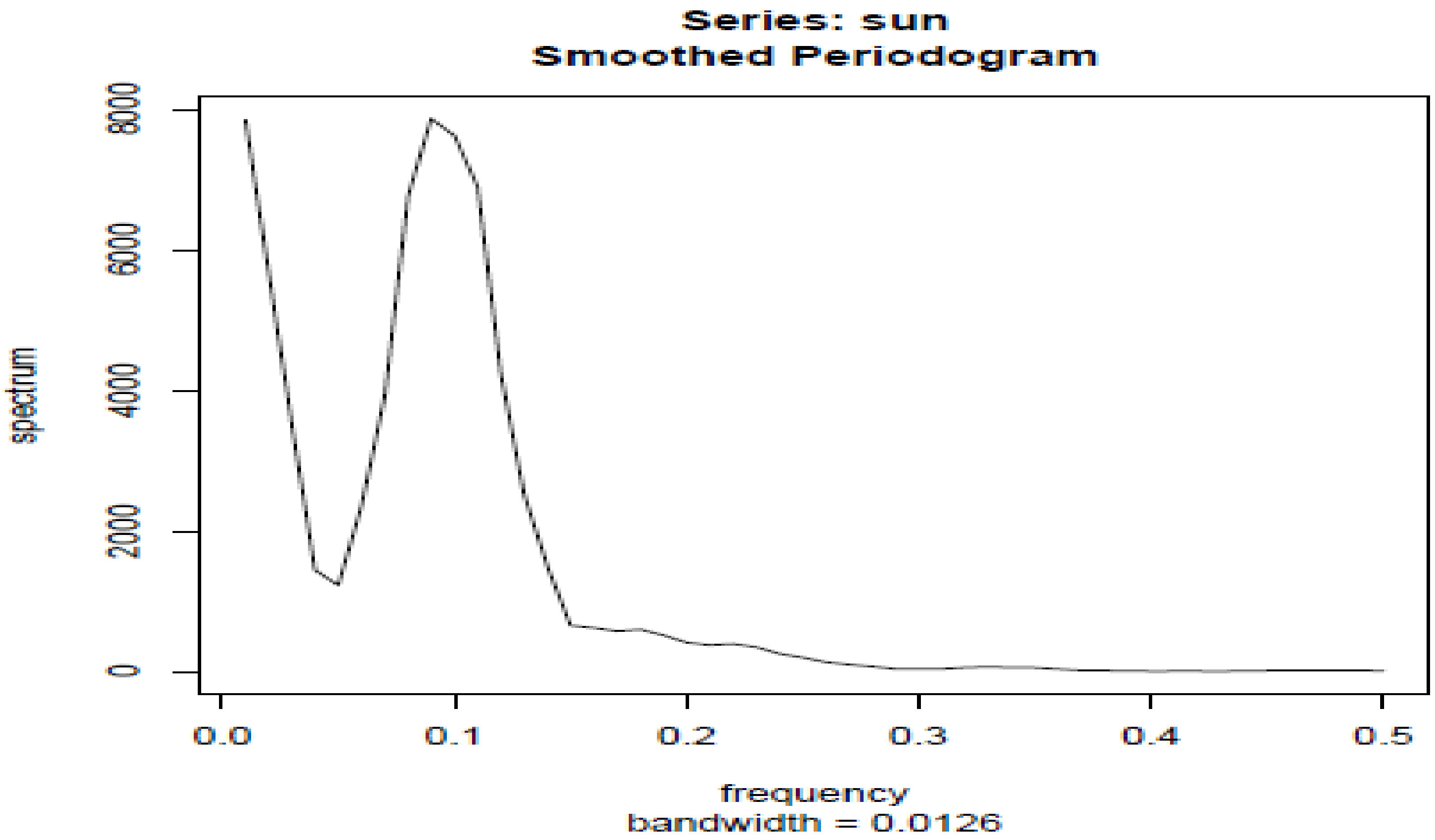
Examples

Periodogram with Smoothing Window of 3



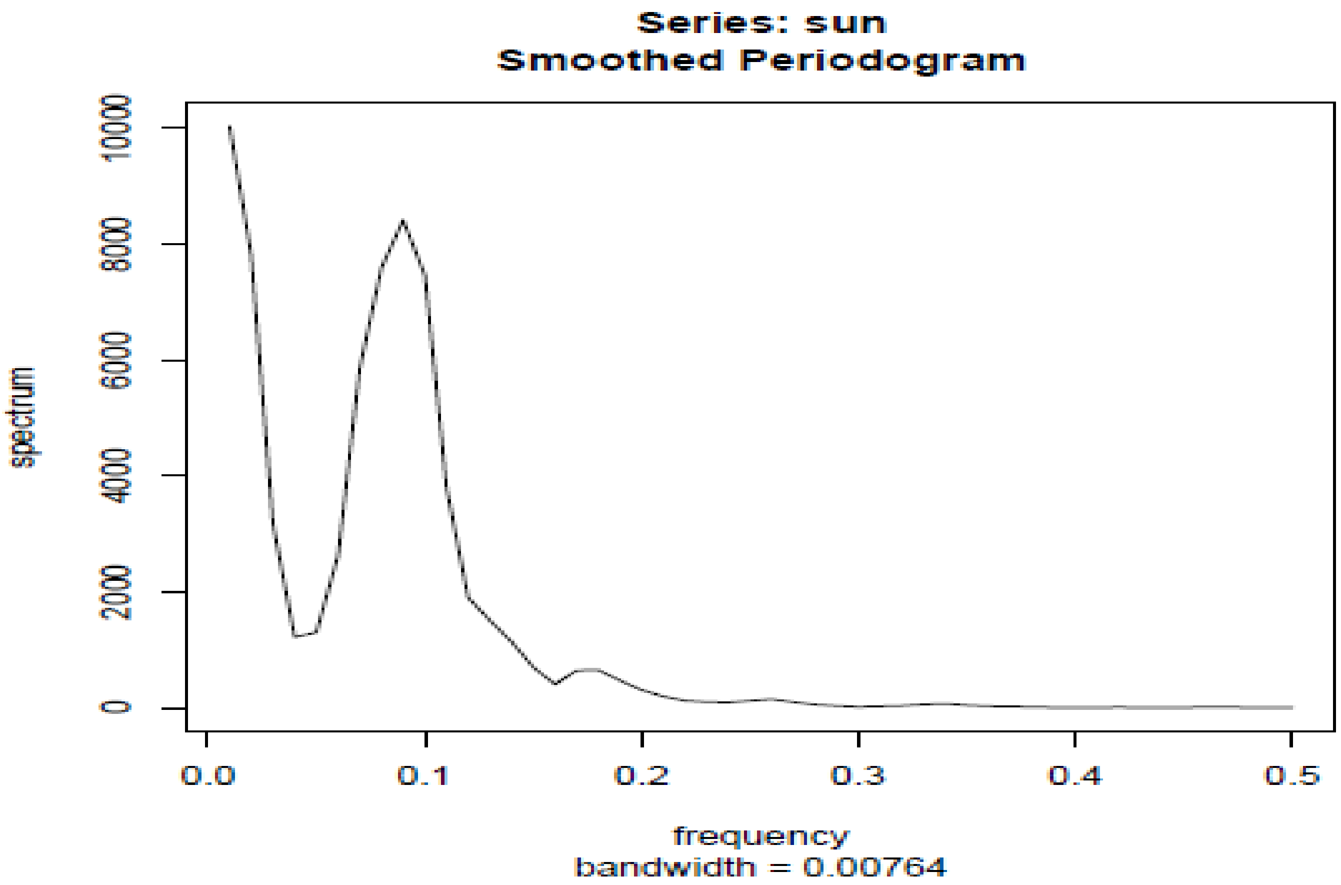
Examples

Periodogram with Smoothing Window of 5



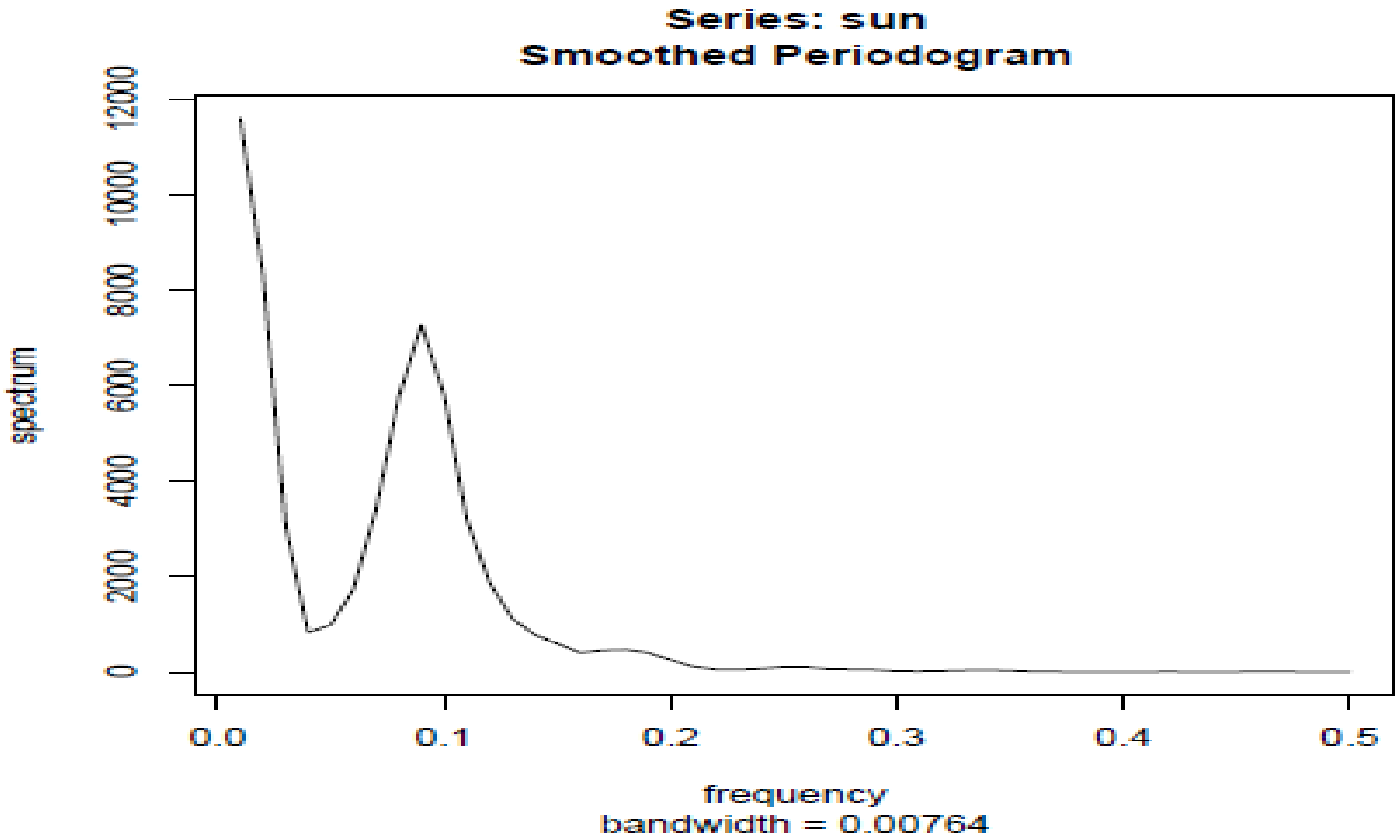
Examples

Periodogram with Smoothing Window of 3 with Some Tapering



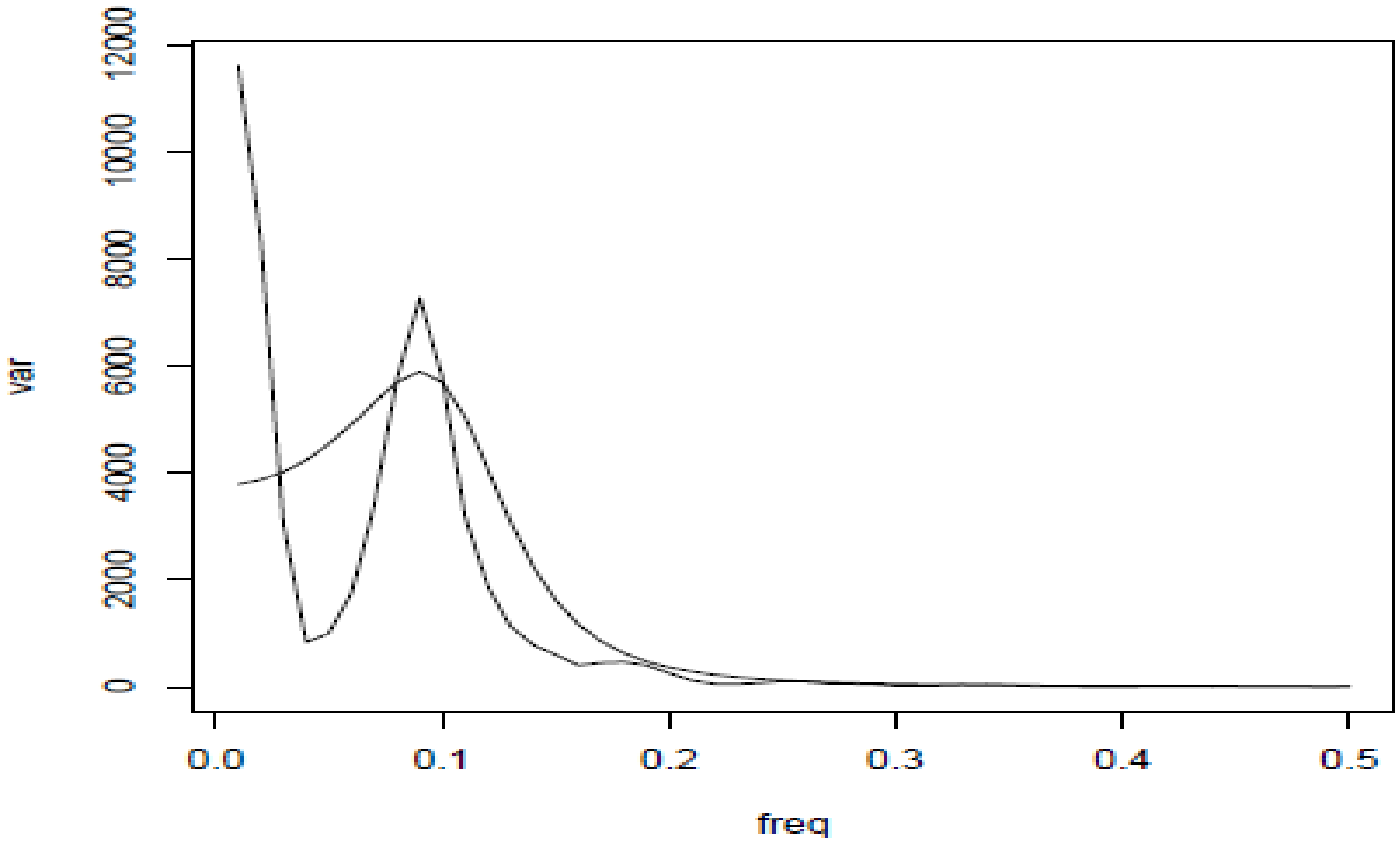
Examples

Periodogram with Smoothing Window of 3 with More Tapering



Smoothed Periodogram with ARMA Spectral Density

The smoothed Periodogram of the sun spot data with the spectral density of the AR(3) model overlayed



Dynamic Fourier Analysis

- What can be done for non-stationary data?
- One approach is to decompose our time series as a sum of a non-constant (deterministic) trend plus a stationary “noise” term:
    - $x_t = \mu_t + y_t$
- What if our data instead appears as a stationary model locally, but globally the model appears to shift? One approach is to divide the data into shorter sections (perhaps overlapping) and
- This approach is developed in Shumway and Stoffer. One essentially looks at how the spectral density changes over time.

Wavelets

- We have been using Fourier components as a basis to represent stationary processes and seasonal trends.
- Since we are dealing with finite data, we must use a finite number of terms, and perhaps one could use an alternative basis.
- Wavelets are one option to accomplish this goal. They are particularly well suited to the same situation as Dynamic Fourier analysis.

Introduction to Spectral Analysis

Spectral Analysis

1. Spectral density: Facts and examples.
2. Spectral distribution function.
3. Wold’s decomposition.

A periodic time series

- Consider
  - $X_t = A\sin(2\pi\nu t) + B\cos(2\pi\nu t)$
  - $= C\sin(2\pi\nu t + \varphi)$,
- where $A,\ B$ are uncorrelated, mean zero, variance $\sigma^2 = 1$, and
- $C^2 = A^2 + B^2$, $\tan\varphi = B/A$. Then
  - $\mu_t = \mathrm{E}[X_t] = 0$
  - $\gamma(t, t+h) = \cos(2\pi\nu h)$.
- So $\{X_t\}$ is stationary.

An aside: Some trigonometric identities

$$\tan\theta = \frac{\sin\theta}{\cos\theta}$$

$$\sin^2\theta + \cos^2\theta = 1$$

$$\sin(a+b) = \sin a\cos b + \cos a\sin b,$$

$$\cos(a+b) = \cos a\cos b - \sin a\sin b.$$

A periodic time series

- For $X_t = A \sin(2\pi vt) + B \cos(2\pi vt)$, with uncorrelated A, B
- (mean 0, variance σ^2), $\gamma(h) = \sigma^2 \cos(2\pi vh)$.
- The auto covariance of the sum of two uncorrelated time series is the sum of their auto covariances. Thus, the auto covariance of a sum of random sinusoids is a sum of sinusoids with the corresponding frequencies:

$$X_t = \sum_{j=1}^{k} (A_j \sin(2\pi v_j t) + B_j \cos(2\pi v_j t)),$$

$$\gamma(h) = \sum_{j=1}^{k} \sigma_j^2 \cos(2\pi v_j h),$$

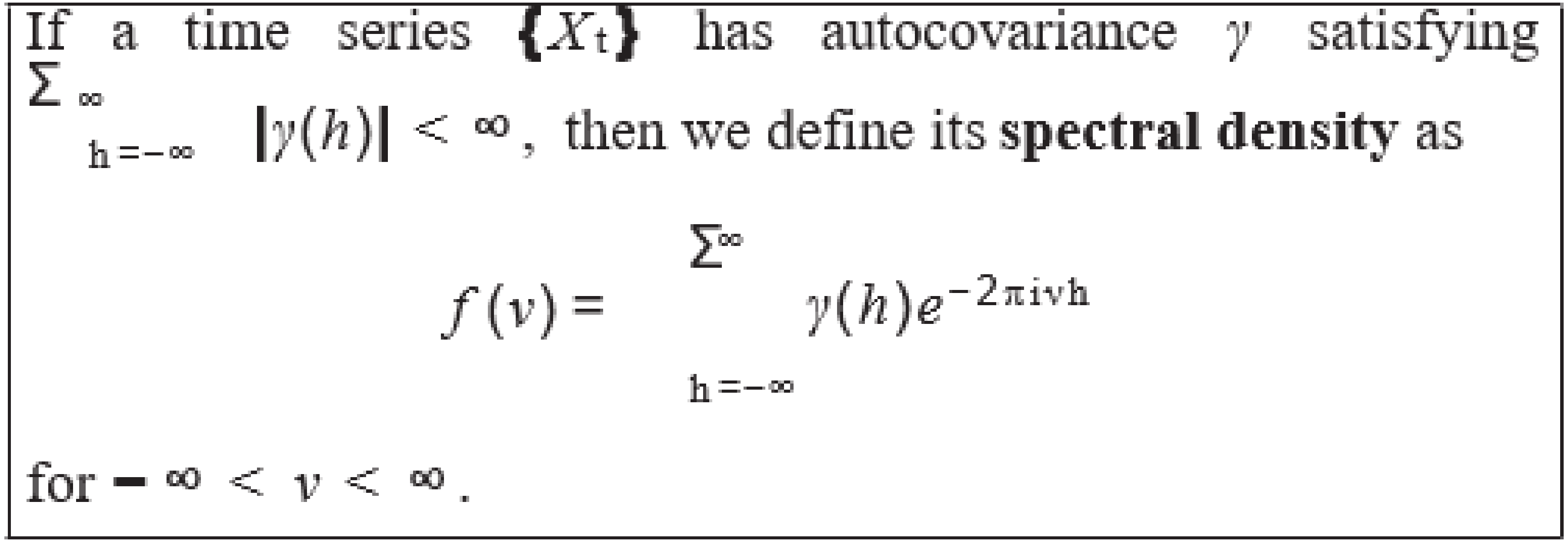
where A_j, B_j are uncorrelated, mean zero, and $\mathrm{Var}(A_j) = \mathrm{Var}(B_j) = \sigma^2$. j

A periodic time series

$$X_t = \sum_{j=1}^{k} \left(A_j \sin(2\pi v_j t) + B_j \cos(2\pi v_j t)\right), \qquad \gamma(h) = \sum_{j=1}^{k} \sigma_j^2 \cos(2\pi v_j h).$$

- Thus, we can represent $\gamma(h)$ using a Fourier series. The coefficients are the variances of the sinusoidal components.
- The spectral density is the continuous analog: the Fourier transform of γ.
- (The analogous spectral representation of a stationary process X_t involves a stochastic integral—a sum of discrete components at a finite number of frequencies is a special case. We won’t consider this representation in this course.)

Spectral density



If a time series $\{X_t\}$ has autocovariance γ satisfying $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, then we define its **spectral density** as

$$f(\nu) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \nu h}$$

for $-\infty < \nu < \infty$.

Spectral density: Some facts

1. We have $\sum_{h=-\infty}^{\infty} \left|\gamma(h)e^{-2\pi i v h}\right| < \infty.$

   This is because $|e^{i\theta}| = |\cos\theta + i\sin\theta| = (\cos^2\theta + \sin^2\theta)^{1/2} = 1$, and because of the absolute summability of γ.

2. f is periodic, with period 1.

   This is true since $e^{-2\pi i v h}$ is a periodic function of v with period 1.

   Thus, we can restrict the domain of f to $-1/2 \leq v \leq 1/2$. (The text does this.)

Spectral density: Some facts

3. f is even (that is, $f(\nu) = f(-\nu)$).

   To see this, write

$$f(\nu) = \sum_{h=-\infty}^{-1} \gamma(h)e^{-2\pi i\nu h} + \gamma(0) + \sum_{h=1}^{\infty} \gamma(h)e^{-2\pi i\nu h},$$

$$f(-\nu) = \sum_{h=-\infty}^{-1} \gamma(h)e^{-2\pi i\nu(-h)} + \gamma(0) + \sum_{h=1}^{\infty} \gamma(h)e^{-2\pi i\nu(-h)},$$

$$= \sum_{h=1}^{\infty} \gamma(-h)e^{-2\pi i\nu h} + \gamma(0) + \sum_{h=-\infty}^{-1} \gamma(-h)e^{-2\pi i\nu h}$$

$$= f(\nu).$$

4. $f(\nu) \geq 0$.

Spectral density: Some facts

5. $\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i v h} f(v)\, dv.$

$$\int_{-1/2}^{1/2} e^{2\pi i v h} f(v)\, dv = \int_{-1/2}^{1/2} \sum_{j=-\infty}^{\infty} e^{-2\pi i v (j-h)} \gamma(j)\, dv$$

$$= \sum_{j=-\infty}^{\infty} \gamma(j) \int_{-1/2}^{1/2} e^{-2\pi i v (j-h)}\, dv$$

$$= \gamma(h) + \sum_{j \neq h} \frac{\gamma(j)}{2\pi i (j-h)} \cdot \left(e^{\pi i (j-h)} - e^{-\pi i (j-h)} \right)$$

$$= \gamma(h) + \sum_{j \neq h} \frac{\gamma(j) \sin(\pi(j-h))}{\pi(j-h)} = \gamma(h).$$

Example: White noise

For white noise $\{W_t\}$, we have seen that $\gamma(0) = \sigma^2$and $\gamma(h) = 0$ for h f= 0

Thus,

$$f(v) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i v h}$$

$$= \gamma(0) = \sigma^2{}_w$$

- That is, the spectral density is constant across all frequencies: each frequency in the spectrum contributes equally to the variance. This is the origin of the name white noise: it is like white light, which is a uniform mixture of all frequencies in the visible spectrum.

Example: AR(1)

For $X_t = \varphi_1 X_{t-1} + W_t$, we have seen that $\gamma(h) = \sigma_w^2 \varphi_1^{|h|}/(1 - \varphi_1^2)$. Thus,

$$f(\nu) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \nu h} = \frac{\sigma_w^2}{1-\varphi_1^2} \sum_{h=-\infty}^{\infty} \varphi_1^{|h|} e^{-2\pi i \nu h}$$

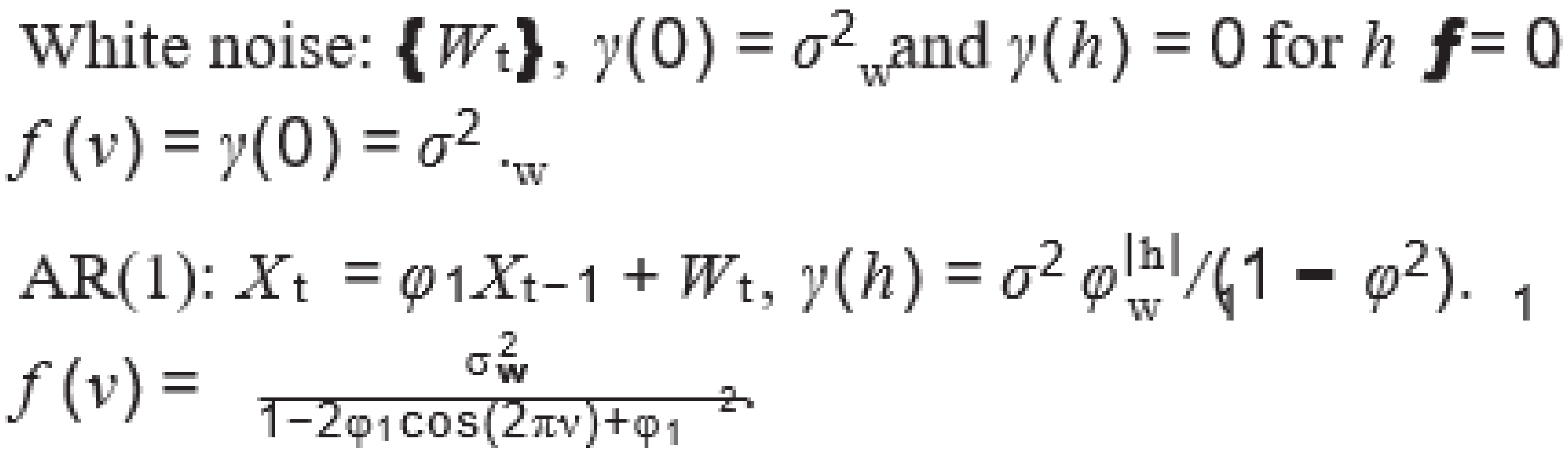
$$= \frac{\sigma_w^2}{1-\varphi_1^2} \cdot 1 + \sum_{h=1}^{\infty} \varphi_1^{h} \cdot e^{-2\pi i \nu h} + e^{2\pi i \nu h} \Sigma \Sigma$$

$$= \frac{\sigma_w^2}{1-\varphi_1^2} \cdot 1 + \frac{\varphi_1 e^{-2\pi i \nu}}{1-\varphi_1 e^{-2\pi i \nu}} + \frac{\varphi_1 e^{2\pi i \nu}}{1-\varphi_1 e^{2\pi i \nu}} \Sigma$$

$$= \frac{\sigma_w^2}{(1-\varphi_1^2)} \frac{1-\varphi_1 e^{-2\pi i \nu} \varphi_1 e^{2\pi i \nu}}{(1-\varphi_1 e^{-2\pi i \nu})(1-\varphi_1 e^{2\pi i \nu})}$$

$$= \frac{\sigma_w^2}{1-2\varphi_1 \cos(2\pi \nu) + \varphi_1^2}.$$

Examples



White noise: $\{W_t\}$, $\gamma(0) = \sigma^2{}_{w}$and $\gamma(h) = 0$ for h $\boldsymbol{f}= 0$

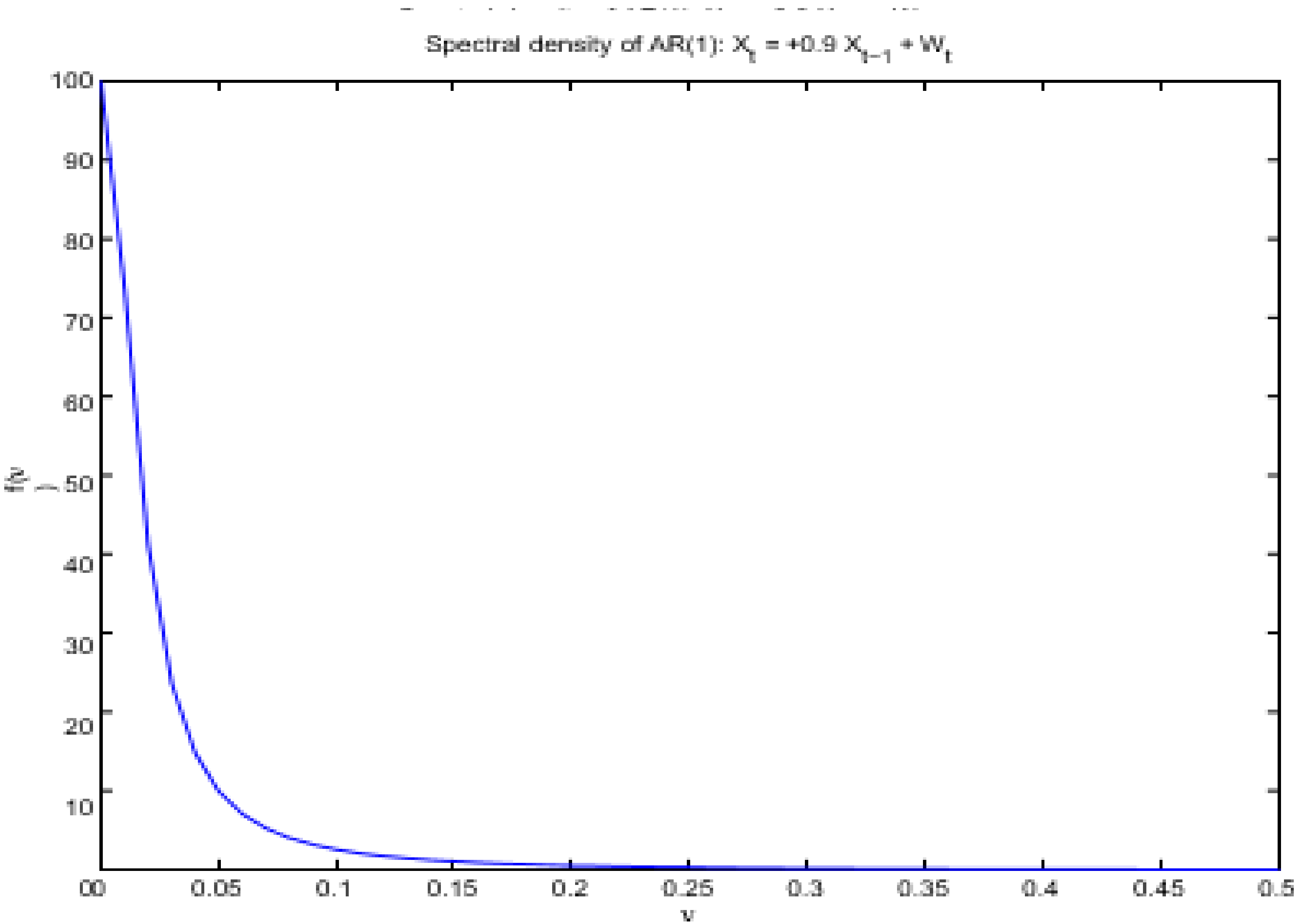
$f(\nu) = \gamma(0) = \sigma^2{}_{\cdot w}$

AR(1): $X_t = \varphi_1 X_{t-1} + W_t$, $\gamma(h) = \sigma^2 \varphi^{|h|}_{w} / (1 - \varphi^2)._{\;1}$

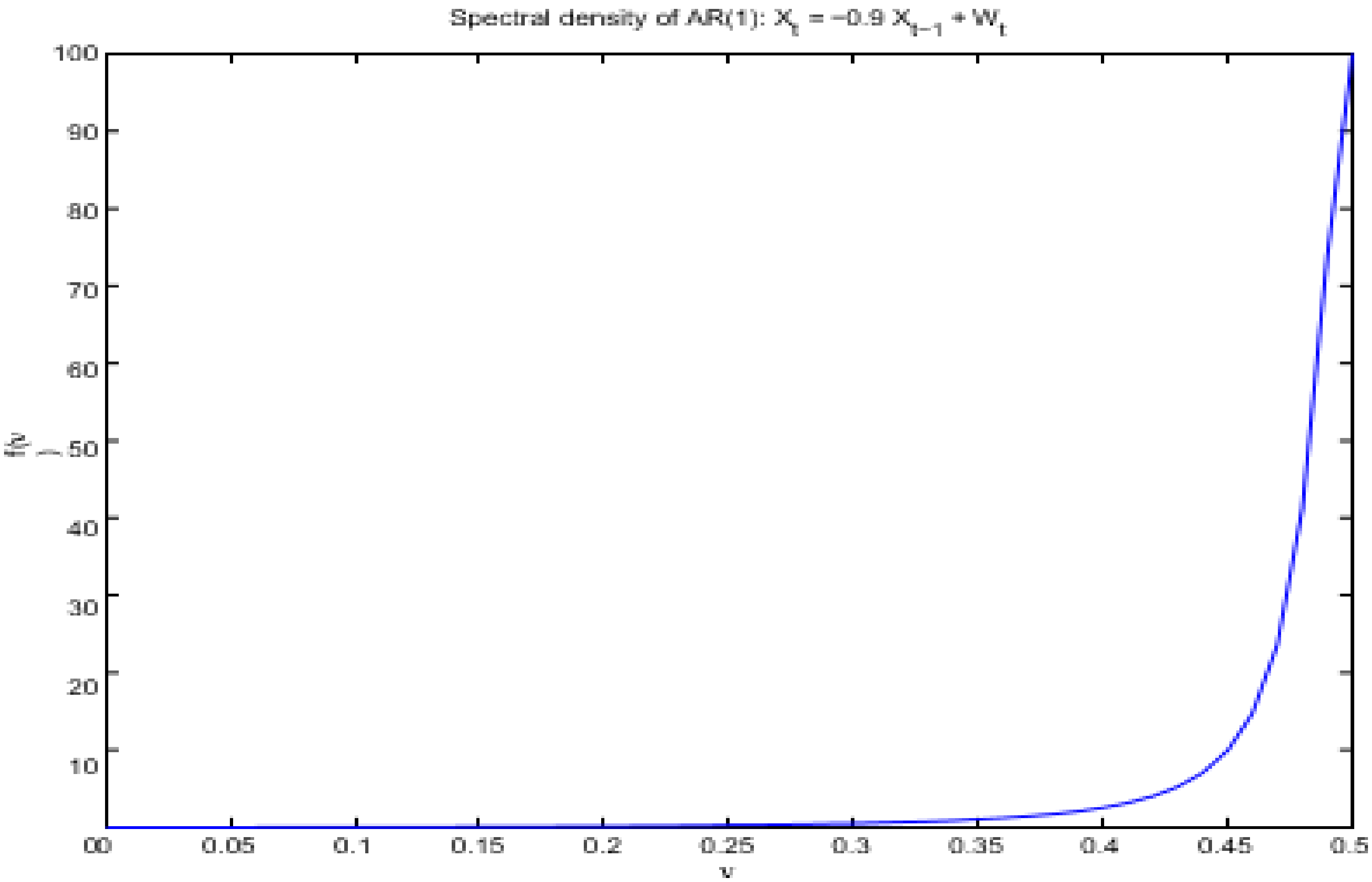
$$f(\nu) = \frac{\sigma_w^2}{1 - 2\varphi_1 \cos(2\pi\nu) + \varphi_1^{\;2}}$$

- If $\varphi_1 > 0$ (positive autocorrelation), spectrum is dominated by low frequency components—smooth in the time domain.
- If $\varphi_1 < 0$ (negative autocorrelation), spectrum is dominated by high frequency components—rough in the time domain.

Example: AR(1)


Spectral density of AR(1): $X_t = +0.9\ X_{t-1} + W_t$
100
90
80
70
60
50
40
30
20
10
f(ν)
00
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
ν

Example: AR(1)


Spectral density of AR(1): $X_t = -0.9\,X_{t-1} + W_t$
100
90
80
70
60
50
40
30
20
10
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
$f(\nu)$
ν

A periodic time series

- Consider
  - $X_t = A\sin(2\pi\nu t) + B\cos(2\pi\nu t)$
  - $= C\sin(2\pi\nu t + \varphi),$
- where $A,\ B$ are uncorrelated, mean zero, variance $\sigma^2 = 1$, and
- $C^2 = A^2 + B^2$, $\tan\varphi = B/A$. Then
  - $\mu_t = \mathrm{E}[X_t] = 0$
  - $\gamma(t, t+h) = \cos(2\pi\nu h).$
- So $\{X_t\}$ is stationary.

A periodic time series

- Consider
  - $X_t = A\sin(2\pi\nu t) + B\cos(2\pi\nu t)$
  - $= C\sin(2\pi\nu t + \varphi),$
- where $A,\ B$ are uncorrelated, mean zero, variance $\sigma^2 = 1$, and
- $C^2 = A^2 + B^2$, $\tan\varphi = B/A$. Then
  - $\mu_t = \mathrm{E}[X_t] = 0$
  - $\gamma(t, t+h) = \cos(2\pi\nu h).$
- So $\{X_t\}$ is stationary.

A periodic time series

- Consider
  - $X_t = A\sin(2\pi\nu t) + B\cos(2\pi\nu t)$
  - $= C\sin(2\pi\nu t + \varphi),$
- where $A,\ B$ are uncorrelated, mean zero, variance $\sigma^2 = 1$, and
- $C^2 = A^2 + B^2$, $\tan\varphi = B/A$. Then
  - $\mu_t = \mathrm{E}[X_t] = 0$
  - $\gamma(t, t+h) = \cos(2\pi\nu h).$
- So $\{X_t\}$ is stationary.

A periodic time series

- Consider
  - $X_t = A\sin(2\pi\nu t) + B\cos(2\pi\nu t)$
    - $= C\sin(2\pi\nu t + \varphi),$
- where $A,\ B$ are uncorrelated, mean zero, variance $\sigma^2 = 1$, and
- $C^2 = A^2 + B^2$, $\tan\varphi = B/A$. Then
  - $\mu_t = \mathrm{E}[X_t] = 0$
  - $\gamma(t, t+h) = \cos(2\pi\nu h).$
- So $\{X_t\}$ is stationary.

Artificial Intelligence for Time-Series and Sequential Decision Making

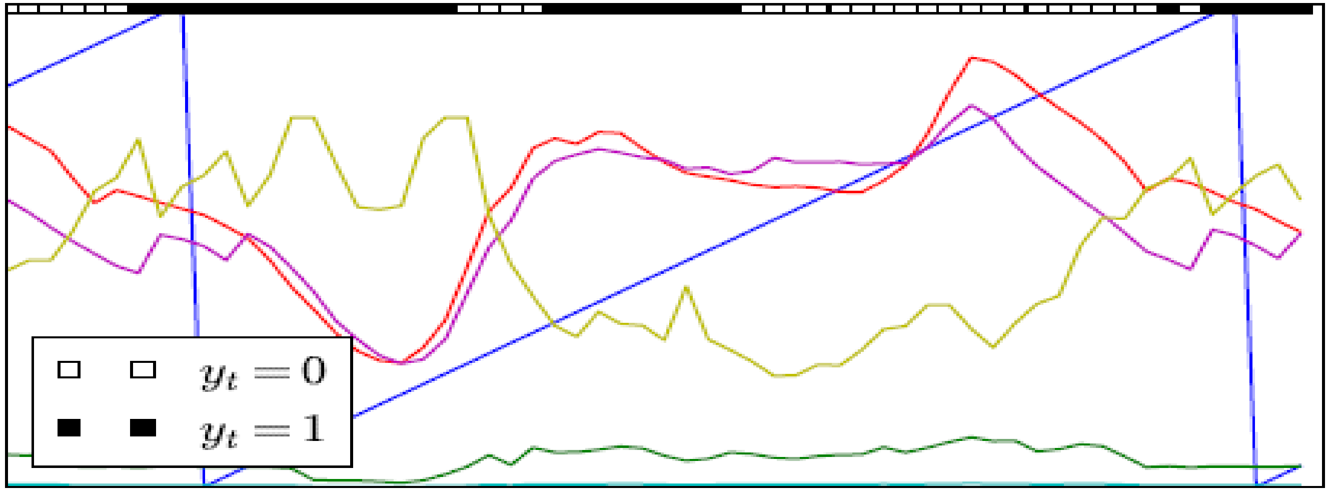
Outline

- Time Series
- Filtering
- Forecasting
- Embedding
- Classifier and Repressor Chains
- Sequential Decision Making

Time Series

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots$$

- Generated by some process $\mathbf{x} \sim p(X\ )$ in the domain we are interested in.
- Measurements may be continuous, $\mathbf{x}_t \in \mathrm{R}^D$ or discrete, $\mathbf{x}_t \in \mathrm{N}_+{}^D$; across time t.
- May be associated with unobserved signal $\mathbf{y}_t$.



Time series $\mathbf{x}_t \in \mathbb{R}^5$ associated with state $y_t \in \{0, 1\}$.

Examples of time series data:

- Electricity demand for a city
- Sensor measurements on equipment in an aircraft

Number of calls to an insurance service

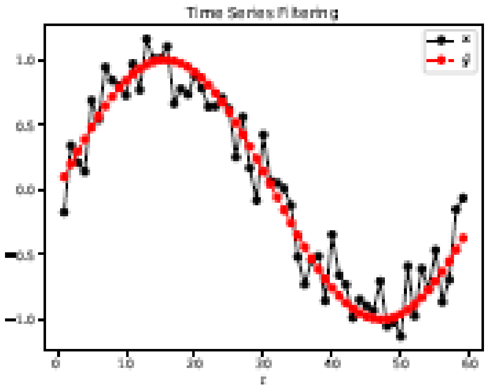
- Light-sensor measurements (and movement through a room)
- Smartphone GPS and signal strength measurements of urban travellers (and their predicted trajectory)
- EEG and ECG signals obtained during sleep

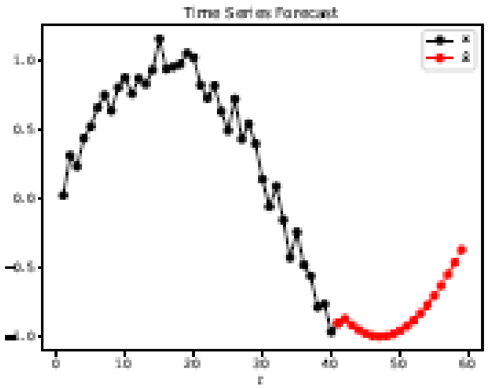
Cellular growth in trees

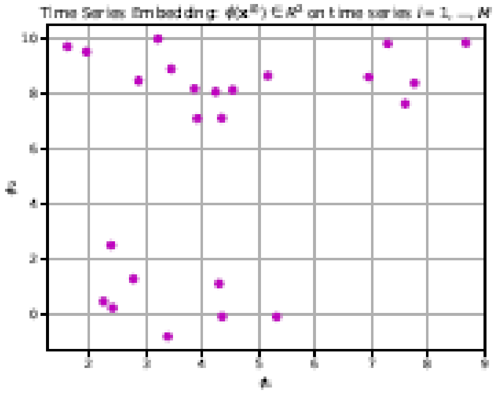
- Environmental measurements (temperature, humidity)

Time Series Tasks

- Filtering (*estimate*) $\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_t$ from observations
- $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t$
- Forecasting (*predict*) $\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots$ from time t. Embedding: Describe a time series $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ as a vector $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_N]$ of fixed length N.







- Clustering
- Classification
- Motif extraction
- Novelty/anomaly detection
- Query by content

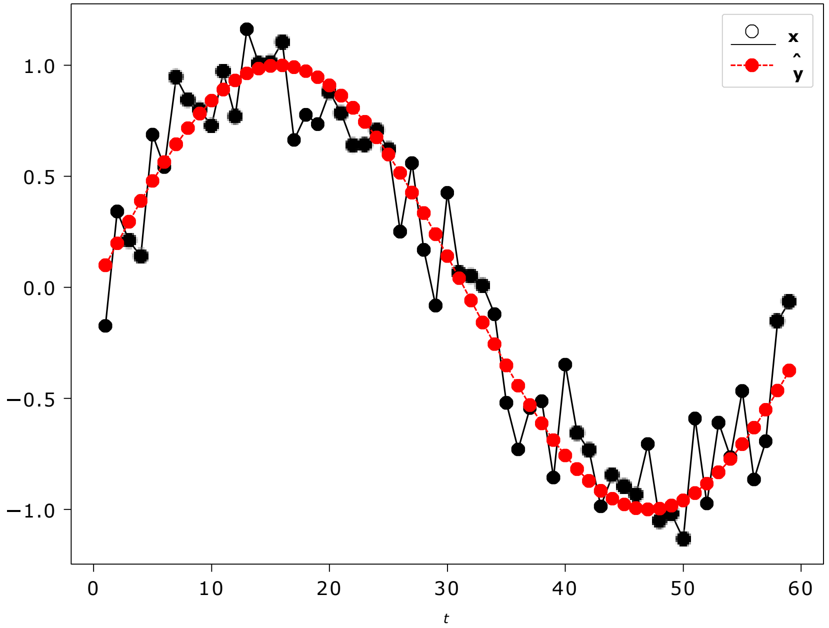
Filtering

Given observations (time series)$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots\}$

we want a model *f* to predict corresponding

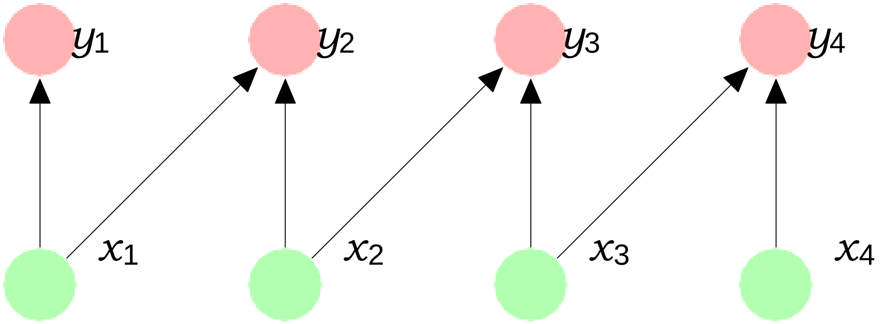
$\{{}^\wedge\mathbf{y}_1, {}^\wedge\mathbf{y}_2, \ldots, {}^\wedge\mathbf{y}_t, \ldots\}$

Time Series Filtering



Traditional Methods

- Finite impulse response (FIR) filter
- Moving average, exponential smoothing (low-pass filters) Kalman filter, particle filters
- ARIMA (Auto-Regressive Integrated Moving Average)

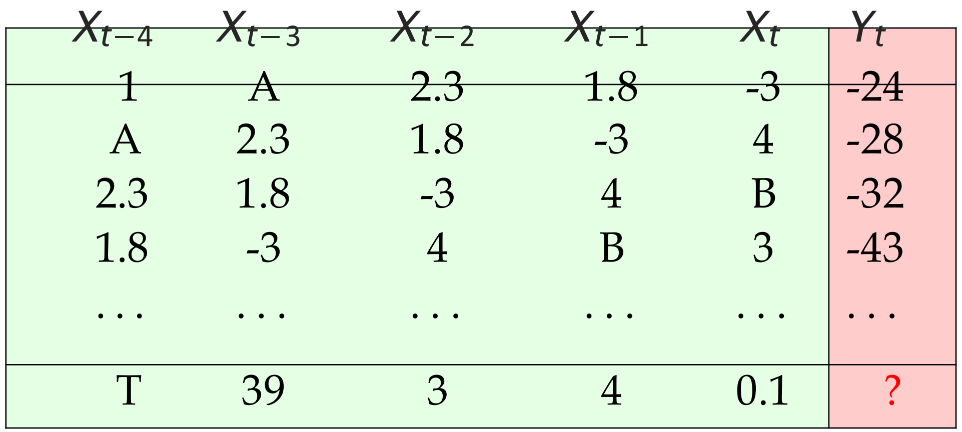


- $y_t = f(w_1 x_{t-0} + \cdots + w_k x_{t-k}) + s_t$

- with some weights $\mathbf{w} = [w_1, \ldots, w_k]$ (window size k). This is a convolution with kernel $\mathbf{w}$.
- Robust and well-understood
- Need to be hand-crafted, calibration by domain expert else not suitable for multiple dimensions; complex problems

Machine Learning for Filtering

- Given training data, we can design a machine learning approach (e.g., artificial neural networks, decision trees, . . .), on



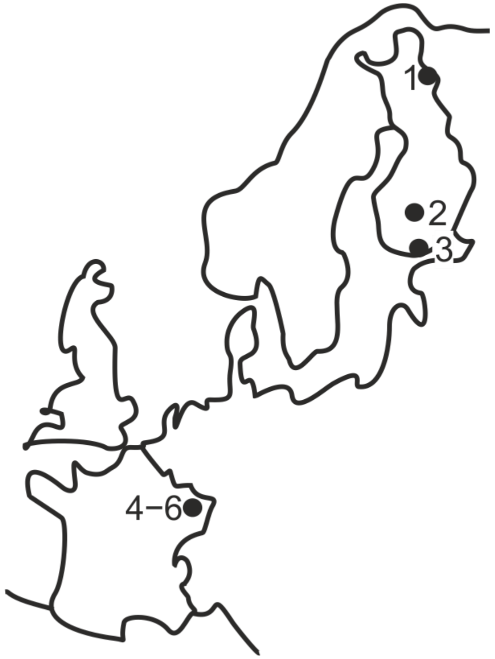
X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}	X_t	Y_t
1	A	2.3	1.8	-3	-24
A	2.3	1.8	-3	4	-28
2.3	1.8	-3	4	B	-32
1.8	-3	4	B	3	-43
. . .	. . .	. . .	. . .	. . .	. . .
T	39	3	4	0.1	?

i.e., model $y_t = f(x_{t-4}, \ldots, x_t; \theta) + s$

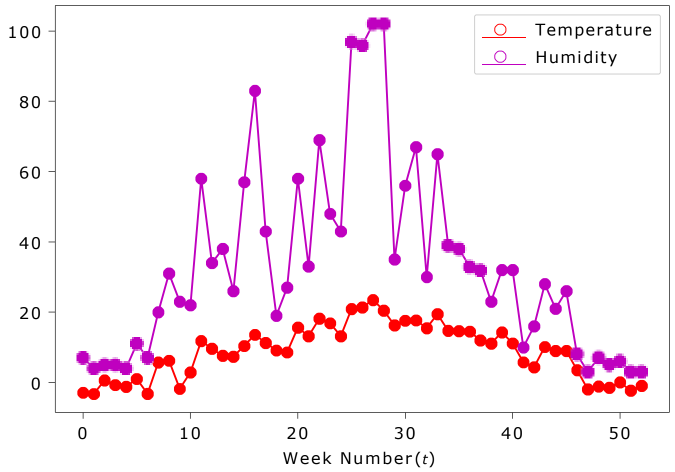
The decision making and interpretation is relegated to the learner.

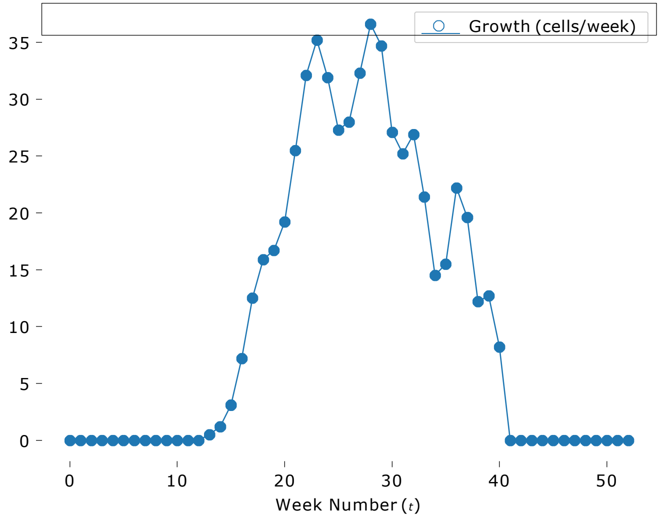
Example: Predicting Cellular Growth in Scots Pine

- 6 sites in Finland and France, of Scots pine
- Interested in modelling cellular growth under different latitude, altitude, . . .
- Models must be carefully crafted, parametrized, and adjusted by domain experts, *per site*.



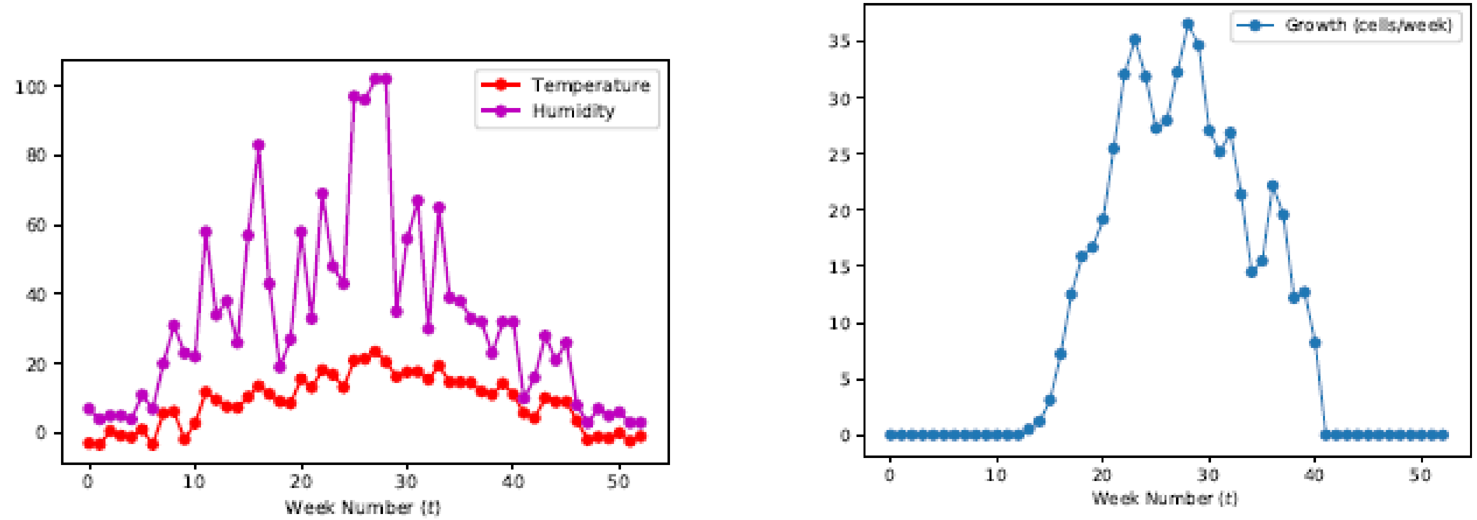
Example: Predicting Cellular Growth in Scots Pine





- Environmental measurements (temperature, humidity, . . .).
- Some cell-growth data (from micro-core samples and counts during growth season) over 3–4 years

Outline



Domain experts were using numerous functions, e.g., growth timing variable (left) and heat sum (right),

e.g., where $\tau_t =$ temperature and week t,

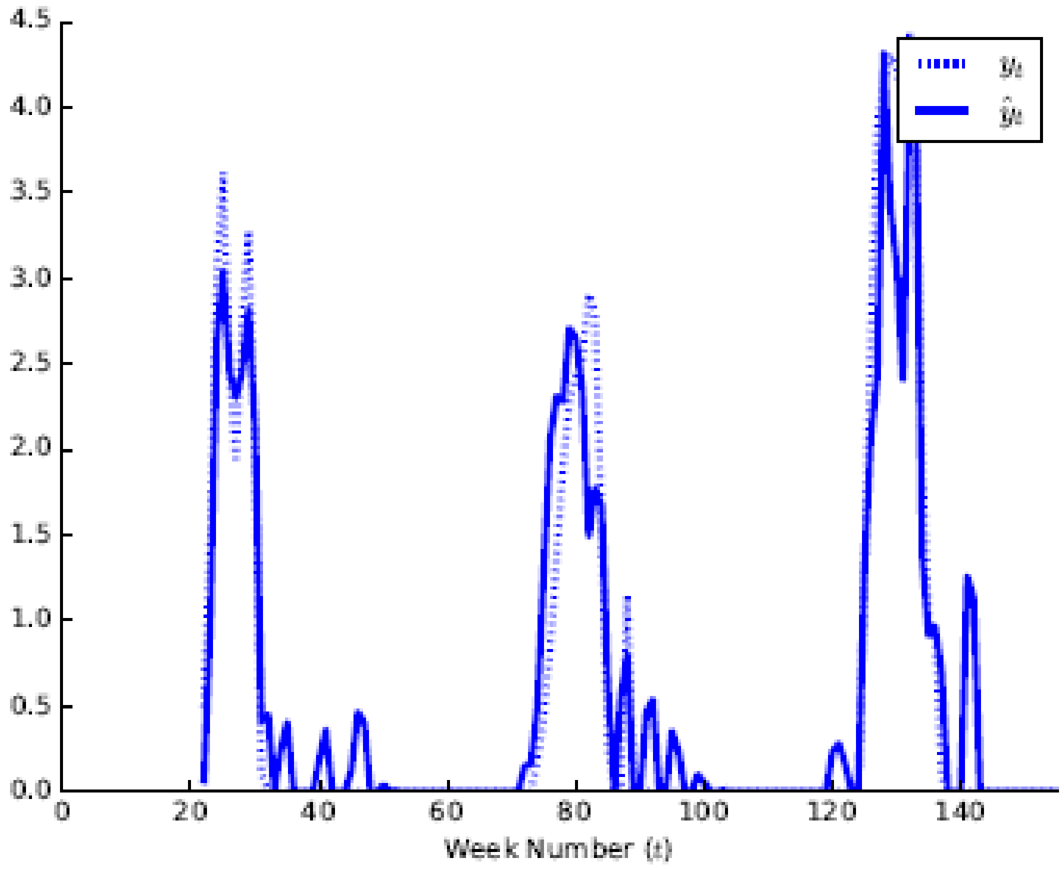
and c, β are per-site parameters.

$$z_t = 1_{\tau_t > c} \quad \Sigma \; \frac{1}{1 + \exp(-\beta\tau_t)} \; \Sigma$$

Assembled into a differential equation

About 4-5 parameters to be hand-selected *per site*

Outline

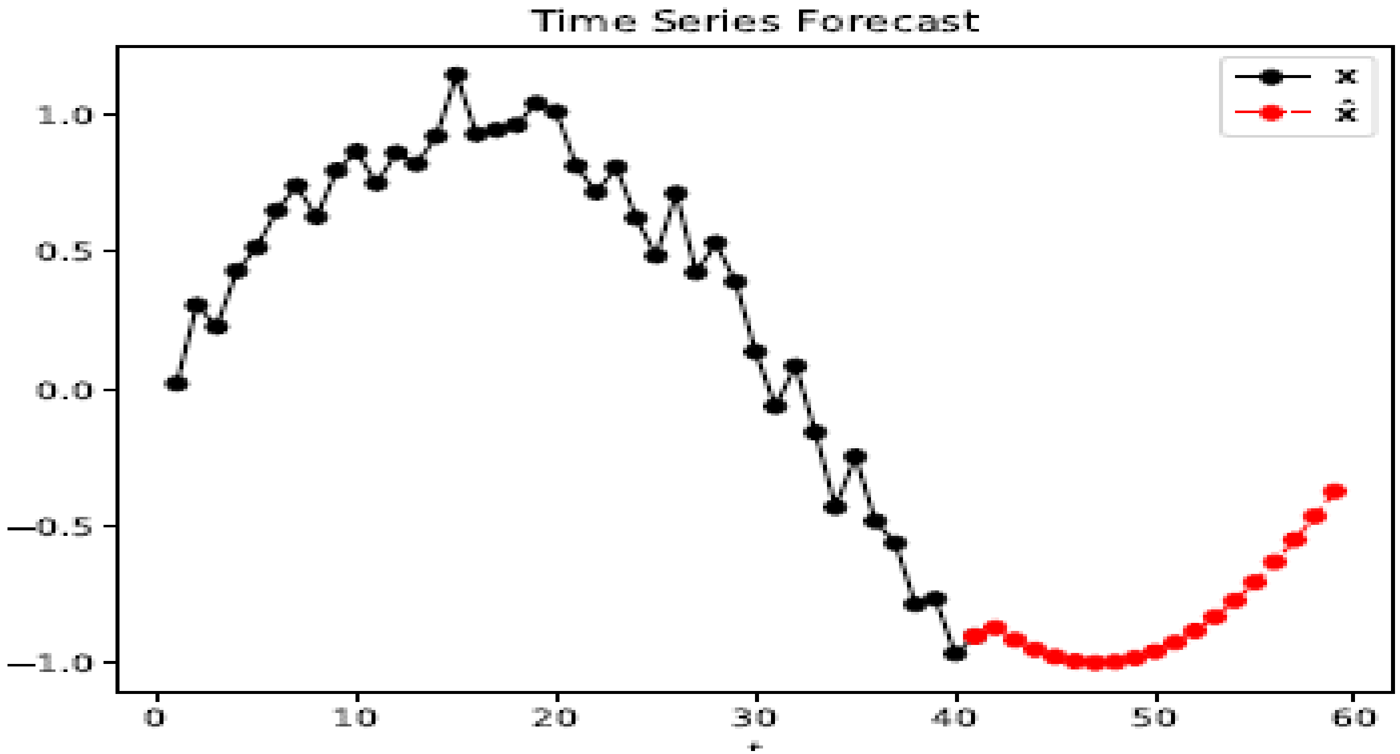


- Data-driven model to parametrize and combine expert-inspired functions, for each site
- Achieved accuracy to within a fraction of a cell per week
- Using decision tree learners, interpretation was possible (e.g., how far back to take into account temperature measurements)

Time-Series Forecasting (Prediction)

Given $\mathbf{x}_1, \mathbf{x}_2, . . . , \mathbf{x}_t$

we want a model f to predict ^$\mathbf{x}_{t+1}$,^$\mathbf{x}_{t+2}$, . . . ,^$\mathbf{x}_{t+A}$



Traditional Methods

- Naive Forecasting (rain today = rain tomorrow) Often effective.

$$\hat{x}_{t+1} = x_t$$

- Moving average (mean of last *k* observations)

$$\hat{x}_{t+1} = \mathbf{w}^\top \mathbf{x}$$

$$\mathbf{x} = [x_{t-(k-1)}, \ldots, x_t], \mathbf{w} = [\tfrac{1}{k}, \ldots, \tfrac{1}{k}].$$

on window

- Auto-regressive linear fit on previous *k* points, and extrapolate.

Machine Learning for Forecasting

- Formulating a data-driven supervised learning problem:

$X_{t-Æ}$	$X_{\ldots}$	X_{t-2}	X_{t-1}	X_t	X_{t+1}
1	A	2.3	1.8	-3	4
A	2.3	1.8	-3	4	B
2.3	1.8	-3	4	B	3
1.8	-3	4	B	3	-7
. . .	. . .	. . .	. . .	. . .	. . .
T	39	3	4	0.1	?

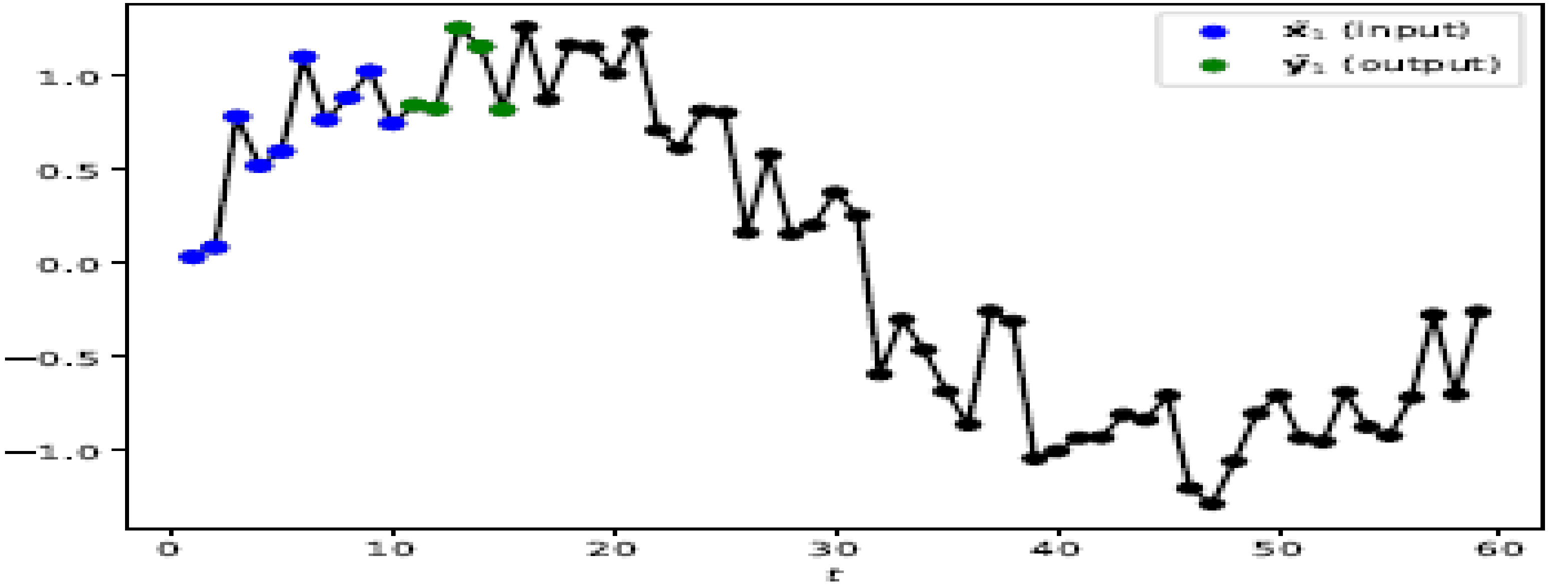
i.e., model $\hat{x}_{t+1} = f(x_{t-4}, \ldots, x_t; \theta)$

(we can plug in $\hat{x}_{t+1}$ and propagate); or estimate a window directly:

$$\hat{x}_{t+1}, \ldots, \hat{x}_{t+k} = f(x_{t-4}, \ldots, x_t)$$

Machine Learning for Forecasting

- Formulating a data-driven supervised learning problem:

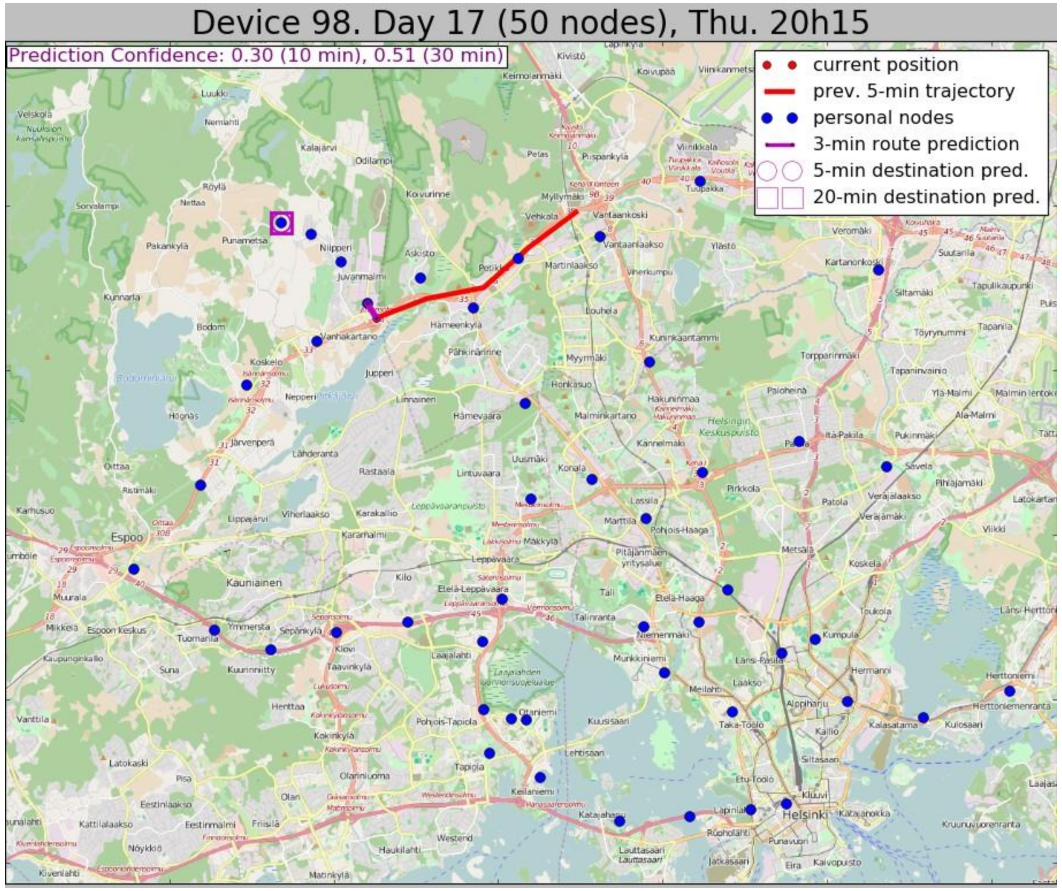


i.e., model $\hat{x}_{t+1} = f(x_{t-4}, \ldots, x_t; \theta)$

(we can plug in $\hat{x}_{t+1}$ and propagate); or estimate a window directly:

$$\hat{x}_{t+1}, \ldots, \hat{x}_{t+k} = f(x_{t-4}, \ldots, x_t)$$

Machine Learning for Forecasting


Device 98. Day 17 (50 nodes), Thu. 20h15
Prediction Confidence: 0.30 (10 min), 0.51 (30 min)
● ● current position
▬ prev. 5-min trajectory
● ● personal nodes
▬ 3-min route prediction
○○ 5-min destination pred.
□□ 20-min destination pred.
Espoo
Kauniainen
Helsinki

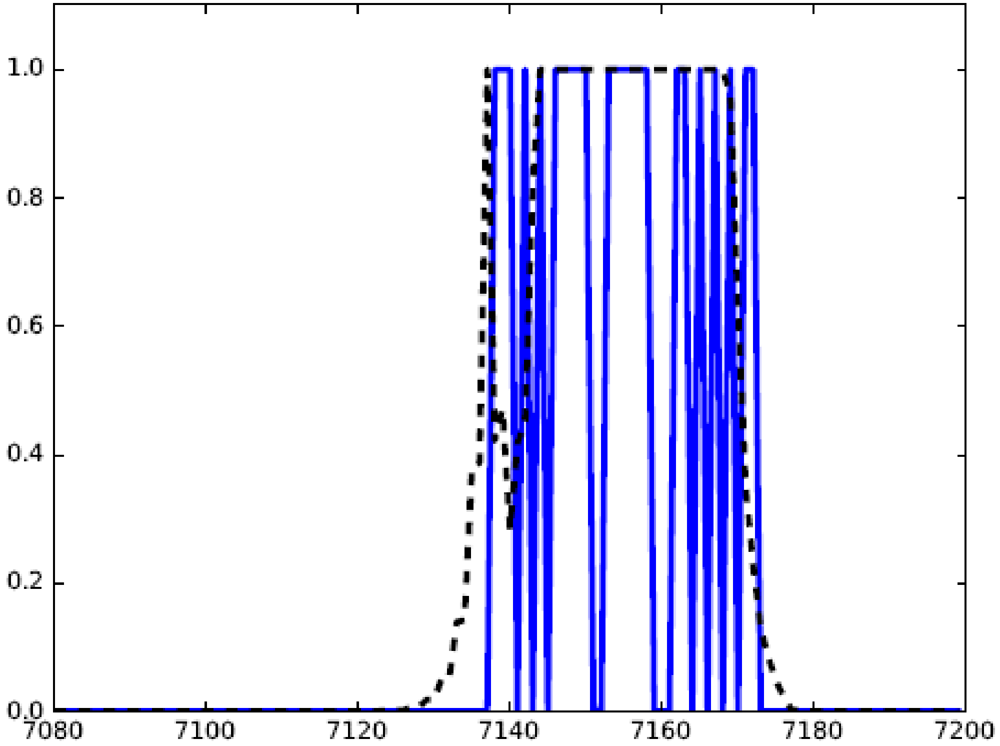
- Collected data of travellers[1]: GPS coordinates, signal strength, battery level, current time, . . .
- Predict future trajectory from current trajectory

[1] All participants volunteered to install App; share data

Work with Jaakko Hollmèn et al. @Aalto University

Example: Predictive Maintenance of Aircraft

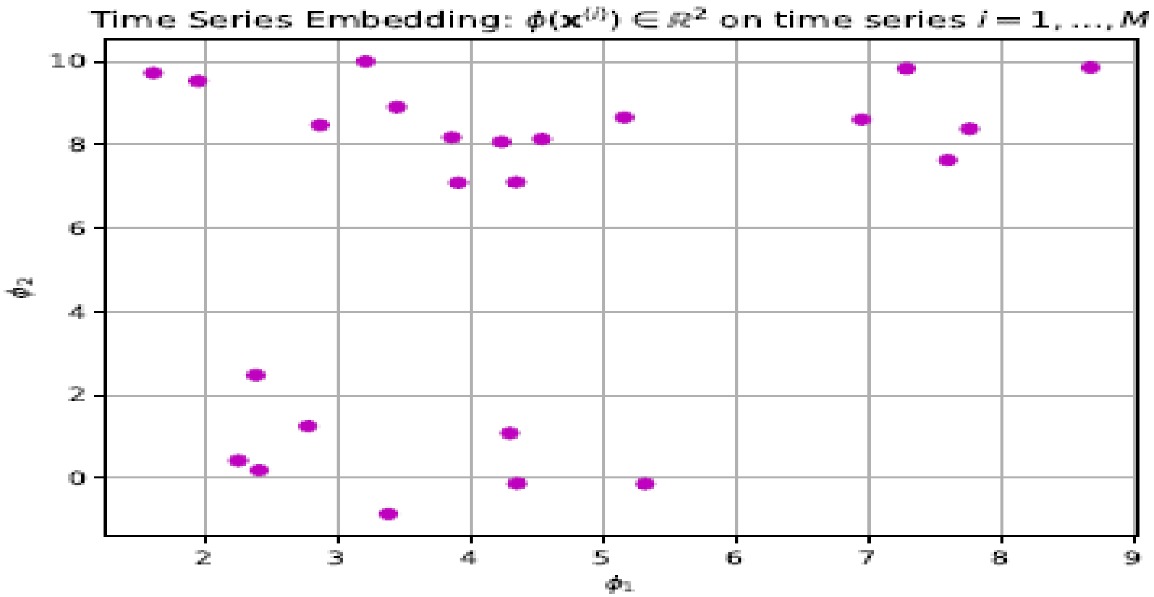
- Sensor readings from aircraft and textual description of observations
- Predict warnings/required replacement of components



Work with DaSciM team @École Polytechnique

Embedding Time Series

We seek to turn variable-length time series $\{\mathrm{x}_1^{(i)}, \ldots, \mathrm{x}_{T_i}^{(i)}\}_{i=1}^{M}$ into fixed-length vectors $\boldsymbol{\varphi}^{(i)} = [\varphi_1, \ldots, \varphi_L]$.


Time Series Embedding: $\phi(\mathbf{x}^{(i)}) \in \mathbb{R}^2$ on time series $i = 1, \ldots, M$

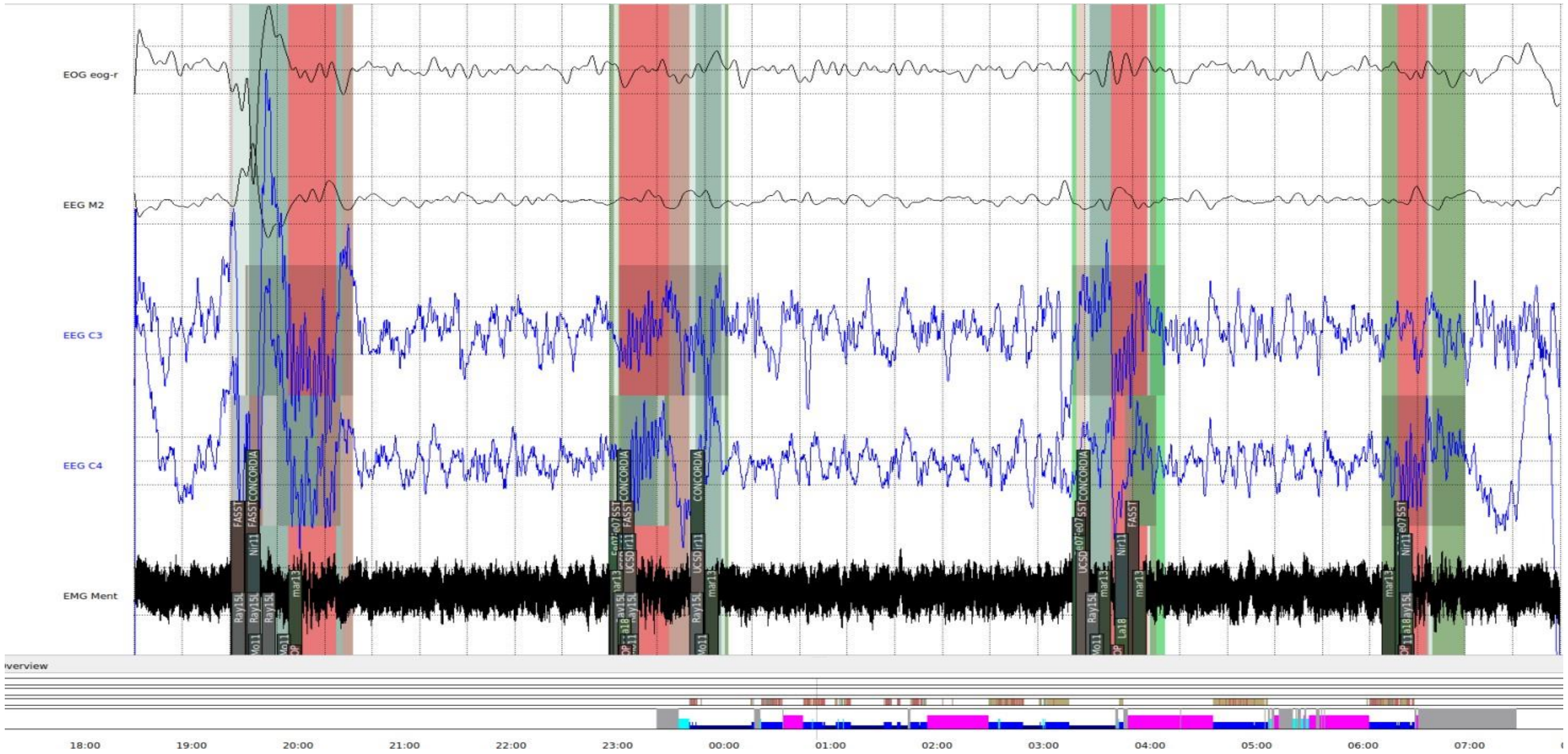
- This lets us compare and cluster time series/look for anomalies, (and classify, if we have the label): measure similarity/distance between $\varphi(\mathbf{x}^{(i)})$ and $\varphi(\mathbf{x}^{(2)})$.

Example: Modelling and Treating Chronic Insomnia

- Goal: (semi-)automate clinical assessment; what kind of insomnia + treatment recommendation.
- Data from patients:
  - Psychological questionnaires (MMPI, CAS) EEG and ECG data overnight
  - Some labels: follow-up tests/questionnaires and *biofeedback* results (some patients found success without pharmaceutical intervention, others not)

- Questionnaire data: can take 'standard' machine learning approach, $f : X \rightarrow Y$, and inspect feature importance, statistical correlation wrt to label variable (extent of insomnia, and improvement); cluster into groups, etc.
- Time-series data: different lengths, contains artifacts, subjects fall asleep at different times, ... How to compare?

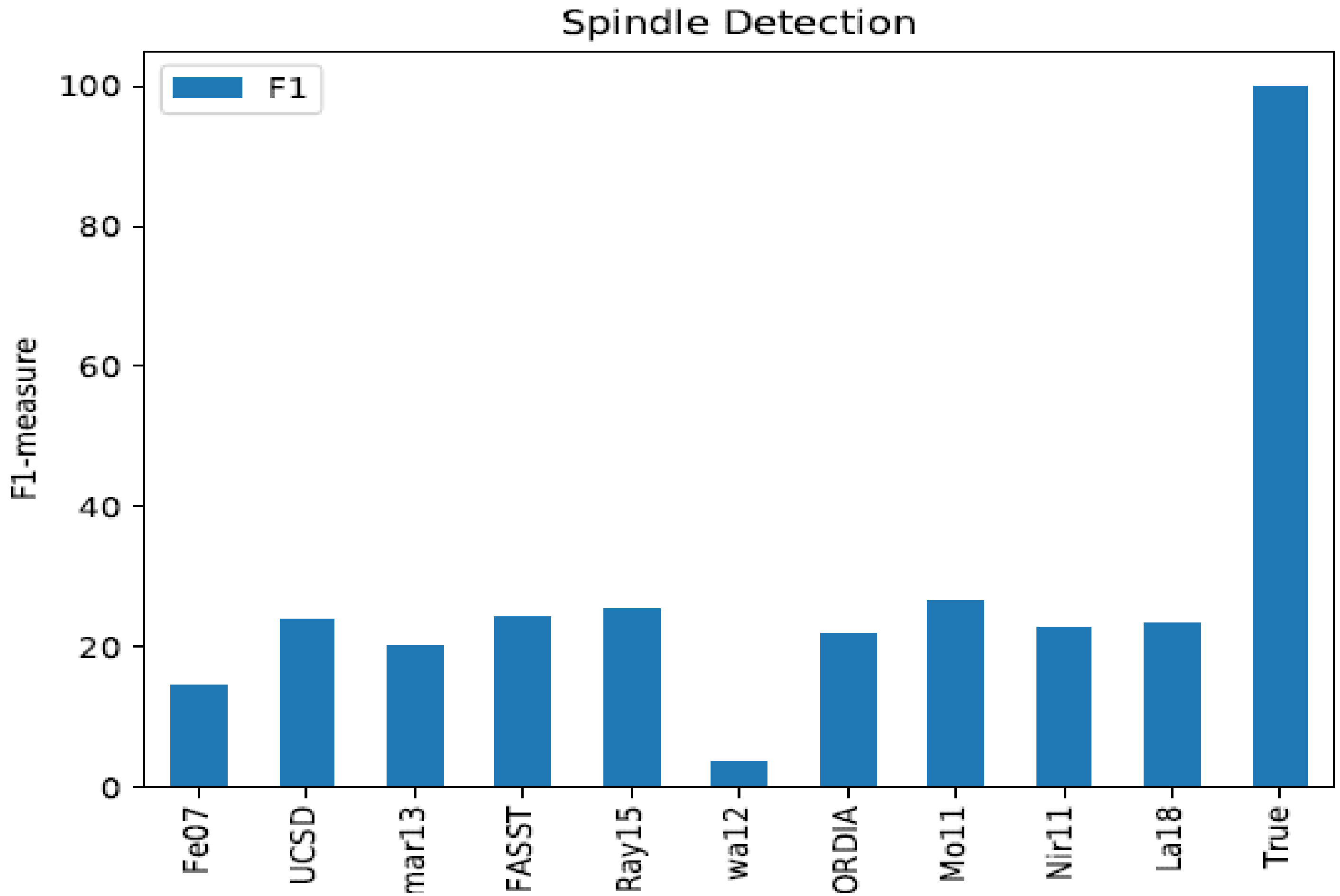
Example: Modelling and Treating Chronic Insomnia



- Certain signals are of interest: Spindles, α-waves, β-waves, . . . Simple embeddings, e.g.,
- $\varphi(\mathbf{x}^{(i)})$ = [spindles/hour, avg freq of spindle]. Detection and labelling by an expert is labour intensive.

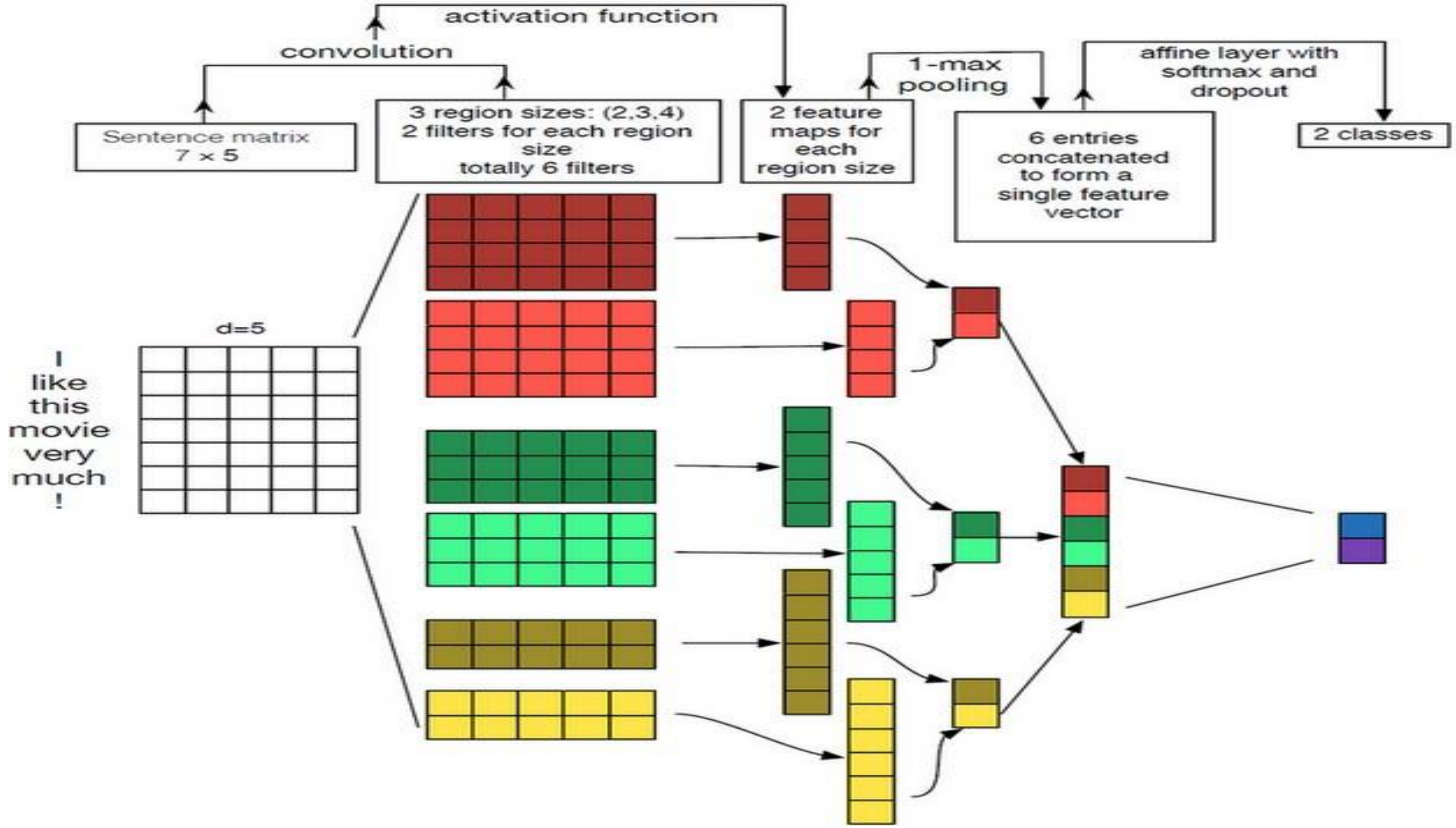
Outline

There exist many rule-based methods, e.g., wavelet analysis But predictive performance is insufficient in many practical settings

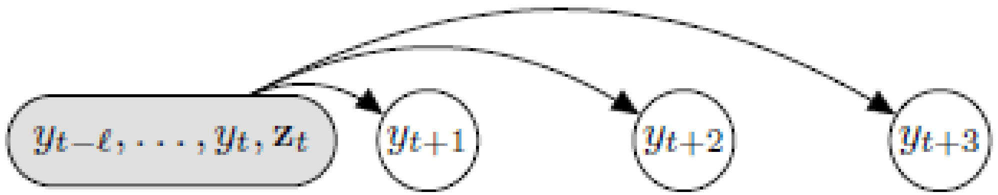


Deep learning

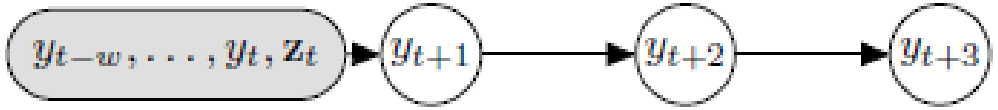
- Many current solutions are inspired by / related to NLP.
- Similar to a 'simple' embedding, but more data-driven.


activation function
convolution
1-max pooling
affine layer with softmax and dropout
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 entries concatenated to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

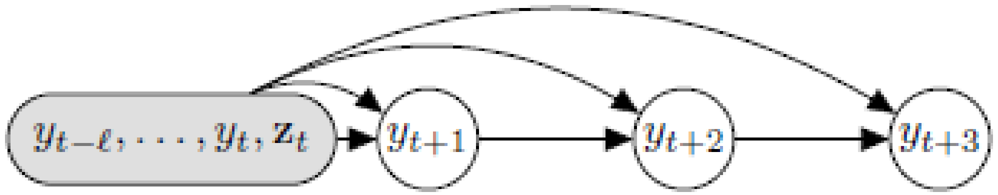
Multi-Step-Ahead Forecasting



Direct

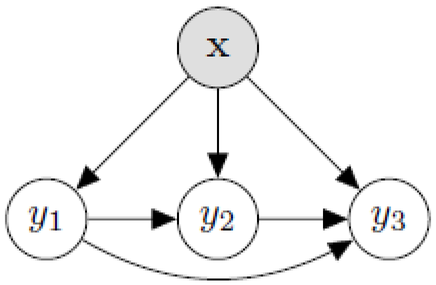


Iterated

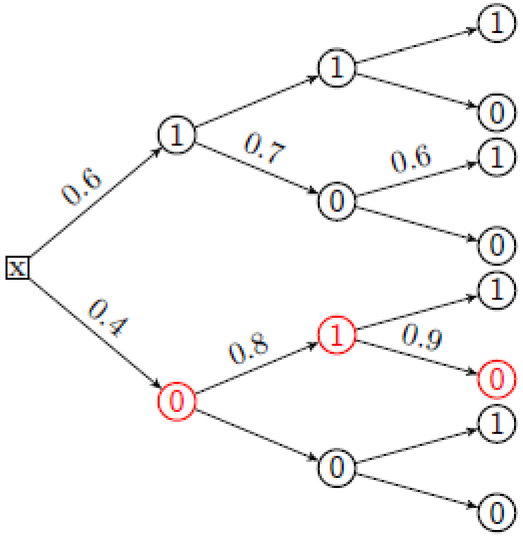


Classifier/Regressor Chain cascade

Classifier Chains



For example, where each $y_t \in \{0, 1\}$



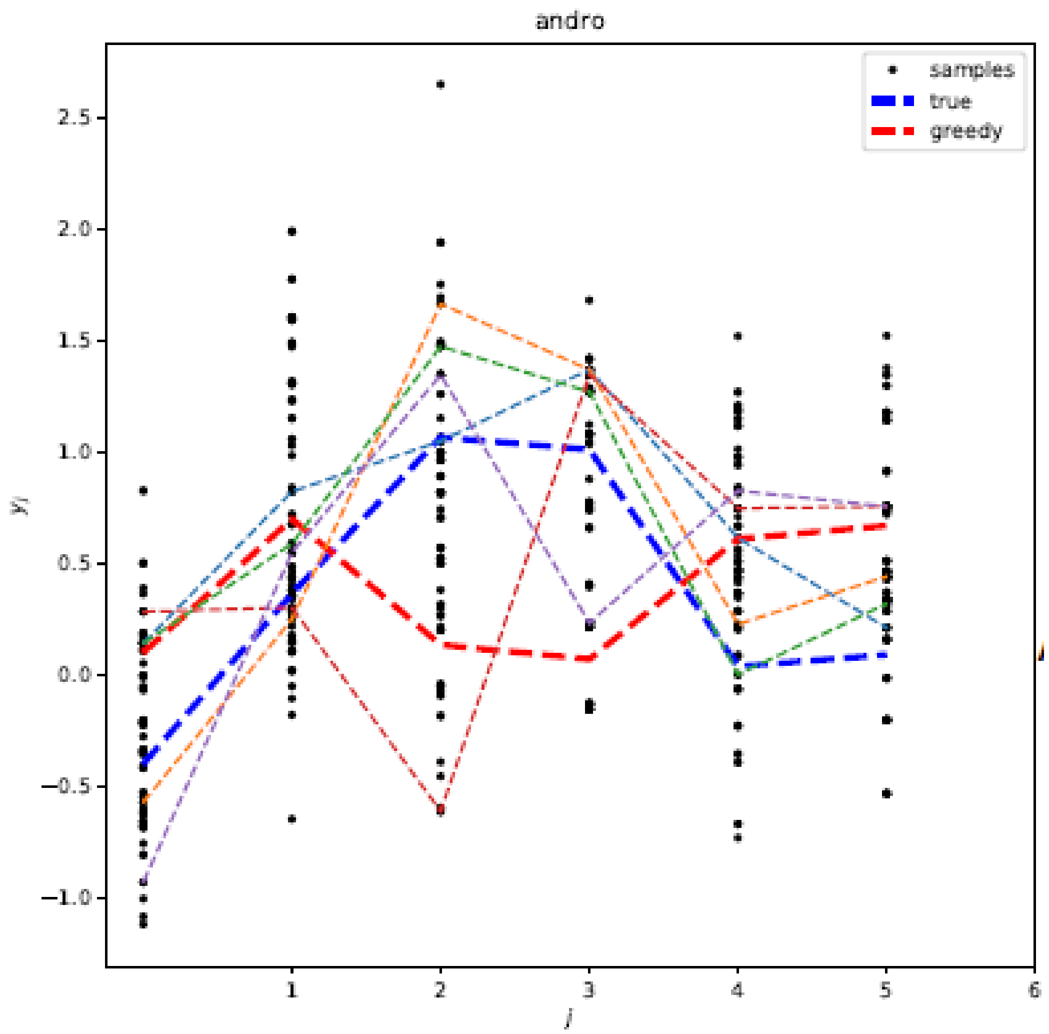
- Predictions become input, across a cascade/chain
- Efficient
- Probabilistic interpretation:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|\mathbf{x}, y_1, \ldots, y_{t-1})$$

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}\in\{0,1\}^3} P(\mathbf{y}|\mathbf{x})$$

- Search probability tree (for best prediction) with AI-search techniques (Monte-Carlo search, beam search, A* search, . . .)
- Explore structure

Regressor Chains



- e.g., where $\mathbf{y} \in R^6$,
  - Sample down the chain
  - $y_{t+1} \sim p(y_{t+1}|y_1, \ldots, y_t, \mathbf{x})$
  - More samples = more hypotheses
  - Consider different loss functions
- Applications:
  - Multi-output regression Tracking
  - Forecasting

One-Step Decision Theory

Under uncertainty, we wish to assign $y^* = f^*(\mathbf{x})$, the best label/hypothesis, $y^* \in$ Y, given $\mathbf{x} \in \mathrm{R}^D$

.Minimizing conditional expected loss

$$f^* = \operatorname*{argmin}_{f\in\mathcal{F}} \underbrace{\sum_{y\in\mathcal{Y}} \ell(f(\mathbf{x}), y)P(y|\mathbf{x})}_{\mathbb{E}_{Y\sim P(Y|\mathbf{x})}[\ell(\hat{y},Y)|\mathbf{x}]}$$

under loss function *A*,which describes our preferences. In the case of 0/1 loss (1 if *y f*= *ŷ*,else 0),

Maximum a Posteriori

$$y^* = \operatorname*{argmax}_{y\in\mathcal{Y}} p(\mathbf{x}|y)P(y) = \operatorname*{argmax}_{y\in\{0,1\}} P(y|\mathbf{x})$$

We can estimate *P* from the training data.

Expected Utility

- An intelligent agent wishes to make a decision to achieve a goal. The decision which involves the least risk. Another way of looking at the problem: utility. A rational agent maximizes their expected utility, not necessarily a simple *payoff* (e.g., amount of money):

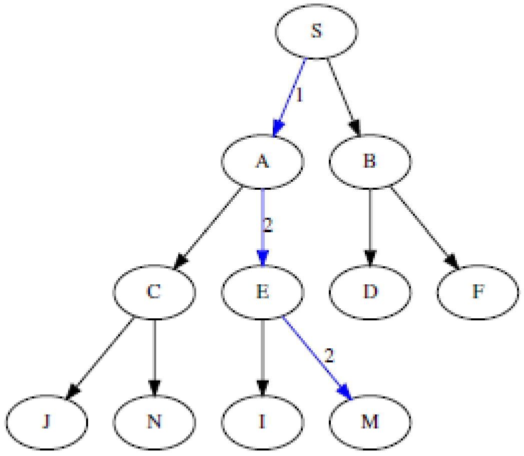
Expected Utility $$U(y) = \sum_{y \in \mathcal{Y}} u(y)p(y)$$

- with satisfaction/utility *u*(*y*) for outcome *y* . Different agents may have different utility functions, even when 'payoff' is the same item. Instead of labels given input, we can deal with actions given evidence and belief.
  - A risk-prone agent will tend to gamble higher stakes A conservative (risk-adverse) agent will not
  - A risk-neutral agent only cares about payoff *y* directly

What about sequential decisions?

In a Deterministic Environment

(e.g., board games – chess, etc.)



The state space, e.g., $s_t \in \{A, B, \ldots, M\}$

An initial state, e.g., $s_0 = S$

A goal state, e.g., $s_t = M$

A set of actions, e.g., $a_t \in \{1, 2\}$

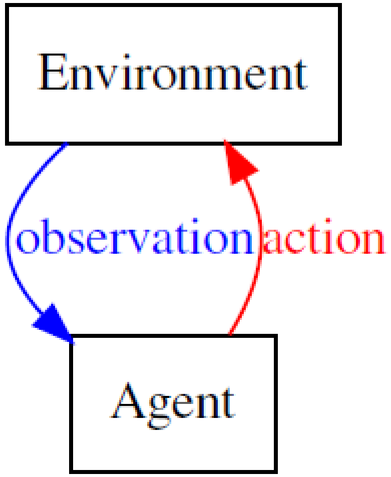
A cost for each branch, e.g., $\mathrm{Cost}(S, A) = 1$

It's just a search! AI-search techniques applicable (DFS, A^*, . . .).

Markov Decision Processes (MDP)

MDPs are models that seek to provide optimal solutions for stocastic sequential decision problems.

MDP = Markov Chain + One-step Decision Theory



Outline

Now we have a model with

$P(s^j|s,\ a)$ transition function

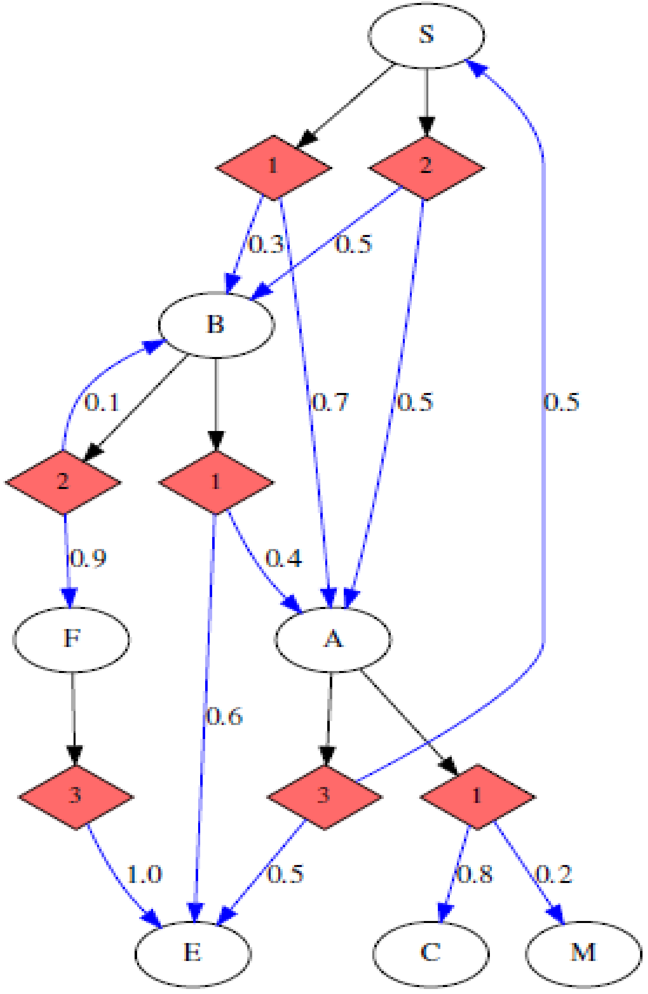
$R(s^j,\ a,\ s)$ reward function

Objective: obtain a policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

which maximizes expected reward:

$$\mathbb{E}[R_0|s_0 = s] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)\right]$$

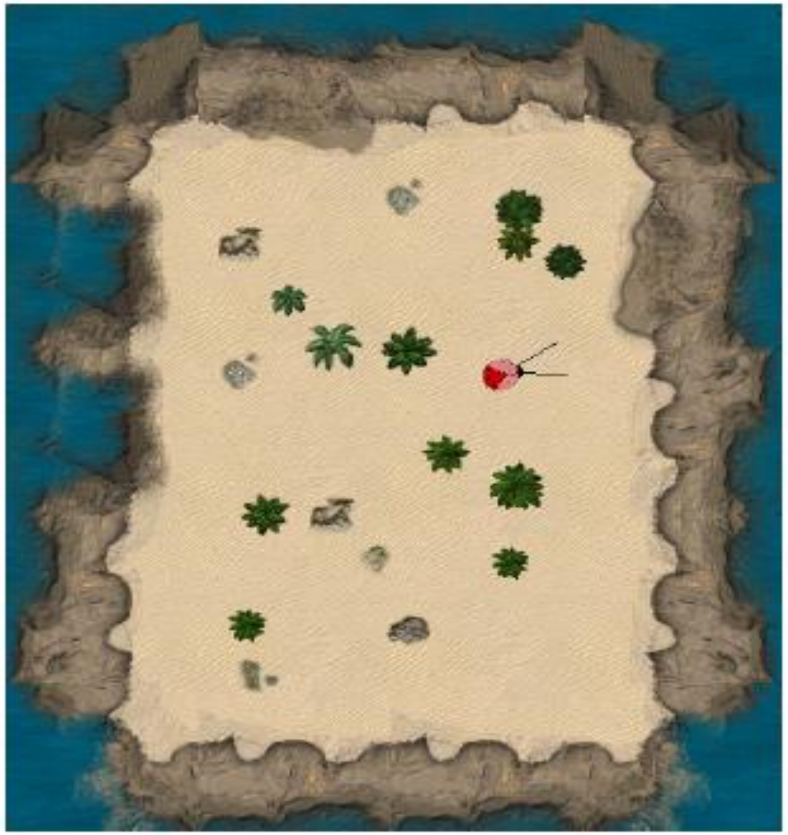


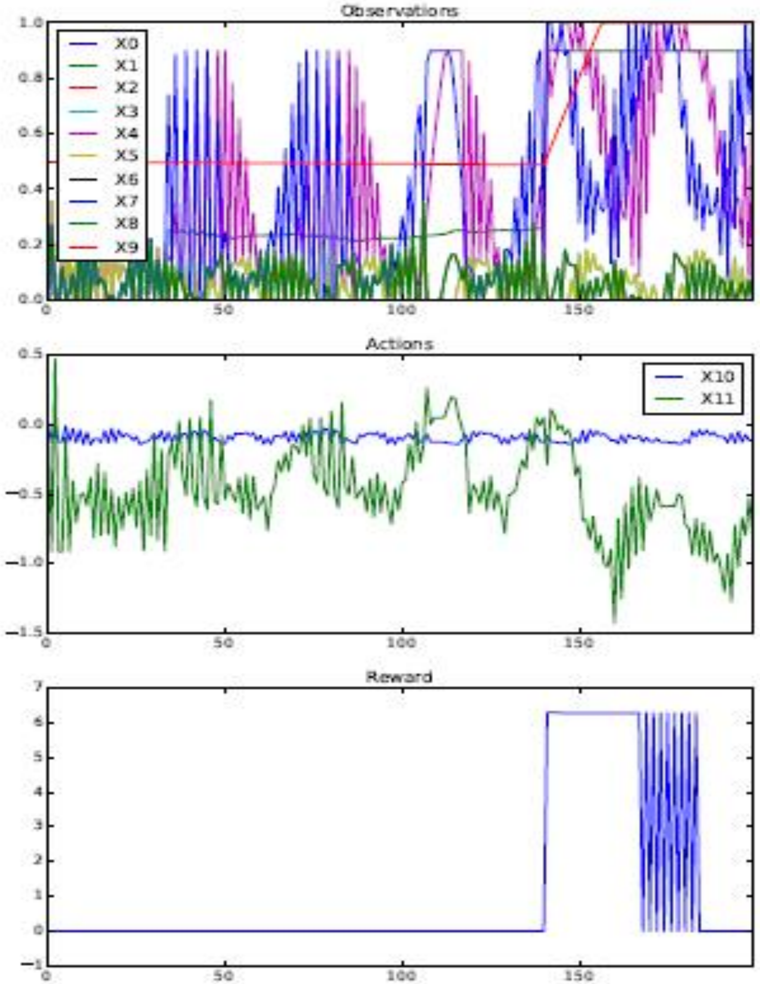
solution can be found via dynamic programming!
Just need the model . . .

Reinforcement Learning

- We don’t have the model!
- Don’t have transition/reward functions.
- No input-output training pairs, just reward signal.
- The agent needs to experiment! Exploration vs exploitation. Deep neural net can learn a model
- . . .over millions of iterations. Emerging applications:
    - Gameplay
    - Robotics (usually trained in simulation) Parameter-tuning, etc. (as a tool)
- Transfer learning is promising

Outline




Observations
1.0
0.8
0.6
0.4
0.2
0.0
X0
X1
X2
X3
X4
X5
X6
X7
X8
X9
0
50
100
150
Actions
0.5
0.0
−0.5
−1.0
−1.5
X10
X11
0
50
100
150
Reward
7
6
5
4
3
2
1
0
−1
0
50
100
150

References

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017

Image Courtesy

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

https://otexts.com/fpp2/stationarity.html
https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics
https://bookdown.org/rdpeng/timeseriesbook/spectral-analysis.html
https://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/15.pdf
https://astrostatistics.psu.edu/su07/fricks_2timeseries07.pdf)
https://blog.octo.com/en/time-series-features-extraction-using-fourier-and-wavelet-transforms-on-ecg-data/
https://jmread.github.io/talks/Time_Series_AI.pdf

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu