Specialization Project

Ulrik Bernhardt Danielsen

September 19, 2022

Contents

Steps:

1. z-scoring (standardization although dividing by standard deviation is commented out)
2. Smoothing by 3d splines
3. Detrending around the smoothed curve (Why do we need to do this? Do we expect trends?)
4. Morlet continuous wavelet transformation
   a) Hypothesis testing against 1st order autoregressive process
   b) Smoothing of power across scales
   c) Scale-averaged wavelet power
   d) Rescaling features
   e) Concatenating trend data and power spectrum
5. Feature vectors were downsampled in time at 1 Hz and pooled acreoss animals and conditions. Where is this done in the code? it
6. PCA, reducing to feature dimension explaining at least 95% of the variance
7. t-SNE
8. Watershed segmentation

Chapter 1

Introduction

1.1 Motivation

The human brain is an incredibly complex structures that researchers have been trying to understand for a long time. One way to gain information about how the brain operates is to study its neurons. Neurons are cells which can communicate with each other through synapses. This communication are electric signals and can be recorded. Source? At Kavli Institute for Systems Neuroscience at NTNU they are interested in relating these neural spike recordings to the behavior in rats. This in turn begs the question of how rats behave. Manually labelling video recordings of rats running around seems a tedious and unfruitful endeavor. Additionally it introduces bias in our prior assumptions of how the rats behave, and which activities they engage in. Thus, a methodology for automatically detecting distinct behaviours is needed.

1.2 Previous work

Is this necessary?

Chapter 2

Theory

2.1 Time series analysis

We define time series as a realization $y_t = \{y_{t_1}, y_{t_2}, \dots, y_{t_n}\}$ of a stochastic process $Y(w,t)$, where $w \in \Omega$, Ω being the sample space, and $t \in \mathbb{Z}$, $\mathbb{Z}$ being the chosen index set Wei 2006. It is an ordered series of random variables which can be described completely by its joint probability function

$$F_{t_1,\dots,t_n}(x_1,\dots,x_n) = \Pr\{y_{t_1} \leq x_1, \dots, y_{t_n} \leq x_n)\}.$$

The mean and variance function of a time series y are defined as

$$\mu_t = E(y_t) \tag{2.1}$$

and

$$\sigma_t^2 = E(y_t - \mu_t)^2.$$

Given two random variables in the series y_{t_1} and y_{t_2}, we define the covariance function and correlation function as

$$\gamma(t_1,t_2) = E[(y_{t_1} - \mu_{t_1})(y_{t_2} - \mu_{t_2})] \tag{2.2}$$

and

$$\rho(t_1,t_2) = \frac{\gamma(t_1,t_2)}{\sqrt{\sigma_{t_1}^2}\sqrt{\sigma_{t_2}^2}}. \tag{2.3}$$

2.1.1 Stationarity

A time series y_t is nth-order stationary if for any shift h and indexes $t_1, t_2, \dots, t_n$ if

$$F_{y_{t_1},\dots,y_{t_n}}(x_1,\dots,x_n) = F_{y_{t_1+h},\dots,y_{t_n+h}}(x_1,\dots,x_n). \tag{2.4}$$

If (2.4) holds for all n, the time series is called *strictly* stationary. We also define a nth-order *weakly* stationary time series y_t if the first n joint moments are finite and time invariant. Specifically we define the second-order weakly stationary, i.e. with constant and time invariant mean function (2.1), and where the covariance function (2.2) is solely a function of the time difference,

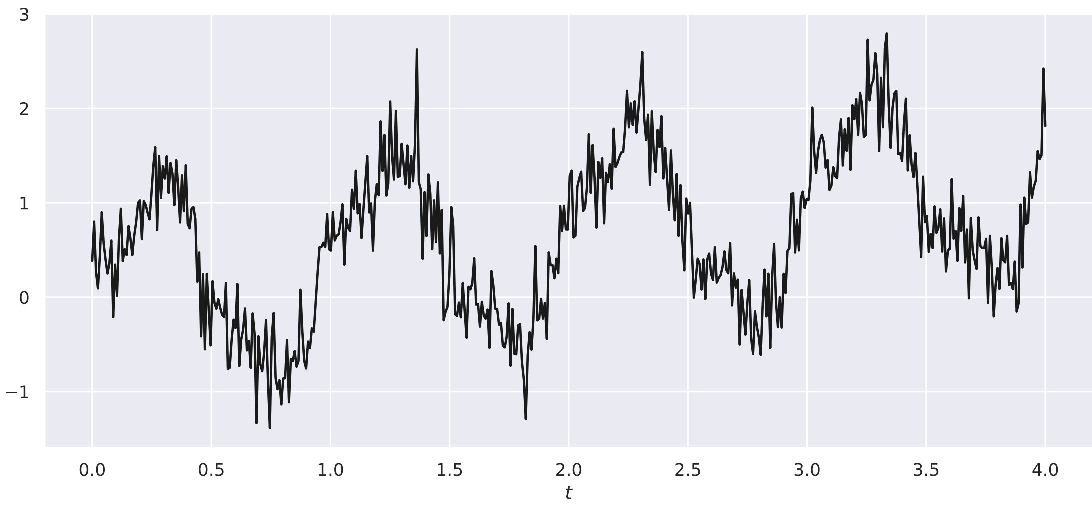

Figure 2.1: Example of a non-stationary time series with $t_0 = 0, t_n = 4$. The discrete data points are connected to better visualize the movement through time.

as *covariance* stationary. When the covariance function between t_1, t_2 can be written as a function of the time difference $h = |t_1 - t_2|$, i.e. $\gamma(t_1, t_2) = \gamma(h) = \gamma_h$, we call it an *autocovariance* function. The same is true for the correlation function (2.3), which when is a function of the time difference is called an *autocorrelation* function (ACF). Figure 2.1 shows an example of a time series. As the mean seem to increase with t it is non-stationary.

2.1.2 Detrending

Many methods for analysing and processing time series requires stationarity Shumway and Stoffer 2017. If the series is non-stationary, we can split the it into one stationary and one non-stationary part called the *trend*. Mathematically we write it as

$$y_t = \mu_t + x_t,$$

where x_t denotes the stationary part and μ_t the trend. The process of finding μ_t and then computing $x_t = y_t - \mu_t$ is called *detrending*. Detecting the trend can be done in many ways, for instance using regression techniques or smoothing. The simplest way is to assume a linear trend, $\mu_t = \beta_0 + \beta_1 t$ and estimate the parameters using least squares. In figure 2.2 the linear regression fit is shown, showing an upwards in the time series.

2.2 Fourier analysis

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of numbers. For simplicity in notation we assume n to be an odd number. It can be shown that the sequence can be represented as a linear combination of complex exponentials

$$Z_t = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} c_k e^{\frac{i2\pi kt}{n}}. \tag{2.5}$$

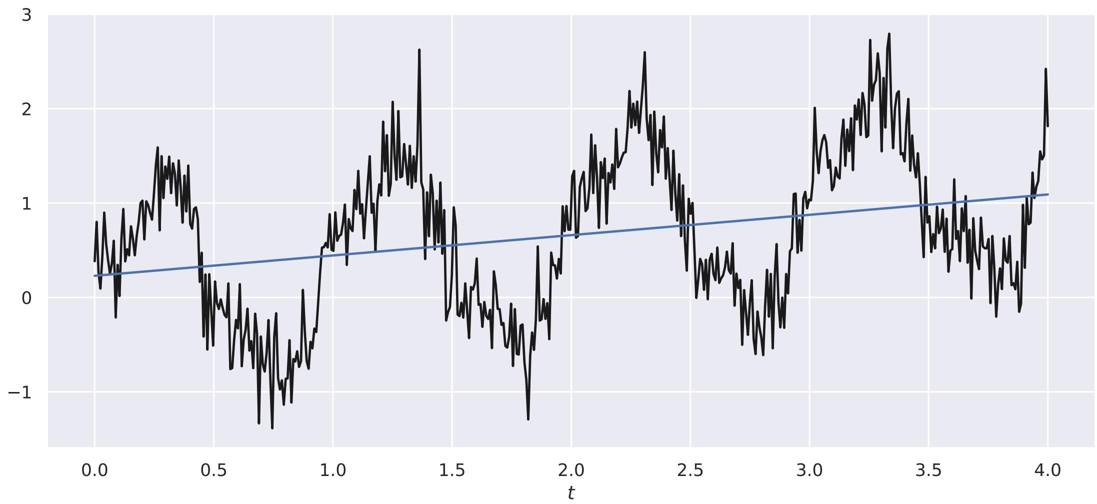

Figure 2.2: Time series from figure 2.1 with an estimated linear trend shown in blue. It is clear that there exists an upward trend in the data.

This comes from the fact that the set

$$\left\{ e^{\frac{i2\pi kt}{n}} \middle| k \in \left[-\frac{n-1}{2}, \frac{n-1}{2} \right] \right\}$$

consists of n orthogonal functions Wei 2006. I.e., that

$$\sum_{i=1}^{n} e^{\frac{i2\pi kt}{n}} e^{-\frac{i2\pi jt}{n}} = \begin{cases} n, & k = j \\ 0, & k \neq j \end{cases}.$$

The coefficients c_k are given by

$$c_k = \frac{1}{n} \sum_{t=1}^{n} Z_t e^{-\frac{i2\pi kt}{n}}.$$

It is clear that (2.5) is periodic with period n, meaning $Z_{t+jn} = Z_t, j = 0, \pm 1, \pm 2, \ldots$. Thus the Fourier series is able to capture periodic sequences. The smallest positive integer n for which $Z_{t+n} = Z_t$ is called the fundamental period, with corresponding fundamental frequency $2\pi/n$. For the components $k = \pm j, j = 1, 2, \ldots, (n-1)/2$ the frequencies are multiples of the fundamental frequency, $w_k = k(2\pi/n)$. The set of frequencies making up the series is called the *spectrum*. As a consequence the coefficients c_k can be viewed as weighting the importance of the contributions for the different frequencies making up the full sequence. This is formalized by the definitions of energy and power,

$$\text{energy} = \sum_{t=2}^{n} Z_t^2 = n \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} |c_k|^2, \tag{2.6}$$

$$\text{power} = \frac{\text{energy}}{n} = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} |c_k|^2. \tag{2.7}$$

Let p_k be the contribution to the power from frequency $k = 0, 1, \ldots, (n-1)/2$. As w_k and w_{-k} corresponds to the same frequency the contribution is given as $p_0 = c_0^2, p_k = 2|c_k|^2, k = 1, \ldots, (n-1)/2$. The values p_k are called the power spectrum of the series.

2.2.1 Discrete-Time Fourier Transform

We have seen that all sequences of length n can be viewed and parameterized as Fourier series with period n. Moving to non-periodic sequences essentially amounts to taking the limit of the series as n approaches infinity. Formally we now let Z_t be a finite discrete function of t, where $Z_t = 0$ when $|t| > M$ for some integer M. Choosing $n = 2M + 1$ the function

$$Y_{t+jn} = Z_t,\ t \in \left[-\frac{n-1}{2}, \frac{n-1}{2}\right],\ j \in \mathbb{Z}$$

is periodic with period n. It's Fourier series is

$$Y_t = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} c_k e^{\frac{i2\pi kt}{n}}.$$

As $Y_t = Z_t$ when $t \in [-(n-1)/2, (n-1)/2]$, and $Z_t = 0$ when $|t| > (n-1)/2$, the coefficients c_k can be written as the infinite sum

$$\begin{aligned} c_k &= \frac{1}{n} \sum_{t=-\infty}^{\infty} Z_t e^{\frac{-i2\pi kt}{n}} \\ &= \frac{2\pi}{n} f\left(\frac{2\pi k}{n}\right), \end{aligned}$$

where

$$f(w) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} Z_t e^{-iwt}.$$

If we now take the limit $Z_t = \lim_{n\to\infty} Y_t$ the summation becomes an integral over the length 2π Wei 2006. This gives the relation

$$Z_t = \int_{-\pi}^{\pi} f(w)e^{iwt}, \quad t \in \mathbb{Z} \tag{2.8}$$

$$f(w) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} Z_t e^{-iwt}, \quad -\pi \leq w \leq \pi, \tag{2.9}$$

where $f(w)$ is called the discrete-time Fourier transform of Z_t.

We call this function $f(w)$ the discrete-time Fourier transform of the sequence Z_t.

2.3 Spectral analysis

2.3.1 Wavelet transformation

Morlet wavelet—a sine wave that is "windowed" (i.e., multiplied point by point) by a Gaussian

2.4 Piecewise polynomials

Suppose we have an interval $[a, b]$ divided into M contiguous subintervals. The connecting edges of the subintervals $a = \xi_0, \xi_1, \dots, \xi_{M-1}, \xi_M = b$ are called knots. On each of the intervals $[\xi_i, \xi_{i+1}], i = 0, \dots, M-1$ we define a polynomial $p_i(t)$. The function

$$f(t) = \begin{cases} p_0(t), & t \in [\xi_0, \xi_1) \\ p_1(t), & t \in [\xi_1, \xi_2) \\ & \vdots \\ p_{M-1}(t), & t \in [\xi_{M-1}, \xi_M] \end{cases}$$

is called a *piecewise polynomial*.

2.4.1 Splines

In the definition of piecewise polynomials no restrictions are made on the polynomials, they are allowed to take any form. As in Quarteroni, Sacco, and Saleri 2010 we define a *spline* $s_k(t)$ of order k on the interval $[a, b]$ as a piecewise polynomial where

$$\begin{aligned} &s_k(t) \in \mathcal{P}^k, \quad t \in [\xi_i, \xi_{i+1}], \quad i = 0, 1, \dots, M-1 \\ &s_k(t) \in \mathcal{C}^{k-1}[a, b]. \end{aligned}$$

I.e., the spline consists of piecewise polynomials of order k and has continuous derivatives up to order $k-1$. A common choice is letting $k = 3$, providing continuous second derivatives over the interval. This is called *cubic* splines, and are often considered sufficiently smooth for function approximations. It is also common to add curvature constraints at the endpoints, $s_3''(a) = s_3''(b)$, arriving at the *natural* cubic splines.

2.4.2 Regression splines

Suppose now we have data points $y_{t_1}, y_{t_2}, \dots, y_{t_n}$ on $[a = t_1, b = t_n]$. A spline of order k with chosen knots at $a = t_1 = \xi_0, \xi_1, \dots, \xi_M = t_n = b$ can be parameterized as

$$s_k(t) = \sum_{i=1}^{M+K} \beta_i h_i(t),$$

where the functions h_i are the truncated-power basis set

$$\begin{aligned} h_j(t) &= t^{j-1}, \ j = 1, \dots, k+1, \\ h_{k+1+l}(t) &= (t - \xi_l)_+^k, \ l = 1, \dots, M-1, \end{aligned}$$

with $(t)_+ = \max\{t, 0\}$ Hastie, Friedman, and Tisbshirani 2017. The parameters β_i can be found using least squares. An example of cubic spline regression are shown in figure 2.3.

2.5 Dimensionality reduction

2.5.1 Principal Component Analysis

2.5.2 t-Stochastic Neighbor Embedding

2.6 Watershed segmentation

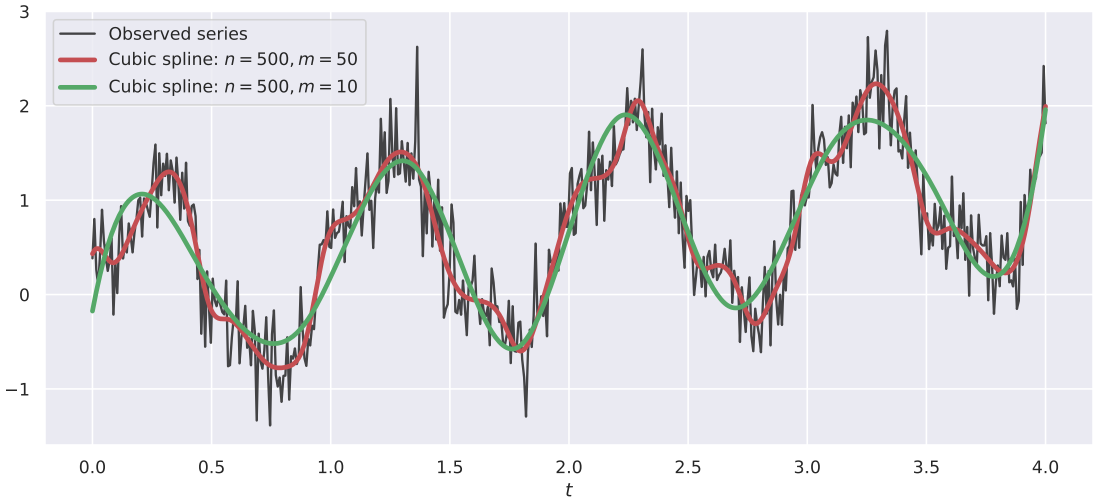

Figure 2.3: Two cubic splines fitted using least square regression on the time series from figure 2.1. Observe that the red spline with 50 knots (including endpoints) fits the data closer than the green spline with 10 knots.

Chapter 3

Methodology

What is done in practice. Discussion of choices made.

3.1 Input data

The starting point of the analysis is a collection of n separate time series of equal length m right?. We denote these $\mathbf{Y} = \{y_1(t), \ldots, y_n(t)\}$, where $t \in \{t_1, t_2, \ldots, t_m\}$.

3.2 Feature extraction

3.3 Manifold embedding

Bibliography

Hastie, Trevor, Jerome Friedman, and Robert Tisbshirani (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction.* 2nd ed. Springer.

Quarteroni, Alfio, Riccardo Sacco, and Fausto Saleri (2010). *Numerical mathematics.* Springer.

Shumway, Robert H. and David S. Stoffer (2017). *Time series analysis and its applications: With R examples.* 4th ed. Springer International Publishing.

Wei, William W.S. (2006). *Time series analysis univariate and multivariate methods.* 2nd ed. Pearson.