

Specialization Project

Ulrik Bernhardt Danielsen

November 15, 2022

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Previous work	3
2	Theory	5
2.1	Time series analysis	5
2.1.1	Stationarity	5
2.1.2	Detrending	6
2.2	Fourier analysis	7
2.2.1	Discrete-Time Fourier Transform	8
2.3	Spectral density estimation	9
2.3.1	Periodogram	9
2.3.2	Sample spectrum	10
2.3.3	Spectral Window	10
2.3.4	Lag window	11
2.4	Time-Frequency Analysis	12
2.4.1	Short-Time Fourier Transform	12
2.4.2	Wavelet Transform	13
2.5	Piecewise polynomials	14
2.5.1	Splines	15
2.5.2	Regression Splines	15
2.6	Dimensionality Reduction	15
2.6.1	Principal Component Analysis	16

2.6.2	t-Stochastic Neighbor Embedding	16
2.7	Kernel Density Estimation	17
2.8	Watershed Segmentation	18
3	Methodology	21
3.1	Feature extraction	21
3.1.1	Detrending	21
3.1.2	Time Frequency Analysis	21
3.2	Manifold embedding	23
4	Discussion	25
4.1	Simulated Data	25
4.2	Benefits of t-SNE	25
4.3	Tuning the perplexity parameter	26
4.4	Tuning the bandwidth matrix	26
Bibliography		29

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 $Y(\omega, t)$, where $\omega \in \Omega$, Ω being the sample space, and $t \in \mathbb{Z}$, \mathbb{Z} being the chosen index set [1]. 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 n th-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 n th-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,

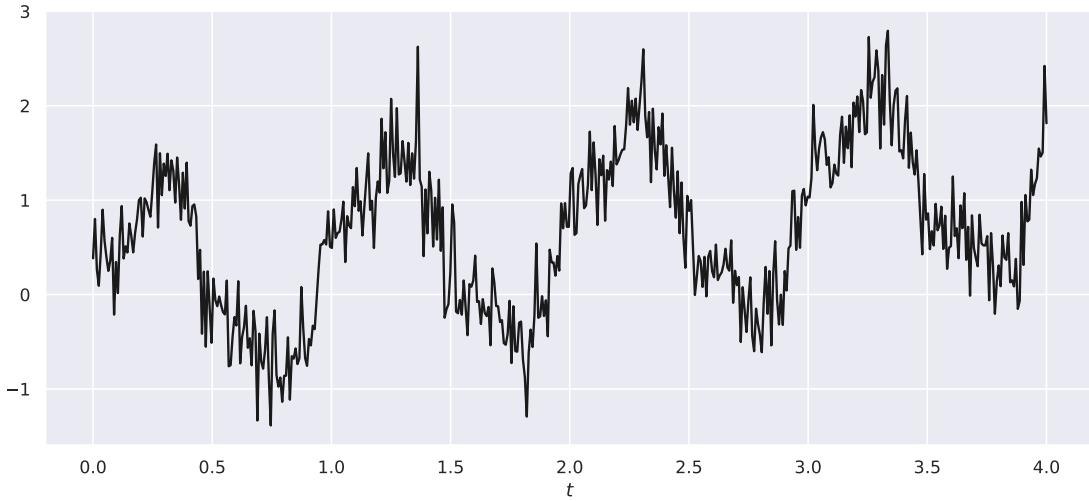


Figure 2.1: Example of a non-stationary time series with $t_0 = 0, t_n = 4$. The discrete data points are connected to better visualize the movement through time.

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.

We also define the common estimates for the mean and covariance functions. They are called the sample mean and sample covariance, and are written as

$$\bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_{t_i}, \quad (2.5)$$

$$\hat{\gamma}_h = \frac{1}{n} \sum_{i=1}^{n-h} (y_{t_i} - \bar{y}_t)(y_{t_i+h} - \bar{y}_t), \quad (2.6)$$

respectively.

2.1.2 Detrending

Many methods for analysing and processing time series requires stationarity [2]. 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.

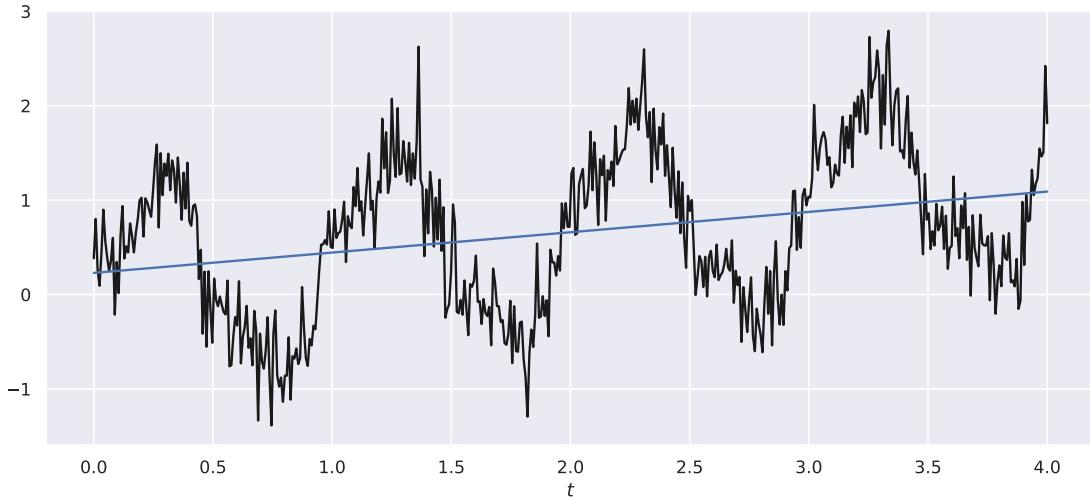


Figure 2.2: Time series from figure 2.1 with an estimated linear trend shown in blue.

It is clear that there exists an upward trend in the data.

2.2 Fourier analysis

Let Z_1, Z_2, \dots, 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

$$Z_t = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} c_k e^{\frac{i2\pi kt}{n}}. \quad (2.7)$$

This comes from the fact that the set

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

consists of n orthogonal functions [1]. I.e., that

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

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 (2.7) is periodic with period n , meaning $Z_{t+jn} = Z_t, j = 0, \pm 1, \pm 2, \dots$. 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 $k = \pm j, j = 1, 2, \dots, (n-1)/2$ the frequencies are multiples of the fundamental frequency, $\omega_k = k(2\pi/n)$. 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,

$$\text{energy} = \sum_{t=2}^n Z_t^2 = n \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} |c_k|^2, \quad (2.8)$$

$$\text{power} = \frac{\text{energy}}{n} = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} |c_k|^2. \quad (2.9)$$

Let p_k be the contribution to the power from frequency $k = 0, 1, \dots, (n-1)/2$. As ω_k and ω_{-k} corresponds to the same frequency the contribution is given as $p_0 = c_0^2, p_k = 2|c_k|^2, k = 1, \dots, (n-1)/2$. 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

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

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

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

If we now take the limit $Z_t = \lim_{n \rightarrow \infty} Y_t$ the summation becomes an integral over the length 2π [1]. This gives the relation

$$Z_t = \int_{-\pi}^{\pi} f(\omega) e^{i\omega t} d\omega, \quad t \in \mathbb{Z} \quad (2.10)$$

$$f(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} Z_t e^{-i\omega t}, \quad -\pi \leq \omega \leq \pi, \quad (2.11)$$

where $f(\omega)$ in (2.11) is called the discrete-time Fourier transform of Z_t . As opposed to the periodic case (2.7) where the periodic sequence was made up by a finite number of frequencies, the non-periodic sequence is an integral over a continuum of frequencies ω . We call $|f(\omega)|$ the spectrum of the sequence, and the function $g(\omega) = 2\pi|f(\omega)|^2$ the energy spectrum. The energy spectrum definition comes from Parseval's relation

$$\sum_{t=-\infty}^{\infty} |Z_t|^2 = 2\pi \int_{-\pi}^{\pi} |f(\omega)|^2 d\omega. \quad (2.12)$$

It is worth noting that the relation in (2.12) only holds when the sequence Z_t is absolutely summable, i.e., that

$$\sum_{t=-\infty}^{\infty} |Z_t| < \infty. \quad (2.13)$$

2.3 Spectral density estimation

Let y_t be a stationary time series where the autocovariance function γ_h from (2.2) is absolutely summable. Then we can write γ_h as a Fourier transform pair

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_h e^{-i\omega h}, \quad (2.14)$$

$$\gamma_h = \int_{-\pi}^{\pi} f(\omega) e^{i\omega h} d\omega. \quad (2.15)$$

It can be shown [2] that the spectrum $f(\omega)$ in (2.14) is real-valued and non-negative. Furthermore, as $\text{Var}(y_t) = \gamma_0$, we get the interpretation

$$\text{Var}(y_t) = \int_{-\pi}^{\pi} f(\omega) d\omega,$$

i.e., that $f(\omega)$ is the contribution to the variance for frequency ω . We often want to locate these important frequencies, and thus an important task is to estimate this spectrum.

2.3.1 Periodogram

Again we consider a time series sample y_1, y_2, \dots, y_n where n is chosen to be odd for simplicity. It can be written as a real Fourier representation

$$y_t = a_0 + \sum_{k=1}^{\frac{n-1}{2}} (a_k \cos(\omega_k t) + b_k \sin(\omega_k t)),$$

where $\omega_k = 2\pi k/n$, $k = 0, 1, \dots, (n-1)/2$ and the coefficients are given as

$$a_0 = \bar{y}_t, \quad a_k = \frac{2}{n} \sum_{t=1}^n y_t \cos(\omega_k t), \quad b_k = \frac{2}{n} \sum_{t=1}^n y_t \sin(\omega_k t).$$

We then define the periodogram as

$$I(\omega_k) = \begin{cases} n a_0^2, & k = 0 \\ \frac{n}{2} (a_k^2 + b_k^2), & k = 1, \dots, (n-1)/2. \end{cases} \quad (2.16)$$

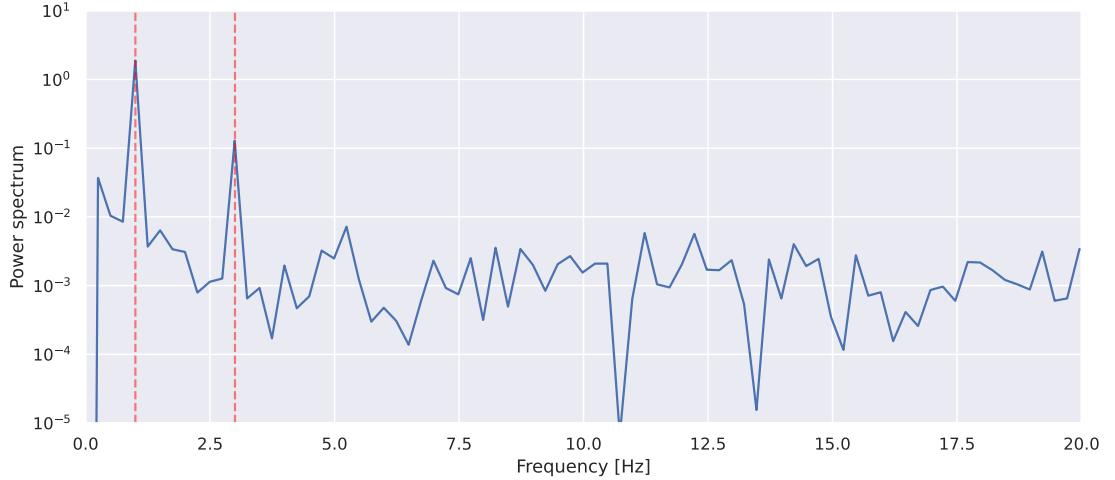


Figure 2.3: Periodogram for the time series plotted in figure 2.1. The time series is detrended assuming a linear trend. Observe that the two highest peaks in the periodogram are at the underlying frequencies 1Hz and 3Hz shown with dashed red lines.

The periodogram is of interest as it has a large value if the frequency ω_k is of importance in the series. A scaled periodogram $\frac{2}{n} I(\omega_k)$ estimates the sample variance of the sinusoid component at frequency ω_k [2]. A periodogram for the time series example in figure 2.1 are shown in figure 2.3. Note that the time series is linearly detrended for the periodogram is computed. The two highest peaks match the underlying generating function which can be revealed to be $\sin(2\pi t) + \frac{1}{4} \cos(6\pi t) + t/3$.

2.3.2 Sample spectrum

One intuitive way of estimating the spectrum is to replace the autocovariance with the sample autocovariance from equation (2.6). I.e., we define the sample spectrum for a realization y_1, y_2, \dots, y_n

$$\hat{f}(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \hat{\gamma}_k e^{-i\omega k}. \quad (2.17)$$

At the Fourier frequencies ω_k it is related to the periodogram through [1]

$$\hat{f}(\omega_k) = \frac{I(\omega_k)}{4\pi}.$$

Although $\hat{f}(\omega_k)$ is asymptotically unbiased, meaning $\lim_{n \rightarrow \infty} E(\hat{f}(\omega)) = f(\omega)$, it lacks consistency in the variance as n tends to infinity, i.e.,

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{f}(\omega_k)) \neq 0.$$

2.3.3 Spectral Window

The fact that the variance does not decrease with the sample size produces quite jagged and noisy spectrum estimations [2]. To account for this we introduce the spectral window which

smooths the spectrum. Mathematically the smoothed spectrum is written as

$$\hat{f}_{\mathcal{W}}(\omega_k) = \sum_{j=-m}^m \mathcal{W}_n(\omega_j) \hat{f}(\omega_k - \omega_j), \quad (2.18)$$

where $\omega_k = 2\pi k/n$, $k = 0, 1, \dots, (n-1)/2$ and m is a function of n , typically $m \ll n$. The value of m decides how many points in the neighborhood of ω_k should be included in the smoothing. Furthermore, the function $\mathcal{W}_n(\omega_j)$ is chosen to have the following properties,

$$\begin{aligned} \sum_{j=-m}^m \mathcal{W}_n(\omega_j) &= 1, \\ \mathcal{W}_n(\omega_j) &= \mathcal{W}_n(-\omega_j), \\ \lim_{n \rightarrow \infty} \sum_{j=-m}^m \mathcal{W}_n^2(\omega_j) &= 0. \end{aligned} \quad (2.19)$$

We view the smoothed spectrum as a weighted average of the sample spectrum (2.17) in a window around the target frequency ω_k . How the weights are distributed in the window is governed by $\mathcal{W}_n(\omega_j)$, giving it the name spectral window. Because of the property in (2.19) we have

$$\begin{aligned} \text{Var}(\hat{f}_{\mathcal{W}}(\omega_k)) &\approx \sum_{j=-m}^m \mathcal{W}_n^2(\omega_j) (f(\omega_k))^2 \\ &= (f(\omega_k))^2 \sum_{j=-m}^m \mathcal{W}_n^2(\omega_j) \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

assuming $f(\omega)$ is approximately constant in the window. We can thus reduce the variance by increasing the points included in the window. However, when doing so we introduce bias.

2.3.4 Lag window

It can be shown [1] that the spectral window forms a Fourier transform pair with a lag window $W_n(k)$, i.e,

$$W_n(k) = \int_{-\pi}^{\pi} \mathcal{W}_n(\omega) e^{i\omega k} d\omega, \quad k = 0, \pm 1, \dots, \pm M.$$

The lag window is a weighting function applied to the sample autocovariance

$$\hat{f}_{\mathcal{W}} = \frac{1}{2\pi} \sum_{k=-M}^M W_n(k) \hat{\gamma}_k e^{-i\omega k},$$

with $W_n = W(k/M)$, W being a bounded even continuous function

$$\begin{aligned} |W(t)| &\leq 1, \\ W(0) &=, \\ W(t) &= W(-t), \\ W(t) &=, \quad t > 1. \end{aligned}$$

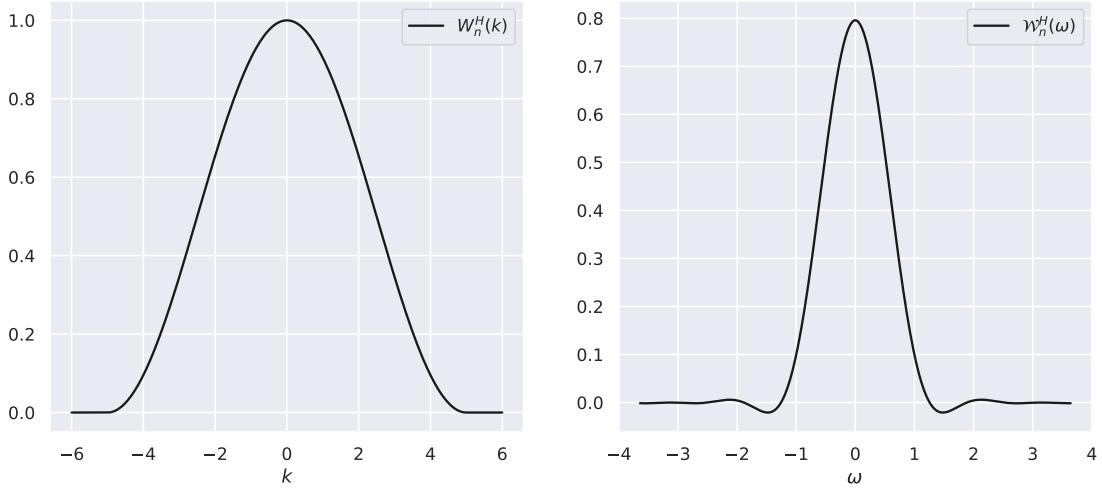


Figure 2.4: On the left the hanning window given by equation (2.20). On the right the corresponding spectral window, i.e., its Fourier transform.

Figure 2.4 shows the popular *hanning* window given by

$$W_n^H = \begin{cases} \frac{1}{2}(1 + \cos(\frac{\pi k}{M})), & |k| \leq M \\ 0, & |k| > M, \end{cases} \quad (2.20)$$

with $M = 5$.

2.4 Time-Frequency Analysis

The methods mentioned so far are only suitable for extracting information about the frequencies of a signal. If the process is stationary this is often all we need. However, if the process is non-stationary and the frequencies change over time, the periodogram is unable to capture this. The original time series contains all the information in the time-scale, and the Fourier transform contains all the information in the frequency scale. We need something in between.

2.4.1 Short-Time Fourier Transform

A simple and intuitive way to gain information about the changes in frequency is to divide the time interval into sections, and compute the Fourier transform on each section. Then we can plot how the periodogram changes over time. If we divide the time interval using a window function this method is called the short-time Fourier Transform (STFT). Mathematically we define it for a discrete sequence $y = y_1, y_2, \dots, y_n$, which is windowed by $W_n(t)$ around time τ , as

$$S_y(\omega, \tau) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} y_t W_n(t - \tau) e^{-i\omega t}. \quad (2.21)$$

The estimated spectrum $|S_y(\omega, \tau)|^2$ is now a function of both time and frequency, and is commonly called the spectrogram [3], which can be plotted to show the changes in frequencies over time.

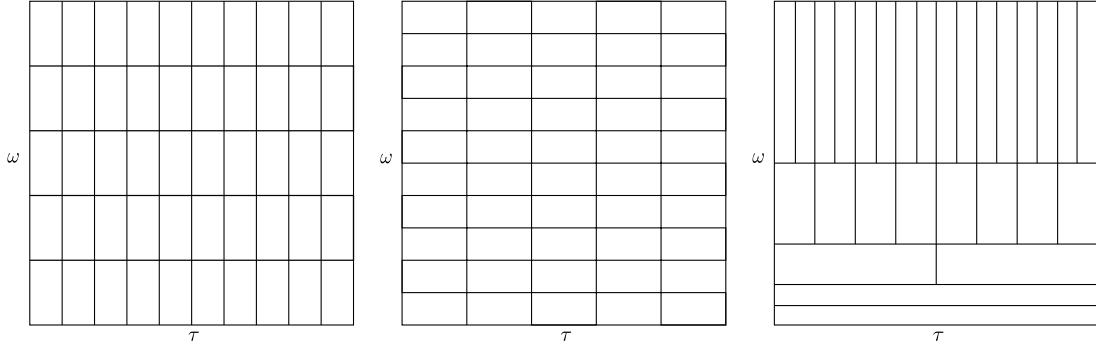


Figure 2.5: *Left/middle*: The figures illustrates the tradeoff in resolution for time and frequency for the short-time Fourier transform from equation (2.21), increasing the resolution in time decreases the resolution in frequency and vice versa. *Right*: Illustrate the corresponding resolution for the continuous wavelet transform from equation (2.22).

Unfortunately the STFT has some severe drawbacks; it can be shown that there exists a limit for the possible precision achieved when measuring both frequency and time simultaneously [4]. There exists a tradeoff in time and frequency resolution, as shown in figure 2.5. This is very intuitive as frequency has to be measured over a certain time period. Decreasing the width of the window function makes detecting smaller frequencies harder, and vice versa. Thus we need to know the approximate frequency scales in the data before performing the STFT, or we can try several window widths. In the case where frequencies exists in the data on several different scales, the STFT in equation (2.21) becomes unsuitable.

2.4.2 Wavelet Transform

The wavelet transform works in a very similar way as the STFT, but circumvents the problem of choosing a fixed width of the window by introducing scaling. Let $\psi(t)$ be a window function (possibly complex valued) which we will call the mother wavelet. One way of scaling the mother wavelet by a factor s is

$$\psi_s(t) = \frac{1}{s^{1/2}} \psi\left(\frac{t}{s}\right).$$

If we also allow a time shift around τ we arrive at the wavelets

$$\psi_{s,\tau}(t) = \frac{1}{s^{1/2}} \left(\frac{t - \tau}{s} \right).$$

This culminates in the definition of the continuous wavelet transform (CWT) of a function f [4]

$$\tilde{f}(s, \tau) \int_{-\infty}^{\infty} \psi_{s,\tau}^*(t) f(t) dt, \quad (2.22)$$

where $\psi_{s,\tau}^*$ is the complex conjugate of the function $\psi_{s,\tau}$.

Again considering the discrete sample y_1, y_2, \dots, y_n with $\delta t = y_{t+1} - y_t$, $t = 1, \dots, n$, its continuous wavelet transform is given by [5]

$$\tilde{f}_n(s, \tau) = \sum_{t=1}^n y_t \psi^* \left(\frac{\delta t}{s} (t - \tau) \right), \quad (2.23)$$

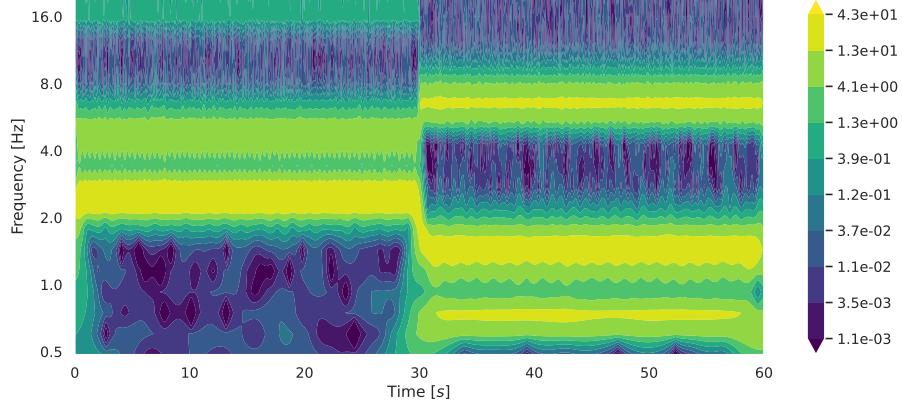


Figure 2.6: An example of a scaleogram of the continuous wavelet transform of function (2.25) with added Gaussian noise.

where τ now is a time index $\tau \in \{1, 2, \dots, n\}$. It is computationally efficient [5] to represent the CWT as the inverse Fourier transform

$$\begin{aligned} \tilde{f}_n(s, \tau) &= \sum_{t=1}^n \hat{y}_k \hat{\psi}^*(s\omega_k) e^{i\omega_k n \delta t}, \\ \hat{y}_k &= \frac{1}{n} \sum_{t=1}^n y_t e^{-\frac{i2\pi k t}{n}}, \\ \omega_k &= \begin{cases} \frac{2\pi k}{n\delta t}, & k \leq \frac{n}{2} \\ -\frac{2\pi k}{n\delta t}, & k > \frac{n}{2} \end{cases}. \end{aligned}$$

As always we are interested in the power spectrum, which for the CWT is defined as $|\tilde{f}_n(s, \tau)|^2$.

There exists many choices for the mother wavelet, and certain important criteria it must satisfy [4]. One popular wavelet is the Morlet wavelet defined as

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2}, \quad (2.24)$$

where ω_0 is a parameter chosen to fit the criteria.

In figure 2.6 a scaleogram is shown for the function

$$f(t) = \begin{cases} 3 \sin(2.4 \cdot 2\pi t) + 2 \sin(4.7 \cdot 2\pi t) + \sin(17 \cdot 2\pi t), & t < 30 \\ \sin(0.7 \cdot 2\pi t) + 2 \sin(1.4 \cdot 2\pi t) + 3 \sin(6.4 \cdot 2\pi t), & t \geq 30, \end{cases} \quad (2.25)$$

with added Gaussian noise. The continuous wavelet transform is computed at 18 frequencies between 0.5Hz and 18 Hz.

2.5 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*.

2.5.1 Splines

In the definition of piecewise polynomials no restrictions are made on the polynomials, they are allowed to take any form. As in [6] 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*.

2.5.2 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

$$s_k(t) = \sum_{i=1}^{M+k} \beta_i h_i(t), \quad (2.26)$$

where the functions h_i are the truncated-power basis set

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

with $(t)_+ = \max\{t, 0\}$ [7]. The parameters β_i can be found using least squares. An example of cubic spline regression are shown in figure 2.7.

2.6 Dimensionality Reduction

Suppose we have n datapoints, each having p numerical features. Mathematically we represent them as the vectors $x_i = (x_{i1}, x_{i2}, \dots, x_{in}) \in \mathbb{R}^p$, $i = 1, 2, \dots, n$, often represented in the $n \times p$ data matrix X , $X_{ij} = x_{ij}$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$. Dimensionality reduction is about finding a low dimensional representation of the data still containing most of its properties. There can be many reasons for performing dimensionality reduction, and often it is performed as a form of feature extraction used before predictive modelling or clustering [7]. If the data is sparse or the features highly correlated, reducing the dimensionality could improve computation speed and remove noise. Reducing the data to two dimensions often allows for easier visualization and interpretability of the data.

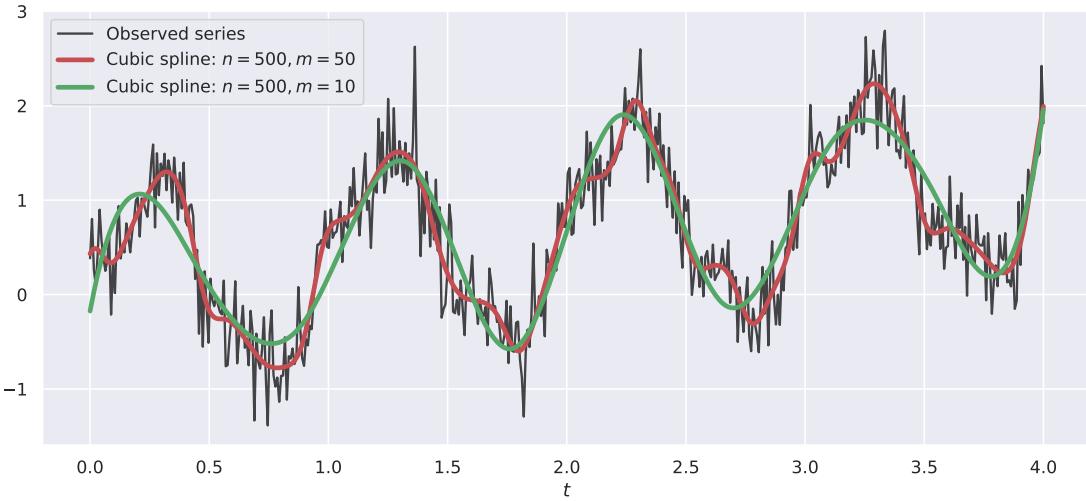


Figure 2.7: Two cubic splines fitted using least square regression on the time series from figure 2.1. Observe that the red spline with 50 knots (including endpoints) fits the data closer than the green spline with 10 knots.

2.6.1 Principal Component Analysis

One of the most commonly used dimensionality reduction methods is called principal component analysis (PCA). There are multiple ways to both derive and interpret the method. Given the $n \times p$ data matrix X we find a sequence of $q \leq p$ orthogonal $p \times 1$ unit vectors v_1, v_2, \dots, v_q , each one chosen such that $Z_j = Xv_j$ has maximum variance. On matrix form we write it as $Z = X V_q$, where the columns of V_q are $v_j, j = 1, 2, \dots, q$. The principal components scores—which represents our new features when applying dimensionality reduction—is given by the q columns in Z . Let S be the sample covariance matrix of X . It can be shown [8] that the vectors v_j that maximises the variance while subject to the mentioned constraints, are the eigenvectors of S , with the corresponding eigenvalue λ_j . Note that $\lambda_i > \lambda_{i+1}, i = 1, 2, \dots, p - 1$. For this reason we need to standardize the columns of X to have mean zero and variance one before applying PCA. The proportion of the total variance explained by the first q principal components is given by

$$\frac{\sum_{i=1}^q \lambda_i}{\sum_{j=1}^p \lambda_j}.$$

Thus when using PCA for dimensionality reduction we set a variance threshold, and keep the number of principal components keeping enough variance.

2.6.2 t-Stochastic Neighbor Embedding

Another dimensionality reduction method specially suited for visualizing data in two dimension, is called t-Stochastic Neighbor Embedding (t-SNE). Again we start with the $n \times p$ data matrix, and wish to embed it into a low dimensional space preserving as much of the original structure in the data as possible. Instead of trying to emulate the Euclidian distances between the data

points x_1, x_2, \dots, x_n , we convert them into conditional probabilities

$$p_{j|i} = \frac{\exp\left\{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right\}}{\sum_{k \neq i} \exp\left\{-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right\}}, \quad i \neq j, \quad p_{i|i} = 0. \quad (2.27)$$

For a point x_i , $p_{j|i}$ is large for nearby points x_j , the magnitude proportional to a Gaussian distribution centered at x_i with variance σ_i^2 [9]. The individual variance σ_i^2 is a tuning parameter to be chosen by the user.

In the low dimensional representation we model the similarities as joint probabilities based on the Student t-distribution with one degree of freedom

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}, \quad i \neq j, \quad q_{ii} = 0. \quad (2.28)$$

It is beneficial to define a joint probability distribution in the high dimensional space as with $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$ [9]. To find the low dimensional embedding Y , t-SNE minimizes the total Kullback-Leibler divergence between the discrete joint probability distributions

$$KL(P||Q) = \sum_{i=1}^n \sum_{j=1}^n p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \quad (2.29)$$

The choice of using the Student-t distribution in equation (2.28) as opposed to a Gaussian (as in the high dimensional space (2.27)), stems from its heavier tails. This allows data points which are relatively dissimilar in the original data, but still close enough to make a difference in the cost function, to be modelled far apart in the low dimensional embedding. I.e., t-SNE focusses on the local structures in the data, and allows clusters to be well separated in the embedding.

Furthermore, we need to choose the tuning parameters σ_i . They relate to the Gaussian distribution modelling the neighbors of x_i . For points with many close by neighbors it makes sense to keep σ_i small, while increasing it is beneficial when there are few neighbors. The algorithm solves this by letting the user choose a perplexity parameter

$$\text{Perp}(P_i) = 2^{H(P_i)}, \quad (2.30)$$

where

$$H(P_i) = - \sum_{j=1}^n p_{j|i} \log_2(p_{j|i})$$

is the Shannon entropy. Then it finds σ_i such that the perplexity satisfies our choice. Van der Maaten and Hinton [9] suggests values between 5 and 50, roughly reflecting the number of neighbors in the algorithm.

2.7 Kernel Density Estimation

Let $x_1, x_2, \dots, x_n, x_i \in \mathbb{R}^p$ be a random sample from a random variable with an underlying unknown density function $f(x)$, which we want to estimate. The general expression for the multivariate kernel density estimator is [10]

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_p \left[H^{-1}(x - x_i) \right], \quad (2.31)$$

where $|H| = |\det(H)|$, H being the $p \times p$ non-singular bandwidth matrix. The function $K_p : \mathbb{R}^p \rightarrow \mathbb{R}$ in (2.31) is called the kernel function, and must satisfy

$$\int_{\mathbb{R}^p} K_p[u]du = 1, \quad \int_{\mathbb{R}^p} uK_p[u]du = 0 \in \mathbb{R}^p, \quad \int_{\mathbb{R}^p} uu^T K_p[u]du = I_p,$$

I_p being the $p \times p$ identity matrix [10]. We can view $\hat{f}(x)$ as a mixture of densities centered at the observations x_1, x_2, \dots, x_n . The bandwidth matrix H needs to be selected by the user. Changing the bandwidth matrix has a large impact on the resulting estimated density, and is often chosen to suit the problem at hand. An example of kernel density estimation is shown in figure 2.8.

2.8 Watershed Segmentation

Watershed segmentation is an image processing technique commonly used to separate objects in an image. Assuming a greyscale image it treats it like a topological map where intense areas are considered to have large altitude [11]. Using an estimated gradient the algorithm finds the local minima in the image, and gradually "flood" the topography with water and mark the lines where the filled basins meet. The watershed segmentation applied to an estimated probability density function is shown in figure 2.8.

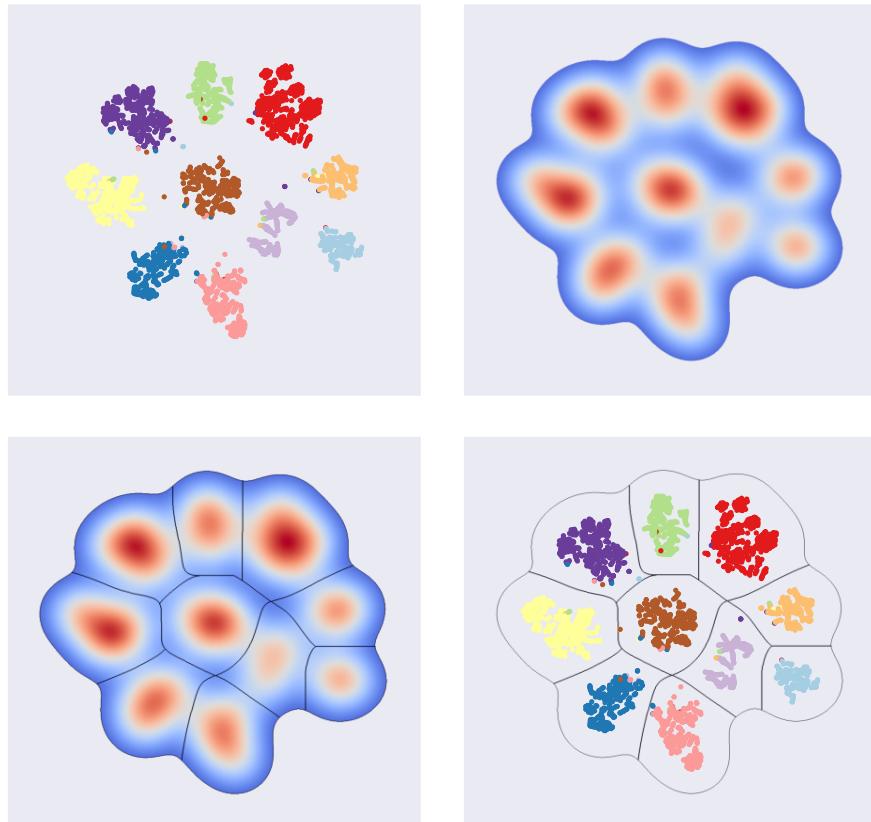


Figure 2.8: Clustering the two-dimensional t-SNE embedding. *Top left:* t-SNE embedding of simulated data, color coded by actual behavior. *Top right:* Kernel density estimation of the t-SNE embedding. *Bottom left:* Watershed segmentation applied to the kernel density estimation. *Bottom right:* Contours of the watershed segmentation with the t-SNE clusters shown inside.

Chapter 3

Methodology

The starting point of the analysis is a set $\mathbf{Y} = \{Y_d : d = 1, 2, \dots, D\}$, where each set Y_d is again a set of independent time series, $Y_d = \{y_1^d(t), y_2^d(t), \dots, y_n^d(t) : t \in \{t_1, t_2, \dots, t_{m_d}\}\}$. Each Y_d represents an animal for which n time series are collected tracking parts of its movements. Recording frequency is the same across animals. The goal of the analysis is to cluster these time points into distinct distinguishable actions.

Some summarization: "This is achieved by..."

3.1 Feature extraction

3.1.1 Detrending

The first step is to analyse each time series separately, modifying them and thus creating $D \times n$ features which we will extract the behavioral information from. Let $y_t = \{y_1, y_2, \dots, y_m\} = y_i^d(t)$ for some $d \in \{1, 2, \dots, D\}$ and $i \in \{1, 2, \dots, n\}$ be one such series.

I don't see the point in centering the data around the mean before the detrending.

We detrend the data as $y_t = s_3(t) + x_t$, $s_3(t)$ being the non-linear trend and x_t the detrended time series. The cubic spline $s_3(t)$ is found using least square regression as in equation (2.26) with equally spaced internal knots $\xi_1, \xi_2, \dots, \xi_M$. We choose the knots to have a fixed frequency, i.e., we choose $\Delta\xi = \xi_{i+1} - \xi_i, i = 2, \dots, M$. Figure 3.1 shows an example of a detrended time series using cubic splines with interior knots every 240th time point.

3.1.2 Time Frequency Analysis

To make the extracted features comparable, we normalize the time series by dividing it by its standard deviation

$$\sigma_x = \frac{1}{m} \sum_{t=1}^m \left(x_t - \frac{1}{m} \sum_{i=1}^m x_i \right)^2.$$

The detrended time series contain information about how the movements vary around the trend. We assume this movement is repetitive in nature, and can be interpreted as a signal. By applying time frequency analysis to the detrended time series, we obtain additional information about how

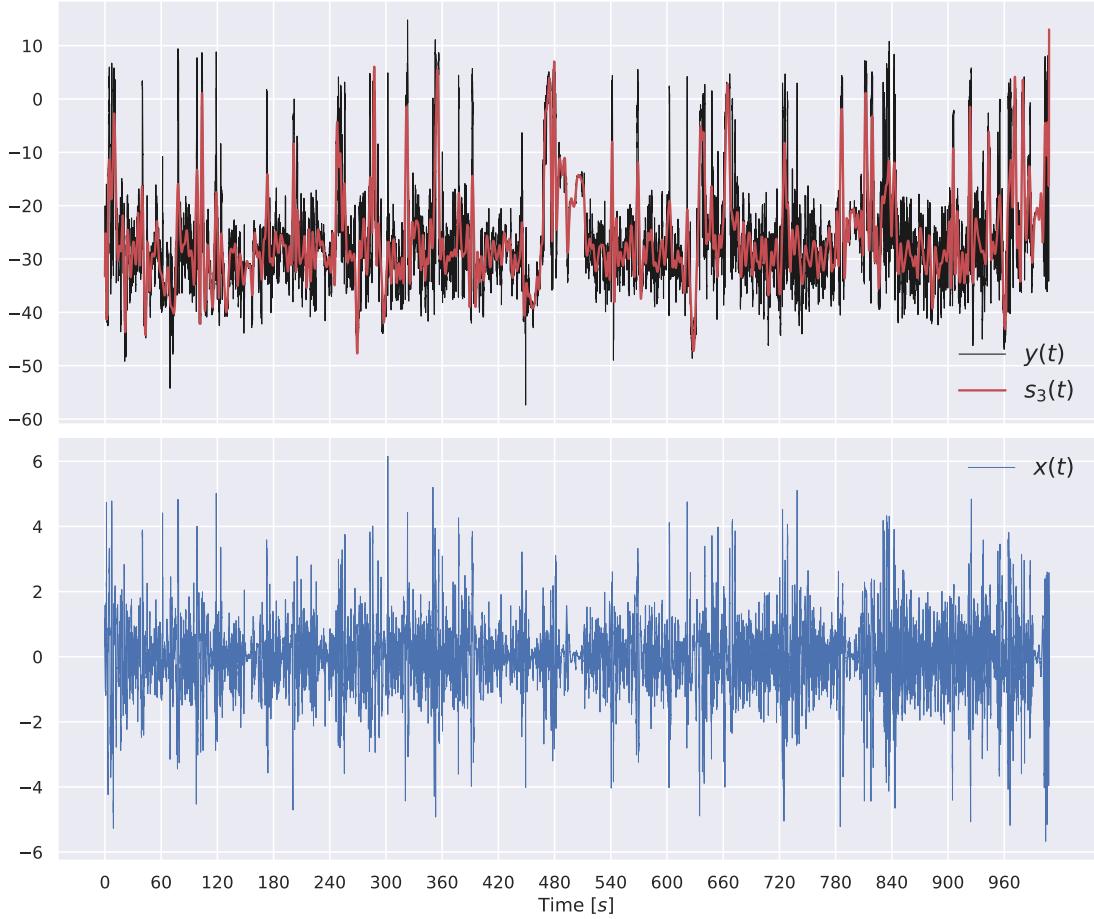


Figure 3.1: Detrended time series using cubic spline regression. The top figure shows the original time series in black with the fitted trend in red. The bottom figure shows the detrended time series.

the frequencies of these movements vary over time. In turn this can contribute to detecting the behavioural clusters.

As we are unaware of at which scales these movements occur beforehand, we opt for a continuous wavelet transform as opposed to the short-time Fourier transform. Let J be the number of predefined frequency scales, ranging from ω_{\min} to ω_{\max} . The scales s_1, s_2, \dots, s_J in equation (2.23) are chosen as fractional powers of two—as suggested in [5]—e.g.,

$$s_j = \frac{2^{(j-1)\delta j}}{\omega_{\max}}, \quad j = 1, 2, \dots, J,$$

$$\delta j = \frac{1}{J-1} \log_2 \left(\frac{\omega_{\max}}{\omega_{\min}} \right).$$

Choice of mother wavelet can for instance be based on expected features in the time series [5]. Using equation (2.23) we compute the discrete continuous wavelet transform $\tilde{f}_m(s, \tau)$, and in

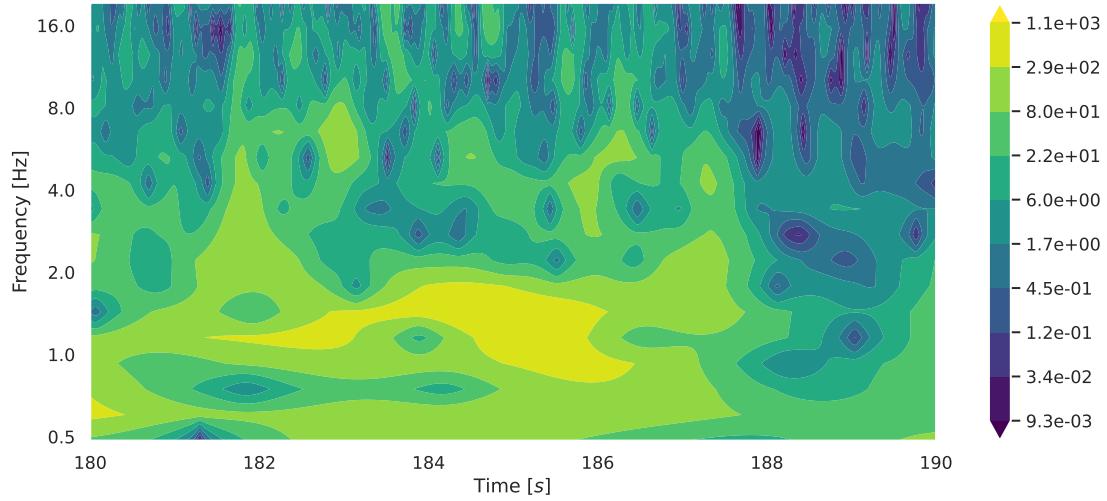


Figure 3.2: Scaleogram of the continuous wavelet transform performed on the detrended time series shown in figure 3.1. The figure shows how the power at different frequency scales changes over time.

turn the scaled scaleogram

$$\frac{1}{s_j} \left| \tilde{f}_m(s_j, t) \right|^2, \quad j = 1, 2, \dots, J, \quad t = 1, 2, \dots, m. \quad (3.1)$$

The scaling is done to keep the power comparable across the varying scales [12]. An example plot of such a scaleogram is shown in figure 3.2.

[There is more done here, square root etc..](#)

3.2 Manifold embedding

After performing the time-frequency analysis on all time series for all animals, we concatenate the results into a new feature vector. For animal d , we have m_d time points, each containing information about the n trend values, together with the $n \cdot J$ frequency powers. The concatenated feature vector thus have dimension $(m_1 + m_2 + \dots + m_D) \times (n \cdot (J + 1))$. To cluster these time points into behaviours, we first reduce the feature dimension. It is logical that there should exist strong correlations between the various frequencies. [Insert discussion about why we do PCA first. Computation time for the t-SNE?](#)

Chapter 4

Discussion

4.1 Simulated Data

To exemplify and visualize the impact of various implementation choices, we make use of simulated data inhibiting the properties we assume to be true for animal behaviour. We simulate one animal with five recorded time series features, each over ten minutes with a recording frequency of 120Hz. We create ten distinct behaviors. For each behavior and feature, the underlying feature is generated as a superposition of four sine waves

$$\sum_{i=1}^4 a_i \sin(\omega_i 2\pi t).$$

The $10 \times 5 \times 4$ frequencies ω are randomly generated uniformly between 0.5Hz and 20Hz. Similarly, the amplitudes a are randomly generated as a lognormal distribution with parameters $\mu_a = 1, \sigma_a = 0.5$. A Gaussian noise is added with $\mu_n = 0, \sigma_n = 0.2$. The length of the behaviors are modelled to follow an exponential distribution with mean $\lambda = 3$. This is equivalent to generate $10 \cdot 60/3 = 200$ time points uniformly over $[0, 600]$, each interval assigned a random behavior. Figure 4.1 shows one such feature, colored by the underlying behaviour at the first 50 seconds.

4.2 Benefits of t-SNE

There are many popular methods for embedding a high dimensional dataset into a lower dimensional space. Principal component analysis is discussed in section 2.6.1, and is a good method to use for variance reduction. It is also commonly used for visualizing data in two or three dimensions. A large caveat is its linearity in the original features, which makes it unable to capture non-linear effects in the data. Another common method is multidimensional scaling (MDS) [7]. It preserves the high dimensional distances as euclidean distances in the low dimensional space. MDS emphasises the global structure in the data, as all points contribute significantly to the loss function. There exists multiple methods extending the MDS methodology, making it more suitable for finding local and non-linear structures in the data. One of these methods is ISOMAP, which first creates a neighborhood graph, on which it computes geodesic distances

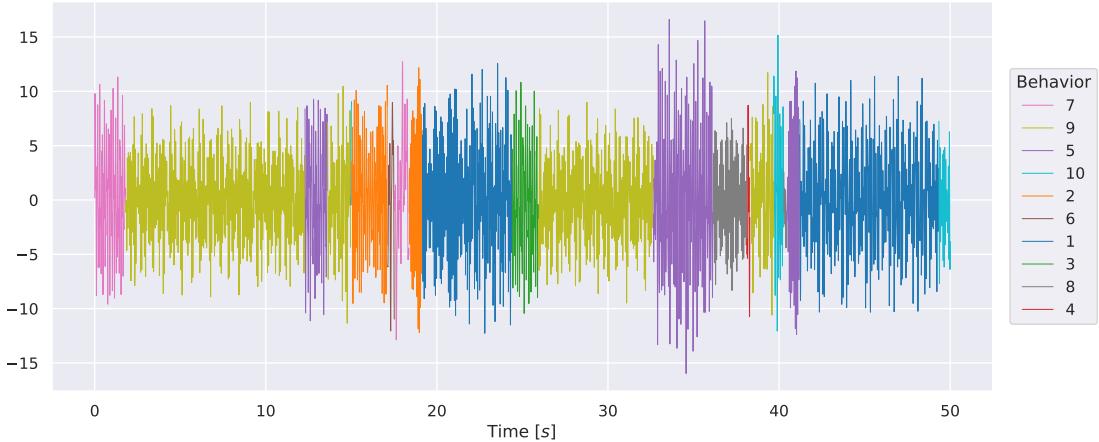


Figure 4.1: Simulated feature vector showing changes in behavior during the first 50 seconds.

between points as the shortest path in the graph [13]. Using these distances it finds the low dimensional embedding via MDS.

As previously stated in section 2.6.2, t-SNE preserves conditional probabilities created from the high dimensional euclidean distances [9]. Due to this it places emphasis on the local structure, and is a natural choice when the goal is to visualize high dimensional clusters in two dimensions. Figure 4.2 illustrates these differences. All methods are able to separate the clusters to some degree, bu t-SNE clearly stands out in its ability to distinguish them.

4.3 Tuning the perplexity parameter

The perplexity parameter from equation (2.30) plays a significant role in the t-SNE embedding [9]. It is closely related to the number of neighbors considered when creating the conditional probabilities (2.27). When the perplexity is small all points contribute "equally little" to the cost function, meaning the algorithm is not able to detect anything. When the perplexity is large it attributes significance to more points, reducing the separation of clusters in the two dimensional embedding. Large datasets should therefore need larges perplexity values. Choosing it between 5 and 50 is suggested. Figure 4.3 shows how the perplexity parameter affects the clustering of the simulated data. Perplexity values 30 and 50 give good results. Note that only 3600 points were used calculating the embedding.

4.4 Tuning the bandwidth matrix

The kernel density estimation (2.31) plays a significant role in final the clustering. It smooths the two dimensional embedding creating local peaks which we interpret as individual behaviors. The degree of smoothing is determined by the bandwidth 2×2 matrix H briefly discussed in section 2.7. Several ways of choosing this parameter have been discussed in literature, most of them based on minimizing the asymptotic mean integrated squared error [10]. As our goal is visualizing clusters, it is better to test several different values, choosing the one which provide the best visualization. To limit the possibilities, we let the scaling be identical in both dimensions, letting H be a diagonal matrix with identical elements.

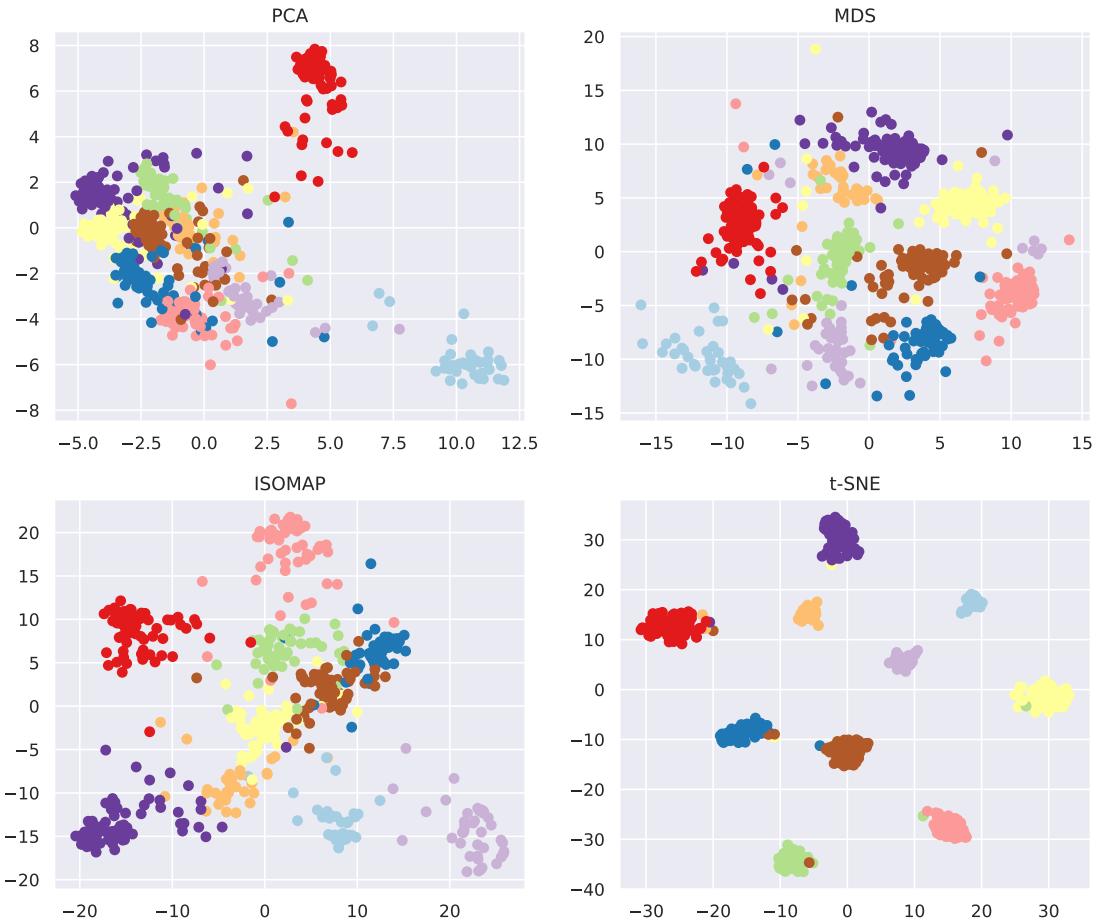


Figure 4.2: Two dimensional embedding of the standardized simulated data using different dimensionality reduction methods. *Top left:* Principal component analysis. *Top right:* Multi-dimensional scaling. *Bottom left:* ISOMAP using $n = 20$ neighbors construction the graph. *Bottom right:* t-SNE with perplexity 30.

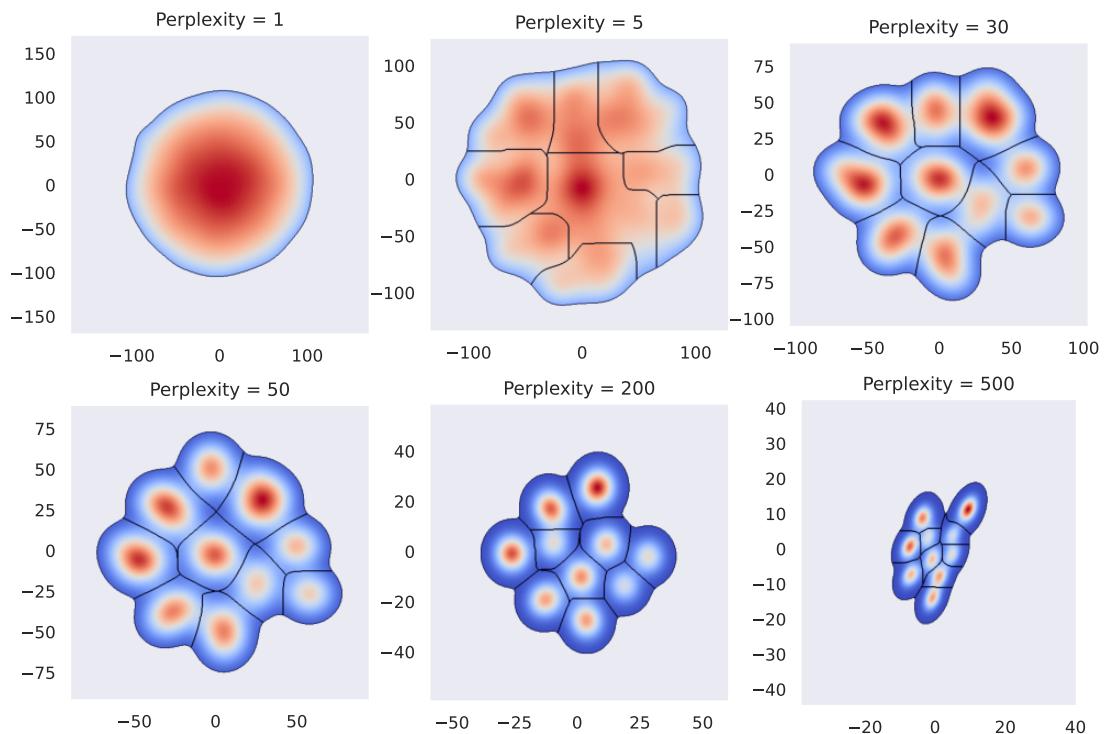


Figure 4.3: Clustering of the simulated data varying the perplexity parameter in t-SNE. The bandwidth for the kernel density estimation is 0.2 for all embeddings.

Bibliography

- [1] W. W. Wei, *Time series analysis univariate and multivariate methods*, 2nd ed. Pearson, 2006.
- [2] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: With R examples*, 4th ed. Springer International Publishing, 2017.
- [3] L. Cohen, “Time-frequency distributions-a review,” *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989. DOI: [10.1109/5.30749](https://doi.org/10.1109/5.30749).
- [4] G. Kaiser, *A friendly guide to wavelets*. Birkhäuser, 1994.
- [5] C. Torrence and G. P. Compo, “A practical guide to wavelet analysis,” *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998. DOI: [10.1175/1520-0477\(1998\)079<0061:APGTWA>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2).
- [6] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*. Springer, 2010.
- [7] T. Hastie, J. Friedman, and R. Tibshirani, *The elements of Statistical Learning: Data Mining, Inference, and prediction*, 2nd ed. Springer, 2017.
- [8] I. T. Jolliffe, *Principal component analysis*. 2nd ed. Springer-Verlag, 2002.
- [9] L. van der Maaten and G. E. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [10] J. S. Simonoff, *Smoothing methods in statistics*. Springer, 1998.
- [11] L. Najman and M. Schmitt, “Geodesic saliency of watershed contours and hierarchical segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1163–1173, 1996. DOI: [10.1109/34.546254](https://doi.org/10.1109/34.546254).
- [12] Y. Liu, X. San Liang, and R. H. Weisberg, “Rectification of the bias in the wavelet power spectrum,” *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 12, pp. 2093–2102, 2007. DOI: [10.1175/2007jtecho511.1](https://doi.org/10.1175/2007jtecho511.1).
- [13] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. DOI: [10.1126/science.290.5500.2319](https://doi.org/10.1126/science.290.5500.2319).