Phenomenology of cortical dynamics: All you need to know about your data

1.1 Introduction

In neuroscience, experimentalists are confronted with a huge amount of data of very different nature. At the same time, given a good model, it is easy to reproduce realistic dynamics mimicking those signals. For example, it is possible to produce the output of a neuron given its input with great fidelity. The simulations obtained by computer scientists also generates a huge amount of data and the resulting signals are very close to those recorded in biology. The similarity of these artificial and natural data suggest that the same methods of analysis should be used. We present in this chapter a collection of tools and techniques which can be used to analyze and classify signals in biological and computational neurosciences.

The first part is an introduction to the common representations of the brain activity that are the spike train, the membrane potential of a neuron and the EEG. The dynamics at the single cell level is characterized by static properties related to the distribution of the membrane potential, spectral properties and firing properties. We also describe more sophisticated measures like based on information theory to manage signals from multiple channels and attractor reconstruction which found applications in the analysis of macroscopic signals. A method based on time frequency analysis is proposed to compress long recordings into a sequence of states and a graph representation of these states and their transitions is provided. In the second part, three classification algorithms are described: K-means, hierarchical tree and self-organized maps and we propose some methods to compare and combine them, thus avoiding the pitfalls inherent to each algorithm. The analysis techniques described in the first two parts are applied, in the third part, to single cell recordings of the ongoing activity in the primary visual cortex of anesthetized cats. Each data sample is represented by 25 parameters and a clusterization in this parameter space gives an optimal partition into 6 clusters. Under visual stimulation, the same cells gathers in the main cluster so that we find more accessible dynamics in ongoing activity than in the evoked activity.

Classes of neuronal dynamics are classically defined by the response of a neuron to a stereotyped electrical stimulation, this study aims at the definition of new classes based on the ongoing and visually evoked activity.

1.2 Temporal signals in neuroscience

The nervous system is considered from Galien to Descartes by an hydraulic analogy with a nervous fluid flowing in the pipes of the nervous system. The electrical nature of the flow in the nervous system was first demonstrated by Luigi Galvani in Bologna at the end of the 18^{th} century. He reported in 1791 that an electrical stimulation of a nerve fiber of a frog could generate a muscle contraction in its leg and, in 1797, he reported that the same contraction could be obtained by pulling to nerve fibers together suggesting the first evidence for animal electricity production. During the 19^{th} century, galvanometers became more and more precise to detect electrical signals and German physiologists, like Emil du Bois-Reymond, could characterize the nervous signals as constituted of short depolarizing events. At the end of the 19^{th}, the physico-chemical mechanisms responsible for this signal were better understood with for example the electro-chemical law giving the potential difference resulting from ion concentrations inside and outside the cell, now known as Nernst potential. With the giant squid axon, Hodgkin and Huxley found, in the 30's, a nerve fiber thick enough to record its activity with a microelectrode clamped to the neuron and this led to their seminal work of the 50's were they described precisely the action potential and proposed the model for its generation. This led to modern electrophysiology were the membrane potential with spiking activity and synaptic events is now recorded in many animal preparation. Using a thicker electrode, the population activity can be recorded and depending on the impedance of the electrode and the filtering of the signal, the recorded activity can reflect the mean depolarization in the dendritic tree or the spiking activity of a set of neurons. By using matrices of such electrodes (MEA), few hundreds of neurons can be recorded at the same time. The Electroencephalogram (EEG) is also a macroscopic signal measuring the spatially averaged activity over a large population of neurons. The whole brain activity can be mapped through an electrode array of 64 or 128 electrodes. The rhythms found in this signal are of special interest for cognitive neuroscience. It can used for assessing the level of consciousness of a subject, to detect precursors of an epilepsy crisis and it also have specific patterns depending on the task the subject is doing. Magnetoencephalogram (MEG) complements EEG by measuring the magnetic field produced by currents running tangentially to the surface of the skull. The obtained signal is easier to localize and less affected by the skull but the measurement must be done in an environment free of magnetic perturbation thus requiring a heavy equipment. More recent techniques to record neuronal activity rely on optical methods. Through calcium imaging the propagation of an action potential can be tracked with fine temporal and spatial resolution. Macroscopic signals obtained from intrinsic optical imaging (IOS) or after the application of a fluorescent dye sensitive to the voltage (VSD) gives a coarse grained picture of the nervous activity in cortical tissues.

In order to analyze the ongoing dynamics in the primary visual cortex of the cat, we will focus on the membrane potential and the intracranial EEG. Those signals are related since EEG signal is an spatial average of the synaptic inputs and collective variations of the membrane potential are correlated with the EEG variations.

1.2.1 Analysis of a spike train

Spikes extraction A temporal trace of the membrane potential V_m recorded at the soma of a cell contains spikes [1] which are short and rare events easily detectable by a human as shown in fig 1.1 and it would bias any processing of the membrane potential. The extraction of these spikes is thus necessary for a simpler description of the membrane potential and a compact representation of the information contained in the spikes.

The spike time is defined as a maximum in second derivative of the membrane potential which correspond to an explosion of the curvature in the trace when the spike is initiated. This maximum is one order of magnitude higher than spurious maxima due to fluctuations in the membrane potential, so that it is easy to detect by requiring to be at least 3 times higher than the standard deviation.

Near the spike time, the shape of the spike can be approximated by a quadratic curve,$V_m(t) = V_m(t_i) + \kappa t^2$ with κ the curvature, or an exponential function, $V_m(t) = V_m(t_i) + e^{t/\Delta}$. An approximation of the spike time precision can be obtained from the curvature,see [38] and [39]:

$$\delta t = \sqrt{\frac{< \delta V >}{< \kappa >}}$$

for $t_i < t < t_s$ where t_s is the time at which V_m reaches the top of the spike and with averages taken over all spikes.

The value of the membrane potential when the spike is initiated is the spiking threshold and the time it takes for the membrane potential to terminate, that is to cross this threshold from top to down, is the spike duration. Spike removal is achieved by interpolating the membrane potential trace between spike initiation and spike termination. In fig 1.1, the interpolation is linear but smoother traces could be obtained by using splines. An efficient way to remove all spikes on a

[1] The mechanism responsible for the generation of those spikes will be detailed in Chapter 2.

membrane potential trace is to calculate the average wave form of the spike and to estimate the spiking threshold and the spike duration on this average spike. The same threshold and the same duration is then used for all spikes in the trace.

Spiking activity. After the spikes have been removed, the spike train and the spike-stripped subthreshold membrane potential (which will be referred as membrane potential for simplicity in the following) can be analysed separately. General methods for the analysis of spiking activity can be found in [40], [41] and more sophisticated methods are described in [42], [43].

The spike train is a vector of spike timings, $\mathbf{t} = (t_i)_{1<i<n}$ of size the number of spikes detected. Actually, knowing whether the absolute value of those times is of special interest is still an open issue but the time between two spike occurrences gives an indication of the level of activity of the neuron. The interspikes interval, $ISI_i = t_i - t_{i-1}$, is used to define the firing instantaneous frequency of the neuron $f_i = \frac{1}{ISI_i}$. The firing rate can be obtained by averaging the spike count over a time window of width τ, $r_\tau(t) = \frac{1}{\tau}\int_t^{t+\tau} \rho(t)dt$ where $\rho(t) = \sum_{1<i<n} \delta(t - t_i)dt$ is the spike train function [2]. When this quantity is averaged over all the time of the recording or on a time window larger than the spike duration, it is called the mean firing rate and when averaged over many neurons it is called the population firing rate. The firing frequency of a neuron depends highly on its cellular type and on the brain area where it is located. In visual cortex, cells fire with an average firing rate around 1 Hz in barrel cortex [3], 5 Hz in the primary visual cortex and 15 Hz for spontaneous activity in higher level areas like motor cortex or prefrontal cortex with up to 80 Hz when it is activated. During a spike, the membrane is insensitive to incoming current so that even when strongly stimulated in artificial conditions, the firing frequency of a neuron is limited at 1000 Hz due to this refractory period of few milliseconds.

Spike trains are digital signals that is series of 0 and 1 and an analog representation of the spike train s is obtained after convolution of a kernel f with the spike train function $s(t) = \sum_{1<i<n} f(t - t_i)$. The commonly used kernels are the exponential kernel, $f_{exp}(t) = H(t)e^{-\frac{t}{\tau}}$, H being the Heaviside function, and the alpha kernel $f_{alpha}(t) = te^{-\frac{t}{\tau}}$. This analog signal provides a realistic approximation of the input current or conductance corresponding to this spike train and, as will be shown in the part "Metrics and measures", it is also used for building spike train metrics.

Spiking regularity The ISI distribution is also useful to quantify the regularity of the spiking activity of a neuron by the coefficient of variation of interspikes intervals $CV = \frac{<ISI-<ISI>>^2}{<ISI>}$ [4]. For a perfectly regular spiking

[2] The Dirac function $\delta(t - t_i)$ is 1 when $t = t_i$ and 0 otherwise.

[3] The barrel cortex is the somatosensory receiving inputs from vibrissae of the rat or mouse

[4] It is thus the ratio $\frac{\text{VarianceofISIs}}{\text{MeanofISIs}}$.

neuron, all ISIs are the same and the CV is 0. For neurons having $CV = 1$, the variance of interspikes intervals is equal to its average. The simplest stochastic process generating spike train with this property is the Poisson process, where ISIs are independents, and it is a commonly used to model irregular trains of events (see Chapter 2), the ISI distribution of a Poisson process follows a Gaussian law. Many cells in the brain fires in a Poissonian fashion, $CV \approx 1$ in the spontaneous regime, but a closer look at the ISIs distribution shows that it is better described with a gamma law [5] than a Gaussian law. A sub-Poissonian ISIs distribution, $CV < 1$, is characteristic of cells having a more regular firing than if its spike train was generated by a poisson process. A supra-Poissonian ISIs distribution, $CV > 1$, is characteristic of cells that tend to fire with bursts of spikes and is found in evoked activity. The slope of the decay in the ISI distribution may also be an important parameter in cells with low frequency spiking because it reflects how rare events occur which is not taken into account in the previously described parameters.

1.2.2 Analysis of a membrane potential trace

Spikes are a major feature of neuronal dynamics but the subliminar activity, that is fluctuations of the membrane potential under threshold, is also very informative. The membrane potential is a very complex signal reflecting the activity of the network in which it is embedded. Bistability of the membrane potential is found in multiple areas of the nervous system. It sometimes result from intrinsic mechanisms like in the Purkinje cells of the cerebellar cortex [44] where it may support information processing or it may collective and rely on network mechanisms, like in the prefrontal cortex where columns have persistent up state during the storage of an object in the working memory. During slow wave sleep those transitions are correlated with EEG variations. The presence of several levels of activity, like an up activated and a down desactivated state, indicates multistability of the network and transient oscillations are a sign of coordinated spiking in the population. The analysis should then be led carefully to detect such events.

Static properties As will be seen in section 4, much of the information about a cell is hidden in its membrane potential distribution. The simplest way to characterize it is to calculate its successive moments of order k relative to the mean μ_{Vm}, $\mu_k = E((V_m - \mu_{V_m})^k$. The Gaussian, used as a reference to compare probability distribution functions, has a finite second order moment and null moments of higher order. It is defined by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

[5] The gamma law is a two parameters (k, θ) probability distribution function defined as follows:

$$f(x, k, \theta) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)}$$

with $x, k, \theta > 0$ and Γ the gamma Euler function.

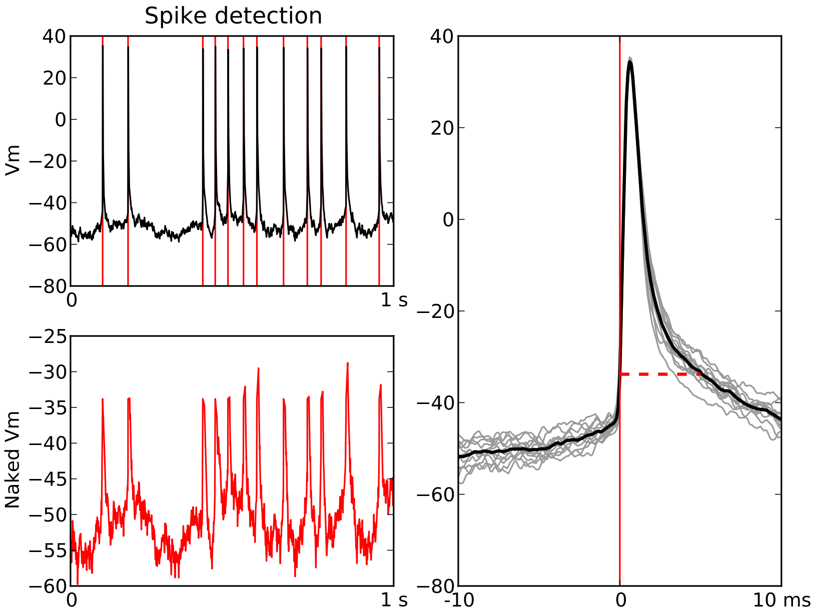

Figure 1.1: **Spike extraction** - (Left-top) Temporal trace of the membrane potential with spike times. (Left-bottom) Trace of the membrane potential after the spikes have been removed. (Right) Average spike for the estimation of the spiking threshold and spike duration.

.

The mean of the membrane potential can be very different from one experiment to another because it depends on many parameters of the experimental preparation. It is usually between -80 and -50 mV [6]. The standard deviation, $\sigma_{V_m} = \sqrt{\mu_2}$, reflects the level of activity in the network. It often depends on the mean μ_{V_m}, there are less fluctuations when a cell is close to threshold than when it is depolarized. The mean and the standard deviation of the distribution are sufficient to fit a Gaussian distribution and the coefficient of regression measures the goodness of the fit. The *skewness*, $\gamma_1 = \frac{\mu_3}{\sigma^3}$ reflects the symmetry of deviations from the mean, it is 0 for a Gaussian distribution. A positive skewness indicates the presence of micro up states as in excitation driven cells and a negative skewness indicates the presence of micro down states as in inhibition driven cells. The symmetry of the distribution can also be checked by using the fitted Gaussian law as a reference and calculating the following coefficients:

$$S_1 = 3\frac{m - \mu_{V_m}}{\sigma}$$

and

$$S_2 = 3\frac{m - med_{V_m}}{\sigma}$$

with m, σ the mean and standard deviation of the Gaussian function and med_{V_m} the median of the empirical distribution. The *kurtosis*, $\beta_2 = \frac{\mu_4}{\sigma^4} - 3$, reflects the sparseness of deviations from the mean, it is 3 for a Gaussian distribution. Distributions with kurtosis greater than 3 are flat and correspond to traces with small and fast fluctuations as would be characteristic of a cell embedded in a very active asynchronous network. Distributions with a kurtosis less than 3 are sharp and corresponds to cells with slow and large deviations from the mean as would be characteristic of a network with low but synchronous activity.

A distribution F is unimodal if there exists a mode m such that F is convex on $[-\infty, m[$ and F is concave on $]m, \infty[$. If the distribution is multimodal that is if it contains more than one peak, the Gaussian distribution is not a good approximation anymore and the distribution can be fitted with a sum of two or more Gaussian laws. For bimodal, the upper peak defines an up state and the lower peak defines a down state. The minimum of the distribution between those two peaks is the threshold separating the up domain from the down domain. Several parameters can be used to characterize deviations from unimodality of a distribution. The distance between an empirical distribution and a test distribution is $\rho(F,G) = sup_x|F(x) - G(x)|$ and the dip of F is $d = inf\rho(F,\mathcal{U})$ where $\mathcal{U}$ is the set of unimodal distributions. A practical way to perform this calculation is described in [45]. The separability is defined from the fit of a sum of two Gaussian functions as

$$Sep = \frac{m_1 - m_2}{2(\sigma_1 + \sigma_2)}$$

[6] The membrane potential is bounded from below by the potassium inversion potential and from above at 0 mV by the Na inversion potential.

with m_1, m_2 the means of the two Gaussian functions ($m_1 > m_2$) and σ_1, σ_2 their standard deviations. The contrast between the two distributions, also called the discretness, is defined as follows from the two Gaussian functions resulting from the fit:

$$Discr = 100 * \sum_i \frac{|G_1(x_i) - G_2(x_i)|}{G_1(x_i) + G_2(x_i)}$$

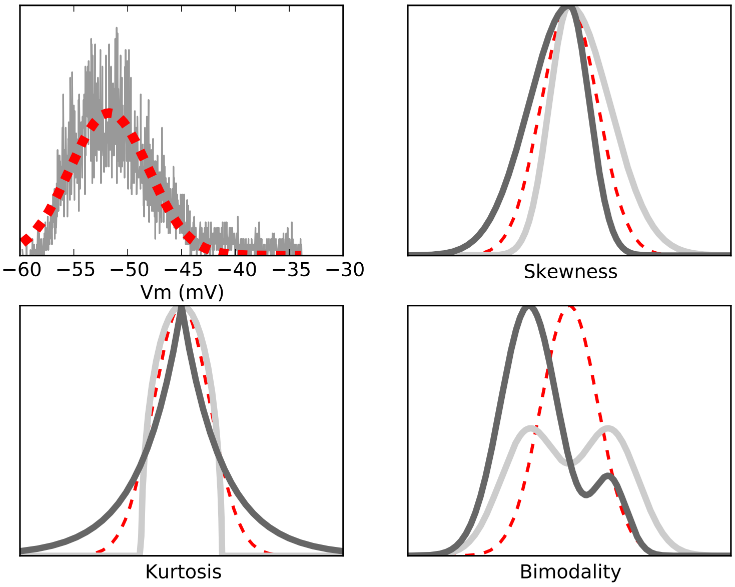

Figure 1.2: **Static properties of V_m** - (Top-left) Gaussian fit for the V_m distribution of cell X. (Top-right) Examples of distributions with positive (dark) and negative (light) skewness. (Bottom-left) Examples of distributions with kurtosis greater than 3 (dark) and less than 3 (light). (Bottom-right) Examples of asymmetric (dark) and symmetric (light) bimodal distributions.

Spectral properties

Autocorrelation Oscillatory behavior of the membrane potential is not detected by the analysis of distribution and transitions between up and down states. There are several possibilities regarding the origin of these oscillations. The whole network can be oscillating in a robust manner at low frequency, this is the case when the brain is in deep sleep, also called slow wave sleep, or when it is in a pathological state like epilepsy. Transient oscillations at higher frequency can also be seen, and are often considered as the propagation of coherent activity among the cell assemblies in which the neuron is embedded. A simple way to detect oscillations in a signal s is to calculate its autocorrelation,

$$R_s(\tau) = \frac{1}{T}\int_0^T s(t)s(t-\tau)dt.$$

A first time constant is given by the extinction rate, τ_e, which can be captured by fitting an exponential function, $e^{-\frac{t}{\tau_e}}$. In the case of cell X, the autocorrelation decreases in a linear fashion. It is still possible to see a slight oscillatory deviation from the linear behavior at $\tau \approx 50ms$, which is close to the average ISI of the spike train.

Power spectral density (PSD) To get more information about the frequency content of the membrane potential fluctuations, it is interesting to calculate the power spectral density and this is done by using the Fourier transform of the signal. The Fourier transform of a signal is

$$\hat{s}(\omega) = \frac{1}{T}\int_T s(t)e^{i\omega t}dt$$

and the PSD is then $S(\omega) = \frac{\hat{s}(\omega)\hat{s}^*(\omega)}{2\pi} = \frac{|\hat{s}(\omega)|^2}{2\pi}$. There exists several efficient methods to compute it like the Fast Fourier Transform which requires the sampling frequency of the signal to be a power of 2 [46]. It is usually represented as a function of the frequency $f = \frac{\omega}{2\pi}$ and in decibels, $S_{dB}(f) = 10log_{10}S(f)$. The PSD is also more easy to interpret when it is smoothed by taking local averages over a short frequency band.

The two features which should be looked at with attention are the local peaks, indicating the oscillatory components of the trace coming from the input temporal structure or from internal properties of the cell, and the slope of the decay in log-representation. Many signals have a power spectrum behaving in a $\frac{1}{f^\alpha}$ fashion and α may give indications about the process underlying fluctuations of the signal. For a white noise, the spectrum is flat and $\alpha = 0$. For a Brown noise, as generated by a Wiener process, $\alpha = 2$ and fluctuations may be associated to a diffusive process. For pink noise, which can be generated by a shot noise process, $\alpha = 1$ and the origin of such fluctuations is still highly debated, a interesting hypothesis is that it could result from a self-organized critical process [47]. For more general Levy processes, α can take fractional value. It was shown in a recent study that different statistics of the visual input lead to different exponent in the scaling of the high frequencies power spectrum [48]. Anyway, these exponents reflecting power scale invariance should be considered with great care because their estimation is very sensitive on the frequency window considered. The PSD of cell X present a peak around 20 Hz and is otherwise nearly flat on the frequency window observed.

Wavelet analysis Fourier analysis describes in a compact manner the structure of temporal fluctuations in a signal but it would fail to detect transient oscillations, a solution can be to calculate the PSD over a time window for each point of time. Continuous wavelet analysis is another way to overcome this problem and to get a spectral representation of the signal at each time, a short introduction to this method is provided in [49] and advanced presentation can be found in [50]. It is gives spectral information at any point of time by convolving the signal with a family of wavelets of different temporal scales as shown on fig 1.3. The Morlet wavelets family, which will be used in the following, is generated by the mother wavelet

$$\Psi_{\omega_0}(t) = \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2} e^{i\omega_0 t}$$

with

$$\Psi_{\sigma\omega_0}(t) = c_\sigma \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2} (e^{i\sigma\omega_0 t} - \kappa_\sigma)$$

where $\kappa_\sigma = e^{-\frac{1}{2}\sigma^2}$ and $c_\sigma = \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2}\right)^{-\frac{1}{2}}$. There is a simple relation between wavelets and their mother, $\Psi_{\sigma\omega_0}(\frac{t}{\sigma}) = \sqrt{\frac{\delta t}{\sigma}}\Psi_{\omega_0}(\frac{t}{\sigma})$, with δt the time step of the signal. The wavelet transform is then $\tilde{s}_t(\omega) = \frac{1}{T}\int_T \Psi_\omega(t'-t)s(t')dt'$. It is actually simpler to use the Fourrier transform of this equation because the convolution becomes a simple multiplication. The Fourrier transform of the mother Morlet wavelet is $\hat{\Psi}_{\omega_0}(\omega) = \frac{1}{\pi^{\frac{1}{4}}} e^{-\frac{(\omega-\omega_0)^2}{2}}$ and the Fourrier transform for the rest of the family can be deduced by using the renormalization $\omega \leftarrow \omega' = \sigma\omega$ and $\hat{\Psi}(\omega') = \sqrt{\frac{2\pi\sigma}{\delta t}}\hat{\Psi}(\omega)$. The inverse FFT then gives the wavelets coefficients in an efficient manner. Transient oscillations appears as bump in the wavelets power spectra represented as a time frequency matrix, such a bump centered around 15Hz can be seen in fig 1.4 at 500ms, and those bumps could be detected automatically by using Gabor filters, see [51].

1.2.3 EEG

The electroencephalogram (EEG) is a very common signal in neuroscience, it can be recorded with an electrode at the surface of the scalp or with an intracranial electrode. As it is an analog signal, it can be processed with the same analysis as was presented for the membrane potential from which spikes have been removed. EEG signals are usually recorded on a longer period of time than the membrane potential with a sampling frequency around 1 kHz whereas the membrane potential is sampled at 10 kHz. Brain rhythms corresponding to different cognitive states can be tracked on this recording. Hans Berger recorded the first EEG signal on his son in 1929. He discovered the α-rhythm, an oscillation around 8 Hz in the occipital region of the brain associated to a rest state with closed eyes. It was further developed to study epilepsy and it is now widely used to measure the level of consciousness of patients or anesthesia depth with what is called the bispectral index. The functional role of these oscillations is

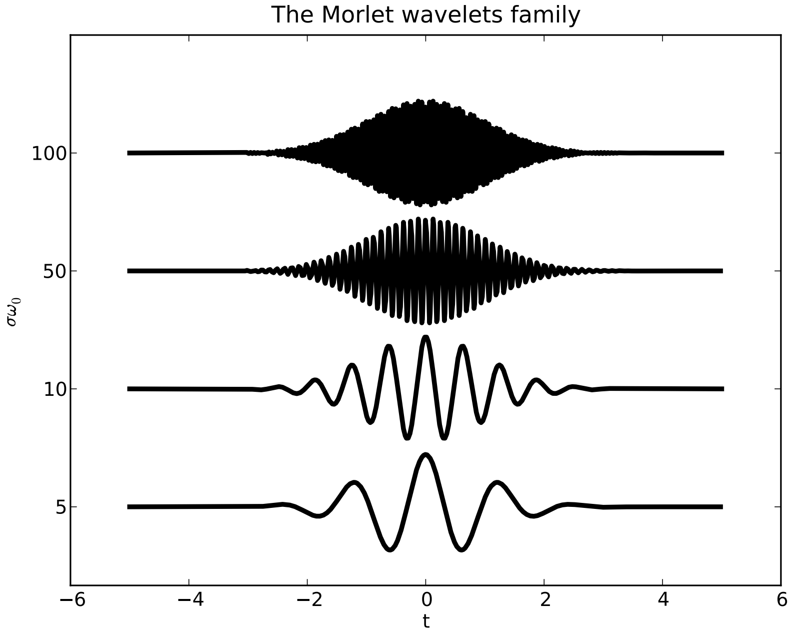

Figure 1.3: **The Morlet family** - Morlet wavelets at different scales.

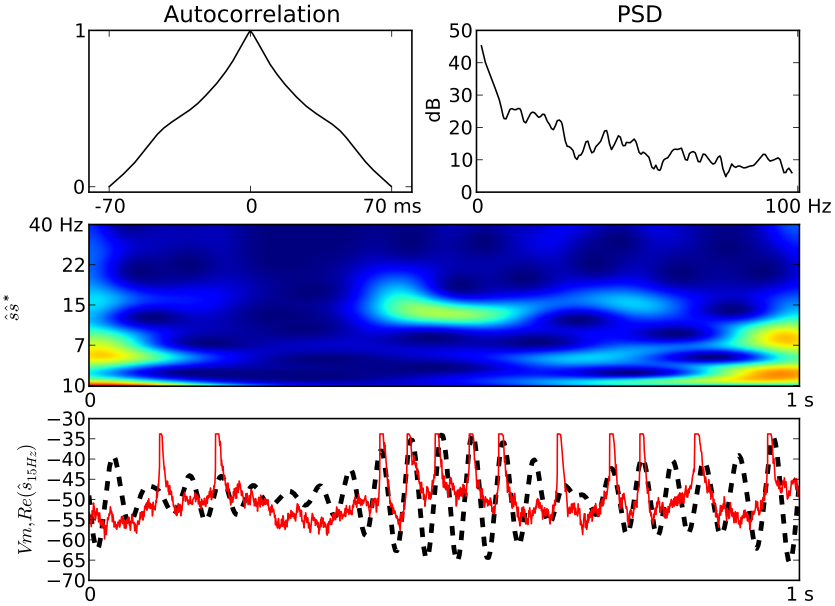

Figure 1.4: **Spectral properties** - (Top-left) Autocorrelation of the V_m trace. (Top-right) PSD of the V_m trace. (Middle) Time-frequency representation of the V_m signal. (Bottom) V_m trace (red) and 20 Hz component of the time-frequency representation (dashed).

still an active topic of research but the low frequency rhythms are usually associated to sleep or pathological states whereas cognitive processing is associated to higher frequency rhythms. The frequency bands can be summarized as:

Name	Frequency band	Functional role
δ	$1 - 3Hz$	Slow wave sleep
θ	$4 - 7Hz$	Memory retrieving
α	$8 - 11Hz$	Resting
β	$12 - 20Hz$	Attention
γ	$> 20Hz$	Perceptual binding of a Gestalt

Recent research in cognitive neuroscience showed the importance of phase synchronisation between electrodes across brain areas [18]. The spatio-temporal structure of correlations between the 64 or 128 electrodes recorded makes it possible to discriminate between conscious and unconscious perception [52], it also reveals the attentional state of the subject [53].

The presence of brain rhythms makes the time frequency analysis particularly useful for EEG signals. For very long time series although, interesting events are difficult to capture and it is also difficult to infer temporal relationships between these rhythms. In the analysis described below, the signal is compressed and a graphical representation of the sequences describes the temporal organization of brain waves.

Example on an artificially generated signal. The artificial EEG Y, shown in fig 1.5, was generated by repeating 3 times the following sequence:

$$\delta \to \beta \to \beta + \gamma \to \theta \to \delta$$

with the γ oscillations are only active near the local maxima of the β oscillation. This sequence of transitions among rhythms and combinations of rhythms can be represented by a graph as shown on Fig??. The aim of the method proposed below is to extract the sequence of rhythms and combinations of rhythms activated and to build the graph corresponding to this sequence based on the time-frequency matrix.

Compression of the time-frequency matrix. The first step is to split the time-frequency matrix into blocks by choosing time and frequency intervals where the cutting are made. Regular sampling of the time at 1Hz enables a precise tracking of rhythms transitions and allows the detection of low frequency oscillations. For the frequency axis, the cutting can be based on the common frequency bands defined in the literature but it can also be adapted to the particular signal by taking frequencies of local minima of the spectrum as frontiers between the frequency bands. In the following, the frequencies are gathered in 4 bands ($b_1 = [1 - 8Hz]$:low frequency,$b_2 = [9 - 19Hz]$: middle frequency,$b_3 = [20 - 40Hz]$: high frequency and $b_4 = [41 - 100Hz]$: very high frequency). The locally integrated power spectral density with sampling window δt is obtained from the wavelet power density W by $L(t, f) = \frac{1}{\delta t} \int_t^{t+\delta t} W(t, f) dt$

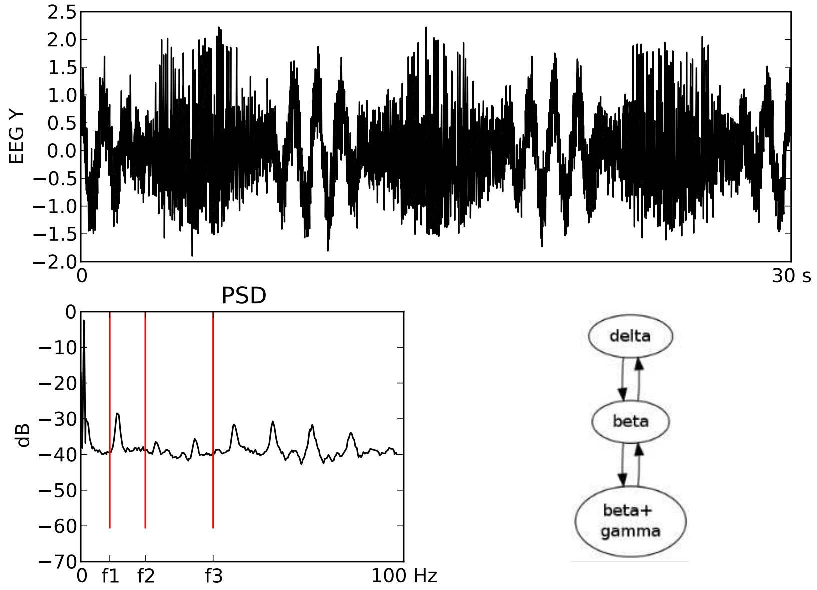

Figure 1.5: **Artificial EEG Y** - (Top) Artificial EEG Y. (Bottom-left) Power spectral density of the signal with limit frequencies of the 4 bands. (Bottom-right) State diagram representing the signal.

and the power density relative to the frequency band i is given by $B_i(t) = \frac{1}{f_{i+1}-f_i}\int_{f_i}^{f_{i+1}} W(t,f)df$. The compressed time-frequency matrix is then

$$\mathcal{C}_i(t) = \frac{1}{\delta t(f_{i+1}-f_i)}\int_t^{t+\delta t}\int_{f_i}^{f_{i+1}} W(t,f)df dt.$$

This compressed matrix will be used to detect transitions in the dynamics. It can also be used for an efficient online sonification of the signal where each frequency band code for a note with intensity given by the matrix values at each time. Transforming neuronal data into sound is useful because the human ear is very good at detecting temporal structure in audio signals.

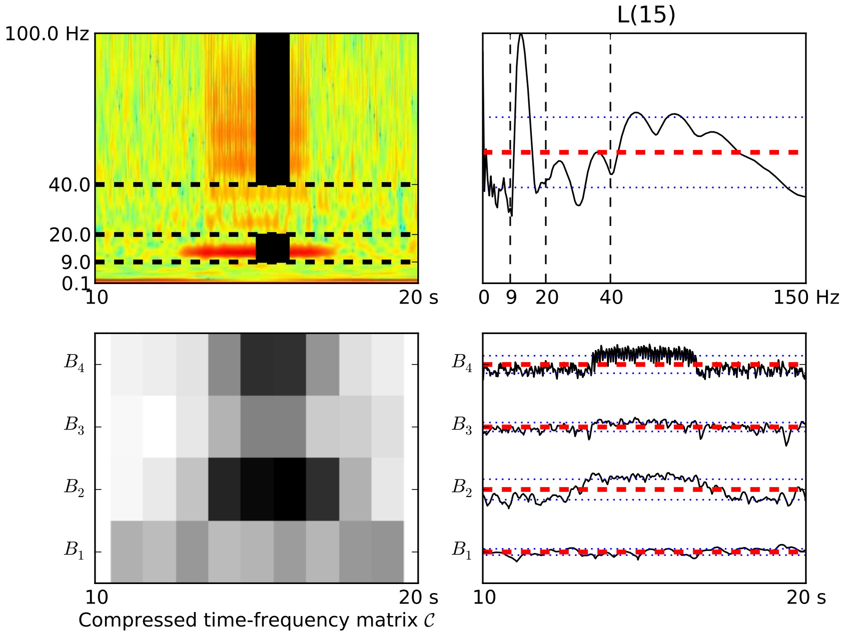

Figure 1.6: **Compression of the EEG** - (Top-left) Time-frequency representation of the signal, shaded areas represent activated bands and dashed lines represent frontiers of the frequency bands. (Top-right) Local power spectral density of the signal at t=15s. (Bottom-left) Compressed representation of the time frequency matrix. (Bottom-right) Dynamics of the integrated power in the four bands.

Definition of the symbols. Each column of the compressed matrix $\mathcal{C}$ provides a compact description of the frequency content of the signal at a time t. An empirical criterion $\theta_\epsilon(b_i,t) = (1-\epsilon)B_i(t) + \epsilon L(t)$ determines if a frequency band b_i is activate at time t by

$$\text{if } \theta_\epsilon(b_i, t) > (1-\epsilon)E(B_i) + \epsilon E(L(t)),\ b_i \text{ is active}$$

. The band is temporally active at t when $\theta_0(t)$ is used as criterion, it has more power density than at other moments of time, and it is spectrally active if $\theta_1(t)$ is used, it has more power density than other frequency band. For intermediate values of ϵ, a frequency band is active depending on its power density relative both to other moments of time and frequency bands. In the EEG Y at $t = 15s$, considering the criterion θ_1, b_2 and b_4 are active. For each column of the compressed matrix, a 4-bits codeword $d_{b_4}d_{b_3}d_{b_2}d_{b_1}$ is formed based on the active bands of the signal. The digit b_i is equal to 1 if the frequency band b_i is active and 0 if it is inactive. The codeword for EEG Y at $t = 15s$ is 1010 and its decimal representation is 10. The same principle could be adapted to an arbitrary number of frequency bands and the codeword representation could be made more efficient by using Huffmann coding [7].

Building of the graph The signal can be represented as a string where each letter is the decimal translation of the codeword (between 0 and 15). The frequency of occurrences f of each letter and of each two letters word are then collected in a dictionary and a test is applied to each two letter word. If $f(ab) > f(a)f(b)$, the word ab is more frequent than it would be if a and b where appearing randomly in an independent way, the transition from a to b will then be reported on the graph. By this way, the graph of fig 1.5 for EEG Y is recovered. The result of this analysis for recorded EEG of 60 s duration is shown on fig 1.6. The detection of N-letters words can be made optimal by using Lempel-Ziv-Welch coding [8]. The graph of fig 1.7 is obtained from an EEG trace of 3 hours by drawing the strongest links. The graph can be used to build a statistical model like a markov chain giving the probability of occurrence of a state given the current. Transitions between brain states can also be represented as trajectories in a low dimensional phase space based on the spectral properties of the signal [54]. It would be interesting to check how these states relate to classes of neurodynamics at the single cell level.

1.3 Metrics and measures.

We consider a dataset $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$. Each data $\mathbf{x_i}$ is a p-dimensional vector representing a neuron recording. The neuron recording can be represented by its membrane potential trace, its spike train or p parameters extracted from those. We list below distances which can be used to evaluate the closeness of two data samples and measures representing the structure of the data set. We first investigate analog signals and then discuss the case of discrete data samples.

[7] Huffman coding is a way to perform loss-less compression of data by building a variable length code based on the probability of occurrences of the source symbols.

[8] LZW algorithm also performs loss-less compression. It is based on the encoding of substrings appearing in the data sequence to be compressed.

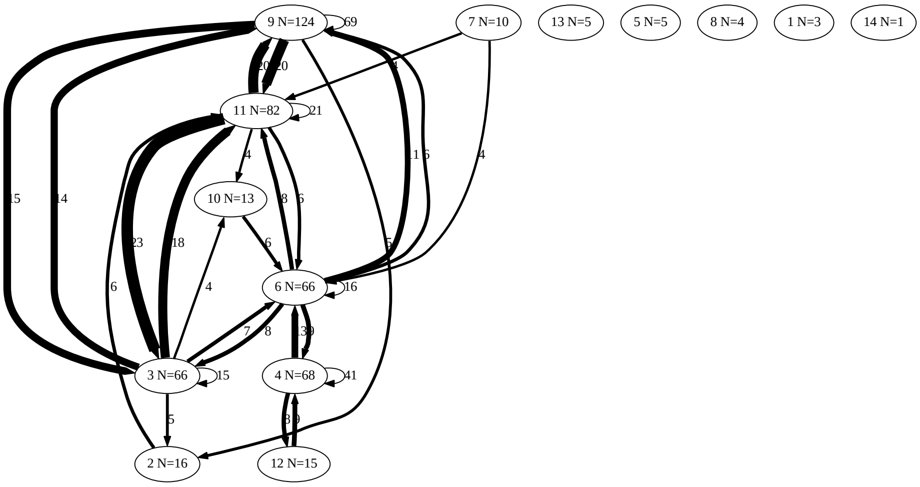

Figure 1.7: **State diagram of an EEG trace.** - Each node of the graph is a state with the number of occurrence N. The thickness of arrows represent the probability of transitions among those states.

1.3.1 Analog signals.

Classical distances. The Minkowski distance between two data samples depends on a parameter q, $d_q(\mathbf{x_i}, \mathbf{x_j}) = (\sum_{k=1}^{p} \|x_{ik} - x_{jk}\|^{q})^{\frac{1}{q}}$. The Euclidian distance is the most natural metric to evaluate the similarity between 2 data samples. It is defined by $d_2(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{k=1}^{p} \|x_{ik} - x_{jk}\|^2}$. The city block distance is also used $d_1(\mathbf{x_i}, \mathbf{x_j}) = \sum_{k=1}^{p} \|x_{ik} - x_{jk}\|$. The distance matrix $D_{\mathbf{X}}$ of the dataset $\mathbf{X}$ is then obtained from the $d_{ij} = d_q(\mathbf{x_i}, \mathbf{x_j})$.

Correlation-based measures. The Pearson correlation coefficients are defined by $r_{ij} = \frac{1}{p} \sum_{k=1}^{p} \frac{(x_{ik} - \bar{x_i})}{\sigma_{x_i}} \frac{(x_{jk} - \bar{x_j})}{\sigma_{x_j}}$. It should not be confused with the covariance matrix, $Cov_{ij} = \frac{1}{p} \sum_{k=1}^{p} (x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})$. Other measures are defined in a similar way. The coherence of two signals is defined by considering the cross-correlation of the their power spectral density. The phase synchrony at specific frequency is obtained by cross-correlating the phase of these two signals at this band obtained from the time-frequency analysis.

1.3.2 Spike trains.

Pearson correlation. The simplest way to evaluate the similarity between two spike trains x_i and x_j is to consider their Pearson correlation coefficient defined similarly as that of a continuous signal. With such a measure, an exact synchrony of the two spike trains is necessary for being similar. For example, if B is just a copy of A with a shift δt greater than the time window used for the analysis, the correlation coefficient of A and B may be zero although the two spike trains are very similar. Other metrics have been developed to avoid such pecularities.

Cost based method The Victor-Purpura distance [55] is based on the number of operations necessary for transforming x_i into x_j. The three basic operations considered are spike addition or deletion both having a cost of 1 and temporal displacement of δt having a cost of $\frac{\delta t}{\tau}$. The time constant τ is a free parameter of the defined distance.

Convolution based method As described above, a filtered version of the spike trains s_i and s_j are obtained by applying exponential or Gaussian kernels with width τ. A distance is then defined by [56]:

$$D^2(x_i, x_j) = \frac{1}{\tau} \int_0^T [s_i(t') - s_j(t')]^2 dt'.$$

For two spike trains differing only by the insertion or deletion of a spike, $D^2(x_i, x_j) = \frac{1}{2}$ and if the only difference is a shift δt of one spike, $D^2(x_i, x_j) = 1 - e^{-\frac{|\delta t|}{\tau}}$. Another similarity measure based on the filtered signals s_i, s_j is the following defined in [57]:

$$S(x_i, x_j) = \frac{\int_0^T s_i(t) s_j(t) dt}{\sqrt{\int_0^T s_i(t) dt} \sqrt{\int_0^T s_i(t) dt}}$$

. In both methods, a narrow width of the kernel makes the distance or similarity measure sensitive to spike jitter whereas with a broader width, the additional or missing spikes are detected.

Parameter free method Other methods for the estimation of (dis)similarity are described in [58]. The ISI-distance method has the advantage of being parameter free. The current interspikes interval is defined by $ISI_i(t) = min(t_{ik}|t_{ik} > t) - max(t_{ik}|t_{ik} < t)$ where t_{ik} is the k^{th} spike of the i^{th} neuron. The ISI-distance between x_i and x_j is then:

$$D(x_i, x_j) = \frac{1}{T} \int_0^T |I(t)| dt$$

with:

$$I(t) = \begin{cases} \frac{ISI_i(t)}{ISI_j(t)} - 1 & \text{if } ISI_i(t) \leq ISI_j(t) \\ -(\frac{ISI_j(t)}{ISI_i(t)} - 1) & \text{else} \end{cases}$$

1.3.3 Information theoretic measures.

Information theoretical measures, as an application field of probability theory, heavily relies on the estimation of the probability distribution of the data samples. As this estimation for finite size samples is often a difficult task, the following describes the concepts used in information theory with random variables and we provide simple application examples to illustrate it. For a more deep treatment of this subject, see [59], and for applications to spike train analysis see [40].

Shannon entropy

Definition and properties The Shannon entropy of a random variable X taking discrete values $\mathcal{X} = [x_0, ..., x_m]$, is $H(X) = \sum_{l=0}^{m} -P(X = x_l)log_2(P(X = x_l))$. H gives a measure of the uncertainty that is the number of yes/no questions it takes to guess the value of the random variable when following an optimal strategy based on the past occurrences of this variable. It is measured in bits and variables with maximal entropy for a given set $\mathcal{X}$ follows a uniform law. The Shannon entropy has the following properties:

- $H(X) > 0$
- $H(X, Y) = H(X|Y) + H(Y)$
- $H(X, Y) \leq H(X) + H(Y)$ with equality if and only if X and Y are independent.

H can be extended to continuous variables with the differential entropy, $h(p) = -\int_{-\infty}^{\infty} p(x)log(p(x))dx$ but classical properties of the entropy do not hold anymore. A more convenient way for the extension to continuous variables is to consider the relative entropy with a reference probability distribution q, also called the Kullback-Leibler divergence: $D_{KL}(p||q) = -\int_{-\infty}^{\infty} p(x)log(\frac{p(x)}{q(x)})dx$ where q is commonly taken as a Gaussian function. The differential entropy of a data sample of N points generated from a multivariate Gaussian law of average μ and covariance matrix $\mathbf{\Sigma}$ is $h(\mathcal{N}(\mu, \mathbf{\Sigma})) = \frac{1}{2}ln((2\pi e)^N|\mathbf{\Sigma}|)$ with $|\mathbf{\Sigma}|$ the determinant of the covariance matrix.

Estimation The estimation of differential entropy of a process is not an easy task because a precise estimation depends on the bin width used for estimating of the probability density. The entropy is thus bounded by $logN_{bin}$, the entropy of a random variable with uniform probability distribution having the same support. In fig 1.8, the entropy of a Gaussian signal at 10 kHz sampling frequency is estimated across time with the number of bins being 3 times the number of points in the signal used for estimation, the result is close to the theoretical value.

For cell X, the entropy of the membrane potential is compared with the entropy of a Gaussian variable with the same mean and variance in fig 1.8.

Before the first spike, the entropy increases linearly close to the behavior of a Gaussian random variable and it drops after the first spike. The entropy then grows at a much slower rate because there is a big part of the range (between -35 and -40 mV) which remains nearly unexplored.

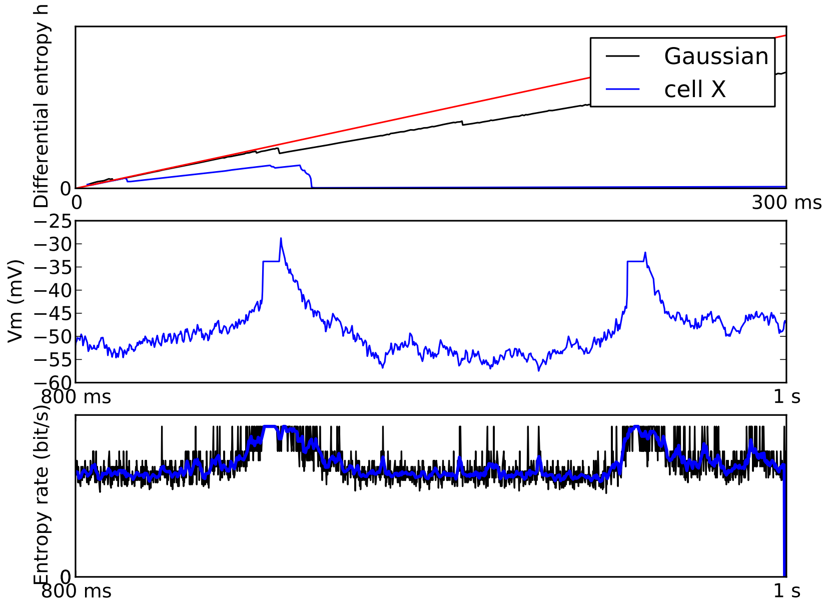

Figure 1.8: **Differential entropy and entropy rate** - (Top) Differential entropy estimated for a Gaussian process (black) and for the cell X (blue). The red line indicates the theoretical value for the Gaussian process. (Middle) 200ms of the V_m trace used for the estimation of the entropy rate. (Bottom) Entropy rate and its coarse grained version for 200 ms of the cell X.

Entropy rate The entropy estimate is difficult to interpret because it often far from its theoretical value and because its range drops drastically after a event like a spike occurrence. The entropy rate, $\frac{dh}{dt}$ is a better way to follows changes in the signal. As shown on fig 1.8, the rising part of a spike is associated with an entropy production and the falling part with entropy destruction.

Fisher information.

We suppose a parameter θ has to be estimated from observations of the random variable X. The likelihood function $f(X, \theta)$ gives the probability distribution of X given θ. The Fisher information is then

$$\mathcal{I}(\theta) = \mathbf{E}[(\frac{\partial lnf(X,\theta)}{\partial\theta})^2|\theta].$$

For sufficiently regular likelihood functions, it can also be written:

$$\mathcal{I}(\theta) = -\mathbf{E}[\frac{\partial^2}{\partial\theta}lnf(X,\theta)|\theta].$$

Applications. Based on this definition of information, the Cramer-Rao bound gives the limit of precision achieved by an unbiased estimator $\hat{\theta}$:

$$Var(\hat{\theta}) = \frac{1}{\mathcal{I}(\theta)}.$$

This theoretical bound can then be used for example to find the interval of confidence of the estimated frequency of an noisy oscillation. The Fisher information is also very important in probability theory because it is used to build a metric in spaces of probability distribution functions which is the starting point of information geometry [60].

Mutual information.

Definition and properties The mutual information between two discrete variables X and Y is defined from the entropy of the marginals and the joint probability distributions $I_2(X,Y) := H(X) + H(Y) - H(XY)$ or equivalently $I_2(X,Y) = H(X) - H(X|Y)$, it is symmetric $I_2(X,Y) = I(Y,X)$. There is no restrictions anymore to extend the definition to continuous variables with probability distributions p_X and p_Y and the integral version is $I_2(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}log(\frac{P_{XY}}{P_X P_Y})$, it is the Kullback-Leibler divergence between the joint law and the product of the marginal laws of X and Y. The mutual information measures the reduction of uncertainty in the estimation of X resulting from knowledge of Y. It is 0 for independent variables and it is $H(X)$ when Y is a copy of X.

Example on a multivariate Gaussian The 3 examples presented on fig 1.9 corresponds to sets of Gaussian variables (**X**,**Y**,**Z**) with the following covariance matrices:

$$A = \begin{pmatrix} .1 & .75 & .75 \\ .75 & .1 & .75 \\ .75 & .75 & .1 \end{pmatrix}$$

$$B = \begin{pmatrix} 1.22 & .7 & 0 \\ .7 & 1.22 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In case A, where all variables are depending on each other, the mutual information is the same for any pair of variables. In case B, where only **X** and **Y** are correlated, the mutual information of (**X**,**Y**) is higher than for other pairs because observations on one of the variables reduces uncertainty about the other. When all variables are independent as in case C, the mutual information should be 0 for any pair but the finite size of the samples introduce a bias.

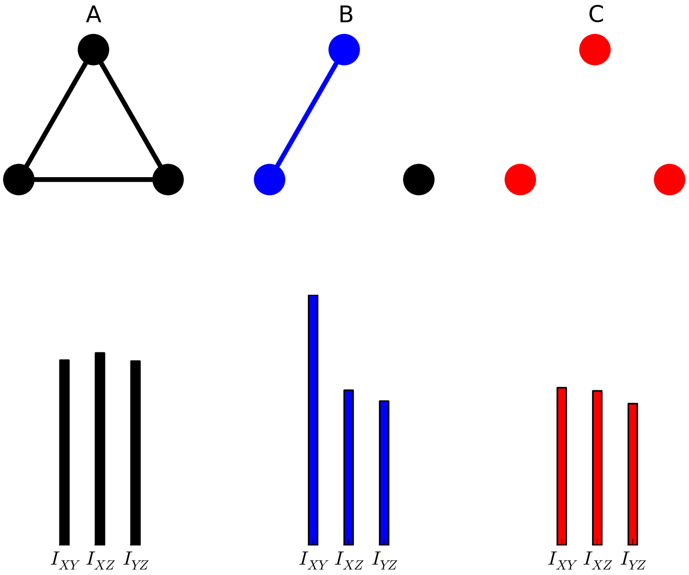

Figure 1.9: **Mutual information** - Mutual information for random processes generated by the multivariate Gaussian processes of covariance matrices A, B and C.

Neuronal complexity.

The mutual information can be generalized into the multi-information of any set of k random variables **X**: $I_k(\mathbf{X}) = \sum_{1<i<k} H(X_i)) - H(\mathbf{X})$, this quantity is also called the integration of the set and it is zero when all variables are independent. The neuronal complexity defined in [61] for a set of N variables

is:

$$\mathcal{C}_N = \frac{1}{N}\sum_{k=1,\dots n-1}(\frac{k}{n}I_N - < I_k >_k)$$

with $< ... >_k$ denoting an average over all the subsets of k elements. An approximation for weakly correlated variables is given in [62]:

$$\mathcal{C}_N = \frac{n+1}{24}(tr(R-I)^2 + tr(R-I)^3)$$

with R the correlation matrix [9] for off-diagonal elements. If the data $\mathbf{X}$ is generated by a coupled Ornstein-Uhlenbeck process [10]

$$d\mathbf{X_t} = \mathbf{X_t}(I-C) + \sigma d\mathbf{W_t}$$

, the complexity should be related to the coupling matrix. The previous approximation gives:

$$\mathcal{C}_N = \frac{n+1}{48}\sum_{i\neq j}(C_{ij}^2 + C_{ij}C_{ji}) + \frac{n+1}{96}\sum_{i\neq j\neq k}3C_{ij}C_{jk}C_{ik} + \frac{n+1}{24}\sum_{i\neq j}C_i i(C_{ij}^2 + C_{ij}C_{ji}).$$

The neuronal complexity is thus related to the decomposition of the structure of the network in loops (first order term), 3-cycles (second order term),...The neuronal complexity thus quantifies how much a system is "more than the sum of its parts", a geometrical interpretation based on a comparison with families of exponential probability distributions can be found in [63]. The neuronal complexity is thus a promising measure for analyzing the huge amount of data arising from neuroscience experiments but it is still difficult to estimate it in an efficient fashion. There has been some recent progress for estimating the entropy of spike trains [64] and computational tools for this estimation are a growing field in neuroscience [65].

1.3.4 Attractor reconstruction

The signal recorded by intracellular electrodes or EEG devices is generated by non linear dynamical systems of high dimension but the effective dimension of the dynamics may be small due to the presence of rhythms. A theorem from Whitney and Takens further developed in [66] showed that for an attractor of effective dimension d, a delay-map in $\mathbf{R}^{2d+1}$ can be built which is qualitatively similar to the original attractor (that is there exists a diffeomorphism transforming one into the other). In this attractor reconstruction the delay and the dimension have to be chosen.

[9] The correlation matrix is composed of 1's on the diagonal and $R_{ij} = \frac{cov(X_i,X_j)}{\sqrt{var(X_i)}\sqrt{var(X_i)}}$

[10] Which can be considered as the linearization the stochastic Wilson-Cowan equations presented in Chapter 2

Optimal delay A rule of the thumb for choosing the delay of an oscillatory pattern is to take $\frac{1}{4}$ of the period. When there is no clear period in the signal, the optimal delay can be chosen as the minimum of the autocorrelation or the maximum of the mutual information $MI(\tau)$ between the signal and its time delayed version.

Correlation dimension A way to estimate the effective dimension of the attractor of a chaotic dynamical has been proposed in the 80's by Grassberger and Proccaccia in [67]. From the N points of the temporal signal $x_i = (y(i), y(i+\tau), y(i+2\tau), ..., y(i+k\tau))$ reconstructed from the original signal, the correlation sum is defined as:

$$C(r) = \frac{2}{N(N-1)} \sum_{i<j} \theta(r - |x_i - x_j|)$$

and the correlation dimension is $D = lim_{r \to 0} \frac{logC(r)}{logr}$ so that the correlation sum behaves as $C(r) \approx r^D$ for small r. This correlation dimension can be calculated for several values of the embedding dimension k and as k increases the optimal embedding dimension is obtained when D reaches a plateau. A public domain software called TISEAN [68] can be used for these calculations. For EEG signals, it has been suggested that the correlation dimension of EEG signals is reduced during sleep and pathological states like epilepsy [69].

1.4 Data classification

The previous sections showed that many parameters can be used to characterize signals corresponding to neuronal activity and that there are several ways to evaluate the similarity between two of those signals. In this section, we consider that some parameters have been extracted from the recordings and we wish to obtain a classification based on the comparison of these parameters. When many parameters are used, it is difficult to perform a efficient classification. This "curse of the dimension" can be attenuated by reducing the dimension of the parameter space.

1.4.1 Preprocessing of the data set

Normalization of the feature space The dataset $\mathbf{X}$ is first normalized by $\tilde{\mathbf{x_i}} = (\frac{\mathbf{x_{ik}} - \overline{\mathbf{x_k}}}{\sigma_{x_k}})_{1<k<p}$. After this operation, all parameters have the same variance 1 and the classification based on this normalized dataset is not affected by the range over which the parameter take values.

Orthogonalization of the feature space (PCA) Principal components of the dataset are extracted using the covariance matrix $C = \tilde{X}^t\tilde{X}$. C is symmetric so it can be diagonalized $C = {}^tP\Lambda P$. Λ is a diagonal matrix where each diagonal term represent the contribution of the corresponding eigenvector to the total variance. Vectors are then reordered from the one with the biggest eigenvalue to the one with the smallest eigenvalue. For the classification of the recordings, the dimension of the parameter space can be reduced by selecting only the M first vectors explaining 90% of the variance.

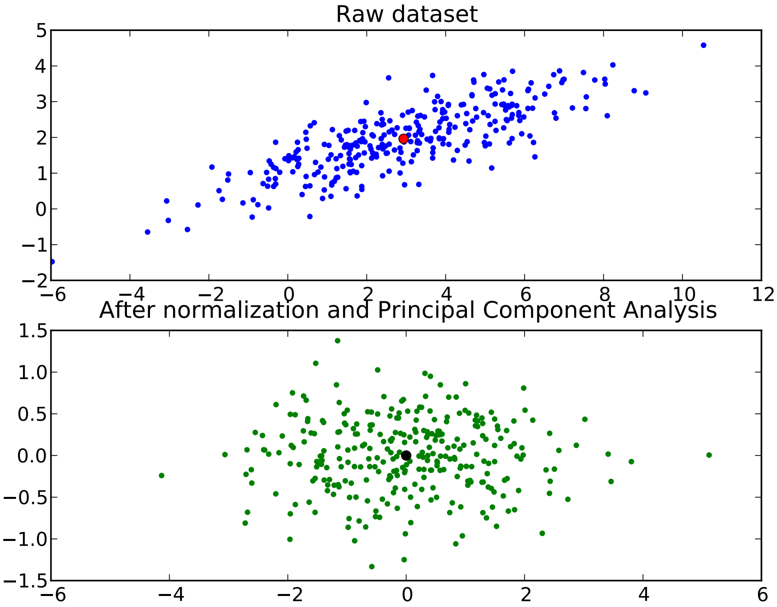

Figure 1.10: **Preprocessing of a 2D-Gaussian dataset (300 points) -** (Top) Raw data set. (Bottom) After PCA, the principal axis of the Gaussian becomes aligned with the horizontal axis.

1.4.2 K-means clustering

Description of the algorithm.

The K-means method is a way to clusterize cells by making an a priori assumption on the number of clusters K [70]. We will discuss possible ways to select seeds and generate partitions of the parameter space. This method is simple and efficient, it is widely used in the scientific community but it also have pitfalls of all unsupervised learning method. A common example of application where it gives a poor result is the Fisher iris data base. We consider **X**, a set of n data points $(x_i)_{1 \leq i \leq n}$ in $\mathbb{R}^p$. The algorithm will partition the points around K centers $(C_k)_{1 \leq k \leq K}$ minimizing a potential function $\phi = \sum_{1 \leq k \leq K} min_{x_i \in C_k} \|x_i - c_k\|^2$. This potential function is monotonically decreasing during the K-means algorithm and it will always terminate because the number of possible partitions is bounded by K^n. Although the clustering procedure will always terminate, finding the globally optimal partition is a NP hard problem [11] and we will discuss possible solutions to approach this global optimum.

Seed selection. Seeds are the initial cluster centroids $(c_j^0)_{1<j<k}$ and the simplest way to select it is to choose randomly K points as seeds with uniform probability law from the data set. This method is standard but better results can be obtained with a careful seeding as shown [71]. The selection of the seeds is the following where $D(\mathbf{x})$ is the distance of the point $\mathbf{x}$ to the closest centroid already chosen:

a Take the first centroid c_1 randomly with uniform probability law.

b Take next centroid among $\mathbf{X}$ with probability $\frac{D(\mathbf{x})}{\sum_{\mathbf{x}\in\mathbf{X}} D(\mathbf{x})^2}$.

c Repeat a and b until K centroids are selected.

For this Kmeans++ algorithm, the potential function is shown to check $\mathbf{E}[\phi] < 8(lnK + 2)\phi_{OPT}$ where ϕ_{OPT} is the optimal partition of the data set.

Iterative procedure. The process unfolds in two steps repeated until convergence is obtained:

1 Attribute each data point to its closest centroid by computing $j_c(\mathbf{x_i}) = min_j d(\mathbf{x_i}, \mathbf{c_j^t})$. The j corresponding to the minimum distance is the cluster id which will be attributed to the cell.

2 Compute the new centroid position $\mathbf{c_j^{t+1}} = (\sum_{x_i k \in C_j^{t+1}} x_{ik})_{1\leq i\leq p}$.

3 Steps 1 and 2 are repeated until successive centroids stay close to each other, $d(\mathbf{c_j^t}, \mathbf{c_j^{t+1}}) < \epsilon$.

where t indicates the iteration of the process.

Selection of a ”correct” partition.

The procedure described above is repeated many times and a criterion has to be defined so that the partition is considered as robust or not too ”bad”. As we already saw the result of the K means depends on initial conditions, so those are changed at each iteration. The distance to used also affect the resulting partition, the Euclidian distance is commonly used for K-means procedure and the city-block distance is used when medians are taken as centroids.

[11] NP hard problems take a very long time to solve when the size of the system grows.

Dunn Index We can consider a partition to be "correct" if its clusters are sufficiently compact and well separated. The relevant measures for this are the radius of the largest cluster, $R_{max} = max_{C_k} \frac{\sum_{x_i \in C_k} \|x_i - c_k\|}{|C_k|}$ and the minimum distance between clusters, $L_{min} = min_{C_i, C_j} \|c_i - c_j\|$. These two quantities are combined in the Dunn index $DI = \frac{L_{min}}{R_{max}}$. Among the partitions generated by the K-means algorithm, we select the one with the highest DI and we stop the selection when there is no improvement in the DI. We also use the DI to detect optimal K for the partition.

Frequency of occurrence of the partition After many runs of the K-means algorithm, the final partition will often be the same. A way to quantify this is to run the algorithm many times and to take the partition which is most often encountered in this process.

Example on a mixture of Gaussian distributions To illustrate the K-means algorithm, we generated artificially two data clouds **A** and **B**. In the data set **A**, 180 points are randomly chosen following a 2D-Gaussian probability distribution function (pdf). The data set **B** is prepared with a mixture of Gaussian probability distribution functions where for each Gaussian, having different means and covariance matrices, 30 points are chosen randomly. The data set **A** lack of any internal structure and the frequency of the most often encountered partition in 10000 repetitions of the K means algorithm as a function of K, the number of centroids considered, decreases montonically in an exponential fashion. In the data set **B**, deviations from this monotonical decrease shows the non-homogeneity of the data set and the drastic drop when K goes from 6 to 7 suggest that the data set can be well represented as a collection of 6 clusters.

1.4.3 Tree Building.

The K-means method to find clusters in a data set is stochastic, because the final partition depends on the initial conditions. All the more, it is a "flat" method because the obtained clusters are disjoint. A clustering procedure is hierarchical if in the resulting partition, each cluster is formed with subclusters, themselves containing subclusters,...If two points are grouped together at a given level, they will stay grouped at higher levels. The natural representation for a data set on which hierarchical clustering has been applied is a tree, also called a dendrogram. There are two possible ways to perform hierarchical clustering, it can be started with every data sample in a singleton cluster and this bottom-up process is an agglomerative tree building, or it can be started with a giant cluster containing all the data samples splitted successively until each cluster contains only one data sample and this top-down process is a divisive tree building, both methods are described in [70] and we here focus on the agglomerative method.

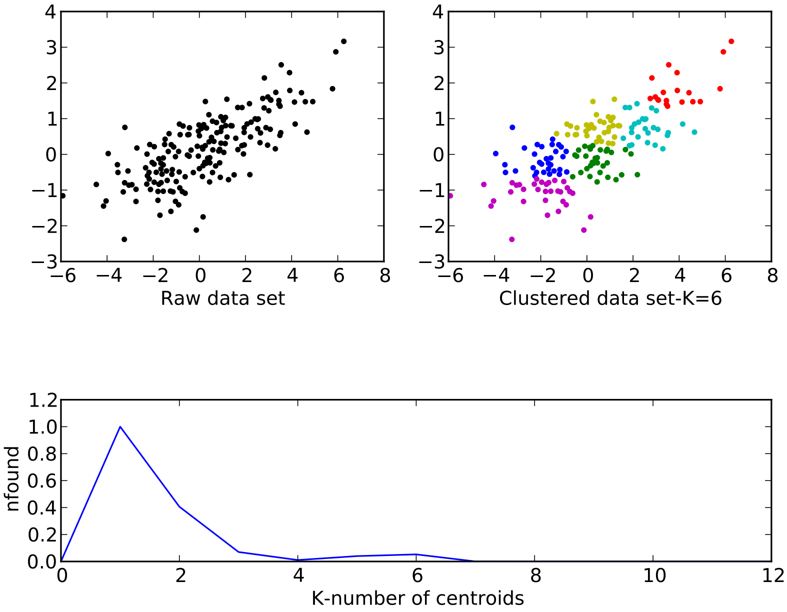

Figure 1.11: **Data set A** - 180 points chosen randomly with a Gaussian pdf.

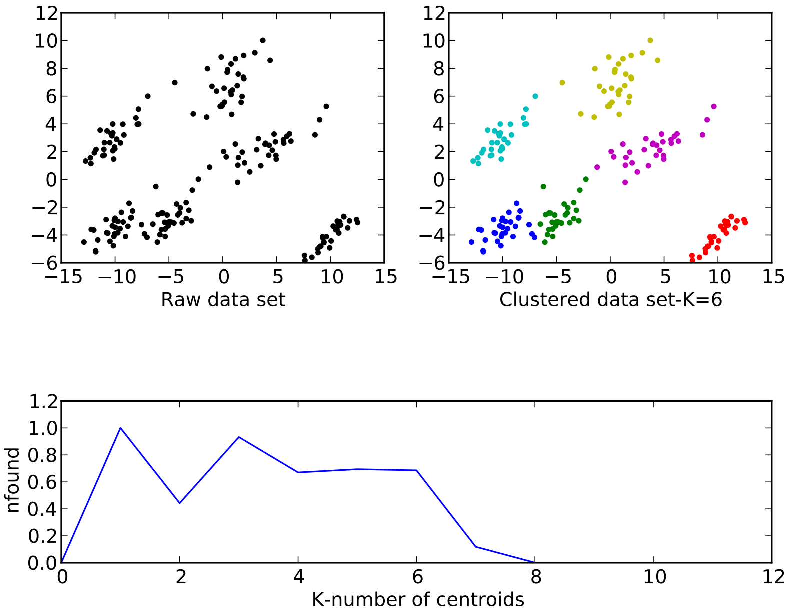

Figure 1.12: **Data set B -** 180 points chosen randomly with a mixture of 6 Gaussian pdfs.

Agglomerative tree building.

Algorithm. In the agglomerative tree building method, every data point is considered as a cluster or node. The process then merges the two closest clusters and iterates until only one cluster remains. A pairwise distance has to be chosen so that we can evaluate the distance between two cluster. The pairwise average-linkage (pal) distance considers the distance between two clusters as the average over all the pairwise distances between elements of the two clusters. The process consists of the following steps:

1 Merge the closest clusters $C_i, C_j = argmin_{i,j} d_{pal}(C_i, C_j)$ into the cluster C_l.

2 Repeat 1 until C_l contains all elements of the data set.

Application to Gaussian mixture. As for the K means clustering, the Tree Building algorithm is applied to the data set A, 180 points randomly chosen from a mixture of 6 Gaussian distributions. By cutting the tree at an appropriate depth, 6 clusters are obtained corresponding to the 6 Gaussian distributions from which the data samples are generated.

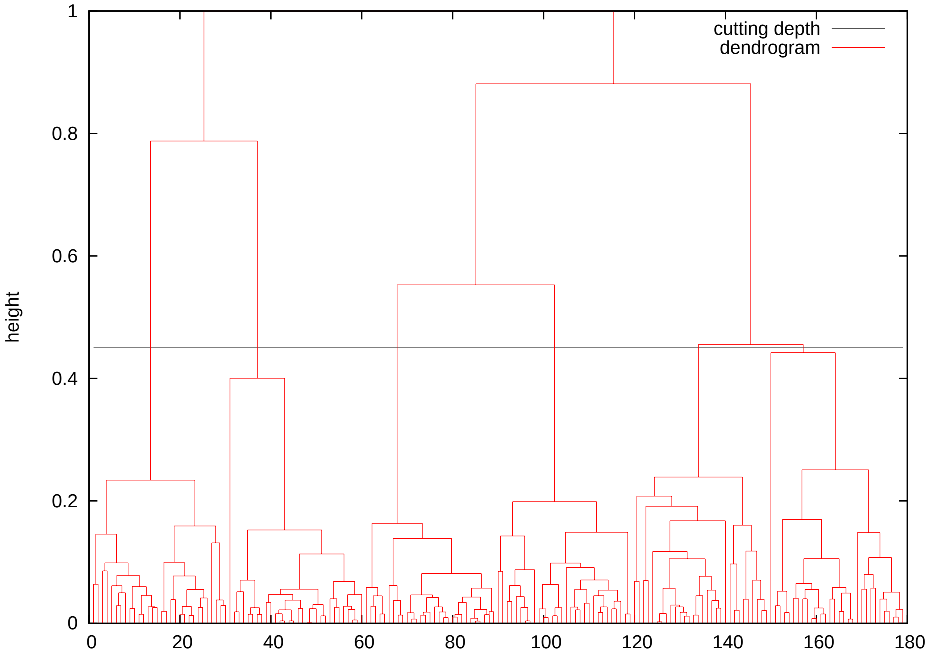

Figure 1.13: **Data set B -** 180 points chosen randomly with a mixture of 6 Gaussian pdfs.

Distance matrix and cutting of the tree

Base of tree At the base of the tree, cells are ordered in a way related to the hierarchical structure of the dataset. This order can be considered as an interesting way to enumerate the data samples. In the distance matrix of fig 1.14, the lower triangle stands for the data samples in their initial order and the upper triangle for the order resulting from the tree building procedure. The 6 clusters are clearly detectable in the upper triangle.

Cutting the tree An arbitrary number of clusters (bounded by the number of cells) can be obtained by cutting the tree at an appropriate depth or more elaborate cutting methods [72]. By splitting the tree into the same number of clusters as resulting from the K means procedure, the two partitions can be compared. A possible way to compare the partitions obtained from two different methods or in two different experimental conditions is to build a matrix H of dimensions (K_1, K_2) where K_1 is the number of clusters resulting from the first method and K_2 for the second. Elements of the matrix are filled as follows:

$$H_{i,j} = |\text{data samples belonging to cluster i by method 1 and j by method 2}|.$$

1.4.4 Kohonen network.

The self-organizing map algorithm is a biologically inspired model used to map data samples $(\mathbf{x_i})_{1<i<n}$ from the input space to nodes $(y_j)_{1<j<k}$ of the feature space. A weight vector $\mathbf{w_j}$ is associated to each node and a dynamic evolution of these weights representing learning, leads to a low dimensional representation of the data samples. The simplified version presented below is often referred to as a Kohonen network [73] and more sophisticated models of self-organizing maps will be described in the Chapter 3 dealing with models of V1 formation. The relaxation time of this dynamic evolution is a free parameter of the algorithm. The feature space is often taken on a 2D regular grid of dimensions (N_x, N_y). In this algorithm, the final result is dependent on the order of presentation of data samples.

SOM algorithm For each data sample presentation, a competition is taking place and the winner dictates the weights evolution dynamics in its neighborhood. The process is as follows after random initialization of the weight vectors:

1 Compute the activation for each node in the feature space $y_k = \sum_{(1<j<p)} |w_{kj} - x_j|$ and select the closest one y^* from the data sample x.

2 Update weights according to the following learning rule:

$$\mathbf{w_k}(i+1) = \mathbf{w_k}(i) + \alpha(i)h_{*k}(i)(\mathbf{x} - \mathbf{w_k}(i))$$

.

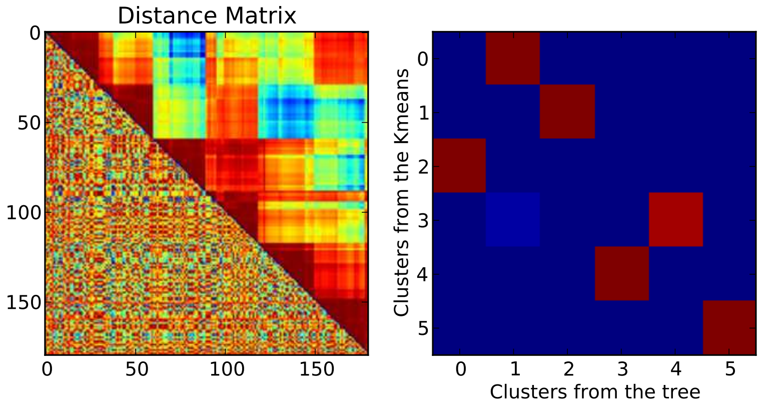

Figure 1.14: **Similarity and comparison of the obtained partitions -** (Left) The lower triangle shows the similarity in a random order and the upper triangle shows the similarity between data samples ordered according to the tree. (Right) The quantity in (i,j) indicates how many cells fall into cluster i with the tree partition and into cluster j with the K means partition.

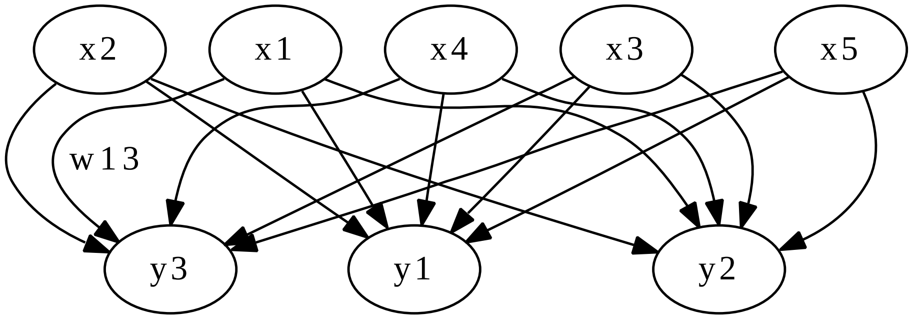

Figure 1.15: **SOM network.** - The network is composed of input nodes x and map nodes y.

3 Renormalize the modified weights so that $\|\mathbf{w}\| = 1$.

The learning rate $\alpha(i)$ is a decreasing function of time depending on the relaxation time τ, a common choice is $\alpha(i) = \tau(1 - \frac{i}{n})$. The neighborhood function $h_{*k}(i)$ is 1 if the node k is closer than the influence radius $R(t)$ from the winner and 0 if it is farther. The influence radius decreases as $R(i) = R_{max}(1 - \frac{i}{n})$ with $R_{max} = \sqrt{N_x^2 + N_y^2}$.

Biological interpretation. The step 1 can be seen as an implementation of a neural field and the step 2 is an example of implementation of a plasticity rule. The algorithm is thus inspired by cognitive theories about assosiative memory and adaptive learning. A more detailed model inspired from the SOM but including recurrent connections will be studied in Chapter 3. The relaxation time τ is a free parameter of the model and it should be adapted to the studied dataset.

Application to the data set B The SOM algorithm is applied to the data set B with $\tau = 0.02$ and 200 iterations on 5x5 grid. As shown on Fig 1.16, the data samples are composed of 6 main clusters with more than 15 data samples and a collection of smaller clusters. An important property of the SOM algorithm is that close points in the data cloud will fall onto close points of the map.

1.4.5 Misclassification and metasimilarity

Each method employed to partition a data set should be related to the question asked by the analyst and its advantages and pecularities should be well understood. On the one side, the K-means method is an easy way to determine the number of clusters into which the data set will be split. A good K can be deduced from the evolution of the Dunn index or the frequency of occurrence of the partition most often encountered. On the other side, it offers no indication about the relation between clusters. The tree building is very nice to visualize the fine hierarchical structure of the data set and to provide an order in accordance to this hierarchical structure but, as a deterministic method, its result is highly affected by outliers. All the more, it is often difficult to know at which depth the tree should be cut to give a good partition. To have an idea of topological relations between clusters, the SOM algorithm makes a very good job but it necessitates a tuning on several parameters (relaxation time, number of iterations, size of the grid).

By employing multiple techniques, the resulting partitions can be compared. In fig 1.14, the cluster C_5 of the partition P_{KM} (obtained from K-means) is the same as the one from the partition P_{TB} (obtained from tree building). The cluster C_0 in P_{KM} becomes the cluster C_2 in P_{TB}. A data sample which is in cluster C_3 in P_{KM} lands at C_1 in P_{TB} whereas its expected destination is C_4.

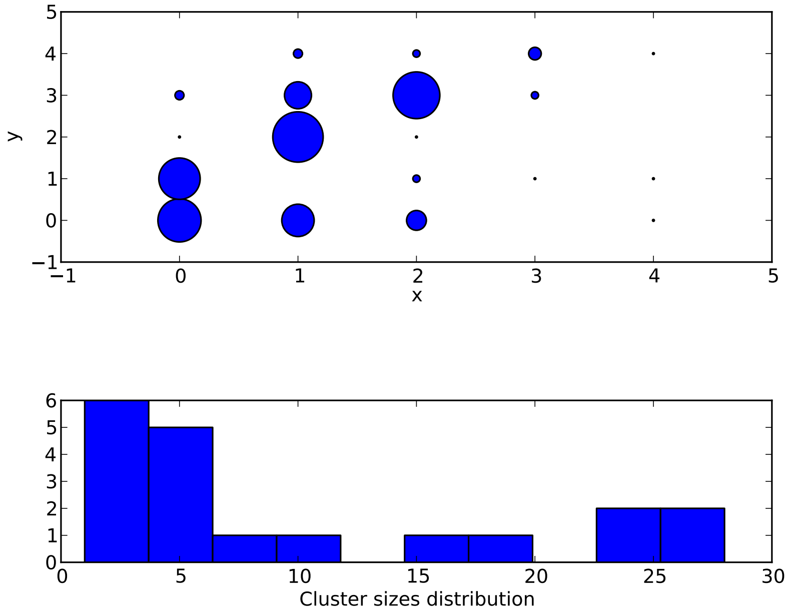

Figure 1.16: **SOM classification of the data set B -** (Top) Resulting 5x5 SOM. (Bottom) Cluster size distribution.

The matrix H thus provides an easy way to detect inconsistent classification. These misclassified points are near the frontiers between different clusters.

The 3 classification presented can be used to formalize a new notion of similarity. The metasimilarity between the data samples i and j is defined by

$$\mu_{ij} = 1 + \delta_{ij}^{KM} + \delta_{ij}^{TB} - d_{ij}^{SOM}$$

with $\delta_{ij}^{KM} = 1$ if the data samples i and j are in the same cluster of P^{KM}, $\delta_{ij}^{KM} = 1$ if the data samples i and j are in the same cluster of P^{TB} and d_{ij}^{SOM} is the Euclidian distance between clusters of the data samples i and j in the SOM normalized between 0 and 1.

If $\mu_{ij} = 3$, the data samples i and j lands in the same cluster whatever the clustering method and in that sense data samples i and j are metasimilar. If $2 \leq \mu_{ij} < 3$, the two data samples are simililar but the cluster to which they belong could be split into subclusters given by the SOM to highlight the difference. If $1 \leq \mu_{ij} < 2$, i and j belongs to different clusters or there may be a misclassification problem for one of the two samples. Finally, if $\mu_{ij} < 1$, the two samples are clearly in different clusters. The metasimilarity thus formalize in a comprehensive way the results of different data classification methods.

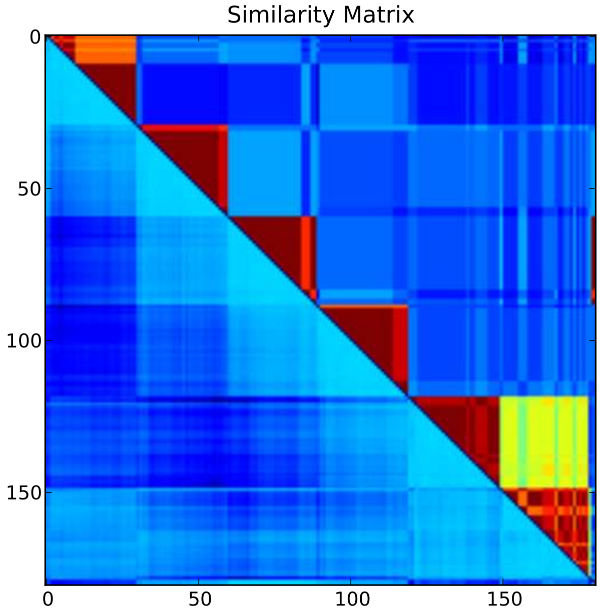

Figure 1.17: **Similarity and metasimilarity** - The lower triangle shows the similarity and the upper triangle shows the metasimilarity between data samples.

1.5 Application to electrophysiological recordings.

We consider the membrane potential recorded with an intracellular electrode in area 17/18 of the anesthetized cat (alfatesin). Some parameters are extracted from electrophysiological recordings to build the data sets. A first data set, Spt150, is composed of 150 data samples from cells recorded during ongoing activity. Another data set, Vis143, is composed of 143 data samples from cells recorded during the presentation of a visual stimulus. The obtained classification is compared to the same pool of cells during ongoing activity, Spt143.

1.5.1 Parameters extracted from electrophysiological recordings.

We show in fig 1.18 the list of 25 parameters with their average and standard deviation over the population.

There are 3 groups of parameters: parameters related to the distribution of the membrane potential, parameters related to the spectral properties of the membrane potential and parameters related to spikes of the cell. In parameters related to the distribution, we find moments of the distributions and coefficients reflecting the asymmetry or the deviation from unimodality. Most of spectral parameters are fractions of the power spectrum integrated over a frequency band. The instantaneous firing is the reciprocal of the interspikes interval whereas the average firing rate is the spike count divided by the recording duration.

1.5.2 The on-going activity data set (150 cells).

Optimal partition. A home made K-means method is applied to Spt150 with values of K from 2 to 9 and with the Dunn index as an optimization criterion. Only the 9 first components of the PCA are necessaty to explain 90% of the variance. The light blue curve of fig 1.19 corresponding to K=6 shows rapid convergence to its optimal Dunn index. Moreover, the optimal value for K=6 (black curve) is higher than the optimal value for K=5 (red curve). This suggest that the partition with K=6 gives a better description than with other values of K. The optimal partition for K=6 is given on fig 1.20 and the successive splittings of the data cloud can be visualized on fig 1.21. Note that the clustering obtained in fig 1.20 is different from that of fig 1.21 for K=6 because k-means++ method was applied to choose initial conditions in the first case whereas it is taken randomly with uniform probability in the second case.

Description of the partition. By a multifactorial decomposition analysis in Matlab, the main relevant parameters for the description of the partition are those related to the distribution of V_m and the frequency content in high frequency. The largest cluster (red) is composed of cells having a Gaussian distribution. The smallest cluster (pink) is composed of cells having a symmetric

Parameter	**Average**	**Standard deviation**
Mean of V_m [mV]	-66.94	3.91
Std of V_m [mV]	4.54	1.60
Skewness of V_m	0.61	0.53
S1	1.38	1.34
S2	0.40	0.40
kurtosis of Vm	0.34	1.42
Dip Test	4.21	12.66
Separability	22.32	15.38
Discretness	82.69	9.81
Regression coefficient from a fit of the distribution with a Gaussian	0.93	0.084
Regression of 2 Gaussians - Regression 1 Gaussian	0.054	0.074
Regression of 3 Gaussians-Regression of 2 Gaussians	0.0029	0.010
Slope coefficient of the Vm PSD	-2.74	0.43
Power ratio	0.93	0.44
Power in the delta band of the V_m PSD [dB]	0.41	15.86
Power in the theta band of the V_m PSD [dB]	30.83	7.79
Power in the alpha band of the V_m PSD [dB]	16.32	5.07
Power in the beta band of the V_m PSD [dB]	12.22	5.10
Power in the gamma band of the V_m PSD [dB]	14.40	11.09
Ratio between maximal autocorrelation and mean autocorrelation of V_m	3.87	2.71
Relaxation time of the autocorrelation [ms]	35.92	24.88
Average firing rate [Hz]	5.78	5.84
Instantaneous firing rate [Hz]	34.40	30.36
Coefficient of variation of interspikes intervals	1.28	0.40
Slope coefficient of the ISI distribution	-0.13	0.50

Figure 1.18: Parameters used for the classification of the cells.

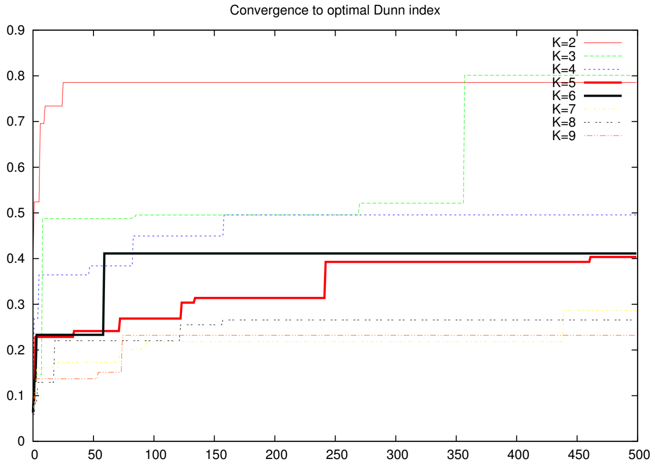

Figure 1.19: **Dunn index for Spt150 -** The evolution of the Dunn index over iterations is plotted for several values of K. The black curve corresponding to K=6 saturates faster and at a higher value than the red curve corresponding to K=5.

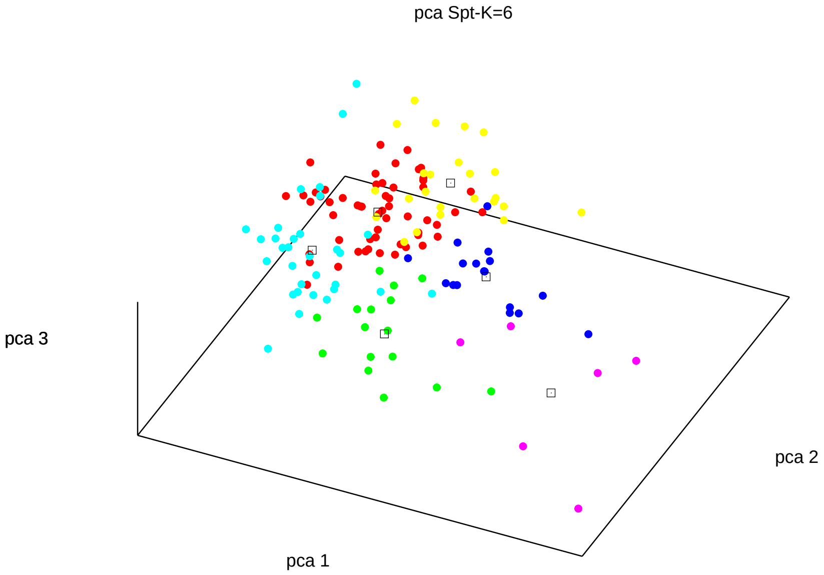

Figure 1.20: **Optimal partition of Spt for K=6-** Each cluster is represented by a different color and centroids are represented by square boxes. Axes are the 3 first principal components.

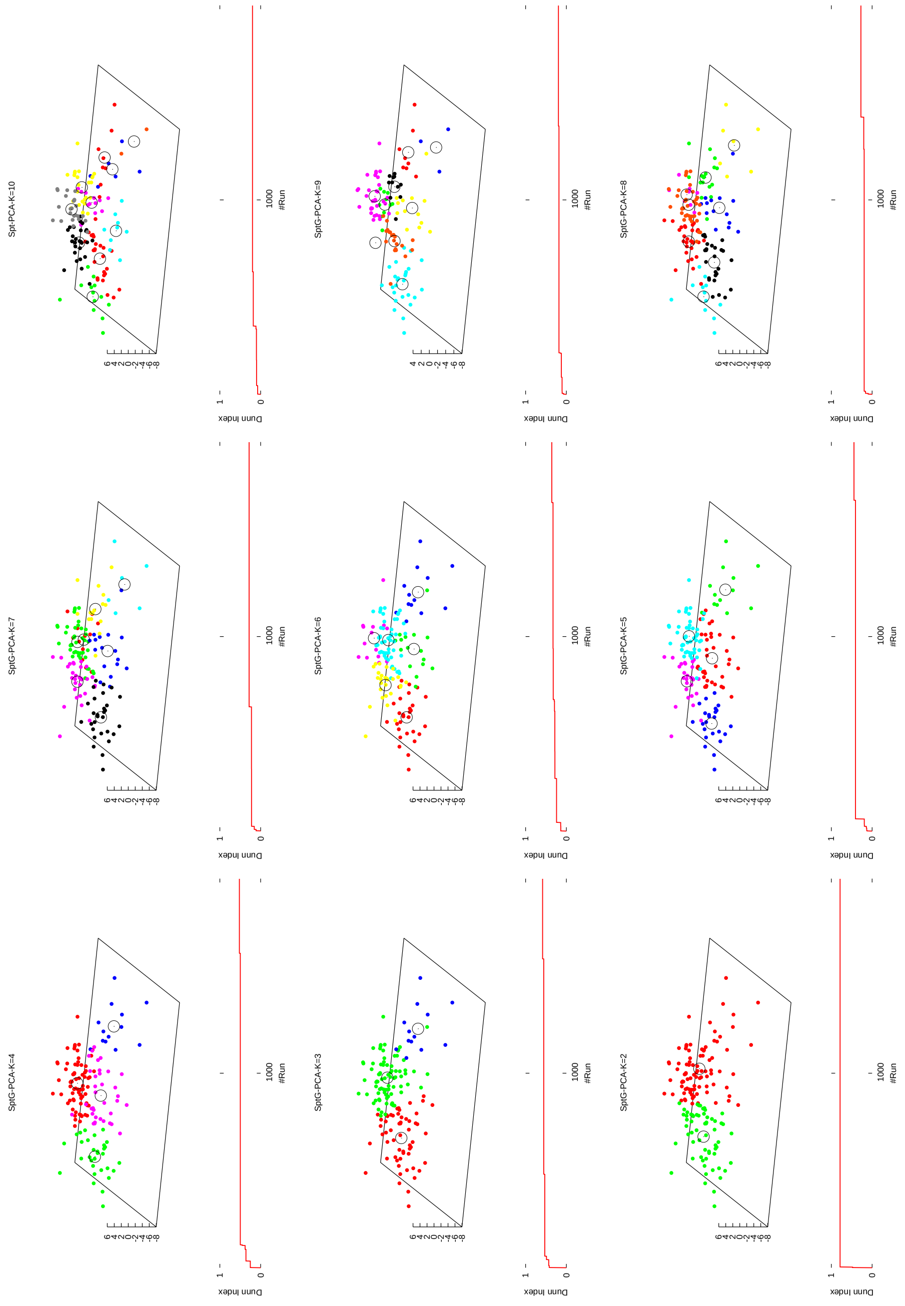

Figure 1.21: **Clustering of Spt150 for different values of K -** For each K, the clusters are represented by different colors in the PCA space. Red curves show the Dunn index evolution over 2000 iterations.

bimodal distribution and up and down transitions are correlated with the EEG fluctuations. The yellow cluster corresponds to very sharp distribution of V_m, that is cells with slow variations of V_m have correlations with the EEG. The deep blue and green clusters correspond to cells having asymmetric bimodal distributions dominated by the up and the down state respectively. The light blue cluster correspond to cells having a broad V_m distribution but only one peak. The data samples which are the closest from the centroids give a good summary of this partition and wo examples of membrane potential for each cluster with the associated EEG signal are shown on fig 1.22.

1.5.3 The visually evoked data set (143 cells).

For the visually evoked data set of 143 cells, Vis143, 11 components of the PCA are necessary to explain 90% of the variance. As can be seen on fig 1.23, the tree obtained for this data set is not well equilibrated because some few cells behaves very differently from the major part of the cells. This can be checked on the H matrix of fig 1.24, with all cells gathering in the fourth cluster of the tree based partition. The metasimilarity matrix describes the partition at a finer scale.

Clusterization of Spt143. The same algorithms were applied to the reduced data set of 143 cells of on-going for which the visually evoked activity is available. The K-means with the frequency of occurrence of the partition as an optimization criterion gives 4 clusters as optimal partition. Similarly to the clusterization of Vis143 cutting the tree in 4 clusters gives a poor result because it gathers most of the cells in a giant cluster.

Visually evoked activity compared to the spontaneous activity In the first three components of the PCA, the standard deviation is 1.25 for the spontaneous activity whereas it is 1.11 for the visually evoked activity. The visual stimulation thus pushes the activity toward the central red cluster of fig 1.27 corresponding to a Gaussian distribution of the membrane potential. There is no creation of a new domain for the dynamics as shown on fig 1.27. The comparison matrix for the clustering of the spontaneous activity data set and the visually evoked data set shows that there is a correspondance between the cluster 3 of the spontaneous activity data set and the cluster 1 of the visually evoked data set and another between cluster 2 of Spt and cluster 3 of Vis. Thus for clusters 2 and 3 of the spontaneous activity, cells don't jump to other clusters but stay close together when a visual stimulus is presented.

1.6 Conclusion

We presented a method to characterize and classify neuronal dynamics. Each classification has its own pitfalls and it is thus necessary to rely on a multi-algorithm approach to obtain a robust classification. We now summarize the classes obtained in the K means partition. The red cluster in fig 1.20 is the

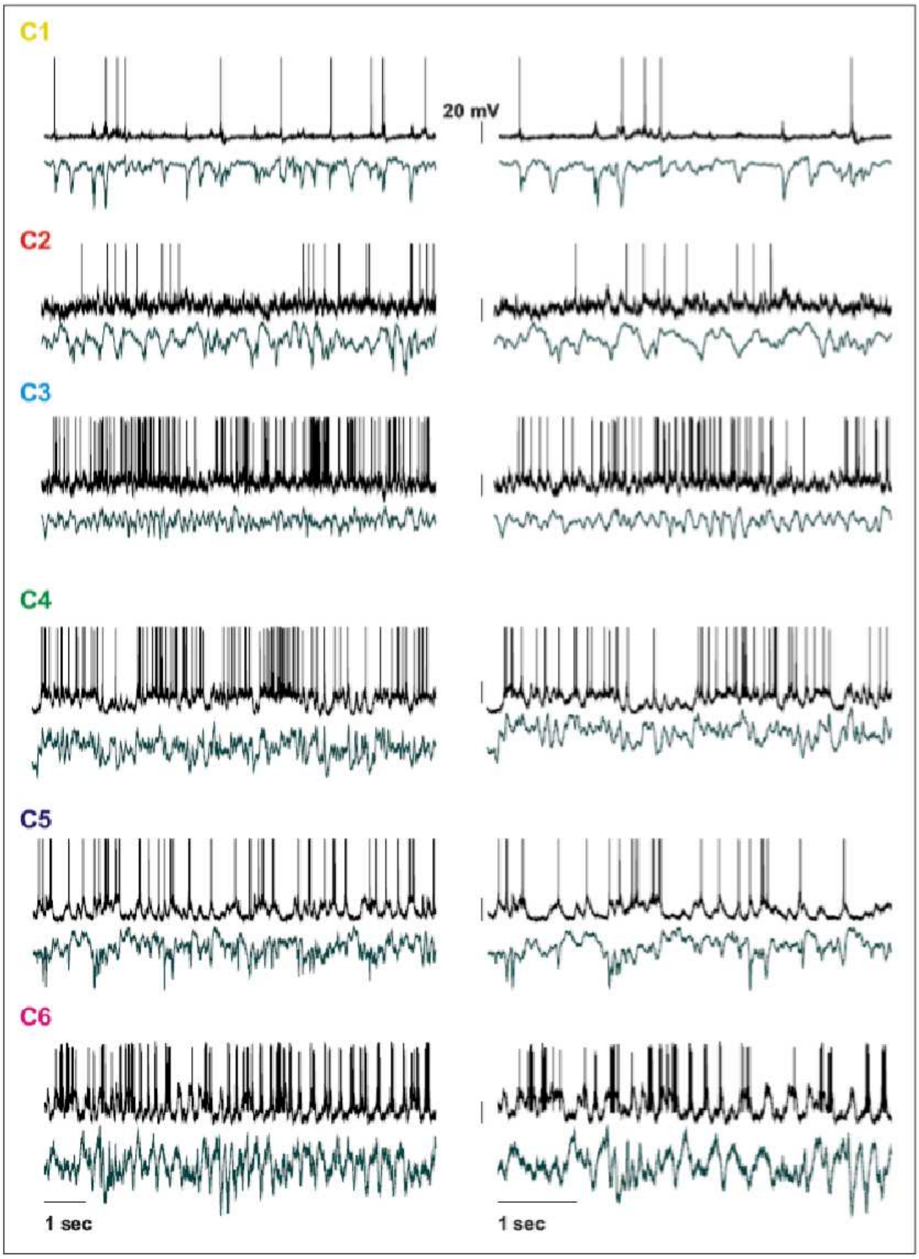

Figure 1.22: **Example of dynamics from the 6 clusters.** - For each cluster, the upper traces are membrane potentials and below is the EEG.

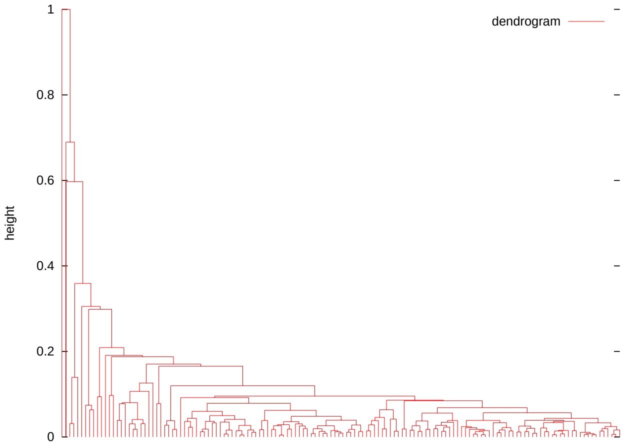

Figure 1.23: **Tree obtained for the 143 cells of the visually evoked activity dataset.**

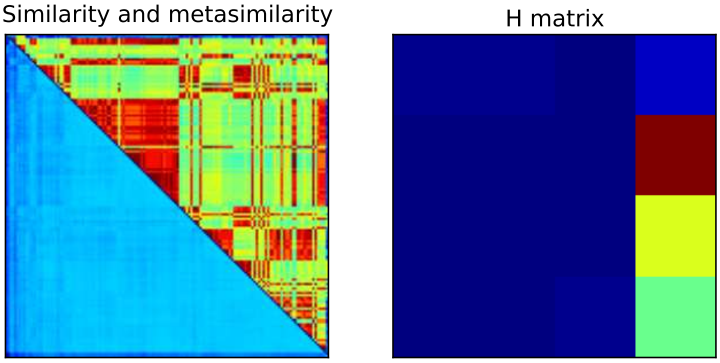

Figure 1.24: **Summary of the clustering algorithms for the 143 cells of the visually evoked activity data set.** - (Left) The lower triangle is the similarity matrix and the upper triangle is the metasimilarity matrix. (Right) H matrix obtained from the K-means partition and the cutting of the tree into four components.

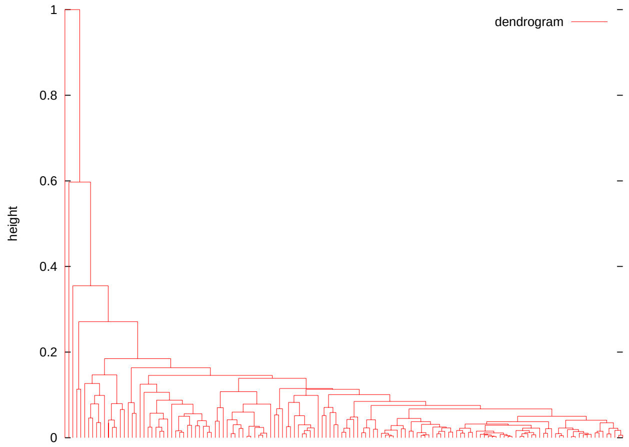

Figure 1.25: **Tree obtained for the 143 cells of the spontaneous activity dataset.**

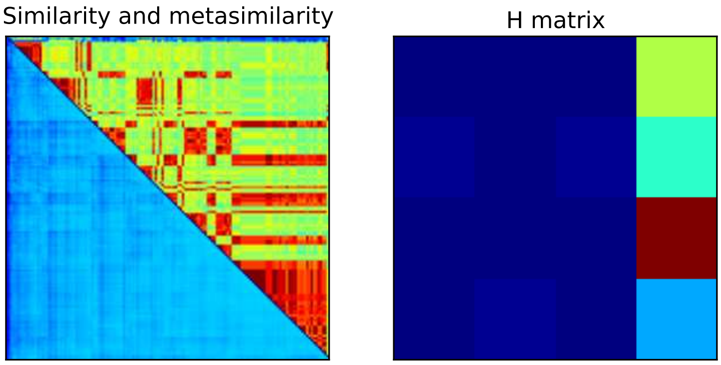

Figure 1.26: **Summary of the clustering algorithms for the 143 cells of the spontaneous activity data set. -** (Left) The lower triangle is the similarity matrix and the upper triangle is the metasimilarity matrix. (Right) H matrix obtained from the K-means partition and the cutting of the tree into four components.

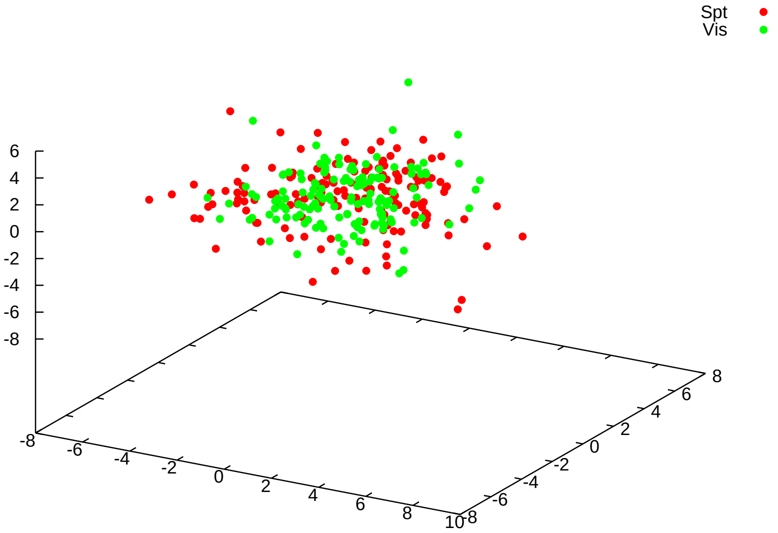

Figure 1.27: **PCA space. -** The on-going activity (red) and the visually evoked activity (green) data sets are projected on the PCAs of the on-going activity data set.

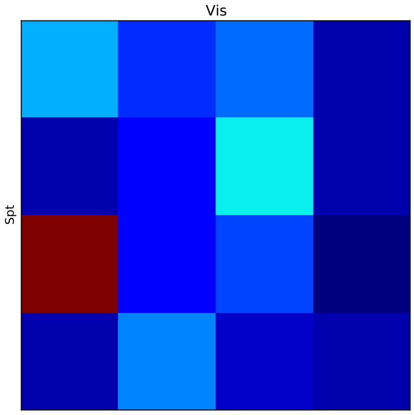

Figure 1.28: **H matrix for Spt and Vis. -** The on-going activity (red) and the visually evoked activity (green) data sets are projected on the PCAs of the on-going activity data set.

biggest cluster of Spt150 with 6 clusters and, during visual stimulation, neuronal dynamics gathers in this central cloud. Cells in this cluster are characterized by a Gaussian distribution of the membrane potential and we suggest that it defines an operating state of the network in which inputs are processed in a fast and efficient manner. Theoretical arguments explaining how such an asynchronous irregular state can be sustained in the network will be given in the next chapter. In the pink cluster, cells have a clear bistable behaviour and transitions between up and down states are correlated with the EEG suggesting a coherent low frequency oscillation at the network level, similar to that observed during slow wave sleep. In the yellow cluster, cells have only very few large deviations from the mean membrane potential suggesting discharge in a synfire mode. An estimation of the number of synchronous inputs generating the synaptic events may be computed from the measurement of the size of these events. Other classes includes cells with micro up or micro down states.

Morphological reconstruction of the neurons could determine whether some of the classes we obtained characterizes a specific cellular type or a cell can access any of the classes. It has been shown, in xylazine-ketamine preparations, that the up state of bistable cell share many similarities with the irregular activity of the awake state [74] so that the same cell could be in the red cluster or the pink cluster depending on the global state of the brain monitored by anesthesia. It would also be interesting to check if classes of neuronal dynamics are related to the states obtained after compression of long EEG recordings and if a cell jumps from a cluster to another during an EEG state transition.

The classes defined above can also be used to investigate how the functional properties of a cell depends on the state of the network in which it is embedded. In a work realized at the UNIC by Nicolas Benech on 118 cells of the data set we studied, it was shown that most of bistable cells have a complex receptive field and that their latency is longer than when cells have monomodal distribution characterizing the operating state. It was shown, in a xylazine-ketamine preparation, that the response to visual stimulation is enhanced when the stimulus is triggered during the up state [75] so that up states may be a cellular analog of attentional facilitation [76].