Classification of Population Activity in Parkinson's Disease

Gustav Röhss, Míriam Vall

VT 2020

Abstract

This project has been concerned with creating methods for feature extraction for the purpose of classification of brain activity in parkinsonian brains. Classification of LFP activity was based on feature vectors produced from the DFT of activity, and a heuristic-motivated visualization of these using the k-Means algorithm for prediction and classification. These feature vectors were also subject to principal component analysis as a further means of feature extraction and classification, highlighting differences in brain regions. Classification of spiking activity was based on spiking rates, joint rate distributions, serial correlation coefficients, power spectral density, and spectral entropy.

Sammanfattning

Vissa korrekta översättningar okända. Detta projekt har syftat att skapa metoder för att utvinna särdrag och att klassifiera hjärnaktivitet i parkinsonska hjärnor. Klassificering av lokal fältpotential baserades på särdragsvektorer producerade från DFT:n av aktivitet, och en heuristiskt motiverad visualisering av dessa användande k-Means algoritmen för förutsägning och klassificering. Dessa särdragsvektorer analyserades även med principalkomponentanalys för att utvinna ytterligare särdrag samt för klassificering, vilket påvisade skillnader i olika regioner av hjärnan. Klassificering av spetsaktivitet baserades på spetstakt, beroende taktfördelning, seriella korrelationskoefficenter, spektral effektdensitet, och spektral entropi.

Contents

1   Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder of movement that is age-related. It affects tens of millions of people worldwide, and the frequency and associated socioeconomic burden of the condition are set to increase as the elderly population grows.

The disease is characterized by poverty of voluntary movements (akinesia), slowness and impaired scaling of voluntary movement (bradykinesia), muscle rigidity and tremor of the limbs at rest (DeMaags 2015).

The hallmark feature of PD is the degeneration of dopamine neurons in the basal ganglia (BG), which is the region of the brain responsible for functions such as learning or movement (Hammond 2007). The BG consists of a few interconnected nuclei: the striatum (STR), globus pallidus (GP), subthalamic nucleus (STN), and substantia nigra (SN). The dopamine deficiency leads to a cascade of functional changes in the basal ganglia circuitry, which are ultimately responsible for the development of the main features of PD.

1.1   Purpose

The purpose of this project is to attempt to find one or several methods for classification of the brain activity in patients with Parkinson's disease.

Furthermore this project aims to, to some extent, use any produced methods to evaluate differences is brain activity of different categories. It is of interest to consider what differences any such methods show when comparing brain activity from different brain regions. It is also of interest to make a similar comparison for the brain activity in different subjects.

1.2   Delimitations

The authors of this report are not well educated or experienced in studying brain activity. The methods produced are mainly means to serve as a strong foundation for further research into deeper understanding of Parkinson's disease. The authors attempt to describe and interpret output produced by the methods, but do so outside of any broader implications the methods have for the further study of Parkinson's disease.

No software used, or produced, either by others or by the authors, is subject to extensive formal verification within the scope of this project. The same is true for the datasets used within the scope of this project. The datasets used are instead assumed to have been produced/recorded to a satisfactory quality for their uses within the scope of this project.

1.3   Research questions

- How can effective methods for classification of brain activity in patients with Parkinson's disease be produced?

- How can such methods be used to distinguish between brain activity from different regions of brains in patients with Parkinson's disease?

- How can such methods be used to distinguish between brain activity from different patients with Parkinson's disease?

2 Background

The most valuable contribution to the clarification of the functional changes occurring in PD has been provided by the animal models of the disease, particularly those applied to rodents and primates (Mallet 2008). In rodents, chronic and acute dopaminergic denervation (obtained by inflicting a selective lesion of the BG that mimics the effects of PD) leads to excessive beta oscillatory activity in the BG's LFP that can be suppressed by treatment (Mallet 2008).

We will attempt to observe such oscillations and synchronization in the dataset we have acquired for this project, which consists of several different measurements of brain activity in dopamine-depleted rodents. The used measurements of brain activity are:

- Local Field Potential (LFP): electrical signals generated in nervous tissues by the summed and synchronous activity of the individual neurons in that tissue.

- Spiking activity (also called action potential), which is a rapid rise and fall in voltage or membrane potential across a neuron's membrane. We have it in the form of background unit activity (BUA) and signal unit activity (SUA).

They were recorded simultaneously from three previously mentioned different regions of the BG; globus pallidus, striatum, and subthalamic nucleus, in sessions of 100 seconds at a sampling frequency of 16000 Hz.

Therefore, for these parts of the BG we dispose of information on LFP and spiking activity; the data for each specific type of recording (e.g. LFP in the GP), for a specific animal and recording session will be called a "channel". These continuous time series provide insight into the synchronous LFP and spike discharges of local neuronal ensembles, and are relevant to us as we aim to study the variation and synchrony of simultaneous brain activity in different regions of the BG and attempt to classify patients into different categories.

2.1 Beta oscillations and neuronal synchronization

Dopamine coordinates neuronal activity in the frequency domain. When controlled by dopamine, beta oscillations (oscillations inside the range of 15 to 30 Hz) in basal ganglia circuits are important for normal movement (Cagnan 2019). However, dopamine loss in the basal ganglia may not only support but also actively promote the emergence of excessively synchronized beta oscillations at a network level, which indicate pathological synchronization between the regions of the BG (Hammond 2007). This synchronization does not appear in non-parkinsonian brains, and it probably reflects a variety of pathophysiological mechanisms in the parkinsonian state.

2.2 Discrete Fourier transform

The Fourier transform, or more specifically, the family of Fourier transforms, are mathematical tools with a long and rich history and many use cases. The Discrete Fourier Transform (DFT) is the version of the Fourier Transform used on discrete points of data (rather than e.g. a continuous function). One use case of the Fourier transform is to transform a function of time into a function of frequency. The DFT can be said to convert data from the *temporal* (time) domain to the *spectral* (frequency) domain.

Specifically, the Fourier transform can be used to approximately decompose a function, or a series, into a large number of waves of different frequencies and amplitudes. This method can be used to approximate the amplitude or power of activity in specific frequencies in a signal made up of waves of many frequencies (MathWorks n.d.).

NumPy, a software library, provides useful tools for usage of the DFT in its' `fft` package, specifically the `numpy.fft.fft` (`FFT`) function (Oliphant 2006–).

`FFT` can also be given an input argument to pad the input array with additional zeroes. The user then receives a *higher fidelity* output; output information for a larger amount of frequencies. This can be shown by the `FFT`-helper function, `numpy.fft.fftfreq` (`FFTFREQ`). The `FFTFREQ` function takes arguments *window length* and *sample spacing*, and returns an array of *unit frequency bin centers*.

The amplitude spectrum for `FFT` output is obtained by taking the absolute values of the output from the `FFT` complex-valued output array, specifically using the `numpy.abs` (`ABS`) function (Oliphant 2006–).

2.3 k-Means

The k-Means algorithm is a clustering algorithm. One noticeable peculiarity of the k-Means algorithm is that the user makes a choice of k, the amount of clusters.

The algorithm works by first randomly generating k initial *cluster mean vectors*. These are vectors with the same dimensionality as the data samples to be clustered. The algorithm then attempts to minimize the *within-cluster sum of squares* of samples assigned to each cluster; the sum of square (Euclidean) distances from the cluster mean vectors, taken over the individual data samples. The algorithm works iteratively.

- Each sample is assigned to the cluster for which the square distance is minimized.

- New cluster mean vectors are created from the mean of the new assignments of samples to clusters.

The iteration ends when the cluster mean vectors no longer change (possibly within some tolerance), or a set amount of iterations is reached (Bruce & Bruce 2017, p258-260).

The *scikit learn* software library has an implementation of the k-Means algorithm in its' `sklearn.cluster.KMeans` module (`KMeans`), and is the implementation used in this project (Pedregosa et al. 2011).

2.4 Principal component analysis

One method for extracting lower-dimensional features from higher-dimensional data is by using principal component analysis (PCA). Specifically, PCA refers to the computation and use of principal components (PCs). For a set of data, a PC is a direction in the space of the data's features along which the data samples are highly variable. It's possible for a linear combination of PCs to describe all samples in a dataset with a great degree of accuracy. If the amount of PCs required for this is lower than the amount of features in the data samples, PCA becomes an effective means of feature reduction for that dataset. The PCs produced can also be used to visualize the data, and the individual components can be interesting for analyzing the data in their own right (James et al. 2017, p374-380).

The *scikit learn* software library enables easy computation of PCs using `sklearn.decomposition.PCA` (`PCA`). It also allows the user to transform members of a dataset into their respective representation under a certain set of PCs. It should be noted that such a representation is often approximate. Furthermore, the user is able to see the *explained variance* and *ratio of explained variance* for each PC produced for a specific dataset (Pedregosa et al. 2011).

2.5 Spiking activity as a simple Poisson process

In neurosciences, the action potentials (spikes) are the main components for the real-time information processing in the brain (Eden 2015). The duration of each spike is very small, about one millisecond. It is consequently quite natural to assimilate a spike to a random event. Therefore we can model a spike train (a sequence of recorded times at which a neuron fires a spike) as a stochastic process (or point process), since they are widely used to represent events that appear to vary in a random manner over time.

A stochastic process may be specified in terms of spike counts. It is often useful, for constructing data analysis methods, to consider point processes in discrete time (Eden 2015). One approach for constructing a discrete-time representation of a point process is to partition the whole recorded time into smaller time windows that are equally sized. It is then possible to express a spike train in terms of discrete counts of spikes fired during successive time windows.

This discrete time description loses information about the precise timing of spikes within a time window. However, in the limit as the discrete-time window size goes to zero, the discrete-time spike counts become as informative as the continuous-time descriptions, and the likelihoods and probability distributions associated with the discrete-time methods converge to their continuous-time counterparts (Eden 2015). In many cases, the analysis methods are easier to develop and implement in discrete-time.

The first naive model that can be used is the most basic stochastic process: a homogeneous Poisson process, which is extensively used in neurosciences due to its simplicity (Gerstner 2014). As it is common in all Poisson processes, the number of spikes in each time window is completely independent of all the others. This model is only based on the firing rate λ of the process, which remains constant over time. But it is precisely for this reason that several features of spike trains cannot be modelled, in particular:

- Their non-stationarity along time.

- The dependence with respect to the previous spikes in the same spike train.

- The dependence with respect to spikes of other spike trains (i.e. neuronal synchronization).

2.6 Inhomogeneous Poisson processes for spike count train modelling

As is likely the case with this project's dataset, the process could be non-stationary as the firing rate could vary depending on the time windows. One of the easiest ways to take into account this non-stationarity is to allow the Poisson process to be inhomogeneous. In this case, the firing rate is now time-dependent and is modelled by a function $\lambda(t)$, which is the spiking rate (or intensity) of the inhomogeneous Poisson process. This function is constituted by the quotients of the spike counts and the time length of the windows.

Assuming that a spike count train is an inhomogeneous Poisson process with time-dependent firing rate $\lambda(t)$:

- For two disjoint time windows I and J, their spike counts $N(I)$ and $N(J)$ are independent.

- For any time window I, $N(I)$ is a Poisson variable.

As a result of the independence assumption, this process still has a memory-less property, by which the probability of spiking at any instant does not depend on occurrences or timing of past spikes; there is no history dependence. However, history dependence is an important component of virtually all neural spiking activity and it must be taken into account if the objective is to identify patterns in said activity (H.Perkel 1967). Therefore, it is necessary to further generalize the point process model by removing the independence assumption to capture the history dependent structure.

2.7 Renewal processes for representing history dependence

The simplest type of point process with history dependence is the renewal process. A renewal process is a point process for which the probability of firing a

spike at any point in time can depend on the occurrence time of the last spike, but not on any spikes before then (H.Perkel 1967).

Joint firing rate scattergrams can be employed in determining serial correlations between spiking intensities, where each point on the diagram then represents a pair of "adjacent" rates.

A quantitative measure of such correlation is provided by the serial correlation coefficient of firing rates, which are statistics based on the joint distributions of the rates. The set of serial correlation coefficients is usually called a serial correlogram; it has sometimes been called "autocorrelogram" or simply "autocorrelation" (H.Perkel 1967).

Given the autocorrelation function $x(n)$, the power spectral density $S(m) = |X(m)|^2$, where $X(m)$ is the DFT of $x(n)$. It is also called power spectrum or noise spectrum. Noise is a limiting factor to all forms of information transmission; in this particular case, information transmission by neurons (Gerstner 2014). An important concept of the theory of signal transmission is the signal-to-noise ratio. A signal that is transmitted at a certain frequency ω should be stronger than (or at least of the same order of magnitude as) the noise at the same frequency (Gerstner 2014). For this reason, the noise spectrum $S(\omega)$ of the transmission channel is of interest. By observing the noise spectrum, the frequencies at which the variance or noise is higher (and lower) are identified.

By normalizing the power spectrum to be seen as a probability density function, and later applying the standard formula for Shannon entropy to it, the spectral entropy of the spike train (or channel) can be obtained. It serves as an entropy related index able to measure part of the structural complexity of the time series, as it quantifies the degree of order and predictability derived by the series' power spectrum (Zaccarelli 2011).

Given the power spectral density $S(m)$, the probability distribution $P(m)$ is:

$$P(m) = \frac{S(m)}{\sum_i S(i)} \tag{1}$$

The spectral entropy H is:

$$H = -\sum_{m=1}^{N} P(m)log_2 P(m) \tag{2}$$

Normalizing:

$$H = -\frac{\sum_{m=1}^{N} P(m)log_2 P(m)}{log_2 N} \tag{3}$$

Where N is the total amount of frequency points. The denominator $log_2 N$ represents the maximal spectral entropy of white noise, uniformly distributed in the frequency domain.

3 Methods

Some methods in this project focused on amplitudes of activity in the spectral domain for the data used in the scope of this project. The reasoning behind this is that previous research on Parkinson's disease has shown that LFP activity in the basal ganglia in the beta-range is abnormally synchronized compared to that of the same activity in subjects unaffected by Parkinson's disease, as was discussed in subsection 2.1 (Cagnan 2019). This suggests that such activity might embed additional useful information for research, and possibly a means for classification.

Apart from focusing on LFP recordings, the analysis of spiking rates also provides insight into the activity and synchronization of the different basal ganglia regions, and can be used to extract relevant features which are apt for classification.

3.1 Spectrum feature extraction

One important method for feature extraction used in this project was the DFT, using `FFT`. The data was first split into uniform-sized (in array length, or equivalently length of time) *epochs*. Software implemented for this end was designed such that the *epoch size* could be varied. Each such epoch was then transformed using `FFT`.

Interpreting `FFTFREQ` for the data used in this project, it takes input epoch size (in number of points), and time between samples (multiplicative inverse of sampling frequency). It should be noted that the "epoch size" argument given to `FFTFREQ` should match the optional argument given to `FFT` which causes zero-padding (resulting in higher fidelity output).

`FFTFREQ` returns an *array index-to-frequency in Hertz* mapping for the output of `FFT` in this context (Oliphant 2006–).

The `ABS` function was used to produce the amplitude spectrum of `FFT` output. `FFTFREQ` was used to find indices in the `FFT` output representing frequencies in a certain range, and selecting such values. This range was implemented to be variable.

Using this process, for each epoch a *feature vector* was produced. Each value in the vector represents an amplitude of LFP activity for a specific frequency, epoch, channel, and session with a specific set of parameters for fidelity, epoch size, and range of frequencies. For ease of reference, these vectors will be referred to as *spectrum feature vectors* (SFVs) in this report.

3.2 k-Means as a visualization heuristic

Initially, k-Means was considered as a means for clustering SFVs in an attempt at classification. Attempting to support this choice with an argument as to why it is an appropriate choice of algorithm for this particular use case proved difficult. However, a heuristic argument can be presented.

The SFVs represent a portion of an amplitude spectrum. The square of an amplitude spectrum is a power spectrum (Oliphant 2006–). The distance between a cluster mean vector (as generated by k-Means) and a SFV can therefor be interpreted as a "difference in amplitude"-spectrum. As such, the square of this distance can be interpreted as a "difference in power"-spectrum. Then, when using k-Means with the SFVs, the "distance" the algorithm will be attempting to minimize is the sum of this "difference in power"-spectrum; or just difference in power.

To reiterate, the presented argument should not be considered extensive for this use case, but rather should be considered a heuristic motivation for its' use as, primarily, a method of visualization.

With this in mind, `KMeans` was used to produce cluster means and assignments for some subsets of SFVs produced. Specifically, a set of SFVs for all channels of a certain session were first produced. A subset of these SFVs were then used as training data for `KMeans`. The resulting `KMeans` model was then used to classify (predict) the entire session-SFV-set. This procedure was repeated for several different sessions. The parameters used for k for `KMeans`, the epochs size, and the range of frequencies included in the SFVs are able to be varied in order to produce extensive results.

3.3 Usage of principal component analysis

In order to better describe the SFVs, PCA was used.

The specific PCs of the SFVs are of interest, as they describe the spectrum-components along which LFP activity vary the most. Analyzing these specific PCs and how their prominence varies in different subsets could highlight key similarities and differences in the LFP activity of different brain regions or subjects.

Should a small amount of PCs prove able to explain a high ratio of variance in the dataset, this would serve as a strong means for feature reduction, which could then be used in further research. The PCs produced would then also describe in a more easily digestible manner the key spectral components of any LFP activity in this dataset, and have implications for using similar methods for classification of other sets of LFP activity, and possibly brain activity and time-signals in general. Should the distribution of SFVs transformed into PC representations under this model be considerably different for different brain regions or animals, this would have implications for classification attempts of LFP activity.

The process for producing PCs was straightforward. Using `PCA` on the entire set of SFVs, the PCs for these sets were produced.

3.4 Poisson process

The original dataset consists of the number of spikes counted during successive time windows (each of them being half a second). For each recorded channel in

different regions of the basal ganglia, which in this case are the globus pallidus and subthalamic nucleus, 393 recorded time windows were available.

The non-stationarity of the spiking rate over time makes it impossible to model the spike count trains as a homogeneous Poisson process with a constant rate; therefore, it was regarded as an inhomogeneous Poisson process.

It is possible to obtain the time-dependent spiking rate function by dividing each spike count by the duration of the time windows, namely 0.5 seconds. We also produced an estimation of the rate probability distribution by plotting a histogram of the time series.

3.5   Renewal process

Because of the fact that even as an inhomogeneous Poisson process the dependence between spikes cannot be accounted for, the spike count trains were also modelled as a renewal process.

With this approach, the history dependence of spike trains was able to be modelled and was used to extract different measurements that are relevant to examine spiking activity. They were obtained in a sequential manner, as one measurement often depends on the one that was acquired before.

Joint firing rate scattergrams were employed to determine serial correlations between spiking intensities. To quantify these serial correlations between spiking rates, the autocorrelation (or serial correlation coefficients) of the spiking rate function was then calculated. Next, the power spectral density was drawn from the autocorrelation measure, by calculating the square of its DFT. Lastly, using the power spectral density values, the spectral entropy of each channel was obtained to measure the complexity of the signals.

4 Results

This subsection is dedicated to presenting results that were produced from the methods described in the previous section.

4.1 Spectrum feature vectors

Produced SFVs can be visualized using the *matplotlib* software library's useful `matplotlib.pyplot.imshow` (IMSHOW) function (Hunter 2007).

Figure 1 and figure 2 show SFVs for channels `gp_lfp1` and `str_lfp10` for sessions `NPR052e.10` and `NPR064.b08`, respectively. For both figures, each column represents a single SFV, generated for a specific epoch of activity, displayed chronologically. For both figures, higher brightness represent higher amplitude of LFP activity in that epoch and frequency. Figure 1 shows 780 epochs, and figure 2 shows 778 epochs. For both of these sessions the sampling frequency is 16 kHz. For both of these sets of produced SFVs, the epoch size in amount of data points (sampled at sampling frequency) is 2048, resulting in an epoch size of 128 ms. For each of these sets of produced SFVs, there are 46 frequencies sampled. Specifically, the input was padded with to a length of 16384 in order to produce a higher fidelity output, and all frequency samples in the ranges of 5 Hz - 50 Hz were selected. This results in 46 equally-spaced frequency samples, the lowest being approximately 5.86 Hz and the highest being approximately 49.8 Hz. The parameters can be considered arbitrary due to easily being changed. In this specific configuration, the frequency range was selected to be somewhat wider than the beta-range of frequencies. The epoch size was chosen to be somewhat longer than the shortest length of a beta-burst (Cagnan 2019). The fidelity was chosen arbitrarily.

These figures are, clearly, a very small subset of the total set of produced SFVs. They also represent only a very small subset of possible configurations of SFVs in regards to epoch size, sampled frequencies, and amount of frequency samples.
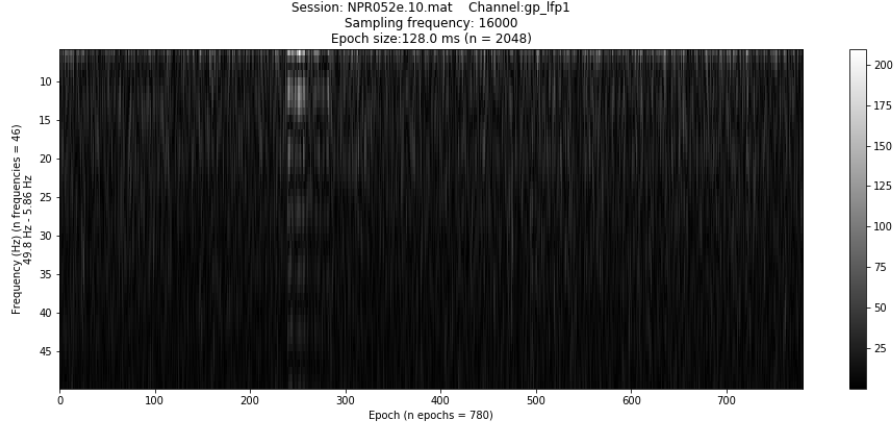
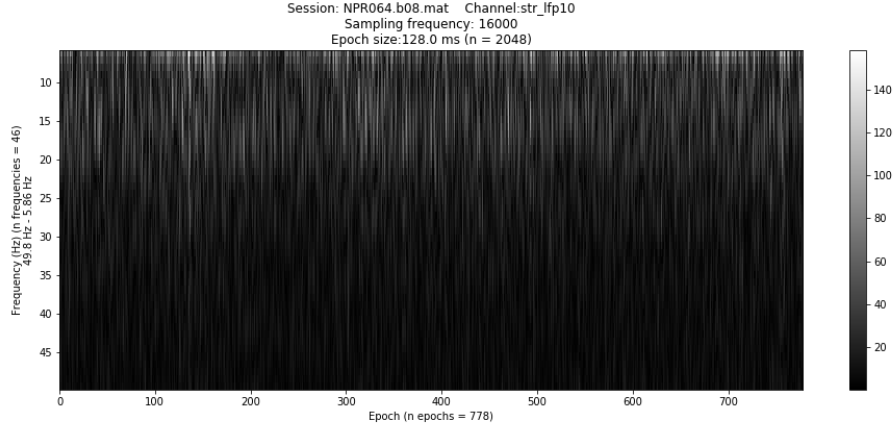Figure 1: SFVs for Globus Pallidus LFP channel of a specific session.

Figure 2: SFVs for Striatum LFP channel of a specific session.

4.2   k-Means-based visualization

To visualize k-Means predictions for a k-Means model trained on SFVs, `IMSHOW` was used once more. Some of the produced figures are figure 3 and figure 4. In order to better relate to the heuristic argument presented in subsection 3.2, the colors for class assignments were chosen to be sorted by the "sum of power spectrum" (sum of squares) of the cluster centers. Lower class index (refer to color bar in either figure) represents lower sum of power spectrum.

In the figures, each row represents the class assignments for a specific channel in the set. The columns represent epochs. Each colored "slice" is the prediction of a single SFV to a class, the result of a trained `KMeans` model. The particular

14

channel names are not shown in these figures, but include both STR and GP channels. The SFVs are identical in production parameters to those presented in subsection 4.1.

The most important thing to note about these figures is how different channels in the same (or close) epochs tend to be given equal class assignments excessively. The authors interpret the class assignments for adjacent epochs to often be "close", meaning that class assignments seem to be followed by slightly higher or slightly lower class assignments (referring to class index), this was however not researched more thoroughly and should be considered an informal observation. Also of note are "streaks" of higher class assignments for specific channels. Examples include for channel 4 at about epoch 700, and channel 12 at about epochs 330-420 (rough estimates), in both figures. There are also "pillars" of excessively similar assignments for all channels at about epochs 140, 290, and 550 (rough estimates), in both figures.
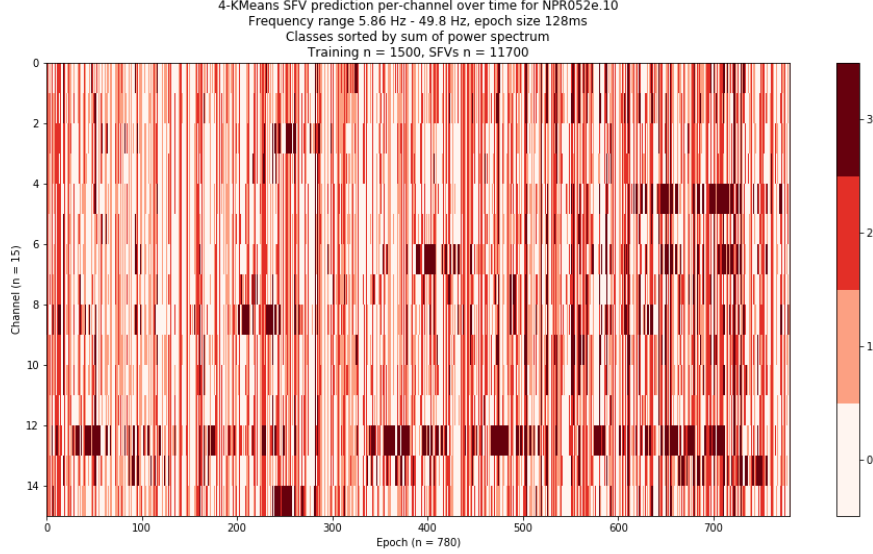
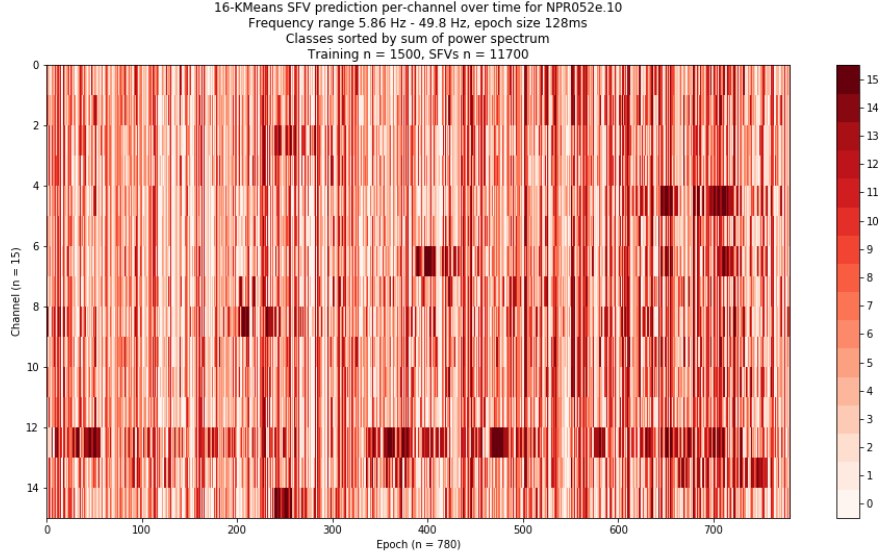Figure 3: 4-k-Means of channels for specific session

Figure 4: 16-k-Means of channels for specific session

4.3 Principal component analysis of spectrum feature vectors

The method outlined in subsection 3.3 was performed using a set of 362,184 SFVs produced with parameters identical to those described in subsection 4.1. Indeed, the SFVs shown in figures 1 and 2 are part of the set used to compute the `PCA` model.

The first thing to consider about the PCs computed for the set of SFVs are the PCs themselves. Figure 5 shows these, as well as their respective explained variance ratios. This figure was produced using the `matplotlib.pyplot.plot` (`PLOT`) function, as were all other figures in this subsection. In this specific case, eight PCs were produced, as this was the points where their cumulative sum of explained variance ratio was above 90%. These PCs can be interpreted as being parts of which a linear combination approximately describe all of the SFVs from which they were computed.
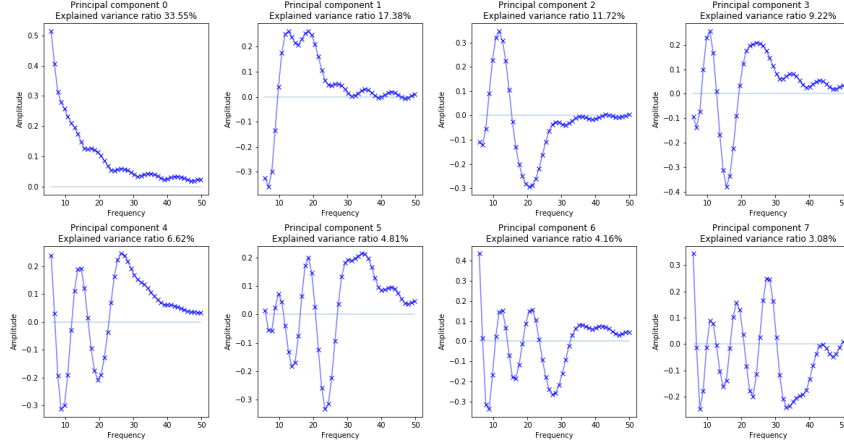
Figure 5: Principal components of SFV-set

Having computed these PCs, the approximate probability distributions of the individual components as they appear in the set of PCA-transformed SFVs are shown in figure 6. In order to better visualize the long tails of these distributions, the probability distributions of the logarithms (base 2) of the individual components are shown in figure 7.
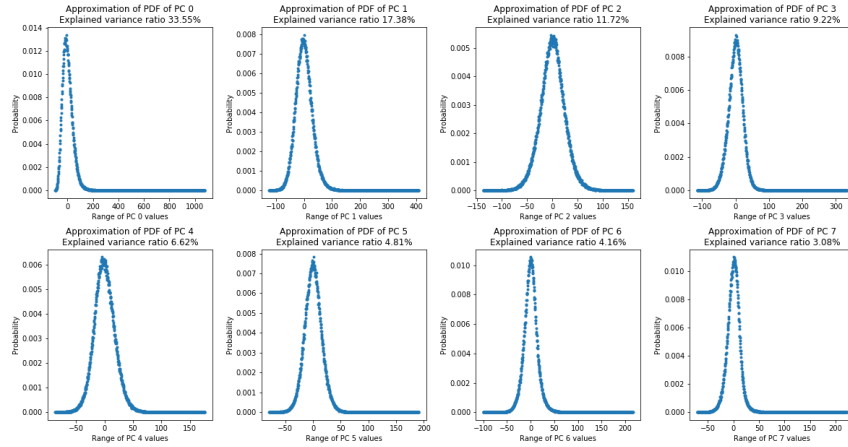

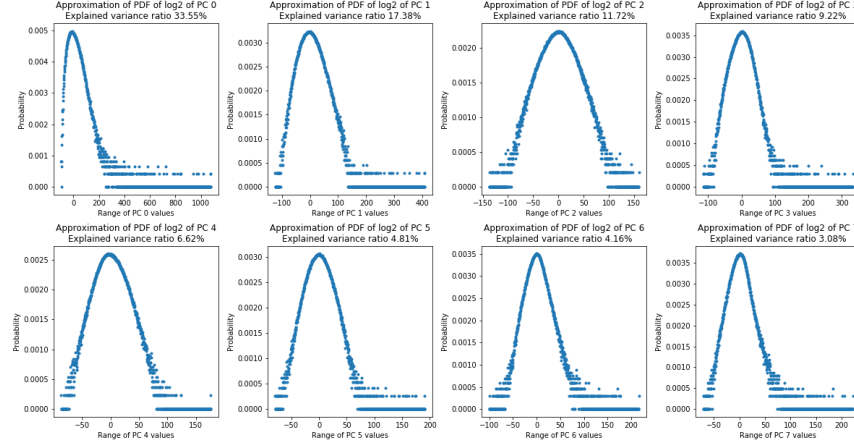Figure 6: Approximate distributions of principal components of SFV-set

Figure 7: Approximate distributions of log_2 of principal components of SFV-set

The information shown in figures 6 and 7 is also shown in figures 8 and 9, except with channels from STR kept separate from those of GP. Notably, GP channels exhibit distributions with more width and higher peaks.
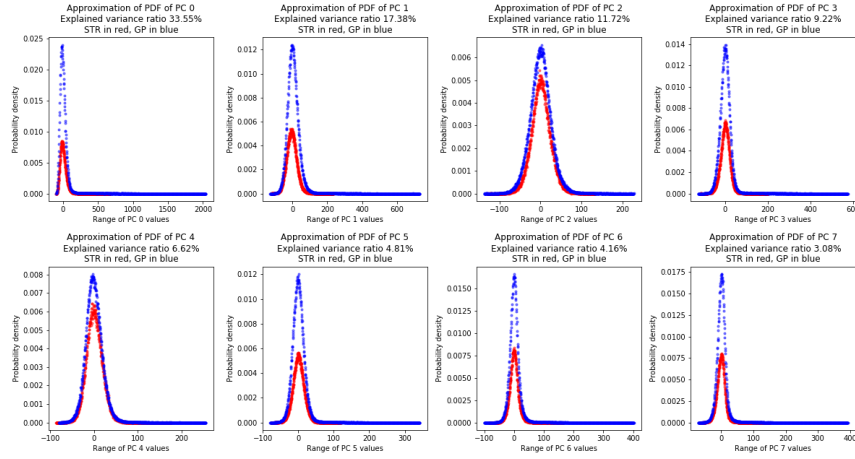

Figure 8: Approximate distributions of principal components of SFV-set, separated by channel type
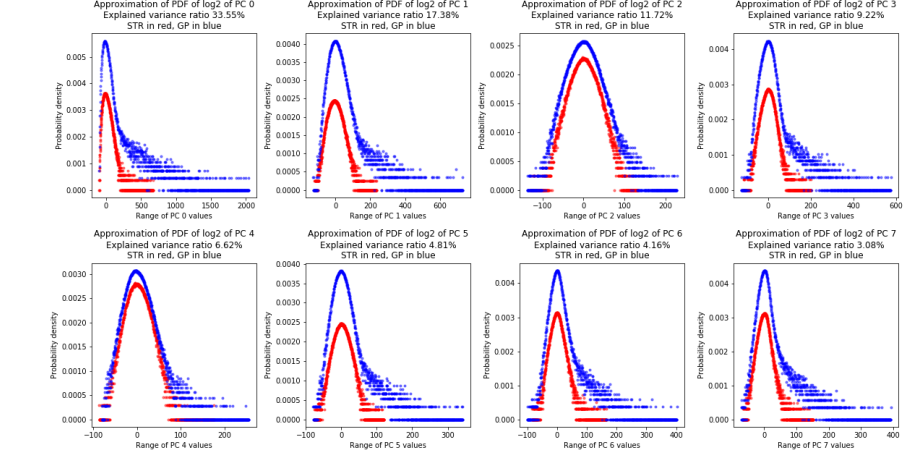
Figure 9: Approximate distributions of log_2 of principal components of SFV-set, separated by channel type

These figures, of course, don't show how these distributions correlate. Tables 1 and 2 show the cross-correlations of PCs of `PCA`-transformed SFVs, for GP and STR channels respectively. Notably, cross-correlation within STR channels is consistently low, while it is consistently considerably higher within GP channels.

	PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
PC 0	1.0	0.43	0.17	0.44	0.23	0.39	0.44	0.5
PC 1	0.43	1.0	0.1	0.26	0.13	0.26	0.26	0.31
PC 2	0.17	0.1	1.0	0.11	0.07	0.08	0.12	0.13
PC 3	0.44	0.26	0.11	1.0	0.16	0.26	0.31	0.35
PC 4	0.23	0.13	0.07	0.16	1.0	0.13	0.16	0.19
PC 5	0.39	0.26	0.08	0.26	0.13	1.0	0.27	0.32
PC 6	0.44	0.26	0.12	0.31	0.16	0.27	1.0	0.35
PC 7	0.5	0.31	0.13	0.35	0.19	0.32	0.35	1.0

Table 1: Cross-correlation of PCs, GP channels, rounded to two decimals.

	PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
PC 0	1.0	-0.05	0.01	0.0	0.0	-0.01	-0.0	0.01
PC 1	-0.05	1.0	0.05	0.08	0.03	0.01	0.04	0.03
PC 2	0.01	0.05	1.0	0.03	-0.02	0.02	0.02	-0.01
PC 3	0.0	0.08	0.03	1.0	-0.0	0.03	-0.02	0.01
PC 4	0.0	0.03	-0.02	-0.0	1.0	0.01	0.01	-0.0
PC 5	-0.01	0.01	0.02	0.03	0.01	1.0	-0.0	-0.01
PC 6	-0.0	0.04	0.02	-0.02	0.01	-0.0	1.0	0.0
PC 7	0.01	0.03	-0.01	0.01	-0.0	-0.01	0.0	1.0

Table 2: Cross-correlation of PCs, STR channels, rounded to two decimals.

4.4   Spiking analysis results

This subsection is dedicated to the results of analysis of spiking rate data, by methods described in All results are produced taking 50 successive time windows of 0.5 seconds each.

4.4.1   Spiking rates

The figures in this sub-subsection were produced using `IMSHOW`. The histograms were obtained with the matplotlib.pyplot.hist function.

For all figures plots, rows represent channels and columns represent time windows. Figure 10 and figure 11 show the spiking rates for the channels in the GP and STN, respectively. Figure 12 shows the spiking rate functions of the channels of both regions (channels 0-6 are from the GP, channels 7-16 are from the STN).
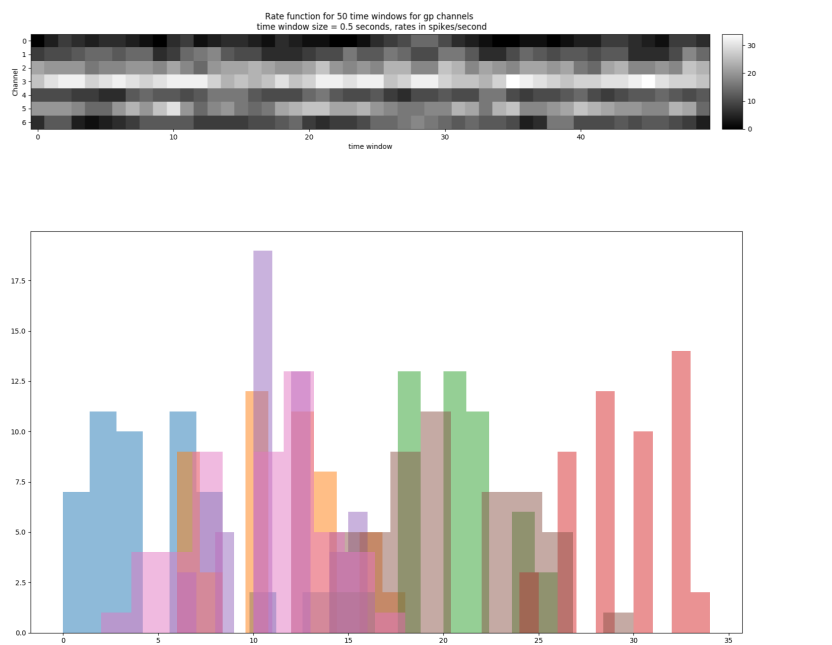
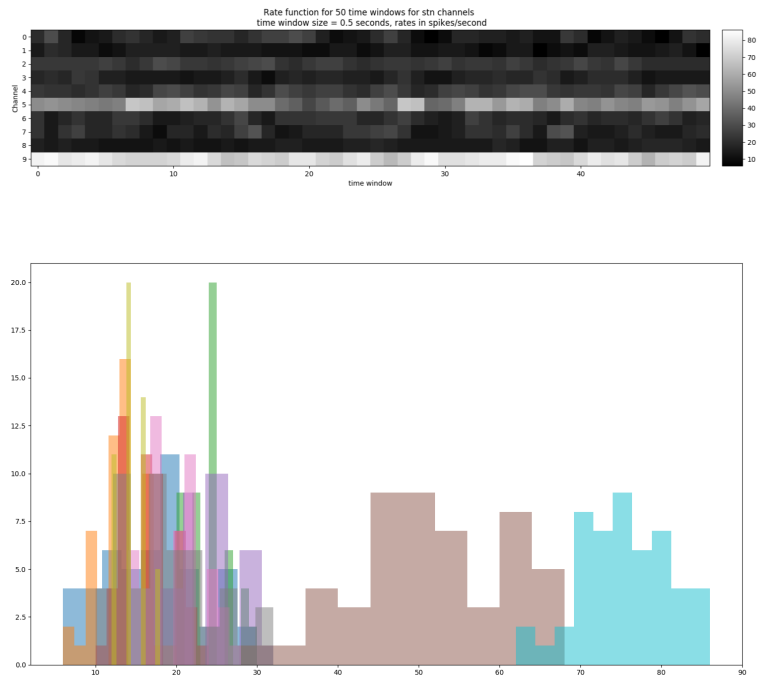Figure 10: Spiking rates of GP channels.
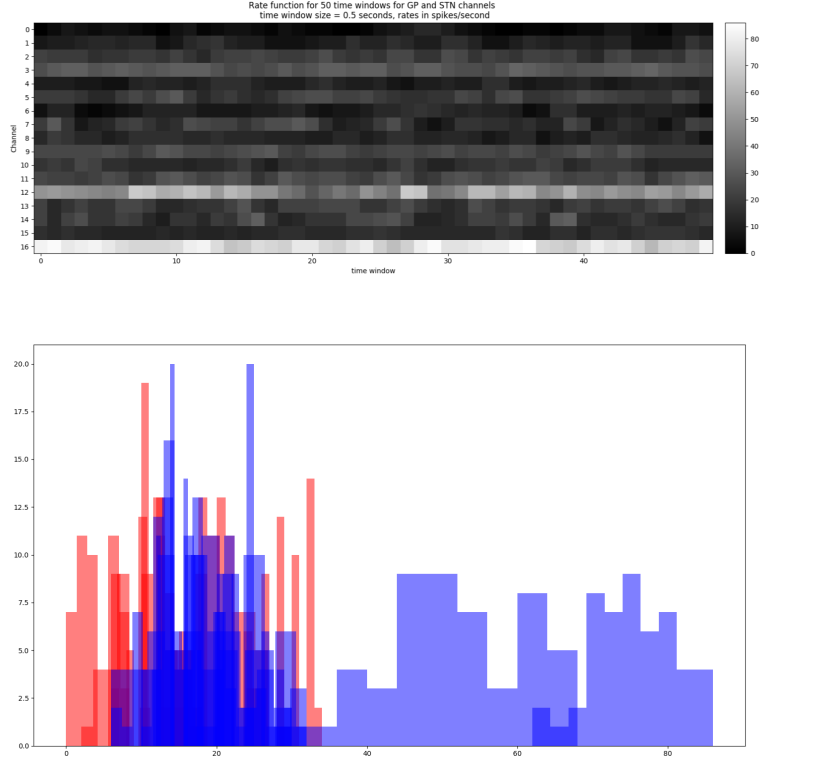
Figure 11: Spiking rates of STN channels.

Figure 12: Spiking rates of GP and STN channels.

4.4.2 Joint rate distributions

The data is displayed in the form of scatter diagrams, in which the sequence of firing rates (represented by the abscissa) is plotted against the same sequence with a lag of one time window (represented by the ordinate). Each point on the diagram then represents a pair of "adjacent" rates.

The hexagonal scatterplots in this sub-subsection were produced with the Seaborn library's `seaborn.jointplot` (Waskom et al. 2020) function, and the normal ones with the `matplotlib.pyplot.scatter` function.

The hexagonal scatterplots are better at showing the density of the overall joint distribution, whereas the normal ones differentiate among channels (in the case they are all from the same region), or among regions (if both regions are plotted together).

Figure 13 and figure 14 respectively show the joint probability distributions of the channels in the GP and STN, whereas figure 15 displays the joint distribution of every channel in both regions. In the non-hexagonal scatterplots, there is one color per channel (in the case of Figure 1 and 2), or red for GP and
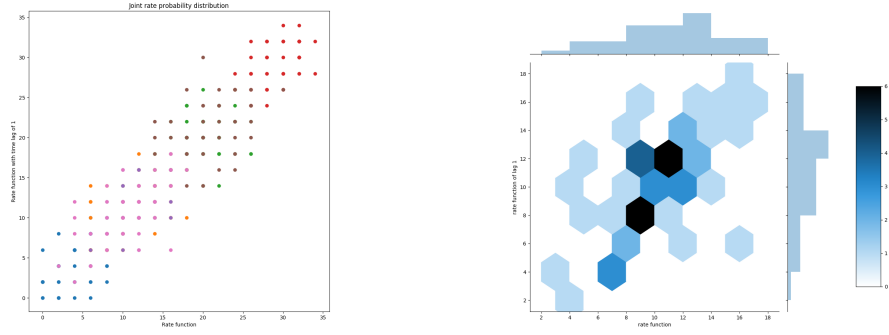
blue for STN (in Figure 3).

Figure 13: Joint probability distribution of the spiking rates of all GP channels and these same sequences with a time lag of one window.
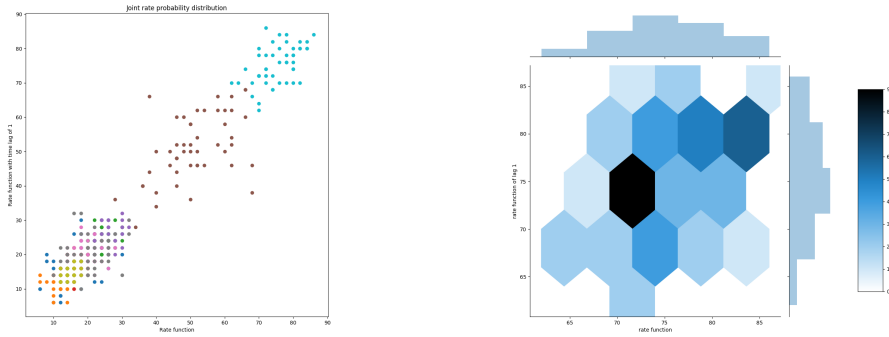

Figure 14: Joint probability distribution of the spiking rates of all STN channels and these same sequences with a time lag of one window.
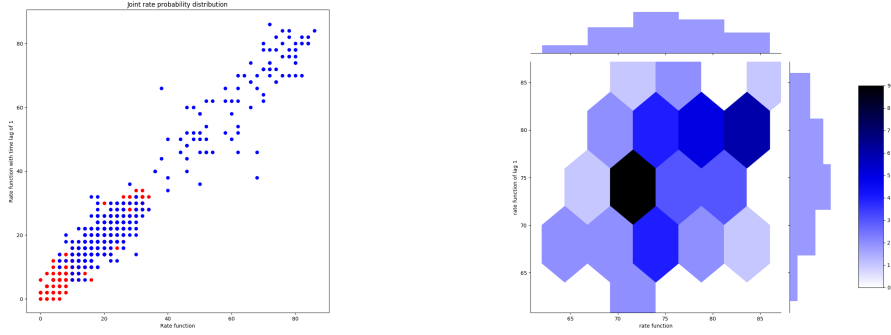
24

Figure 15: Joint probability distribution of the spiking rates of all channels in the GP and STN regions and these same sequences with a time lag of one window.

4.4.3 Serial correlation coefficients

All plots in this sub-subsection were produced using the `IMSHOW` function. The correlation coefficients are normalized to the range *[-1,1]*. The closer the value of the coefficient is to zero, the less correlation there is between the variables. Figure 16 and figure 17 display the serial correlogram for one single spike count train (channel) in the GP. Figure 18 shows the same information for all channels. Channels 0-6 are from GP, channels 7-16 are from STN Rows represent channels and columns represent time windows.
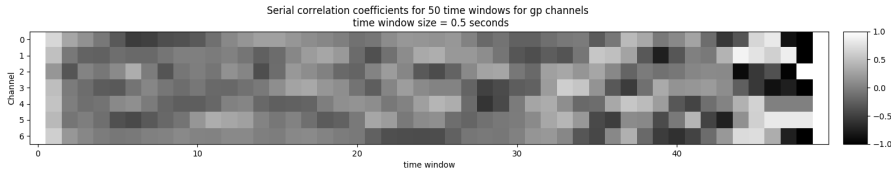

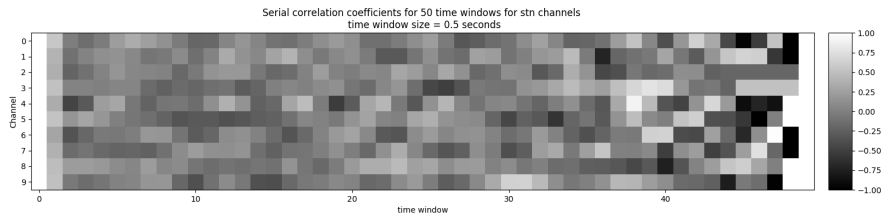Figure 16: Serial correlograms of all channels in the GP.

Figure 17: Serial correlograms of all channels in the STN.

25
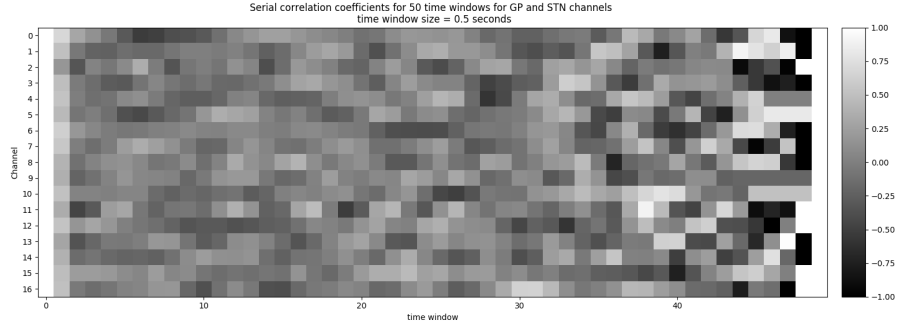
Figure 18: Serial correlograms of all channels in the GP and STN.

4.4.4 Power spectral density

All plots in this sub-subsection were produced with the `PLOT` function. Figure 19 and figure 20 show the power spectral density channels in the GP and STN (one color per channel). Figure 21 shows the power spectrum of both regions (red for GP, blue for STN). The original spike data was recorded with a sampling frequency of 1 Hz, so the maximum frequency that is observable is 1 Hz.

Figure 19: Power spectral density of all channels in the GP.

Figure 20: Power spectral density of all channels in the STN.
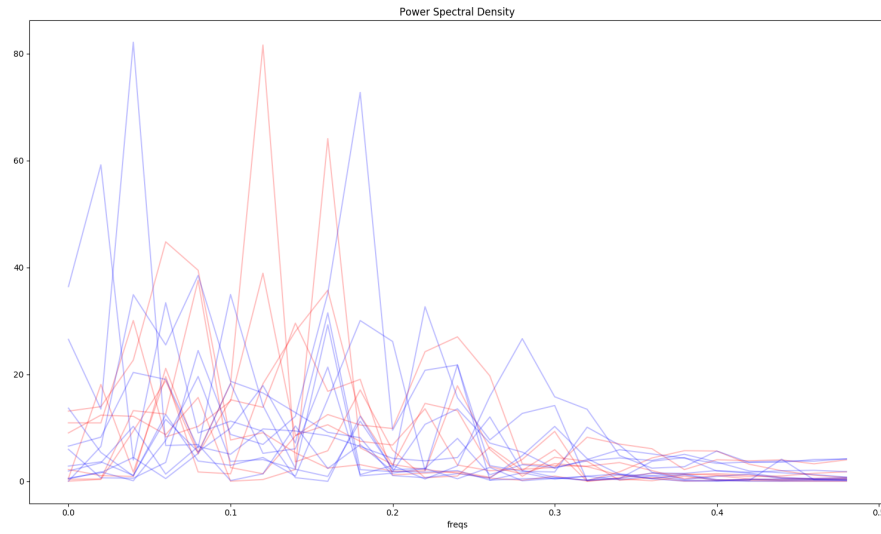
Figure 21: Power spectral density of all channels in the GP and STN.

4.4.5 Spectral entropies
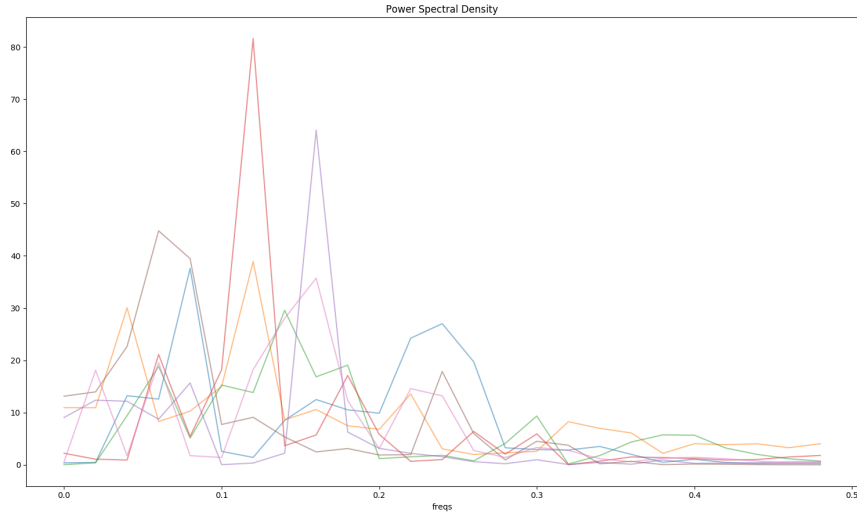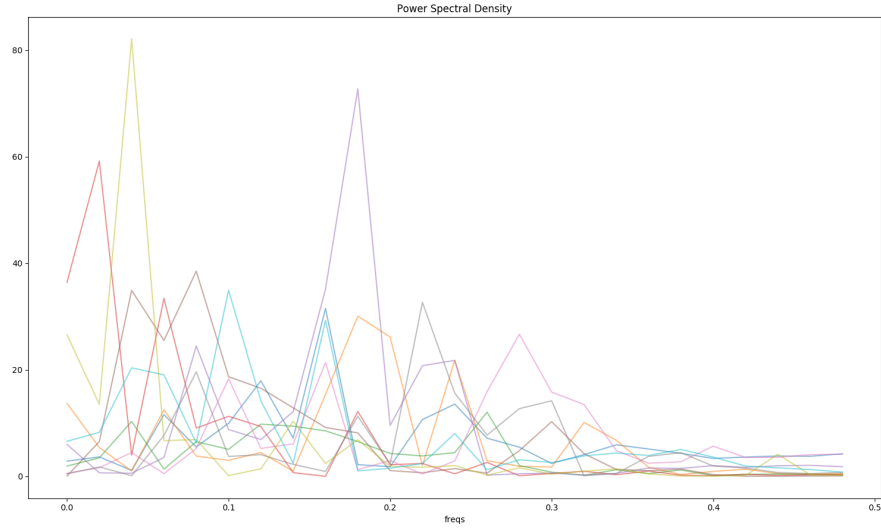
Table 3 show a list of spectral entropies of select channels (channels 0-6 from the GP, channels 7-16 from the STN).

	Spectral Entropy
Channel 0	0.793132
Channel 1	0.898098
Channel 2	0.827477
Channel 3	0.662636
Channel 4	0.619199
Channel 5	0.740590
Channel 6	0.778742
Channel 7	0.903068
Channel 8	0.804082
Channel 9	0.852935
Channel 10	0.638416
Channel 11	0.719677
Channel 12	0.752660
Channel 13	0.863742
Channel 14	0.798566
Channel 15	0.594042
Channel 16	0.832410

Table 3: Spectral entropies of select channels.

5 Discussion

This section is dedicated to discussion and interpretation of the produced results in the context of the research questions for this project.

5.1 LFP activity based methods of classification

This subsection discusses the methods described in subsections 3.1, 3.2, and 3.3, with respective results in subsections 4.1, 4.2, and 4.3.

The authors believe that SFVs, as described in this report, are an appropriate means of feature extraction in the context of this report. This is not intended to be a proclamation of a groundbreaking discovery. Analyzing the spectrum of brain activity has previously shown interesting results (Cagnan 2019), and SFVs are just a generalized approach to describe the process of extracting such spectra, taking fidelity, epoch size, and range of spectra into account. Indeed, SFVs or similar methods have likely been described or used previously with a different name in previous research, not necessarily in this field of research. SFVs also provide a simple mean of visualizing LFP activity in a human-understandable way.

To a layman in the study of neuroscience, the k-Means visualization heuristic applied in this project seems reasonable. The results produced show LFP activity being strikingly similar for different areas of the basal ganglia when compared simultaneously. Furthermore, the "outliers" present in figures 3 and 4 make intuitive sense, as one would expect outliers in any dataset. This intuition, to reiterate, is that of a layman. The authors imagine their chosen method and produced results may have serious flaws. The authors, however, consider these results to at least be a reason to consider more sophisticated methods to attempt to cluster SFVs. Such methods fall outside both the scope of this project, and the expertise of its' authors.

The authors believe that the produced PCs, or more specifically the method by which they were produced, show great potential for similar methods in the use of classification. As discussed, the authors consider SFVs a reasonable means of feature extraction, and PCA proved effective in further investigating these. A much smaller set of PCs than the number of features of their parent SFVs can describe them with great accuracy. These PCs could then reasonably be interpreted, in the quite literal sense, as components making up LFP activity. The PCs themselves could lend themselves to study and interpretation by more experienced scholars. The rather extreme discrepancies between the nature of the distributions of these PCs when comparing GP channels to STR channels (see figures 6, 7, tables 1, 2) could also lend insight into the workings of these different brain regions. Notable are the consistently lesser amplitudes of activity in STR, and the consistently much lower cross-correlation of PCs in STR. While the PCs described in figure 5 are present almost independently in STR channels, they have clear correlation in GP channels. Interpreting this is beyond the experience of the authors.

5.2 Spiking rate based methods of classification

The model used to analyse spikes provides a set of measurements that can be used as a reference to make distinctions in a broad collection of spiking activity data recorded as time-series, from various inter-related perspectives that focus on the independence between consecutive spikes.

With regard to the spiking rates, each region seems to have a specific range of rates that remain constant across channels, except for two individual channels in the subthalamic nucleus that show independent activity, with significantly higher values. On the grayscale plots, a not entirely explicit but moderately identifiable firing pattern for each channel can be seen.

The joint probability scatter diagrams may provide the best insight on how successive spiking rates relate and depend on each other. When it comes to different regions, they can be very clearly distinguished from each other in their common diagram. Although the serial correlation coefficients allow a quantification of the time-dependent correlation of rates, they are not so useful to identify any inherent patterns in the scattergrams; identifying (and quantifying in terms of density) clustered structures in such scattergrams is instead a better solution.

Departures from independence between spiking rates are reflected in the symmetries of the scatter diagram itself. The diagrams show an overall upward trend, which indicates positive correlations between successive rates. That loosely means that short intervals between consecutive spikes tend to be followed by short ones, and long intervals by long ones. A downward trend would mean the opposite.

Not taking into account all the available time windows (393 in total) when making the calculations is more convenient to see more specific dissimilarities among channels. Drawing results from a smaller time frame, for example 50 windows, allows to identify channels that are more "silent" than others at a certain range of time, as can be seen in the scattergrams that differentiate each channel with a colour as well as the hexagonal ones that quantify the density distribution.

The values of the serial correlation coefficients give proof of the original dataset not being apt to be represented by a Poisson process; if it were, due to the complete independence of every Poisson variable from all the others (the value of a neuron's firing rate would give no information as to the likelihood of its firing rate at any other point in time), the autocorrelation histogram of the channels would have been completely flat. From the serial correlograms it can be deducted that it is not such case. A similar thing would have happened with their power spectral densities: the power spectrum of a Poisson spike train is flat at all frequencies.

Based on the power spectrums of the autocorrelation of the channels, an evident synchronization between the globus pallidus and subthalamic nucleus can be observed; in spite of a few different power "peaks", both regions' power spectrums vary together. The estimated frequencies with the highest amount of variance (or noise) are between the values [0, 0.05] and [0.1, 0.2], and are common across the channels although each region can be distinguished from

the other. The highest observed frequency is smaller than 0.5 Hz, which is the half of the sampling frequency (1 Hz). Referring to the implications of noise on signal transmission theory, a high noise at a certain frequency is a limiting factor to information transmission between neurons at that frequency.

The spectral entropy values are then useful to quantify the general level of disorder or noise of each channel, based on the power spectrums. Since the results are normalized to be shown in the range [0,1], it can be affirmed that all channels have moderately high spectral entropies going from approximately 0.6 to 0.9, which indicates their complexity. Nonetheless, most of them remain moderately far from having a value of 1, which would belong to white noise; they still have some level of predictability. The subthalamic nucleus channels show overall higher values of spectral entropy than those in the globus pallidus, which indicates that in spite of the synchronization between these regions, their neuronal behaviour is different, perhaps due to different levels of excitatory activity.

6   Conclusion

This section is not at the same level of completion as previous sections.
    The methods produced for classifying Parkinson's disease based on LFP activity show potential. While results shown have often been comparing e.g. categories of channels, or the activity within a single session, much of this can be generalized. Larger datasets could be used, and subjects could be compared instead of channel categories (and vice versa). More extensive and sophisticated attempts at interpreting the use and results of the methods developed could be undertaken, and additional regions of the brain could be considered. It's likely that many additional avenues for further research could be explored, that are outside the imaginations of the authors.
    The authors believe that the methods and results developed based on LFP activity in the basal ganglia for the purpose of classification of Parkinson's disease show definitive potential for future research.

7 References

References

Bruce, P. & Bruce, A. (2017), *Practical Statistics for Data Scientists*, O'Reilly Media, Inc.

Cagnan, H. e. a. (2019), 'Temporal evolution of beta bursts in the parkinsonian cortical and basal ganglia network', *PNAS* **116**(32), 16095–16104.

DeMaags, G. (2015), 'Parkinson's disease and its management', *Pharmacy and Therapeutics* **40**(8), 504–510, 532.

Eden, U. (2015), 'Neural spike train analysis 1: Introduction to point processes'.

Gerstner, W. (2014), 'Neuronal dynamics: From single neurons to networks and models of cognition'.

Hammond, C. (2007), 'Pathological synchronization in parkinson's disease: networks, models and treatments', *Trends in Neurosciences* **7**, 357–364.

H.Perkel, D. (1967), 'Neuronal spike trains and stochastic point processes', *Biophysical Journal* **7**, 391–418.

Hunter, J. D. (2007), 'Matplotlib: A 2d graphics environment', *Computing in Science & Engineering* **9**(3), 90–95. `https://matplotlib.org/index.html` [051220], `https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.plot.html` [051220], `https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.imshow.html` [051220].

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017), *An Introduction to Statistical Learning*, Springer.

Mallet, N. (2008), 'Parkinsonian beta oscillations in the external globus pallidus and their relationship with subthalamic nucleus activity', *Journal of Neuroscience* **28**(52), 14245–58.

MathWorks (n.d.), 'Fourier transforms'. `https://se.mathworks.com/help/matlab/math/fourier-transforms.html` [032520].

Oliphant, T. (2006–), 'NumPy: A guide to NumPy', USA: Trelgol Publishing. `https://numpy.org/` [051320], `https://numpy.org/doc/1.18/reference/routines.fft.html` [051320] `https://numpy.org/doc/1.18/reference/generated/numpy.fft.fft.html` [051320], `https://numpy.org/doc/1.18/reference/generated/numpy.fft.fftfreq.html` [051320], `https://numpy.org/doc/1.18/reference/generated/numpy.absolute.html` [051320].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830. `https://scikit-learn.org/` [050520], `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html` [051120], `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html` [051120].

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Swain, C., Miles, A., Brunner, T., O'Kane, D., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C. & Brian (2020), 'mwaskom/seaborn: v0.10.1 (april 2020)'. `https://doi.org/10.5281/zenodo.3767070` [051520].

Zaccarelli, N. (2011), 'Order and disorder in ecological time-series: Introducing normalized spectral entropy', *Ecol. Indicat.* . `https://www.sciencedirect.com/science/article/abs/pii/S1470160X11002147?via%3Dihub`.