BE 521: Homework 6

Spike sorting

Spring 2015

58 points

Due: 3/19/2015 11:59 PM

Objective: Detect and cluster spikes

Homework Policy

1. Piazza should be used for peer discussion for all questions related to course material. Please also use Piazza to contact teaching staff for all questions. TA's will be available to help during office hours and occasionally on Piazza.
2. Submit LaTeX write-up (pdf) and Matlab code to Canvas as pennkey_hwx.pdf, .m before listed deadline.
3. Assignments will be returned electronically on Canvas.
4. Collaboration is encouraged but individual write-ups are required. Please list any collaborators. Honor code will be strictly enforced. Note: submitted code is routinely passed through a plagiarism checker.
5. Late Policy: 5% per day. **No homework is accepted after the 5th late day.** (e.g. If originally due Tuesday, 11:59PM, last day to turn in is Sunday, 11:59 PM).

Overview

In this homework, you will do some basic spike sorting using two different datasets. The first (`I521_A0006_D001`) is from a crayfish neuromuscular junction, a good model for human central nervous system synapses[1]. Specifically, the data contains two simultaneous recordings: an extracellular recording from the third nerve (channel `nerve`) of a crayfish abdominal ganglion, which contains six spontaneously active motor neurons, and an intracellular recording from the superficial flexor muscle (channel `muscle`) innervated by this nerve. You will attempt to discern relationships between the classes of spike waveforms you extract from the motor nerve trace and elicited potentials seen in the muscle fiber recording.

Then, you will revisit a human intracranial EEG recording (`I521_A0006_D002`) and use some of the techniques you've learned in class to build a more automated spike sorter.

Note: While spikes may have positive and negative deflections, we will only focus on positive spikes on this homework for simplicity. In other words, the only spikes we are interested in are those surpassing a positive threshold.

[1]The sampling rate of this data is 2000 Hz, which is adequate for this homework's instructional purposes but usually inadequate for real spike sorting, which often uses sampling frequencies on the order of 20 kHz.

1 Spike Detection and Clustering (37 pts)

In this section, you will explore some basic filtering and spike thresholding to ultimately compare spike clusters you pick out by eye to those selected by an automated algorithm.

1. You can assume that the nerve samples have already been low-pass filtered. Here you will high-pass filter in order to remove signals like slow local field potentials and 60 Hz power line noise. Create a 4th order *elliptic filter* with 0.1 dB of ripple in the passband, a stopband 40 dB lower than the peak value in the passband, and a passband edge frequency of 300 Hz (see Matlab's `ellip` function and make sure your give the edge frequency in the correct normalized form). The statement to create this filter (defined by the filter coefficients `b` and `a`) should look something like

   ```
   [b,a]=ellip(n,Rp,Rs,Wp,'high')
   ```

   Clearly specify the denominator and numerator coefficients obtained for your filter function. (2pts)

2. Using the `filter` function and `filtfilt` function, obtain two different filtered outputs of the nerve signal.

   (a) In a 2x1 subplot, plot the first 50 ms of the unfiltered nerve signal in the top subplot; in the bottom subplot, plot the `filter` output in blue and the `filtfilt` output in red. Use a potential range (y-axis) of -20 to 50 millivolts. (4 pts)

   (b) How is the unfiltered signal different from the filtered signal? What is different about the two filtered (red and blue) signals? (2 pts)

   (c) Briefly explain the mathematical difference between the two filtering methods, and why one method might be more advantageous than the other in the context of spike detection? (4 pts)

3. Using a spike threshold of +30 mV, calculate the index and value of the peak voltage for each spike in the **filtered** nerve signal. Use these values to plot the first 2.5 seconds of the nerve signal with a red dot above (e.g. 10 mV above) each spike. (Hint: Plot the entire length of the nerve signal with all the spikes marked but then restrict the x-axis using `xlim` to [0, 2.5] seconds, which will allow you to pan to other parts of the data.) (4 pts)

4. Under the assumption that different cells produce different action potentials with distinct peak amplitudes, decide how many cells you think were recorded (some number between 1 and 6). You may find it helpful to zoom in and pan on the plot you made in question 1.3. You may also find it useful to plot the sorted peak values to gain insight into where "plateaus" might be. (No need to include these preliminary plots in the report, though.) Use thresholds (which you well set manually/by eye) to separate the different spikes. Make a plot of the first 2.5 seconds similar to that in 1.3 except now color the spike dots of each group a different color (e.g., `'r.'`,`'g.'`,`'k.'`,`'m.'`).(6 pts)

5. Use Matlab's k-means[2] function (`kmeans`) to fit k clusters (where k is the number of cells you think the recording is picking up) to the 1D data for each spike.

[2]Clustering, like k-means you are using here, is a form of unsupervised learning.

(a) Using the same color order (for increasing spike amplitude) as you did for the thresholds in question 1.4, plot the spike cluster colors as little dots slightly above those you made for question 1.4. The final figure should be a new plot of the nerve voltage and two dots above each spike, the first being your manual label and the second your clustered label, which (hopefully/usually) should be the same color. (4 pts)

(b) Which labeling, your manual ones or the ones learned by clustering) seem best, or do they both seem just as good? (Again, panning over the entire plot may be helpful.) (2 pts)

6. In this question, you will test the hypothesis that the muscle potential responses are really only due to spikes from a subset of the cells you have identified in the previous two questions. First, plot the first 2.5 seconds of the muscle fiber potential and compare it with that of the nerve. Observe the relationship between spikes and the muscle fiber response. (No need to include this plot and observation in your report.)

   Now, calculate the maximum muscle fiber potential change[3] in the 25 ms[4] window after each spike (with the assumption that spikes without any/much effect on the muscle fiber potential do not directly innervate it).

   (a) Using the cell groups you either manually defined or found via k-means clustering (just specify which you're using) again with different colors, plot a colored point for each spike where the x-value is the spike amplitude and the y-value is the muscle potential change. (6 pts)

   (b) Does this plot support the hypothesis that the muscle fiber responses are only due to a subset of the cells. Explain why or why not. (3 pts)

2 Multivariate Clustering (22 pts)

In this section, you will explore similar methods for spikes sorting and clustering but with a different dataset, the human intracranial data in `I521_A0006_D002`, which is a larger dataset of the same recording you saw in `I521_A0001_D001` of Homework 1.

1. Using a threshold six standard deviations above the mean of the signal, detect the spikes in the signal. In addition, extract the waveform from 1 ms before the peak to 1 ms after it with peak value in the middle. (You will end up with a matrix where each row corresponds to the number of data points in 2 ms of signal minus 1 data point. Use the closest integer number of data points for the $\pm$ 1 ms window.)

   (a) Plot the waveforms of all the spikes overlaid on each other in the same color. (4 pts)

   (b) Does it looks like there is more than one type of spike? (1 pt)

2. For each spike, represent the waveform by its principal components. Use the `pca` command in Matlab. Intuitively, principal component analysis finds the coordinate system that most reduces the variability in your data.

[3] max voltage - min voltage

[4] Note that this 25 ms window is somewhat ad hoc and is just what seems reasonable by eye for this data. It implies no underlying physiological time scale or standard.

(a) Run principal component analysis on all the spike waveforms and represent your data with the top two principal components. Make a scatterplot of your data in this principal component (PC) space. (3 pts)

(b) Each PC also has an associated eigenvalue, representing the amount of variance explained by that PC. This an output of the `PCA` command. Plot the principal component vs the total variance explained. What is the variance explained if you include the top two principal components? (3 pts)

(c) Does it look like there is more than one cluster of spikes? (1 pt)

3. Use the same `kmeans` function as you used before to cluster the spikes based on these two (normalized) features (the waveforms represented by the top two PCs). You will use a slight twist, though, in that you will perform k-medians (which uses the medians instead of the mean for the cluster centers) by using the `'cityblock'` distance metric (instead of the default `'sqEuclidean'` distance). Make a plot similar to that in 2.2.a but now coloring the two clusters red and green. (3 pts)

4. Make a plot similar to 2.1 but now coloring the traces red and green according to which cluster they are in. Overlay the mean of the waveforms in each cluster with a thick black line (use the parameter `'LineWidth'` and value `'4'`). (3 pts)

5. In 3-4 sentences, what is one danger of using the clustering techniques in this homework? (Hint: consider the parameter k) (4 pts)