PURDUE UNIVERSITY

Weldon School of Biomedical Engineering

Biomedical Signal Processing
BME 595 - Fall 2021
Problem Set 2

This problem set accounts for 8% of your final grade. Please submit your solutions through **Brightspace** as a single .ipynb file or .pdf file containing typed explanations, plots, and any code that has been requested explicitly. Keep the text as brief as possible. Each problem lists the plots and results that need to be included in your solutions report. Please label the axes of all plots with appropriate units. You need not include any code that is not explicitly requested, but no problem if you do.

1. Consider the problem of measuring the auditory brainstem response (ABR) to brief sounds, a procedure that is used in newborn hearing screening in hospitals across the country. Two electrodes are affixed to the baby's head; one on the high forehead, and another behind the ear to which sounds are played. The voltage difference between those electrodes are recorded for about $3-4$ minutes, while about 1000 repetitions of the sound are played to the ear. The $\sim 3-4$-min-long recording is then bandpass filtered, and broken down into 1000 segments, one per presentation, in a *time-locked manner*. The segments are then averaged together to obtain the neural response to sounds. Here, we will carry out these processing steps using the data contained in the file `eegdata.mat`. The file contains the raw recording in a variable called `raw`, the sampling rate `fs`, and a variable called `clicks` which contains the sample numbers of the raw recording at which a brief sound was initiated.

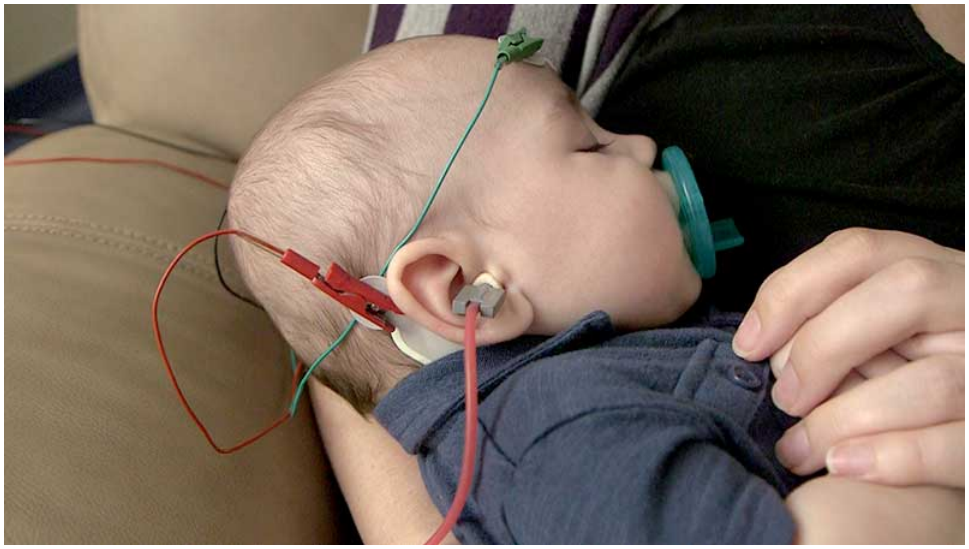


Figure 1: A typical set up for an ABR-based newborn hearing screening. Image courtesy of Boystown National Research Hospital.

   (a) The first step of processing is bandpass filtering from 70 to 3000 Hz. Use MATLAB's `fir1` or `scipy.signal.firwin` to construct the filter. How would you choose the filter length so that filter cutoffs are about 30 Hz sharp (i.e., the filter gain should drop to near zero levels by around

$70-30=40$ Hz on the low-frequency side and $3000+30=3030$ Hz on the high-frequency side respectively.)? Note that the sampling rate for the data is in the variable `fs`.

(b) Calculate and plot the magnitude transfer function $H(f)$ of the filter from part (a) to check that the filter indeed has a bandpass shape with the right cut-offs. Also check that the filter gain is $\leq 10\%$ of the passband gain for $f \leq 40$ and $f \geq 3030$ Hz. Include this plot in your report.

(c) Now let's apply the bandpass filter to the data. The first prominent positive peak in the brainstem response is called the wave-I and is thought to originate from the auditory nerve. Let's say that we are interested in measuring the latency of wave-I, i.e., the length of time between the sound initiation and the peak wave-I. Given that we are interested in the latency, apply the band-pass filter using an appropriate strategy. What strategy would you use, and why?

(d) After filtering, break down the bandpass filtered data by extracting 12 millisecond-long segments of the data going from -2 ms to $+10$ ms from the onset of each sound. Make an array called `epochs` with the segments. The `epochs` array should have size $1000 \times N$, for 1000 segments and N samples, where N is the number of samples in a duration of 12 ms.

(e) Estimate the ABR by averaging the `epochs` array across the 1000 segments. Assuming the noise is uncorrelated from one presentation of the sound to the next in the raw data, by what factor would the signal-to-noise ratio (SNR) be improved in the averaged data compared to the raw data?

(f) Plot the ABR as a function of time relative to the onset of sounds[1]. The time vector in this case, should go from -2 ms to $+10$ ms.

(g) From the plot, identify the time index (i.e., latency) of the first prominent positive peak. The latency is a clinically useful parameter to diagnose hearing loss or anomaly. A latency of more than 2.5 ms would be a red flag. Does the data given to you raise a red flag in this way?

2. In this problem, we will explore active noise cancellation. As discussed in class, let's say we have a noisy signal measurement $x(t) = s(t) + n(t)$, and an independent reference measurement that captured just the noise source $n_{ref}(t)$. Active noise cancellation involves estimating a filter $h(t)$ whose output gives you an estimate of the noise unknown $n(t)$ in $x(t)$. Then, an estimate of the clean signal can be constructed as $\hat{s}(t) = x(t) - h(t) \star n_{ref}(t)$, where $\star$ denotes convolution.

(a) Create a (MATLAB or Python) function called `activefilter` that takes three inputs: an array `x`, an array `nref`, and a filter length `p`. This function should implement the active filter estimation using the matrix notation introduced in class and return two variables as output: an estimated filter `h`, and an estimate of the clean signal `shat`. Include the code for your function along with your report.

(b) **It is good practice to test any signal processing code you write using simulated data where the "correct answer" is known.** To test your active filter, create a one-second long 10 Hz sine wave signal at a sampling rate of 8000 Hz. Create a white noise array $n_{ref}(t)$ (with variance 1) of the same length as as the sine wave. Create a randomly realized filter $h_{true}(t)$ of length 25 (samples) also using random gaussian draws from $\mathcal{N}(0, 0.2)$. Add noise

[1] Although the problem was motivated to you from a newborn hearing screening perspective, the data given to you is from an adult subject. If you happen to compare newborn ABR waveforms from the literature to the results that you get, they will look somewhat different for this reason.

to the sine wave by adding filtered white noise noise, i.e., $h_{true}(t) \star n_{ref}(t)$, to the sinusoid. Use your function from part (a) to construct a filter estimate `h`, and a clean signal `shat`. Plot the true and the estimated filters together. Similarly plot the noisy signal and `shat` together. Include these plots in your report. Is the active filtering code working?

(c) Next, we will apply the active filter to an active noise cancelling headphone scenario. Download the file `noisyspeech.mat` from Brightspace. The file contains three variables:
`x` - A speech signal corrupted with traffic noise
`nref` - A reference of just the traffic noise, and
`fs` - The sampling rate of the data.
In practical applications of active filtering, we do not know the what a suitable filter order (p) would be *a priori*. Instead, the filter order too has to be estimated from the data. A suitable method to determine the filter order is to try different increasing values, and stop when adequate performance is reached. For this problem try filter orders of $\{8, 16, 32, 64, 128,$ and $256\}$. Based on how these filters work, which filter order would you pick (let's call this the "best filter")? Plot the best filter (not the cleaned signal) along with a plot of the 256-point filter. Include this plot in your report and comment on the differences between the two filters.

(d) Listen to your filtered signal. Does it do a good job of noise cancellation?

(e) Does a simple subtraction of n_{ref} from x work for the noisy speech data?

3. How would you use the `activefilter` function that you created in problem 2 to do just a regression-based noise reduction? For example, let's say that you have data x from 20-subjects for how much weight they lost over a year while on a certain diet. However, some of them also exercised regularly over the year which contributes to weight loss. Now, if you are given how much each person exercised n_{ref}, and if it is reasonable to assume that exercise contributed linearly to weightloss, we can "regress out" the effect of exercise from x, to leave a cleaner estimate of the weightloss $\hat{s}$ that can be attributed to just being on the diet. Data for a certain diet is given in the file `dietdata.mat`. The file contains two variables: `weightloss` - The amount of weight lost by each of 20 subjects in pounds (negative numbers mean weight gained), and `exercise` - The amount of exercise each subject performed during the year on diet (in hours per week).

(a) What is the average weight loss across the 20 subjects without taking exercise into account?

(b) Use the `activefilter` function to perform a regression to remove the "noise" in the weight loss measurement using the exercise data as reference. What is the regression filter coefficient? This number can be interpreted as the amount of weight loss attributable to unit exercise (i.e., weight loss from 1 hour/week exercise).

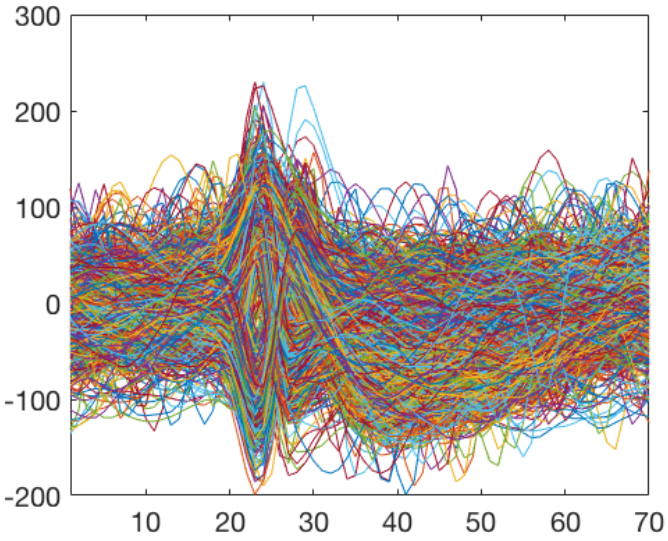
(c) What is the average weight loss after accounting for exercise using linear regression? Does it seem like the diet is helping?

4. Here, we will explore the use pf PCA for dimensionality reduction with application to *neuronal spike sorting*. Often neurophysiological recordings are done with electrodes places extracellularly in the brain tissue such that a given electrode can pick up spikes from more than one neuron. However, accurate interpretation of the data requires that responses of single neurons need to be examined and quantified. In order to do this, each measured spike will need to be "sorted" according to which neuron it originates from. The main clue that allows us to distinguish spikes from different cells and sort them is the shape (i.e., waveform) of the spike itself – two spikes from the same neuron have similar shapes whereas spikes from different neurons look (more) different. The catch is that

the shape, when just drawn out as a time waveform is a high dimensional quantity (e.g., if you have 70 time samples that make up the shape of the waveform, then each spike is a 70-dimensional quantity). If the dimensionality is reduced to small number while retaining key information about the shape, then sorting becomes a lot easier. We will explore this in this problem.

The file `SpikeSorting.mat` contains two variables - A continuous time series called `voltage`, and a vector `spikes` containing the times at which spikes were detected using a peak picking algorithm.

(a) There were 3298 spikes that were detected (as can be verified by looking at the shape of the `spikes` variable). Use the spike times to extract the spike waveforms each of length 70 samples in a variable called `waveforms`. `waveforms` should be of shape 3298×70. Note that the spike times given to you are such that you can just take the 70 samples including and after the spike time of each spike, i.e., you don't need to do anything else to center the spike peak over the 70 samples. Plot these spike waveforms on top of each other. Include the plot in your report. It should look something like this:



(b) Because a spike waveform is a high-dimensional quantity, we shall attempt to reduce the dimensionality by using PCA. Thus, we will consider each spike example as a point in 70 dimensional space with $\{\delta(t),\ \delta(t-1),\ \ldots,\ \delta(t-69)\}$ as axes. Now we want to determine the orthogonal set of directions of maximum variability across the 3298 points. To do that we first need to calculate the covariance of the data in the orignal 70 axes. Estimate the 70×70 covariance matrix of the data by treating each timepoint as a separate axis (or variable), and each spike example as a different observation of those 70 variables. Note that the data given to you are 16 bit integers. Please convert them to double precision (you could use MATLAB's `double` function, or `numpy.ndarray.astype` in Python) before any computations to reduce rounding errors.

(c) Plot the covariance matrix as an image with hotter colors indicating positive numbers and cooler colors indicating negative numbers. Include this image in your report. In this image, you should see one bright spots on the diagonal – what does this bright spot mean? In addition to the bright spot on the diagonal, you should see some lighter hot and cold bands extend away from the bright spot on the diagonal – what do these mean? Hint: Think about what a stereotypical spike shape looks like.

(d) Extract the two directions of maximum variability by calculating the eigenvectors and choosing the ones corresponding to the largest two eigenvalues. Plot these two eigenvectors (let's call them $\underline{q}_1$ and $\underline{q}_2$) on top of each other with different colors. Include this plot in your report.

(e) Project each spike example along each of the two eigenvectors. This gives you a 2D summary of the data (rather than the orignal 70D spread). Plot each spike as a point in this 2D plane, i.e., plot the points (a_k, b_k) on a 2D plane where a_k is the projection of the k^{th} spike waveform onto the first eigenvector and b_k is the projection of the same k^{th} spike onto the second eigenvector. Note that $k \in \{1, 2, \ldots, 3298\}$. Include this plot in your report.

(f) Based on the 2D plot, how many neurons do you think there are in the mix?

(g) Cluster the points manually, i.e., choose some threshold along the horizontal and vertical axes to separate the clusters and subdivide the 3298 points into how many ever clusters you think there are. For example, if you find that all points of a certain cluster have a_k values above 0, then extract that cluster by picking all points that meet that threshold. Replot the plot from part (f) but with the points in different clusters plotted with different colors, i.e., visualize the clustering as a color code.

(h) Plot the full spike shapes (i.e., the original 70 dimensional shape) of all the points that belong to any one of the clusters (Let's call this cluster1). Include this plot in your report. Do the spike shapes look homogenous, i.e., like they all are small variations of the same shape?

(i) Calculate the average $\bar{a}$ of the a_k values, and the average $\bar{b}$ of the b_k values for just the points in cluster1. What numbers do you get? Plot the trace $\bar{a}\underline{q}_1 + \bar{b}\underline{q}_2$ on top of the previous plot of spike shape examples of cluster1 from part (h). Include this plot in your report. Based on this plot, comment on the effectiveness or lack thereof of the dimensionality reduction.

(j) Repeat the plot from part (i) for the other clusters beside cluster1. Include one plot per cluster. Based on these additional plots, comment on the effectiveness or lack thereof of the dimensionality reduction.