18-698 / 42-632
Neural Signal Processing
Prof. Byron Yu Spring 2023

Problem Set 7

This problem set is due on **Thursday, April 20, 11:59pm**.

Be sure to show your work and include all **Matlab code** and plots. **Any Matlab questions without code will receive no credit.**

If you have questions, please post them on the Piazza Q&A webpage, rather than emailing the course staff. This will allow other students with the same question to see the response and any ensuing discussion.

Please submit your work as a **single PDF file** on Gradescope, which is linked from Canvas. When preparing your solutions, please complete each problem on a **separate page**. Gradescope will ask you select the pages that contain the solution to each problem.

Submissions can be written in LaTeX or they can be handwritten and photocopied using a scanner or smartphone camera. Handwritten work should be clearly labeled and legible.

The Matlab code referred to in the following problems can be found on Canvas under "Files → Data sets → ps7_code.zip". The dataset is posted under "Files → Data sets → ps7_data.mat". When you load the .mat file, you will find the following variables:

`Spikes`: a 31×552 matrix of spike snippets[1], where `Spikes(:,n)` is the nth snippet ($n = 1, \ldots, 552$). Values are in μV.

`InitParams`: a structure containing initialization parameters for a Gaussian mixture model, with the following fields

- `mu` is a 2×8 matrix, where the kth column is the initialization of the kth cluster center, $\boldsymbol{\mu}_k$ ($k = 1, \ldots, K$).
- `Sigma` is a 2×2 covariance matrix. Assume that all cluster covariances Σ_k are initialized to the same covariance matrix.

In Problem Sets 5 and 6, we clustered the 31-dimensional spike waveforms directly. Here, we will first use PCA to project each 31-dimensional waveform down to a two-dimensional feature space. Then, we will perform model selection in the feature space.

[1] The neural data have been generously provided by the laboratory of Prof. Krishna Shenoy at Stanford University. The data are to be used exclusively for educational purposes in this course.

1. Automatic feature selection using PCA

   Treat each snippet as a point $\mathbf{x}_n \in \mathbf{R}^D$ $(n = 1, ..., N)$, where $D = 31$ is the number of samples in each snippet and $N = 552$ is the number of detected spikes.

   (a) **(5 points)** Plot the N raw spike snippets in a "voltage vs. time" plot, as in Figure 7(a) of the Lewicki paper.

   (b) **(15 points)** Apply PCA to all N spike snippets. Plot the eigenvector waveforms corresponding to the three largest eigenvalues, as in Figure 7(b) of the Lewicki paper. Label the waveforms by color (1st: red, 2nd: green, 3rd: blue).

   (c) **(5 points)** Plot the square-rooted eigenvalue spectrum, as in Figure 7(c) of the Lewicki paper. Look for an elbow in the eigenvalue spectrum. How many dominant eigenvalues are there?

   (d) **(10 points)** Create a scatter plot of the PC1 score versus the PC2 score, where each point corresponds to a spike snippet, as in Figure 7(d) of the Lewicki paper. How many distinct clusters do you see in the plot?

2. Model selection for number of clusters

   In Problem 1(d), each spike snippet is represented as a point in a two-dimensional feature space found by PCA. We will now perform clustering using a Gaussian mixture model (GMM) in this two-dimensional space. For fitting a GMM using EM, you can either use your own code from Problem Set 6 or our code (`func_GMM.m`).

   As discussed in class, the PC directions are only unique up to a sign difference. Before doing this problem, please check that your answer to Problem 1(d) has the same sign orientation as in ps7_1d.jpg, since we've chosen the EM initialization parameters to be sensible for this sign orientation.

   (a) **(40 points)** Compute the cross-validated likelihoods for a GMM applied to the two-dimensional PCA projections for $K = 1, \ldots, 8$. For each value of K, perform four-fold cross-validation by dividing the dataset into four equally-sized partitions (i.e., fold 1 has $n = 1, \ldots, 138$; fold 2 has $n = 139, \ldots, 276$; $\ldots$).

   Initialize EM using the parameters in `InitParams`. To initialize the $\boldsymbol{\mu}_k$, use the first K columns of `InitParams.mu`. To initialize the Σ_k, use the same `InitParams.Sigma` for each cluster. To initialize the π_k, use $1/K$ for each cluster.

   Plot the cross-validated likelihoods versus K. What is the optimal value of K? (Hint: See Section 4.6 in the Lewicki paper for an explanation of why the optimal value of K may be larger than you expected.)

   (b) **(15 points)** For each value of $K = 1, \ldots, 8$, create a separate plot with:

- the data scatter of PC1 score versus PC2 score, as in Problem 1(d) (this will be the same for each value of K)
- a one-standard-deviation ellipse for each cluster ($k = 1, \ldots, K$) based on Σ_k centered at $\boldsymbol{\mu}_k$. Use the parameters obtained from the first cross-validation fold. To plot ellipses, use `func_plotEllipse.m`.

(c) **(10 points)** For $K = 3$, plot the canonical spike waveform corresponding to each cluster center in a "voltage versus time" plot. This will involve projecting the two-dimensional $\boldsymbol{\mu}_k$ out into the 31-dimensional space. Use the $\boldsymbol{\mu}_k$ from the first cross-validation fold.