G80.3042.002 – Fall 2007
Statistical Analysis and Modeling of Neural Data

Homework 4

Due: Wednesday, 19 Dec 2007

Your results should be in the form of a MATLAB file (typically, the filename should have an extension of `.m`). Email your solutions to `eero@cns.nyu.edu` and `bijan@cns.nyu.edu`.

1. **Mutual Information of Gaussian variables**.

   A well-known normative theory for early sensory processing is that it is designed to optimize the mutual information between stimuli and responses, which we showed in class is equal to: $I(s,r) = H(r) - H(r|s)$, where H is the entropy.

   (a) Derive the entropy of a multi-variate (i.e., vector-valued) Gaussian distribution. It might be easiest to first write the scalar case.

   (b) Imagine you have two neurons, and you can approximate their noise with additive uncorrelated Gaussian random variables, each of variance σ_n. Assume also their mean responses across a particular ensemble of inputs are Gaussian distributed, with covariance matrix C_s that has values σ_s on the diagonal, and α on the off-diagonal. Write down the mutual information, as a function of the parameters $\{sigma_n, \sigma_s, \alpha\}$.

   (c) How does the mutual information vary with the correlatedness of the two responses? In particular, plot the mutual information as a function of α over the range $[-1, 1]$. Where is the minimum? Why?

2. **Spike sorting**

   Two elementary clustering algorithms are the k-nearest neighbor algorithm and the k-means algoritm. In this problem, you will use these algorithms to sort extracellular recorded action potentials.

   Data format: The data is organized in one data structure `Sp` with two fields, `Times` and `Waveforms`. `Sp.Times` contains the times of all the spikes in the recording in units of `ms`. `Sp.Waveforms` contains the event minimum-aligned threshold-crossing waveforms with one event per row. The minimum of each waveform is at bin 6.

   The data set contains `68098` events recorded over a 15 minute interval. The threshold for spike detection was chosen so that not all these events are waveforms from action potentials. Therefore sorting these waveforms into clusters will, at a minimum, give a multiunit cluster that contains activity from many neurons combined with background neural noise. Other clusters may also be present. These clusters contain action potentials that may be isolated from the multiunit activity and each other.

   (a) Use PCA to project the event waveforms into a 3-D subspace. How much variance of the original data does this subspace contain? Plot projections of the events into this subspace. First, plot all the data in the recording, then divide it up into `10000` event sections and plot the data within each section. Are there any differences between the sections? Describe them. Hint: You can use the command `plot(x,y,'.','Markersize',0.5)` to make it easier to see the points.

(b) Implement the k-means algorithm and use it to sort all the data points in the recording. The steps are to choose the number of clusters, initialize the centers of each cluster, assign each event to the cluster with the nearest mean, recompute the centers of each cluster and iterate until convergence. Plot your results as cluster plots to visually evaluate the performance of the algorithm. Comment on your results. How do you decide how many clusters to seed? How many events in each sorted cluster satisfy the requirement for refractoriness?

(c) Repeat this clustering for each `10000` event subsection. Comment on your results. What evidence is there that the number of clusters changes over the course of the recording? How well is refractoriness achieved using the clusters derived in this way?

(d) The nearest-neighbors algorithm is hard to apply to this data set because there are so many points. Use k-means to overcluster the data and selectively merge clusters using the nearest-neighbors algorithm on the clusters instead of the original data points. This is one way to avoid so many computations. Work with the Euclidean distance in the 3-D subspace, choose an appropriate criterion (distance or number of clusters) to stop merging and use only the distance to the nearest neighbor. Compare your results with the results of using straight k-means.