Quantitative Methods in Neuroscience
(NEU 466M)
Homework 8
Due: Friday April 17 by 5 pm in NHB 3.128

In this assignment, we will apply Bayes' rule and Maximum Likelihood Estimation with the dual aims of getting familiar with using this machinery and also learning how it can yield interesting results. *General guidelines:* Read through each complete problem carefully before attempting any parts. Feel free to collaborate in groups of size 2-3, but always note the names of your collaborators on your submitted homework. For graphs: clearly label your axes and use good color and symbol choices. Print out your matlab code (in the form of a script file). For derivations you're asked to do 'by hand' (in other words, analytically, using paper and pencil) feel free to turn in handwritten or typed-out work.

1) **Hypothesis assessment with Bayes' rule.** Very broadly speaking, there are two types of neurons in the cortex: principal neurons and interneurons. Interneurons fire at higher rates than the rate of principal neurons: suppose the mean rate of interneurons is 100 Hz, while the mean rate of principal cells is 40 Hz. Your extracellular electrode is very close to a neuron and in the first 5 ms, you observe a spike. What is the probability you are recording from an interneuron versus a principal neuron, assuming both types of neurons are equally numerous? (In your formulation, let H_I, H_P be the hypotheses that you are recording from an interneuron or a principal cell, respectively, and assess the probability that H_I is true.) Next, suppose that interneurons are 4 times more numerous than principal cells. Now what is the probability you are recording from an interneuron? Use Bayes' rule to obtain these probabilities.

2) **Maximum likelihood rate slope estimation (MLE) for linear-rate Poisson and Gaussian neurons.**

    a. In class, we wrote down a model of neural activity as a Poisson process with rate $\lambda = \theta x$, where x is some stimulus and θ is some unknown proportionality constant (slope). We analytically derived the maximum likelihood estimate $\hat{\theta}_{ML}$ of the unknown slope θ from data-samples $\{(x_1, r_1), \cdots, (x_N, r_N)\}$, and found that it is:

    $$\hat{\theta}_{ML} = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} x_i}. \qquad (1)$$

    Let us now do the same if the probability of neural activty is Gaussian: the mean rate λ varies linearly with the stimulus, $\lambda = \theta x$, with unknown θ. Then the probability of (analog) responses $\mathbf{r} = (r_1, \cdots, r_N)$ for stimulus values $\mathbf{x} = (x_1, \cdots, x_N)$ is given by $p(\mathbf{r}|\mathbf{x}, \theta) = G(\mathbf{r} - \theta\mathbf{x}, \sigma^2) = \prod_{i=1}^{N} G(r_i - \theta x_i, \sigma^2)$, where

$$G(r_i - \theta x_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r_i - \theta x_i)^2/2\sigma^2}.$$

$p(\mathbf{r}|\mathbf{x}, \theta)$ is the likelihood of θ. Maximize the logarithm of the likelihood with respect to θ, to derive the ML estimate $\hat{\theta}_{ML}$ of the unknown slope for this linear-response model with Gaussian noise.

b. You now have two different ML-optimal estimates for the slope of a linear response, assuming the noise in the response is Poisson or Gaussian, respectively. Compare and contrast these two expressions. Look back at your notes on the linear least-squares regression estimate of the slope and compare that expression with these two results. The comparison reveals that linear regression is the optimal solution for fitting lines through data, if the data are Gaussian-distributed about their mean values! It also implies that linear least-squares regression need not be not the optimal solution (in a maximum likelihood sense) if the noise in the response variable around its mean is not Gaussian. In c.), we will explicitly explore this.

c. Download the dataset `linearneuron1.mat` (contains stimulus x and responses r). Use our derived expressions for $\hat{\theta}_{ML}$ for both the linear Gaussian and linear Poisson neuron ML estimates, to obtain two estimates of the slope. In Matlab, plot the data as a scatterplot of r versus x, and also plot a line through it with the two estimated slopes. Which is a better fit? To understand why, collect all data with the same stimulus value and plot a histogram of the responses. Do this for the different stimulus values. Does the distribution look more Gaussian or Poisson? For each stimulus value, obtain the mean of all responses and the variances of all responses. Plot the mean versus the variance for all stimulus values. Now can you tell if the data are Poisson or Gaussian? Finally, add a linear least-squares regression fit to the data: how does it compare?

d. Do the same as [c.], but for the dataset `linearneuron2.mat`.

3) **Maximum likelihood rate estimation (MLE) for constant-rate and linear-rate Bernoulli neurons.**

a. Assume that the firing rate of a neuron is $\lambda = \theta$, a constant value, and spikes are generated as a Bernoulli process (with probability of spiking $p = \lambda$). Suppose you observe a set of spike outcomes $\mathbf{r} = (r_1, \cdots, r_N)$, with $r_i \in \{0, 1\}$, where 1 represents a spike and 0 a non-spike. Write down the likelihood $p(\mathbf{r}|\theta)$, then maximize its logarithm with respect to θ to obtain the ML estimate of θ given the observations.

b. Next, suppose the neural response rate varies linearly with the stimulus $\lambda = \theta x$, and neural spikes are Bernoulii. with probability λ. Write down the likelihood $p(\mathbf{r}|\mathbf{x}, \theta)$. Trying to derive the maximum likelihood value of θ analytically, like you did in [a.] above by setting the derivative of the log-likelihood to zero, gives an equation that does not yield a simple expression for θ.

   Thus, we will approach this problem numerically. Download and run the matlab file `linearneuronBernoulli.m`. This generates a dataset with a specific value θ^*. In Matlab, plug the fixed dataset values into the expression for the likelihood function, while varying the θ, and thus plot the likelihood score as a function of θ. The θ value corresponding to the peak in this likelihood function plot is your numerical ML estimate $\hat{\theta}_{ML}$ of θ^*. Indicate both $\hat{\theta}_{ML}$ and θ^* in this plot, so we can see how close the estimate obtained from N data-samples is to the true underlying value. Repeat what you just did, after first setting the number of data-points to $N = 400$ in `linearneuronBernoulli.m` and generating a new, larger data-set. Compare the error between the true θ^* and your estimate.

c. Bonus (extra credit): Exploring how estimation error varies with the number of data-points: Repeat what you did above in [b.] but vary N between 50 and 500, and plot your squared estimation error versus N. How does squared error vary with N? It may be helpful, for seeing the overall pattern, to do multiple trials for each N and then average the squared error across trials to get one "mean squared error" for that N. Finally, it is often helpful to plot such results on a log-log plot. Why? Try this.