ECE 5960/6960 BME 5900/6900 – Fall 2022
Mathematical Tools for Neural Data Analysis and Modeling

Homework 4

(courtesy of Profs. Simoncelli and Landy, NYU)

Due: November 18, 2022

See the course web site for submission details. For each problem, show your work - if you only provide the answer, and it is wrong, then there is no way to assign partial credit! And, please don't procrastinate until the day before the due date... *start now*!

1. **Bayes' rule and eye color.** A male and female chimpanzee have blue and brown eyes, respectively. The brown-eyed allele can be denoted as a capital B, whereas the blue-eyed allele can be represented as a lowercase b. Assume a simple genetic model in which the gene for brown eyes is always dominant (so that the trait of blue eyes can only arise from two blue-eyed genes, but the trait of brown eyes can arise from two brown-eyed genes, or one of each). You can also assume: i) the probability of the mother being BB is 50% and the probability of her being Bb is 50%; and ii) the *a priori* probability that each of the four gene configurations is equally probable. For each question, provide the math, and explain your reasoning.

   (a) Suppose you observe that they have a single child with brown eyes. What is the probability that the female chimp has a blue-eyed gene?

   (b) Suppose you observe that they have a second child with brown eyes. Now what is the probability?

   (c) Generalizing, suppose they have N children with brown eyes... express the probability, as a function of N.

2. **Poisson neurons.** The Poisson distribution is commonly used to model neural spike counts:

$$p(k) = \frac{\mu^k e^{-\mu}}{k!},$$

   where k is the spike count (over some specified time interval), and μ is the expected number of spikes over that interval.

   (a) We would like to know what the Poisson distribution looks like. Set the expected number of spikes to $\mu = 6$ spikes/interval then create a vector `p` of length 21, whose elements contain the probabilities of Poisson spike counts for $k = [0...20]$. Since we're clipping the range at a maximum value of 20, you'll need to normalize the vector so it sums to one (the distribution given above is normalized over the range from 0 to infinity) to make the vector `p` represent a valid probability distribution. Plot `p` in a bar plot and mark the mean firing rate. Is it equal to μ?

   (b) Generate samples from the Poisson distribution where each sample represents the number of spike count ranging from 0 to 20. To simplify the problem, use a clipped Poisson vector `p` to write a function `samples = randp(p, num)` that generates `num` samples from the probability distribution function (PDF) specified by `p`. [Hint: use the `rand`

function, which generates real values over the interval $[0...1]$, and partition this interval into portions proportional in size to the probabilities in `p`]. Test your function by drawing 1,000 samples from the Poisson distribution in (a), plotting a histogram of how many times each value is sampled, and comparing this to the frequencies predicted by `p`. Verify qualitatively that the answer gets closer (converges) as you increase the number of samples (try 10 raised to powers $[2, 3, 4, 5]$).

(c) Imagine you're recording with an electrode from two neurons simultaneously, whose spikes have very similar waveforms (and thus can't be distinguished by the spike sorting software). Create a probability vector, `q`, for the second neuron, assuming a mean rate of 4 spikes/interval. What is the PDF of the observed spike counts, which will be the sum of spike counts from the two neurons derived from `p` and `q`? [Hint: the output vector should have length $m + n - 1$ when m and n are the lengths of the two input PDFs. This is because the maximum spike count will be bigger than the maximum of each respective individual neuron.]

Verify your answer by comparing it to the histogram of 1,000 samples generated by summing two calls to `randp` (choose a big enough number of samples!).

(d) Now imagine you are recording from a neuron with mean rate 10 spikes/interval (the sum of the rates from the neurons above). Plot the distribution of spike counts for this neuron, in comparison with the distribution of the sum of the previous two neurons. Based on the results of these two experiments, if we record a new spike train, can you tell whether the spikes you have recorded came from one or two neurons just by looking at their distribution of spike counts? Comment about the reason why based on the intuition behind Poisson distribution.

3. **Multi-dimensional Gaussians.**

(a) Write a function `samples = ndRandn(mean, cov, num)` that generates a set of samples drawn from an N-dimensional Gaussian distribution with the specified `mean` (an N-vector) and `cov`ariance (an NxN matrix). The parameter `num` should be optional (defaulting to 1) and should specify the number of samples to return. The returned value should be a matrix with `num` rows each containing a sample of N elements. (Hint: use the MATLAB function `randn` to generate samples from an N-dimensional Gaussian with zero mean and identity covariance matrix, and then transform these to achieve the desired mean/cov. Recall that the covariance of $Y = MX$ is $E(YY^T) = MC_XM^T$ where C_X is the covariance of X). Please use mean $\mu = [4, 5]$ with $C_X = [9, -5; -5, 6]$ to sample and scatterplot 1,000 points to verify your function work as intended.

(b) Now consider the marginal distribution of a generalized 2-D Gaussian with mean μ and covariance Σ in which samples are projected onto a unit vector $\hat{u}$ to obtain a 1-D distribution. Write a mathematical expression for the mean, $\hat{\mu}$, and variance, $\hat{\sigma}^2$, of this marginal distribution as a function of $\hat{u}$ and check it for a set of 48 unit vectors spaced evenly around the unit circle. For each of these, compare the mean and variance predicted from your mathematical expression to the sample mean and variance estimated by projecting your 1,000 samples from part (a) onto $\hat{u}$. Stem plot the mathematically computed mean and the sample mean (on the same plot), and also plot the mathematical variance and the sample variance.

(c) Now scatterplot 1,000 new samples of a 2-dimensional Gaussian using μ and C_X in part (a). Measure the sample mean and covariance of your data points, comparing to

the values that you requested when calling the function. Plot an ellipse on top of the scatterplot by generating unit vectors equi-spaced around the circle, and transforming them with a matrix as in part (a) to have the same mean and covariance as the data. Try this on three additional random data sets with different means and covariance matrices. Does this ellipse capture the shape of the data?

(d) How would you, mathematically, compute the direction (unit vector) that maximizes the variance of the marginal distribution? Compute this direction and verify that it is consistent with your plot.

4. **Analyzing and simulating experimental data.** An international coffee conglomerate recruits you to characterize the neuropsychology underlying their customers' adoration of pumpkin spice. You devise a blood-oxygen level dependent (BOLD) FMRI pilot experiment in which you present one of two classes of odorants to an individual while monitoring the activity of three key voxels located in the amygdala, a structure known to be associated with emotional responses. The file `experimentData.mat` contains: a $(N \times 3)$ matrix `data`, where each row is the BOLD response of the three voxels on a given trial relative to some baseline; and a $(N \times 1)$ vector `trialConds` indicating the experimental condition of each trial. Condition 1 are trials in which you present an odorant selected randomly from a library of possible control odorants, and condition 2 are trials in which the trade secret pumpkin spice odorant is presented.

   (a) Before doing anything quantitative with your data, it is always good practice to visualize it. First, determine how many trials of each trial condition were completed. Display this information as a 2-bin histogram with each bin representing each of the two possible trial conditions, and their heights representing their respective trial counts. Next, plot a 3D scatter plot of the recorded responses, with each point color-coded according to its associated trial condition (use the function `scatter3` in Matlab and be sure to label your axes). Describe your data qualitatively using this figure. Is there a noticeable difference between the two trial conditions? What geometric shape are these 'response clouds', and what distribution would you use to model them?

   (b) Quantify the response statistics of each individual trial condition. Calculate the means of each response cloud, as well as their respective covariance matrices. Compute the covariance matrices of each response cloud using matrix multiplication (remember to center the data first). Verify your calculation is correct by comparing with the output given by the `cov` function. How do the covariance matrices compare (are they similar at all or wildly different)?

   (c) Next, compute the SVD of each covariance matrix. Plot the three singular vectors originating from the center of each response cloud and scale their amplitude by the square root of the singular values. Relative to how similar the covariance matrices were before computing their SVD, how do each trial condition's respective set of singular values compare? Describe what this tells us about the relationship between the two trial conditions and, more fundamentally, the relationship between the three voxels across conditions.

   (d) A powerful method to validate a model is by *generating* (i.e. simulating) new data matching your quantitative description of the real data, and then comparing them with real data. Create a function

   ```
   simResponses = odorExperiment(numTrials1,numTrials2)
   ```

where `numTrials1` and `numTrials2` are the number of trials in a simulated experiment for condition 1 and 2, respectively. `simResponses` is a $(N \times 3)$ matrix containing simulated responses of each of your 3 voxels during $N =$ `numTrials1` + `numTrials2` trials. [Hint: use `ndRandn` from the previous problem]. Plot the simulated and real responses in the same figure (use subplots if you wish) to compare the two. Is your simulated response data a good characterization of the real amygdala voxel responses?