PSYCH-GA.2211/NEURL-GA.2201 – Fall 2018
Mathematical Tools for Neural and Cognitive Science

Homework 5

Due: 30 Nov 2018
(late homeworks penalized 10% per day)

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. **Dueling estimators**. In this problem, we use simulation to compare three estimators of the mean of a Normal (Gaussian) distribution.

   (a). First consider the *average*, which minimizes the sum of squared deviations, and is also the Maximum Likelihood estimator. Generate 10,000 samples, each of size 10, from the Normal(0,1) distribution (a 10x10000 matrix). Compute the average of each of the 10,000 samples. Plot a histogram of the resulting estimates (use 50 bins, and set the plot range to [-2.3,2.3]). What shape should the histogram have (explain why)? What is the (theoretical) variance of the average of 10 values drawn from a univariate Gaussian (derive this)? Is the empirical variance of your 10,000 estimates close to this?

   (b). Now consider the *median*, which minimizes the sum of absolute deviations. Compute the median of each of the 10,000 samples, and again plot a histogram. What shape does this one have? Compare it to a normal distribution using the function `normplot`, which plots the quantiles of a sample of data versus the normal quantiles (known as a Q-Q plot: if data are normally distributed, the points shuld fall nearly on a straight line.) Does the distribution of estimated values deviate significantly from a Normal distribution? Specifically, compare the Q-Q plot for the median estimator to that for the mean from part (a).

   (c). Finally, consider an estimator that computes the average of the minimum and maximum over the sample (as shown in class, this one minimizes the $L_\infty-\mathrm{norm}$). Again, compute this estimate for each of your 10,000 samples, plot the histogram, and examine and comment on the Q-Q plot, just as in part (b).

   (d). All three of these estimators are unbiased (because of the symmetry of the distribution), so we can use variance as the sole criterion for quality. Generate a new set of 10,000 samples, this time of dimension 256. Apply each estimator to sub-matrices of samples of size $\{8, 16, 32, 64, 128, 256\}$, and compute the variance of each estimator for each. Plot these (on a single log-log plot), along with a line showing the theoretically-computed variance of the average estimator. Does the variance of all three estimators converge at the same rate ($1/N$)? How much larger is the variance of the median estimator than the average estimator? How large a sample would you need for the average and median estimators to achieve the same variance as the average-extrema estimator (from part (c)) on samples of size 256?

2. **Bayesian inference of binomial proportions.** Poldrack (2006) published an influential attack on the practice of "reverse inference" in fMRI studies, i.e. inferring that a cognitive process was engaged on the basis of activation in some area. For instance, if Broca's area was found to be activated using standard fMRI statistical-contrast techniques, researchers might

infer that the subjects were using language. In a search of the literature, Poldrack found that Broca's area was reported activated in 103 out of 869 fMRI contrasts involving engagement of language, but this area was also active in 199 out of 2353 contrasts not involving language.

(a) Assume that the conditional probability of activation given language, as well as that of activation given no language, each follow a Bernoulli distribution (i.e., like coin-flipping), with parameters x_l and x_{nl}. Compute the likelihoods of these parameters, given Poldrack's observed frequencies of activation. Compute these functions at the values `x=[0:.001:1]` and plot them as a bar chart.

(b) Find the value of x that maximizes each discretized likelihood function. Compare these to the exact maximum likelihood estimates given by the formula for the ML estimator of a Bernoulli probability.

(c) Using the likelihood functions computed for discrete x, compute and plot the discrete posterior distributions $P(x \mid data)$ and the associated cumulative distributions $P(X \leq x \mid data)$ for both processes. For this, assume a uniform prior $P(x) \propto 1$ and note that it will be necessary to compute (rather than ignore) the normalizing constant for Bayes' rule. Use the cumulative distributions to compute (discrete approximations to) upper and lower 95% confidence bounds on each proportion.

(d) *Are these frequencies different from one another?* Consider the joint posterior distribution over x_l and x_{nl}, the Bernoulli probability parameters for the language and non-language contrasts. Given that these two frequencies are independent, the (discrete) joint distribution is given by the outer product of the two marginals. Plot it (with `imagesc`). Compute (by summing the appropriate entries in the joint distribution) the posterior probabilities that $x_l > x_{nl}$ and, conversely, that $x_l \leq x_{nl}$.

(e) *Is this difference sufficient to support reverse inference?* Compute the probability $P(language \mid activation)$. This is the probability that observing activation in Broca's area implies engagement of language processes. To do this use the estimates from part (b) as the relevant conditional probabilities, and assuming the prior that a contrast engages language, $P(language) = 0.5$. Poldrack's critique said that we cannot simply conclude that activation in a given area indicates that a cognitive process was engaged without computing the posterior probability. Is this critique correct? To answer this, compare the Bayes factor (probability of language vs. not language) after taking Poldrack's data into account, compared to before having done so.

3. **Bayesian estimation.** Tina and Perri are looking for Nikhil in a very large one-dimensional shopping mall. Location is specified by a coordinate X. They know that, all else being equal, Nikhil prefers to be near the center of the shopping mall at location 50. He has a prior Gaussian distribution centered on 50 with variance 40. The only clue they have is a coffee cup of a brand that only Nikhil drinks that they find at location X=30. The coffee cup is cold and Nikhil has wandered off. Based on the location of the coffee cup, the likelihood function of his location is a Gaussian distribution with mean X=30 and variance 100.

   (a) Explain how you would frame this problem as a problem in Bayesian estimation, using appropriate terminology. What is Nikhil's posterior distribution? Draw his prior, likelihood and posterior distributions on a single plot. (Rather than `normpdf`, compute Gaussian probabilities from the formula for the Gaussian distribution.) What is the variance of the posterior?

   (b) The coffee cup was not that cold after all. Nikhil's likelihood function has mean X=30

but with a smaller variance of 20. Redo part a. Describe what happened to the posterior distribution. Has it moved? Does the change make sense?

(c) What would the posterior distribution in (a) be if the prior had been uniform (and, thus, the posterior proportional to the likelihood). What would the variance of this distribution be? Compare this variance to that of the posterior in (a). How does the inclusion of prior information affect the variance?

4. **Signal Detection Theory.**
Consider an experiment where a moving-dot visual stimulus is presented to a subject. The difficulty of detecting the motion is varied by changing the *coherence* of the moving dots. At zero coherence, the dots move randomly, and at 100% coherence, all of the dots move to the right. Let's say we want to decode whether the stimulus is random or is moving to the right. Consider a single neuron that has a firing rate that is approximated by a Gaussian. The neuron fires with a mean of 5 spikes/s in response to a 0% coherence noisy stimulus and 8 spikes/s for 10% coherence, and has a standard deviation of 1 spikes/s for both stimuli. Our decoder will work as follows: set a threshold t. If the firing rate $r \geq t$ then we report "right". If $r < t$ we report "no direction".

   (a) For the "no coherence"' stimulus, generate 1000 trials of the firing rate of the neuron in response to these stimuli (i.e., draw 1000 random samples from a Gaussian with $\mu = 5$ and $\sigma = 1$). Since we cannot have negative firing rates, set all rates that are below zero to zero. Now do the same thing for the 10% coherence stimulus. On the same figure, plot the histograms of the firing rates for each stimulus type.

   (b) The success of the decoder (assuming this model of Gaussian noise) is determined by two things, the separation of the mean firing rates and the standard deviation of the neuron. From class, we know that this is captured in the measure known as d'. Calculate d' for this task and pair of stimuli.

   (c) Select various thresholds t and for each threshold calculate the hit and false-alarm rates using your sample data from (a). What threshold would you pick based on this curve to maximize the percentage-correct of the decoder, assuming that 0% and 10% coherence stimuli occur equally often. Plot this threshold as a point on the ROC curve and as a vertical line on your histogram from part (a). Next, suppose that 10% coherence stimuli occur 75% of the time. Determine and plot the threshold that maximizes percentage correct for this new prior.

   (d) Consider now a neuron with a more "noisy" response so that the mean firing rates are the same but the standard deviation is 2 spikes/s instead of 1 spike/s. What is the new value of d'. Recompute and plot the optimal (maximum accuracy) thresholds for this noisy neuron for both the 50-50 and 75-25 priors.