PSYCH-GA.2211/NEURL-GA.2201 – Fall 2014
Mathematical Tools for Cognitive and Neural Science

Homework 4

Due: 26 Nov 2014
(late homeworks penalized 10% per day)

Save the solutions to each numbered problem as sections of a file called `hw4.m` in a folder called `hw4_Lastname`, along with additional files containing any functions you create. Send a zipped copy of this folder as an attachment in an email message with these attributes:

`To: catherio@nyu.edu, asr443@nyu.edu`
`Subject: Math Tools HW4`

Don't wait until the day before the due date...

1. **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 in some contrast, 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., active or not with some probability as in the coin-flipping example in class). Compute the likelihoods of Poldrack's observed frequencies of activation as functions of the possible values of their respective Bernoulli probability parameters x_l and x_{nl}. 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. Compare these to the exact bounds computed using Matlab's `betacdf` or `betainv` functions.

   (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?* Using the estimates from part (b) as the relevant conditional probabilities, and assuming the prior that a contrast engages language, $P(language) = 0.5$, compute the probability $P(language \mid activation)$ that observing

activation in this area implies engagement of language processes. Is Poldrack's critique correct? How confident should you be in implicating language if you observe activity in Broca's area?

2. **Multi-dimensional Gaussians.**

   (a) Write a function `samples = ndRandn(mean, cov, num)` that generates a set of samples drawn from a multidimensional 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. Your function should use the MATLAB function `randn` to generate samples from an N-dimensional Gaussian with identity covariance matrix, and then modify these appropriately. (Hint: Recall that the sample covariance is $1/N \cdot X^T X$. If $1/N \cdot X^T X$ is I then $(XY)^T(XY)$ has sample covariance $1/N * Y^T Y$. Therefore you need to multiply samples X by a matrix Y such that $Y^T Y = cov$. You can find such a matrix using the `svd` of `cov`).

   (b) Test your function by plotting 1000 samples of a 2-dimensional Gaussian (choose an arbitrary nonzero mean and nonzero covariance). 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 that traces out points that are two standard deviations away from the mean, according to the covariance matrix. Does this ellipse capture the shape of the data?

   (c) Now consider the generalized marginal distribution of your 2-D Gaussian in which samples are projected onto a unit vector $\vec{u}$ to obtain a 1-D distribution. Write a mathematical expression for the mean and variance of this marginal distribution as a function of $\vec{u}$ and check it by choosing some arbitrary (random) vectors $\vec{u}$, and comparing the mean and variance predicted (using the equations presented in class) to the sample mean and variance estimated by projecting your 1,000 samples onto $\vec{u}$.

3. **Dueling estimators**. Consider two estimators of the parameter μ in the Normal(μ,σ^2). The first (which is also the maximum likelihood estimator) is the mean of a sample from the distribution, but a second possibility is the median of the sample. Lets evaluate each of them. For convenience, we set $\mu = 0$ and $\sigma^2 = 1$ and we will see how well each of the estimators does at estimating ahe true value of μ.

   (a). Generate 10,000 samples, each of size 10, from the Normal(0,1) distribution (as a 10x10000 matrix). Compute the average of each of the 10,000 samples. Call this vector est1. Plot a histogram of est1 (use roughly 50 bars, and use `axis` to set the horizontal axis limits to (-2,2) and the vertical axis to convenient values).

   (b). You expect the histogram in (a) to look Normal (Explain why). What is the standard deviation of the means (std(est1))? You should be able to calculate theoretically what the standard deviation of the average of a sample with size 10 from Normal(0,1) should be. What is it? How does your calculation compare to std(est1)?

   (c). Once again, generate 10,000 samples, each of size 10, from the Normal(0,1) distribution (as a 10x1000 matrix). Compute the median of each of the 10,000 samples. Call this vector est2. Histogram est2 with the same axis limits (and with 50 or more bars). You have no reason to expect the distribution to look Normal. Does it? Use the function `normplot` to check your intuition. (This function plots the quantiles of a sample of data versus the normal quantiles – a so-called Q-Q plot. When data are normally distributed, the points

shuld fall nearly on a straight line.) Comment on the results: do the points deviate from what is expected from normality, and if so, what does this mean in terms of the distribution?

(d). These estimators are both unbiased, so our criterion for quality will be their standard deviation (equivalently, their variance). What is the standard deviation of the medians (std(est2))? Which has a higher standard deviation, the mean or the median? Can you see it in the histograms? Compute the ratio of the estimate of the standard deviation of the mean to the estimate of the standard deviation of the median (you could use the theoretical value of the standard deviation of the mean instead of the estimate if you like).

(e). You know how the standard deviation of the mean varies with sample size. Adjust the sample size so that the standard deviation of the mean with the new sample size is close to the standard deviation of the median with sample size 10, and verify that it works. Moral: You can get the same performance out of the better estimator with fewer data points!

4. **Gaussian estimation.** Catherine is looking for Alex in a very large one-dimensional shopping mall. Location is specified by a coordinate X. Catherine knows that, all else being equal, Alex 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 Catherine has is a coffee cup of a brand that only Alex drinks that he finds at location X=30. The coffee cup is cold and Alex 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 Alex'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. Alex'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?