Mathematical Tools for Neuroscience (NEU 314)
Princeton University, Spring 2016
Jonathan Pillow

Homework 7: Maximum likelihood estimators & Bayesian inference

Due: Tuesday, April 19, 9:59am

Problems (#1-3) involve paper-and-pencil mathematics. Please submit solutions either as physical copies in class (if you write the solutions out long-hand), or send them as pdf if you prepare solutions using latex or other equation formatting software. (See `https://www.overleaf.com/` if you'd like help getting started with latex).

1. **General linear model: exponential regression**

   Suppose we have a neuron that responds linearly to contrast x with slope parameter θ (i.e., just like examples we saw in class), but the noise is governed by an exponential distribution:

   $$P(y \mid x) = \alpha e^{-\alpha y} \tag{1}$$

   where $\alpha = \theta x$ is the parameter of the exponential distribution (which in this case corresponds to 1 over the mean of the distribution).

   Derive the maximum likelihood estimator for θ based on a dataset of stimulus-response pairs $\{(x_i, y_i)\}$. (You should be able to find a closed form expression, just like in the Poisson and Gaussian regression problems we examined in class).

   Note this is an example of a *general linear model* — not "generalized", since there's no nonlinearity between the linear stage and the noise.

2. **ML estimator for binomial distribution**.

   The binomial distribution Binom(n, p) describes the probability distribution over the number of heads from n independent coin flips, where each coin has probability p of turning up heads. We can write:

   $$X \sim \text{Binom}(n, p) \tag{2}$$

   to indicate that X is a random variable with a binomial distribution with n trials (often referred to as "n Bernoulli trials") and parameter p. This is equivalent to saying that X has probability mass function:

   $$P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{3}$$

   You can read this equation as saying "The probability that X takes on the value k given n and p is equal to... *(math expression on the right)*". The term $\binom{n}{k}$ in front is the binomial coefficient, given by $n!/(k!(n-k)!)$, which counts the number of different ways you can get k heads from n flipped coins, each of which occurs with probability $p^k(1-p)^{n-k}$. This term is there to make the distribution sums to 1—we will mostly ignore it in everything that follows because it does not involve p.

(a) Derive the maximum likelihood estimator for p given an observation of k heads from n coin flips. To do this: write out the log-likelihood, $\log P(k \mid n, p)$, take the derivative w.r.t p, set it equal to zero, and solve for p. (Show the steps in your derivation!).

(b) You observe three binomial random variables that take on values k_1, k_2, and k_3, from n_1, n_2, and n_3 coin flips. What is the maximum likelihood estimator for p given all three observations (assuming that all coins flipped had the same probability of "heads")?

3. **MAP estimator.** Now, let's take a Bayesian approach and incorporate some prior information. Let's say you feel reasonably confident that the coins you're using are biased towards heads, and you would like to take this into account when estimating p. To incorporate this belief, you place a prior distribution over p, specifically a *beta* distribution, which is given by:

$$P(p) = \tfrac{1}{B(\alpha,\beta)}, p^{\alpha-1}(1-p)^{\beta-1}. \tag{4}$$

The $B(\alpha, \beta)$ term in front (which denotes the Euler beta function) is a normalizing constant, whose job is just to make sure the density integrated from $p = 0$ to $p = 1$ is 1. We can ignore it for most everything we do here.

The beta distribution is a nice prior to use with binomial observations (technically known it known as the *conjugate prior* for the binomial distribution) because the product of prior and likelihood has the same "nice" tractable form as the prior distribution. In practical terms, the beta prior behaves like "extra" observations of coin flips that occurred before the experiment. Placing a Beta(α, β) prior on p is equivalent to having observed $\alpha - 1$ heads and $\beta - 1$ tails before the experiment. A Beta$(1, 1)$ prior is the uniform (or "flat") prior distribution—you can verify this by looking at the form of the distribution above (which involves $p^{\alpha-1}$ and $(1-p)^{\beta-1}$).

(a) Let's say you wish to incorporate your prior beliefs by using a prior over p given by $P(p) = \text{Beta}(3, 1)$. (As explained above, this is equivalent to saying you have observed 0 tails and 2 heads before starting the experiment). Derive the *maximum a posteriori* estimate of p, which is the maximum or most likely value of p under the posterior distribution. The posterior posterior distribution is (according to Bayes' rule) equal to the the product of the (binomial) likelihood and (beta) prior, divided by a normalizing constant:

$$P(p \mid k) = \tfrac{1}{P(k)} P(k \mid p) P(p) \tag{5}$$

(We have dropped the number of coin flips n from the expressions above, since it is easier to recognize Bayes' rule without it.)

You should repeat the basic logic of what you did in problem 2a above: take the log of the posterior (ignoring the $P(k)$ term because it does not involve p), differentiate with respect to p, set equal to zero, and solve for p.

(b) Now derive the MAP estimator for p in the general case of a binomial $P(k \mid n, p)$ likelihood and a Beta(α, β) prior. What happens to the estimate if we increase α and β to very large values and $\alpha = \beta$? When do the likelihood and the prior have "equal influence" in determining the posterior?

4. **Application: Bayesian inference for binomial proportions.** Now let's try implementing the estimators we derived above in matlab.

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 fMRI contrast (i.e., experiment), researchers concluded 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 in each experiment is given by a Bernoulli distribution $P(k \mid p_l) = p_l^k(1-p_l)^{1-k}$, where p_l is the probability of activation and k is 0 (no activation) or 1 (activation). Assume the probability of activation given no-language also follows a Bernoulli distribution $P(k \mid p_{nl}) = p_{nl}^k(1-p_{nl})^{1-k}$, where p_{nl} is the probability of activation when there is no language.

(Note: the binomial distribution arising from only one coin flip is more commonly known as the Bernoulli distribution; we obtain a binomial distribution Binom(n, p) by summing up n independent Bernoulli random variables Ber(p), each of which has probability p of coming up heads.)

Compute the likelihood functions of Poldrack's observed counts of activation as functions of their respective Bernoulli (or binomial) probability parameters p_l and p_{nl}. Compute both likelihood functions at the values `p`=[0:.001:1] and plot them.

(b) Find the values of p that maximize the two discretized likelihood functions. Compare these to the exact maximum likelihood estimates given by the formula for the ML estimator of a binomial probability (as you derived above).

(c) Now let's compute the posterior distribution over p_l and p_{nl}. From Bayes' rule, these are given by

$$P(p_l \mid k_l, n_l) \propto P(k_l \mid n_l, p_l)P(p_l) \quad \text{and} \quad P(p_{nl} \mid k_{nl}, n_{nl}) \propto P(k_{nl} \mid n_{nl}, p_{nl}P(p_{nl}) \qquad (6)$$

Use a uniform prior distribution $P(p) \propto 1$ on each p parameter (which, as we saw above, corresponds to a Beta$(1, 1)$ prior). In practice, you can just normalize each likelihood function so that it sums to 1, since the prior has no influence on shape of the posterior.

Now, find the posterior mean, $\mathbb{E}[p|k, n] = \int_0^1 pP(p|k, n)\,dp$, which is also known as the *Bayes' least squares estimate* for each p parameter. (This is just numerically finding the mean of a distribution, something we did in homework problem set 5!) Do these estimates differ from the maximum likelihood estimates computed above? Why or why not?

(d) Find 95% confidence intervals for p_l and p_{nl} using the posterior distribution. (These intervals are formally known in Bayesian statistics as *credible intervals*.) These intervals are defined by the 0.025 and 0.975 quantiles of the posterior distribution, that is, the points on the x axis where exactly 2.5% and 97.5% of the probability mass is located to the left. Based on these intervals, can you confidently rule out that p_l is different from the estimated value of p_{nl}? What about vice versa?

(e) *Bayes rule again: Is the difference in proportions sufficient to support reverse inference?* Using the estimates of p_l and p_{nl} from part (b) as the relevant conditional probabilities of activation given language or no-language, and assuming the prior that a contrast engages language is $P(language) = 0.5$, compute the probability $P(language \mid activation)$ that observing activation in this area implies engagement of language processes. Is Poldrack's critique correct?