G80.3042.002 – Fall 2007
Statistical Analysis and Modeling of Neural Data

Homework 1

Due: 10 Oct 2007

Your results should be in the form of a MATLAB file (typically, the filename should have an extension of `.m`). Email your solutions to `eero@cns.nyu.edu` and `bijan@cns.nyu.edu`.

1. **Cross-validated rate function**.

   (a) Write a function that generates zero or one spike (i.e., a binary or *Bernoulli* process), with a probability that is sigmoidal with respect to its input variable:

   $$p(x) = (\text{erf}(2x) + 1)/2$$

   over the interval [-1, 1].

   (b) Generate two data sets (one for training, one for testing) by generating 300 random input values $x \in [-1, 1]$, computing the probability of spiking for each, and then generating (zero or one) spikes according to that probability.

   (c) Estimate the rate function as a piecewise constant function, by binning the data with a fixed binwidth, and computing the rate for each bin that maximizes the likelihood. Show that this rate is simply the number of spikes divided by the total number of samples (for each bin).

   Plot the average log likelihood of the training data as a function of binwidth. It's easiest to do this by dividing the interval [-1 1] into N equal bins, and testing various values of N from 1 up to (say) 50. You need not compute this for all N's (i.e., as N gets larger, you can sample at larger intervals).

   (d) Now compute the average log likelihood of the *testing* data as a function of binwidth (using the rate function fitted to the training data). You should do this 10 times for each N, using a randomly selected subset of half of the testing data, and averaging the result. Plot the result on the same graph as the training likelihood. Discuss.

2. **LNP fitting**.

   (a) Write a function `expLNP` that takes as input a stimulus block (dimensions $N \times T$), a linear weighting vector of length N, and a mean firing rate, and generates a spike train that is the response of an LNP model with an exponential nonlinearity. Write a second function `sigLNP` that does the same thing, but uses a sigmoidal nonlinearity of the form $A\text{tan}^{-1}(k \cdot x)/\pi + 1/2$. You'll need to adjust A in order to match the desired mean firing rate.

   (b) Examine the convergence of the STA for both functions. With mean firing rate set to 0.1 spikes / timebin, generate input sequences of stimulus dimension $N = 10$, and different lengths T, and plot the cosine of the angular error $k \cdot \hat{k}/(|k| \cdot |\hat{k}|)$. Show that in both cases, the convergence goes as $1/\sqrt{(}T)$. Which LNP model produces better STA estimates? Why?

(c) Use matlab's `fminunc` function to solve for the kernel k by maximizing the likelihood of the data for the two models (assuming the nonlinearity is *known*). How do errors of these estimates compare to those of the STA (at a single intermediate value of T)? Why?

3. **Renewal process model**. For this problem and the next, you will work with an experimental data set recorded from area LIP of an awake, behaving monkey. This data set contains simultaneous recordings of two neurons as a monkey performs a delayed reach-and-saccade movement to a peripheral target.

   Data format: The data is organized into two data structures, `Baseline` and `Delay` and each data structure contains two fields, `Spike1` and `Spike2`. `Baseline.Spike1` is a cell array containing spike times for 74 trials during the baseline period before the peripheral target is illuminated. `Baseline.Spike2` contains spike times for the second neuron. `Delay.Spike1` and `Delay.Spike1` contain spike times for 9 trials for each neuron during the Delay period when the target is in the response field. Spike times are given in milliseconds.

   (a) Estimate the inter-spike interval density for spikes from each neuron during Baseline and Delay. Use a fixed bin-size histogram estimate. Minimize the cross-validation score for density estimates to determine the bin-size. How do you average across multiple trials in constructing your histogram estimates?

   (b) The gamma function $f(\tau; \lambda, k)$ is often proposed as a model for the probability density of inter-spike intervals, τ.

   $$f(\tau; \lambda, k) = \frac{\tau^{k-1}}{\Gamma(k)} \lambda^k exp(-\lambda\tau)$$

   k is the shape parameter and λ is the rate parameter.

   Based on your histogram estimates above, it should be clear that a single gamma distribution is not a good model for the probability density of these ISIs. Mixture models are a more flexible class of model that represent probability densities as a sum of multiple probability densities each with their own probability weighting, called the mixture coefficient. This allows us to have each observation of the ISI to potentially come from one of multiple densities, each with a different probability. Write down a probability density model for the ISIs in terms of a mixture of two gamma distributions. What are the free parameters of this model?

   (c) Under the assumption that the spike trains are being generated by a renewal process, the inter-spike intervals are independent and identically distributed. Write down the log likelihood function of the observed ISIs for the mixture of gamma distribution pdfs under the renewal process assumption.

   (d) Fitting mixture models is tricky because you don't know the mixture coefficients when maximizing the log-likelihood (or equivalently, minimizing the negative log-likelihood). Here, we will simplify this problem. Assume that the two gamma mixtures are sufficiently well-separated that one gamma distribution can be fit by considering only the ISIs less than 8 ms in duration and one gamma distribution can be fit by considering only the ISIs greater than 8 ms in duration. Under the assumption, what is the mixture coefficient for each of these two components for each neuron during Baseline and Delay?

(e) Fit each gamma mixture component, given the ranges we are considering, by minimizing the negative log-likelihood of our observations for a single gamma distribution. Do not simply use the `gamfit` function in Matlab. Explicitly take the derivative of the log-likelihood function with respect to the shape and rate parameters and set it to zero. Show your work. (Hint: The derivative of the $\Gamma(k)$ is the `Digamma` function). Using `gamfit`, what are the shape and rate parameters for each of the two components in the mixture model for each neuron in each period?

(f) Simulate inter-spike intervals using the full mixture model and plot histogram estimates of the resulting data. You can do this using the `gamrnd` function in Matlab. Directly compare the histogram estimates of the simulated data and the experimental data by plotting them on the same axes. Comment on your results. What features of the data has this procedure captured and what features remain? What can we say about the activity of neurons during Baseline and Delay?

4. **Spectral estimates for point processes**

(a) Write a program to construct a (primitive) spectral estimate. Your program should calculate $dZ_N(f)$ for the spikes frm each trial using the discrete Fourier transform and calculate the spectrum by averaging $|dZ_N(f)|^2$ across trials. Use your program to estimate the spectrum for both neurons in the Baseline and Delay periods. Comment on how the spectral estimates vary with the number of spikes used in the estimator through changing the number of trials used in your estimate.

(b) Write a program to construct an estimate of coherence between two point processes. Your program should use the estimates from part (a) to construct the cross-spectrum and normalize by the spectrum of each process. Estimate the coherence for the pair of neurons in both the Baseline and Delay periods.

(c) Take your simulations of a renewal process from above and estimate the spectrum and coherence of your model of the data using a mixture of gamma distributions. Remember to convert the interval representation in your model to the spike times for the spectral estimates. Directly compare the spectrum and coherence of the model data and the experimental data by plotting them on the same axes. Comment on your results. Do you think the renewal process model captures all the spectral properties of these data?