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

Homework 3

Due: Friday, 7 Dec 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. **Optimal linear estimation**.

   Consider a population of 11 Poisson neurons, which respond independently to a scalar input variable with Gaussian tuning curves centered at locations [-5:5].

   (a) Assume that the width (standard deviation) of the tuning curves is $\sigma = 1$. Compute the Optimal Linear Estimator (OLE) via regression. Specifically, draw 100 samples of input uniformly from the range [-5, 5]. For each of these, draw 10 sample population responses. Solve for the linear weight vector that minimizes the squared error between the weighted sum of responses and the input value. Generate a separate test data set and compute the (cross-validated) mean squared error.

   (b) Examine the behavior of this estimator as a function of σ. Specifically, for a set of sigma's spanning the range [0.25 ... 3], compute the optimal estimator (as above), and compute the mean squared error (as above). Plot the MSE as a function of sigma. What is the best choice of sigma?

2. **Constructing confidence intervals**.

   In this problem, you will use the bootstrap to construct confidence intervals for the spectrum of broad-band extracellular potentials and will compare them with the asymptotic intervals assuming a Gaussian process. You can use these confidence intervals to make inferences about the underlying signals and properly understand the behavior of your data.

   Data format: The data is organized in two data structures `Raw1` and `Raw2` each with the field `Baseline`. These contain the broad-band data during an instructed delay task from two of the simultaneously-recorded electrodes that you analyzed for Homework 2. `Baseline` contains 500 ms of data during the baseline interval before the peripheral cue is presented. The data is organized in 500 ms per trials with one trial per row. The signals are sampled at 20 kHz.

   Software: You have been given a matlab function to estimate smoothed spectral estimates, `dmtspec.m`. Place this program in the path of your matlab session. To estimate the the spectrum of a Raw signal during the Baseline interval with W Hz smoothing for each trial use the command `spec1 = dmtspec(Raw1.Baseline,[.5,W],2e4);`. This will return an array of spectra one for each trial from DC to 10 kHz. `help dmtspec` will return some documentation. When working with spectral power, always take the `log` before analyzing the signal further. it i.e. work with `log(spec1)`.

   (a) The asymptotic distribution for a spectral estimate is:

$$\nu \hat{S(f)}/S(f) \sim \chi^2_\nu.$$

where ν is the number of degrees of freedom in the estimator. For the multitaper estimator used by `dmtspec`, ν is $2\mathtt{Ntr}(2NW-1)$ where `Ntr` is the number of trials in the average and N is the duration of the analysis window in seconds *i.e.* 0.5 .

Estimate the spectrum during the baseline interval using all the trials for `Raw1` and `Raw2` and $W = 2Hz$. What frequency bands are corrupted by artifacts? Use the asymptotic distribution to construct 95% confidence intervals for each signal on the same axes. Are there any frequencies for which these confidence intervals do not overlap? Increase the smoothing to $10Hz$ and above and repeat the analysis. What is the effect of smoothing on the confidence intervals? How much smoothing is necessary to resolve significant differences? Hint: You can use `chi2inv` to get the critical values for χ^2_ν.

(b) The bootstrap is a method for estimating the variance of a statistic that we can use to construct confidence intervals. Instead of averaging all the data available in the dataset to get a single estimate of the quantity of interest, the bootstrap subsamples the data, with replacement, to generate many different estimates of the estimator. You can then estimate the variance of the estimator by estimating the variance of the bootstrap samples. We call this variance the bootstrap variance. The standard error of the estimator, σ_b is simply the square root of the bootstrap variance.

The simplest bootstrap estimate for a confidence interval constructs the Normal interval for the estimator.

$$\hat{S(f)} \pm z_{\alpha/2}\sigma_b$$

This estimate is accurate if the distribution of estimator is close to Normal. Estimate the bootstrap variance for spectrum of Baseline interval activity in each signal with $10Hz$ smoothing and use this to construct 95% Normal intervals. Compare these confidence intervals with the χ^2_ν intervals above. How do they differ? Why do they differ? Plot a graph to support your claim.

(c) The bootstrap can also be used to estimate the distribution of the estimator. We call this estimate the empirical distribution function of the estimator, in contrast with the asymptotic distribution function of the estimator given by the χ^2_ν distribution above. We can determine bootstrap percentile intervals directly from this distribution by taking the upper and lower $\alpha/2$ quantiles.

Estimate the 95% bootstrap percentile confidence intervals for the spectrum of Baseline interval activity for each signal using $10Hz$ smoothing. Compare with the asymptotic and Normal intervals from above. Now estimate 99% intervals. How many bootstrap samples do you think need to estimate these intervals as they become increasingly conservative? Perform an analysis to justify your claim.