EN.580.439/639: Models of The Neuron
Homework #5: Modeling Spike Train Data of a Place Cell

Due: Upload to Blackboard by 11/10/2017 @ 10 am
using usual submission format with final write-up and code

- Part 1: Constructing a statistical model
  Part 2: Evaluating a statistical model (HW #6)

Models are used to explicitly propose a relationship between random signals and any other measured signals including stimulus values. Generally speaking, statistical models can be useful to summarize and track changes to this relationship, and can include physical models (eg. Hodkin-Huxley-type networks) that more concretely postulate mechanisms by which the relationship arises from biology.

Focus Questions:

What is a statistical model?
How do I specify a statistical model for my spiking data?
How do I choose the model parameters?

Take home point:

The statistical model specifies the probability of observing some number of spikes in any time interval given (i.e. conditioned on) related covariates, such as stimulus values, or the spiking history.

1: Plot raw data

```
>> load glm_data.mat
>> who
```

Your variables are:

spiketimes	Time stamps on spike events
x_at_spiketimes	Animal position (x) in m at corresponding times stamps
y_at_spiketimes	Animal position (y) in m at corresponding times stamps
T	Time
spikes_binned	spike data in 33 ms bins
xN	Normalized position (x)
yN	Normalized position (y)
vxN	Normalized velocity (v_x)
vyN	Normalized velocity (v_y)
r	Movement speed = sqrt(v_x^2+ v_y^2)
phi	Movement direction = atan2(v_y,v_x)

Use MATLAB to visualize the spiking activity as a function of time and then as a function of the animal's position. Include plots in your HW solutions with appropriate labels and comment on what you see.

2: Choose a model form

The model is a function that specifies the probability of observing a spike in each bin. Let dN_k be the number of spikes in the k^{th} interval of length Δ milliseconds (we can always choose a small enough Δ such that the number of spikes in any bin is at most 1). dN_k is a function of the parameters of the model (θ) and values of other signals of interest, i.e. covariates, ($x^{(1)}$, $x^{(2)}$,…, $x^{(N)}$) that are postulated to affect the current probability of spiking, such as our stimulus and the occurrence of previous spikes:

$$\Pr(dN_k \mid \theta, x^{(1)}, \ldots, x^{(N)}) = f(dN_k, \theta, x^{(1)}, \ldots, x^{(N)})$$

An example of a covariate would be the intensity (dB) of a sound stimulus is a covariate to the spiking rate of auditory nerve fibres.

Let's explore one way to specify this probability. Define the function λ_k (spikes/ms) that can be used to calculate the instantaneous probability that a spike will occur. We can introduce parameters and covariates into the model by defining λ_k in terms of θ and $x^{(1)}$,…,$x^{(N)}$.

For this exercise, we will assume the bin size is set small enough that the covariates are approximately constant within the bin, so that λ_k will be approximately constant over each bin. We assume the number of spikes that arrive in the k^{th} bin is distributed as a Poisson distribution with mean and variance $\lambda_k\Delta$.

✎ **Write down $Pr(dN_k \mid \ldots)$ as a Poisson distribution with mean spike count $\lambda_k\Delta$.**

To make parameter estimation easier, we restrict the relationship between λ_k, θ, and $x^{(i)}$ to follow the generalized linear model (GLM) framework for a Poisson distribution. This just means that

$$\log(\lambda_k) = \beta_0 + \sum_{j=1}^{N} \beta_j x_k^{(j)}$$

✎ **Pick a set of at least two covariates (remember, this can include any function of the variables above) and write an equation that linearly relates $\log(\lambda_k)$ to your covariates. Visualize the model by plotting the predicted (λ_k) as a function of two of your covariates.**

Note: glm_part1.m is already set up with a set of covariates and the visualization. If it's clear to you which covariants it's using you can just use it's covariants.

3: Write down the data likelihood

The data likelihood is the conditional probability of observing the sequence of spike counts. We are conditioning on the values of the covariates from (2) as well as the GLM parameters:

$L(\theta) = Pr(dN_1,\ldots,dN_k \mid x^{(1)},\ldots, x^{(N)},\theta)$, where the observed data is held fixed.

✎ Write down an equation for $L(\theta)$ as a function of the dN's, x's and θ. Use your results from the previous two questions.

4: Choose model parameters to maximize the data likelihood

We would now like to make a choice of model parameters that would have most likely generated the observed data. This is accomplished by choosing the parameter estimate $\hat{\theta}$ that maximizes the data likelihood:

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Our problem of estimating model parameters has become a question of optimizing a function $L(\theta)$ over a possibly unconstrained space θ. There are several methods to approach optimization. Because we chose models built on the GLM framework, this optimization is computationally simplified.

✎ **Use the Matlab function glmfit to find the parameters of your model that maximize L(θ).**

The script `glm_part1.m` is set up to fit and visualize a GLM for the spiking activity as a function of the rat's position.

```
>> glm_part1
```

This outputs a plot of the raw spiking data and another plot of the model spiking rate, both as a function of the animal's position.

✎ **The firing rate seems localized to a small spatial region. Create a set of covariants that for which the model is a gaussian-shaped "hill" of activity. Instead of writing your own script, modify the GLM in the script `glm_part1` to include other covariates and functions of covariates. Modify line 20 by including your new covariates in the model fit and modify line 30 by writing in the equation for λ.**

B

Different covariant

State the final values of your parameters and comment on your GLM. That is, how does each covariate influence λ_k? You can use the visual of your model that you create below in 5 to comment.

For small parameter vectors, it is often useful to plot L(θ) as a function of θ around the parameter values returned by glmfit, to confirm that glmfit indeed maximized the data likelihood. The shape of the likelihood around the maximum likelihood estimate determines the uncertainty about the estimates from the data.

5: Visualize the model

✎ **Use the equation for λ you wrote down in step 3, plot the intensity as a function of your covariates. Compare this graph against a plot of the signal values at which spikes occurred. This will serve as a visual confirmation of the plausibility of the model in explaining the relationship between the observed spiking and your covariates.**

6. Investigate other covariates besides position of rat.

The `script glm_part2.m` is set up to visualize models of spiking as a function of velocity.

```
>> glm_part2
```

Currently, this simply outputs occupancy normalized histograms of spiking as a function of the velocity covariates. Examining these histograms can provide insight into possible spiking models. **Comment on these histograms.**

✎ **Modify the script `glm_part2` (see commented lines) to fit GLM models to each of these variables, and then plot the GLM fits along with the occupancy normalized histograms. What do these model fits suggest about the relationship between spiking and velocity?**

For next time:
In the next homework, we will investigate how to quantitatively evaluate the quality of the model you developed here.