Exam 2: The Monkey's Paw

36-402, Spring 2015

Due at 11:59 pm on Monday, 13 April 2015

SCIENTIFIC BACKGROUND: Nerve cells (or "neurons") communicate and process information by transmitting little electrical impulses to each other, called "spikes"[1]. Many neurons use "rate codes", where the number of spikes they produce in a short period of time encodes information either about some aspect of the world the organism is sensing, or about how the organism is acting or is going to act.

For example, when very fine electrodes are inserted into certain motor-control regions of the brains of monkeys, so that neuroscientists can record from individual neurons, some cells are found to encode the direction in which the monkey intends to move its hand. Specifically, a neuron has a preferred direction vector $\vec{b}$, and the when the monkey intends to move its hand with velocity $\vec{v}$, the average number of spikes over a short interval is $a + \vec{b} \cdot \vec{v}$, plus or minus some amount of noise. A neuron which behaves like this is said to show "directional tuning", and $\vec{b}$ is its "preferred direction".[2]

The data set `neur.csv` is based on an experiment during which the neuroscientists recorded simultaneously from 96 directionally-sensitive neurons in a monkey's motor region, each cell having a different preferred direction. That is, each neuron i will have its own $\vec{b}_i$ and its own intercept a_i. During each trial, the monkey was to move its hand in one of eight directions, spread evenly around a circle. Each row of the data frame represents 100ms, and so the entries in the data frame are the number of spikes produced by each of the 96 neurons spiked during each time interval.

In this exam, you will both fit a model which derives from this "directional tuning" idea, and consider alternative multivariate models.

1 Specific Problems

1. Explain how this model for spiking is, or is related to, a factor model. Your explanation should indicate how a, $\vec{b}$ and $\vec{v}$ are related to the factor loadings and factor scores, and the number of factors.

[1] Because of how they look in a plot of voltage against time.

[2] For more on such models of neural coding, see, for example, §3.3 of P. Dayan and L. F. Abbott, *Theoretical Neuroscience* (MIT Press, 2001).

2. Fit a factor model with the number of factors you determined is appropriate from problem 1. For each neuron, report its preferred direction. (Since there are a large number of neurons, it would probably be best to report this visually.)

3. Based on your fitted factor model, report an estimate of the intended direction $\vec{v}$ at each time point. (Again, this should probably be reported visually.) The experiment had distinct breaks between trials where the monkey stopped moving in one direction and started moving in another, random direction; can you work out, approximately, where these breaks occurred?

4. Suppose that instead of recording intended velocities in the usual (x, y) coordinates, we used coordinate axes which were rotated 30 degrees counter-clockwise from the usual ones. Show that this would amount to multiplying the intended-velocity vector $\vec{v}$ by $\begin{bmatrix} \cos \pi/6 & -\sin \pi/6 \\ \sin \pi/6 & \cos \pi/6 \end{bmatrix}$. Explain what effect, if any, this would have on the preferred-direction vector $\vec{b}$ of each neuron. Explain how this difference in coordinate systems could, or could not, be detected in your factor analysis of the data. In particular, what would this change of coordinates imply for the interpretation of your factor score estimates and factor loadings?

5. Try fitting a three-cluster mixture model. Why might three clusters, specifically, be reasonable? Which model predicts better, the factor model or the three-cluster mixture model?

   Note: if using the `mixtools` package, you might find it easier to use the function `npEM` to fit a non-parametric mixture model than to use `mvnormalmixEM` to fit a Gaussian mixture model, since the observable variables are discrete counts rather than continuous. Fitting such a mixture model to the full data may take as much as a couple of minutes, so allow plenty of time for debugging and any computation-intensive procedues.

6. Try fitting an eight-cluster mixture model. Why might eight clusters be reasonable? Which model predicts best? (See previous note.)

You are welcome to consider other models for this data as well, but for full credit you must answer all these questions about these models.

2 Formatting Instructions and Rubric

Your main report should be a humanly-readable document of at most 10 single-spaced pages, including figures. It should have the following sections:

INTRODUCTION describing the scientific problem and the data set, possibly including *relevant* summary statistics or exploratory graphs. (Do not include EDA just to have EDA.)

SPECIFIC PROBLEMS answering the questions set above, but avoiding the check-list, itemized format in favor of continuous text, with a logical succession of sentences and paragraphs. (Writing coherently is more important than following the order of the questions.)

CONCLUSIONS summarizing what you have learned from the data and models about whether the directional-tuning model is really a good description of how these neurons encode motion.

You may assume that the reader has a general familiarity with the contents of 401, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set.

Numerical results Numerical quantities should be written out to appropriate precision, i.e., neither more nor fewer significant digits than appropriate.

Code *All* statistical results must be supported by appropriate code, or they will receive no credit. ("Show your work.") The ideal would be to use R Markdown, or knitr+LaTeX, to embed all computations in a humanly readable document, and submit both the knitted version and the source[3] As a second best, it is acceptable to submit a PDF document containing all text and figures, and a separate .R file, containing all supporting computations, clearly labeled via the comments so that it is easy to see which claims or results go with which pieces of code.

Rubric

As usual, this describes the ideal.

Words (10) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (5) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

[3] See examples at `http://yihui.name/knitr/demos/`, and the useful chunk options like `echo` at `http://yihui.name/knitr/options/`; also the examples in the solutions to exam 1.

Code (15) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. All code runs, and the Markdown file knits (if applicable). The main text of the report is free of intrusive blocks of code, which are used only when a specifically-computational point is being made, or when code is actually the clearest way of describing a point.

Specific Problems (25) All specific problems posed in §1 receive clear, well-written and correct answers. The answers show, and convey, a real grasp of the mathematical basis of the models being manipulated, and how quantities in the model are related to the underlying scientific questions about neural coding of movement.

Inference and Uncertainty (15) The actual estimation of model parameters or estimated functions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty (such as confidence intervals or standard errors).

Comparisons (15) All comparisons between models are done in a statistically valid way: if in-sample, they are accompanied by an explanation of why this particular in-sample comparison will not lead to over-fitting; if out-of-sample, there is a clear description of the generalization process being performed. The execution of comparisons is technically correct, and their results clearly described. The extent to which comparisons provide either clear or ambiguous evidence about which models fit better is made plain to the reader, and is carried through to the ultimate conclusions.

Conclusions (15) The substantive questions about neural coding are all answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.