Cogs 209: Scientific Data Analysis and Statistical Learning

Guidelines for mini-projects

To make the material directly relevant and applicable, we will emphasize project-based learning.

- Collaborative science: Work together in teams with complementary expertise. Practice collaboration skills, including both leadership and contributory roles
- Emphasis on rigorous methods, sharp questions, drawing concrete conclusions
- Draw research questions/topics from students' own work and interests
- Practice giving and receiving critical feedback (peer reviewing), and responding to peer reviews

To make the course useful for your needs, we will incorporate projects based on students' own data and research questions. If you are an advanced grad student, you may have your own data (published or unpublished) which you would like to analyze. Early-stage students may seek data from previously published papers, e.g. from your lab or from a public data source.

Groups: You will sign up to be in a group of 3-4 students (sign up in Canvas). Groups should include no more than 1 undergraduate.

Proposal: Each group will propose a mini-project for the class to work on. An example project proposal/template is included at the end of this document. Your mini-project proposal should be a 1-2 page (500-1000 words) documents including these sections:

a. **Title**

b. **Research question**. Try to make the topic/question as specific as possible. You must specifically explain the motivation/rationale — "why" is this question important/interesting, not just "what" is the question

c. **Data/materials.** Provide a "manageable" dataset that can be used to address the topic. You may need to curate the data a bit, to put it into a format that can be readily used for analysis.

d. **Course impact/relevance.** List at least 1 of the course topics (e.g. regression, classification, hypothesis testing, model selection, etc.) which your project addresses.

e. **Outcome(s)**. Provide examples of expected outcomes, e.g. analysis results or findings that may result from analyzing the data.

Projects: Your group will select two of the projects from the class to complete during the quarter. You may select your own project, but this is not required. Your group will submit a writeup of each project following this format:

1. **Introduction and hypotheses**

   Explain the motivation/significance of the project, and any hypotheses that the group aims to address

2. **Methods**

   This is the most important section. Clearly define the methods, emphasizing computational and statistical methods that are used to address the research questions.

3. **Results - model comparison**

   Your results should include at least two different models or analyses of the same data, with an objective comparison/evaluation of the relative merit of the two approaches. Note that the two approaches can be very similar, e.g. fitting a regression with 2 explanatory (independent) variables vs. 3 variables.

4. **Results - visualization**

   Include at least one figure to visualize your results. Your figure should be fully labeled and self-explanatory, with axis labels and a figure legend.

5. **Discussion**

   Discuss the findings, lessons learned, and potential future directions

Peer reviews: you will be required to give "peer review" style feedback to at least two other groups on one of their projects.

Overview of mini-project schedule

- Week 1: Form groups, register your group on Canvas (under People > Mini-project groups) (due: April 4, midnight)
- Week 2: Submit a draft mini-project proposal (due April 11)
- Weeks 3-5: Complete mini-project #1; turn in your project report by May 2
- Weeks 7-9: Complete mini-project #2; turn in your project report by May 23
- Week 5-10: Submit two peer reviews (due June 1)
- Week 11 (finals week): Revise one of your projects and respond to peer reviews. Turn in the final report by June 11.

Project roles

- Each of the five submitted assignments (proposal, projects 1 & 2, peer reviews, revised/final project) should have a designated "First author" (leader)
- Every group member should be the first author on at least one of the submissions
- The first author takes leadership responsibility for that stage of the work, however all students should contribute

Cogs 260 Mini-Project Proposal [example]

Lead Author: Eran Mukamel

Co-authors: ….

Title: **Classifying neurons by type**

Research question.

The brain is a complex biological network, including dozens of distinct neuronal and non-neuronal cell types. Although many differences between cell types in terms of their morphology, connectivity, physiology, and gene expression are known, we do not have standardized methods for assigning cells to specific classes. Single cell RNA sequencing (scRNA-seq) can measure the expression level of every gene, in each of many thousands of individual cells. We propose to use scRNA-seq data to test methods for classifying neurons based on the expression of small groups of marker genes. We hypothesize that non-parametric methods such as k-nearest neighbors classifiers may be able to accurately predict some cell types based on a small subset of 1-10 marker genes. We further hypothesize that such classifiers will work better for distinguishing broad classes (e.g. excitatory vs. inhibitory neurons) than they will for narrowly defined sub-classes (e.g. Layer 6b excitatory neurons). This would be useful for experimental studies which seek to identify cells based on the expression of single genes or small combinations of genes.

Data/materials.

We will use scRNA-seq data from the mouse primary motor cortex (reference: Yao et al. Nature 2022). Single cell RNA-seq was performed on a sample of ~5,000 neurons. The data is available in a Dropbox folder. The folder contains the following files:

File name	Format	Description
Exon_counts.csv.gz	Comma separated values (CSV), gzipped (compressed)	Data table with 5210 rows (cells) and 45,769 columns (genes) Each row corresponds to a single cell (“sample”), with a unique sample_id (first column). The remaining columns contain the “exon counts,” i.e. the number of mRNA molecules that were detected which map to that gene.
Mouse_MOp_markergenes.csv	Csv file	Data table with 5210 rows 16 columns This is a subset of the “exon_counts.csv.gz” file, containing a selected list of 15 genes which have strong cell type specific gene expression. It also includes cell cluster labels, i.e. the assignment of every cell to a cluster of cells with similar gene expression profiles. The clusters correspond to (putative) cell types.
sample_metadata.csv	csv	Metadata about each cell

cluster.membership.csv	csv	Cluster assignment for every cell

One consideration for using these data is that the “clusters” are very fine-grained (there are ~110 different cell types). It may be useful to “coarse-grain” the clusters, combining clusters into larger groups.

Course impact/relevance. This project connects with the topics of classification and cross-validation. The data could also be used to explore dimensionality reduction, logistic regression, and other topics.

Outcome(s). The outcome of this project will be a comparison of classifiers trained on the gene expression data. In particular, we may compare KNN classifiers with different values of K. We would generate a plot showing the classifier performance (fraction of incorrectly classified cells) as a function of the k value. It would be important to compare the performance for both training and test data (e.g. using cross-validation).

Another potential outcome could be a comparison of classification performance using different subsets of genes. We could adopt a subset-selection procedure to choose the genes or combinations of genes that allow the highest accuracy of classification. The results of this type of study would include a plot showing the classification performance for a fixed k-value with different subsets of genes.