

## Homework 3

(Due 05/04/2021, 17:00)

### Instructions:

1. Prepare a report (including your answers/plots) to be uploaded on Moodle.
2. The report should be typeset (for lengthy derivations, the solution can be scanned and embedded into the report).
3. Show all the steps of your work clearly.
4. Unclear presentation of results will be penalized heavily.
5. No partial credits to unjustified answers.
6. Use Matlab or Python for computations.
7. Return all Matlab/Python code that you wrote in a single file.
8. Code should be commented, code for different HW questions should be clearly separated.
9. The code file should NOT return an error during runtime.
10. If the code returns an error at any point, the remaining part of your code will not be evaluated (i.e., 0 points).

Question	Points	Your Score
Q1	50	
Q2	50	
TOTAL	100	

### Question 1. [50 points]

Blood-oxygen level dependent (BOLD) responses of a neural population in human visual cortex are provided in the file `hw3_data2.mat`. This file contains a variable `Yn` that represents 1000 response samples. There is another variable `Xn` that represent 100 regressors that may explain the responses. For all parts, the proportion of explained variance ( $R^2$ ) should be calculated as the square of Pearson's correlation coefficient between measured and predicted responses. Answer the questions below.

**a)** Use the ridge regression method to fit regularized linear models to predict noisy BOLD responses as a weighted sum of given regressors. Perform 10-fold cross-validation to tune the ridge parameter ( $\lambda \in [0 \ 10^{12}]$ ) based on model performance. (Hint: Vary the ridge parameters logarithmically.) Note that for  $\lambda = 0$ , the model obtained with ridge regression is equivalent to the OLS solution. For each cross-validation fold, do a three-way split of the data: select a validation set of 100 contiguous samples, a testing set of 100 samples (that immediately precede the validation set assuming circular symmetry), and a training set of length 800 samples. Fit a separate model for each  $\lambda$  using the training set. Find  $R^2$  of each model on the testing set. Separately estimate  $R^2$  of each model on the validation set. Plot the average  $R^2$  across cross-validation folds, measured on the testing set as a function of  $\lambda$ . Find the optimal ridge parameter  $\lambda_{opt}$  that maximizes average  $R^2$ . Find the model performance by calculating the average  $R^2$  across cross-validation folds, measured on the validation set for  $\lambda_{opt}$ . Plot  $R^2$  curves obtained on testing and validation data for all  $\lambda$  values. Interpret your results.

**b)** Determine confidence intervals for parameters of the OLS model from **part a** (i.e., the model obtained for  $\lambda = 0$ ). Generate bootstrap samples from the 1000 samples in the original data (resample both the regressors and the responses the same way). Perform 500 bootstrap iterations, and refit a separate model at each iteration. Plot the mean and 95% confidence intervals of the parameters in the same graph. Identify and label on your plots, the model regressors which have weights that are significantly different than 0 (at a significance level of  $p < 0.05$ ).

**c)** Determine confidence intervals for parameters of the regularized linear model from **part a** (i.e., the model obtained for  $\lambda_{opt}$ ). Generate bootstrap samples from the 1000 samples in the original data (resample both the regressors and the responses the same way). Perform 500 bootstrap iterations, and refit a separate model at each iteration using  $\lambda_{opt}$  found in **part a**. Plot the mean and 95% confidence intervals of the parameters in the same graph. Identify and label on your plots, the model regressors which have weights that are significantly different than 0 (at a significance level of  $p < 0.05$ ). Compare the results to those in **part b**.

## Question 2. [50 points]

A series of neural response measurements are provided in the file `hw3_data3.mat`. Answer the questions below to examine the relationship between these measurements. Provide plots whenever possible.

**a)** Responses from two separate populations of neurons are stored in the variables `pop1` and `pop2`. We would like to examine whether the mean responses of the two populations are significantly different. The first population contains 7 neurons, whereas the second population contains 5 neurons. Using the bootstrap technique (10000 iterations), find the two-tailed p-value for the null hypothesis that the two datasets follow the same distribution. (Hint: If the two datasets come from a common distribution, is there any need to separate them?)

**b)** BOLD responses recorded in two voxels in the human brain are stored in the variables `vox1` and `vox2`. We would like to examine whether the voxel responses are similar to each other, by calculating their correlation. Using the bootstrap technique (10000 iterations), find the mean and 95% confidence interval of the correlation. Find the percentile of the bootstrap distribution, corresponding to a correlation value of 0. (Hint: Should you resample `vox1` and `vox2` independently or identically?)

**c)** Note that estimation of confidence intervals and hypothesis testing are dual problems. For the dataset examined in **part b**, use bootstrapping (10000 iterations) to simulate the distribution of the null hypothesis that two voxel responses have zero correlation. Find the one-tailed p-value for the two voxel responses having zero or negative correlation. Compare this to the result in **part b**. (Hint: Resample the datasets to break apart the correlation between them.)

**d)** The average BOLD responses in a face-selective region of the human brain have been recorded in two separate experiments. The responses of this region to building images (1st experiment) and face images (2nd experiment) are stored in the variables `building` and `face` for 20 subjects. Assume that the same subject population was recruited in both experiments. Use bootstrapping (10000 iterations) to calculate the two-tailed p-value for the null hypothesis that there is no difference between the building and face responses.

**e)** Repeat the exercise in **part d**, but this time assuming that the subject populations recruited for the two experiments are distinct. Use bootstrapping (10000 iterations) to calculate the two-tailed p-value for the null hypothesis that there is no difference between the building and face responses.