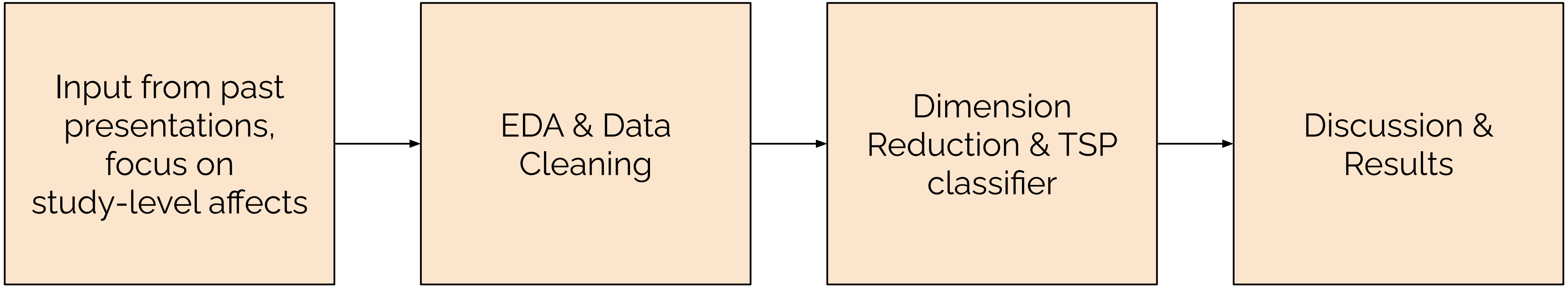
Analysis on pCR Binary Response Data

STAT 117 Final Project, Miela Foster

Roadmap



Input from past presentations, focus on study-level affects

EDA & Data Cleaning

Dimension Reduction & TSP classifier

Discussion & Results

Past Presentations

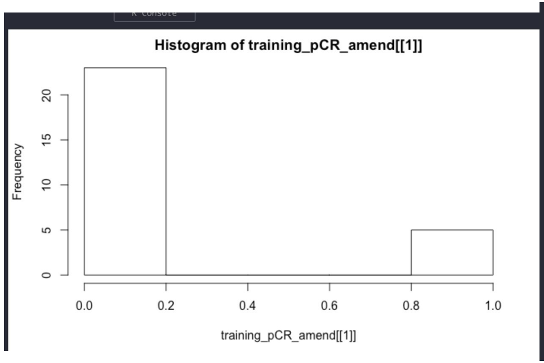
- High degree of variability between studies (size, number of genes, number of participants, types of genes)
- Feature Extraction Methods
- Training and Test Set Split

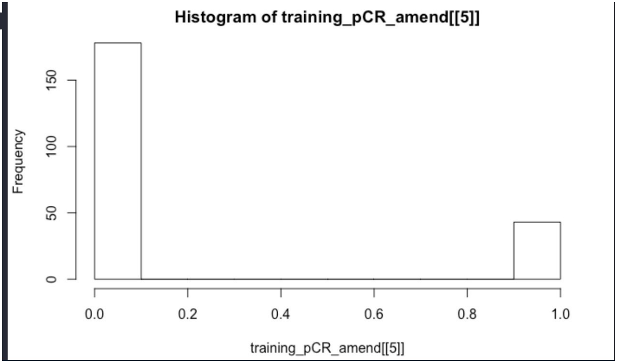
My approach: focus on individual studies in order to control ambiguity of results, keep consistent training and test split, focus of decreasing size of studies to reduce potential overfitting (feature extraction), use TSP for familiarity

EDA & Train/Test Split

- Remove 3 studies because of missing data
- Random number generator to split the train and test set
  - **10 in training set, 3 in test set**

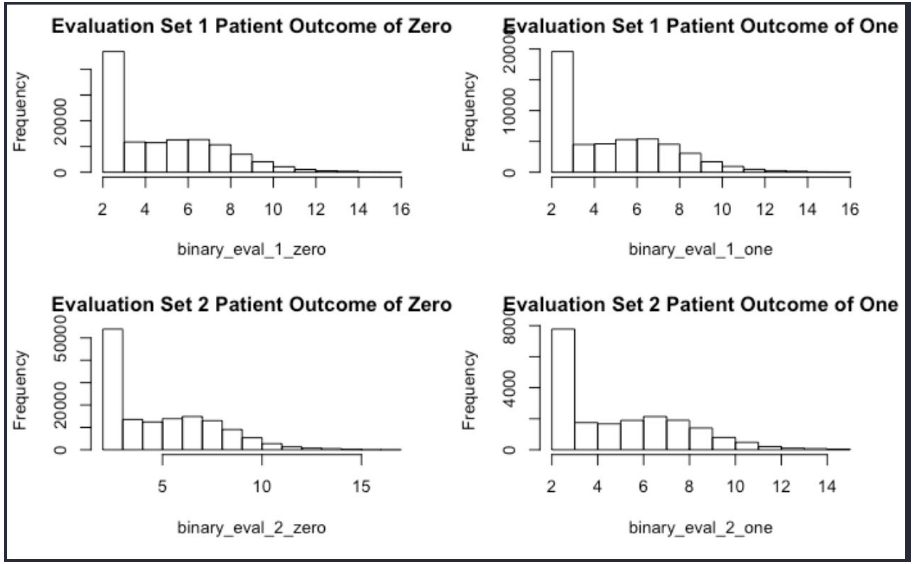
- Evaluate the proportion of each response in each training set
- Interested in looking at the rate of each outcome





Further Investigate the Studies with Variability in Outcome

- What is the distribution of gene expression for each type of outcome
- Focus on studies 7 (eval 1) and 9 (eval 2) because they had most variability in binary outcome
- No clear discernibility based on distribution



Evaluation Set 1 Patient Outcome of Zero

Evaluation Set 1 Patient Outcome of One

Evaluation Set 2 Patient Outcome of Zero

Evaluation Set 2 Patient Outcome of One

Extended Interest in the Correlation between Data

Use **Canonical Correlation** to investigate the highest degree of relatedness between data (note linear relationships)

*use same example sets, study 7 and study 9 *

Association Between 0 Binary Outcome in 1st Dimension:

0.95300842

Association Between 1 Binary Outcome in 1st Dimension:

0.9481966

Association Between Binary Outcomes in 1st Dimension:

0.9412148

High association in the data, and outcomes within each study even more related. Points to feature extraction and dimension reduction.

K-means & TSP classifier

- K-means will help cluster data, allow for feature extraction
- Used TSP classifier for familiarity (krange = 6)

Paired each training and test set based on the number of participants in the study. Tried to match as close as possible. Interested in studying training sets with high variability (1, 10, 7 , 9)

Study 1 Paired with Testing 3

Study 9 Paired with Testing 1

Study 10 Paired with Testing 2

Study 7 Paired with Testing 1

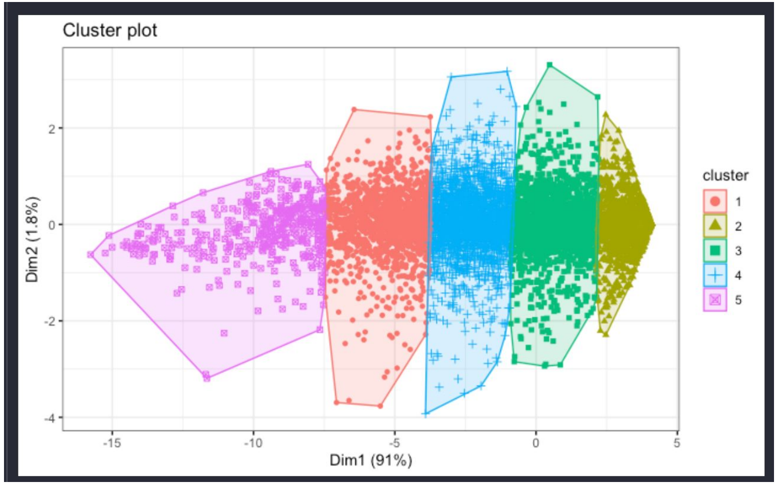
Method: Perform K Means on 4 training sets, 5 clusters. Extract dataset based on the cluster grouping, run TSP classifier for each cluster number, and extract AUC from testing set assigned to that training dataset

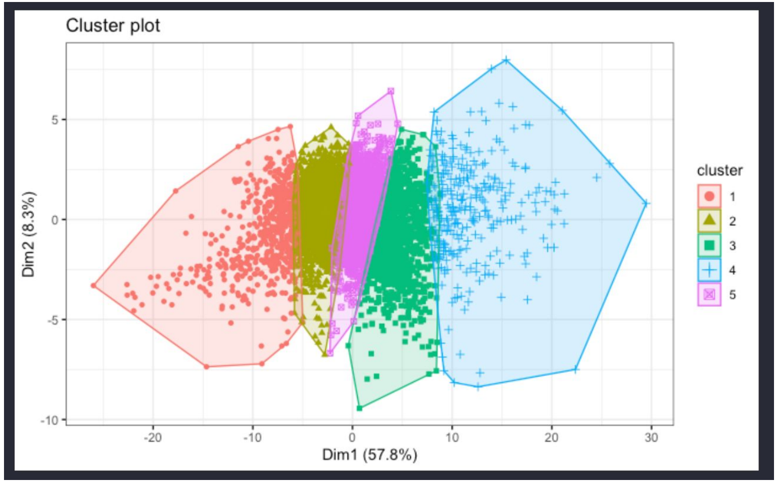
Results from K-Means/TSP

study 1 and study 7 results shown

Study 7 Baseline: 0.5538	Cluster 1 AUC: 0.5115	Cluster 2 AUC: 0.55385	Cluster 3 AUC: 0.756923 **(barely beat the baseline!!)**	Cluster 4 AUC: 0.50385	Cluster 5 AUC: 0.5038
Study 1: Baseline: 0.5	Cluster 1 AUC: 0.714	Cluster 2 AUC: 0.5188	Cluster 3 AUC: 0.7321	Cluster 4 AUC: 0.661	Cluster 5 AUC: 0.5357

Why did Cluster 3 Perform well?





- Cluster 3 SD for Study 7 : **1.05 (second lowest)**
- Cluster 3 SD for Study 1 : **1.027 (second lowest)**
- Study 1 Cluster 3 Size: 1682
- Study 7 Cluster 3 Size: 1780

Discussion

- Is k-means robust enough for feature extraction?
  - PCA/SVD, other clustering algorithms, Manifold Learning
- Did my method help remove study to study issues?
- What might be the results when I try my algorithm on all the interesting training sets I noted?
- Why cluster 3?
- Should I focus on participant size at all?
- Concerns about the test/train split
- Concerns about how correlated our data is