





Predicting the Gestures

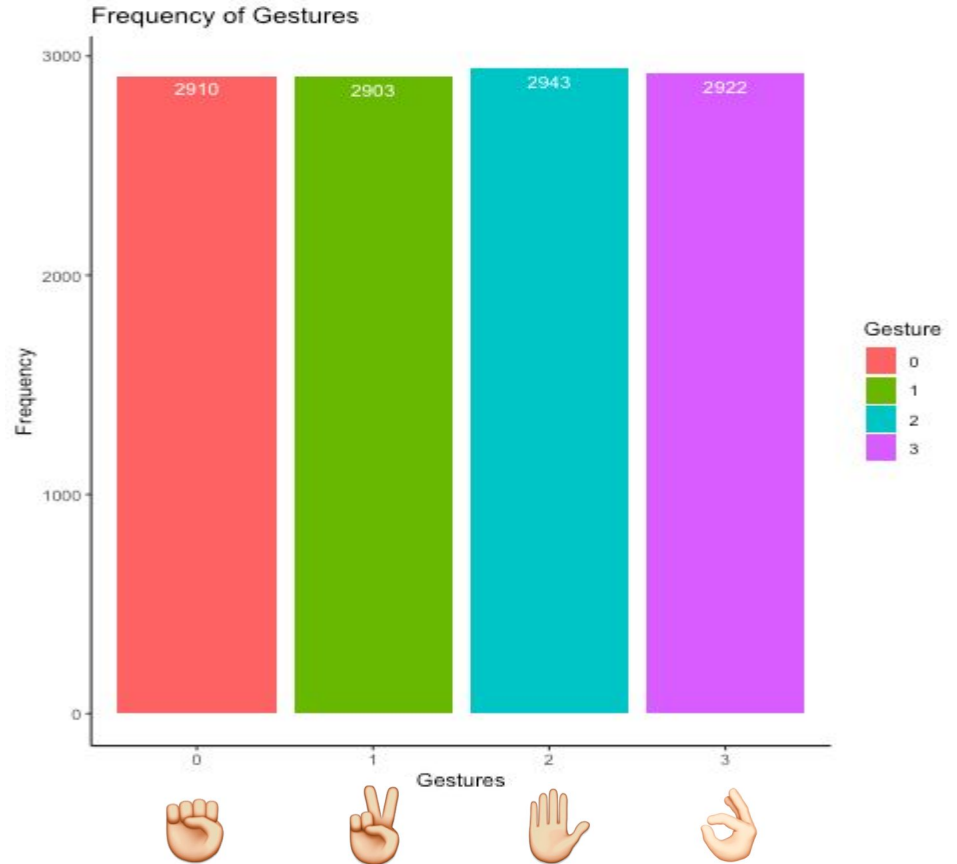
Team 14: Chen Xie, Xun Wang, Xinye
Jiang

Data and Motivation

- Source: <https://www.kaggle.com/kyr7plus/emg-4#0.csv>
- The dataset is about a prosthetic control system recording human hand muscle activity corresponding to four different hand gestures.
 - 0:  1:  2:  3: 
- Dimension: 11678*65
 - 8 consecutive readings of all 8 sensors, so 64 predictors
- Motivation: predict hand gestures depending on data from the sensors

Response Variable

- Balanced Data



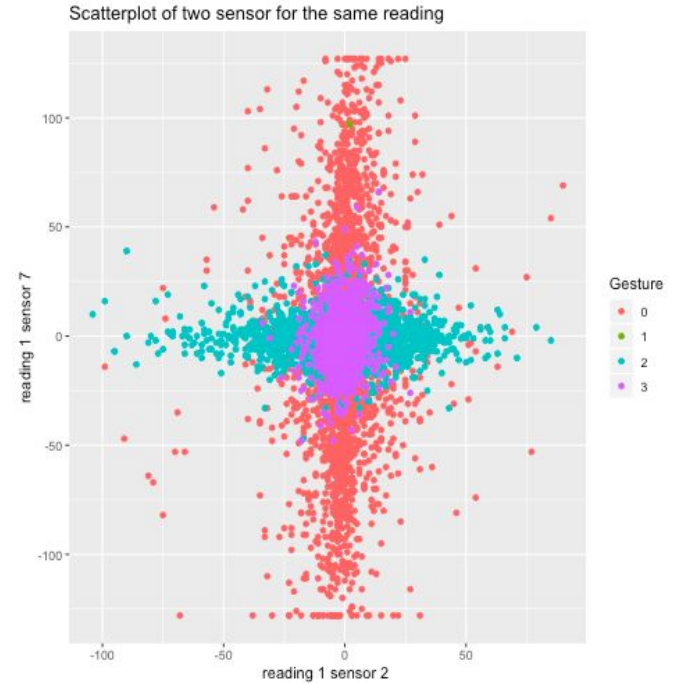
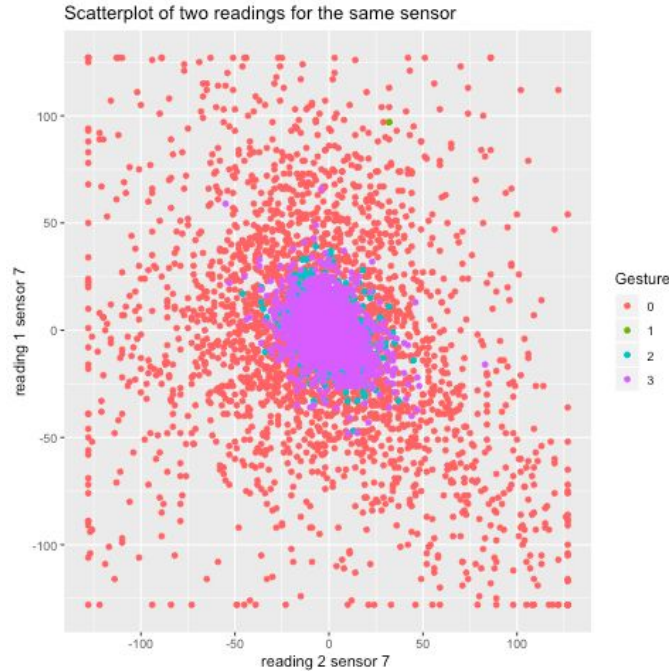
Data Description

Variables	Type	Description
V65	Categorical	Response: gesture classes (rock-0, scissors-1, paper-2, ok-3)
V1-V8	Continuous	Reading 1 Sensor 1-8
V9-V16	Continuous	Reading 2 Sensor 1-8
V17-V24	Continuous	Reading 3 Sensor 1-8
V25-V32	Continuous	Reading 4 Sensor 1-8
V33-V40	Continuous	Reading 5 Sensor 1-8
V41-V48	Continuous	Reading 6 Sensor 1-8
V49-V56	Continuous	Reading 7 Sensor 1-8
V57-V64	Continuous	Reading 8 Sensor 1-8

64 predictors: too many?

- Correlation:
 - Within reading? Within sensor?
- PCA?
 - Structure of variables
- Variable Selection?
 - Guess: Which reading? Which sensor?

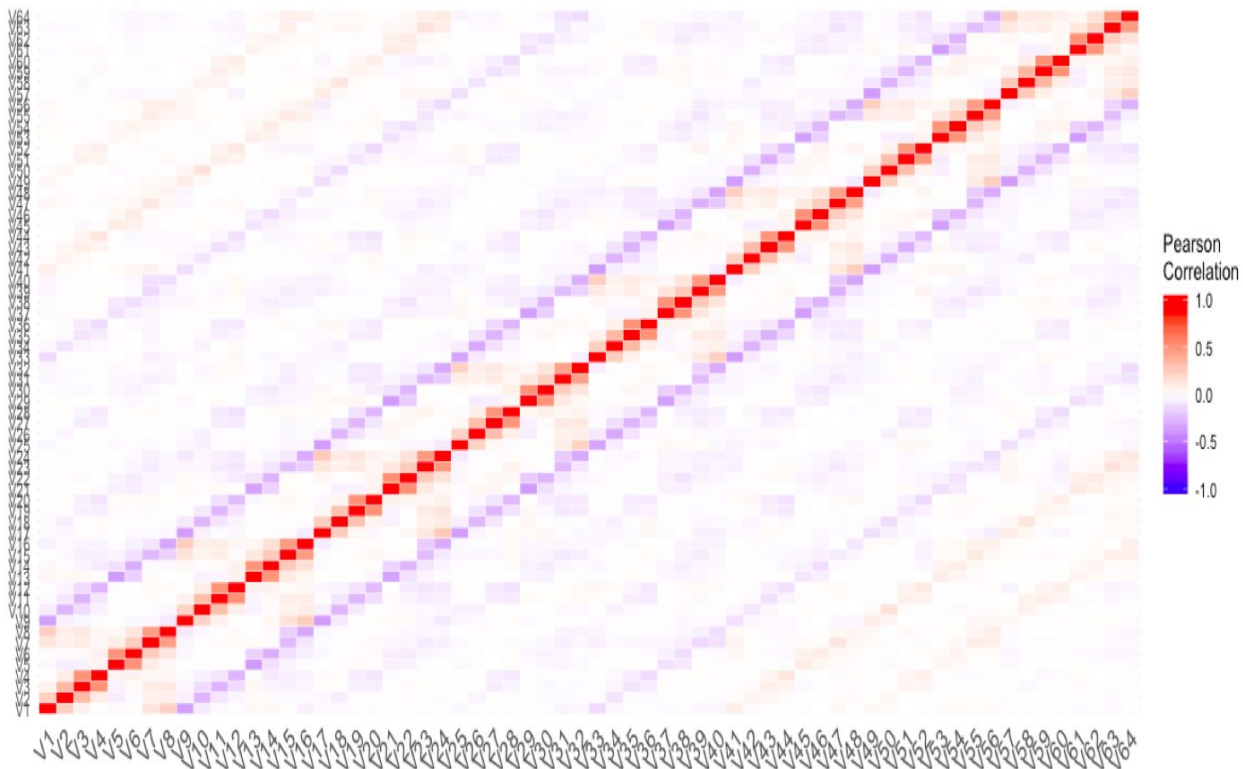
PCA?



- High dimension
- No obvious structure
- Does not work well

Correlation

Correlation (Overview)



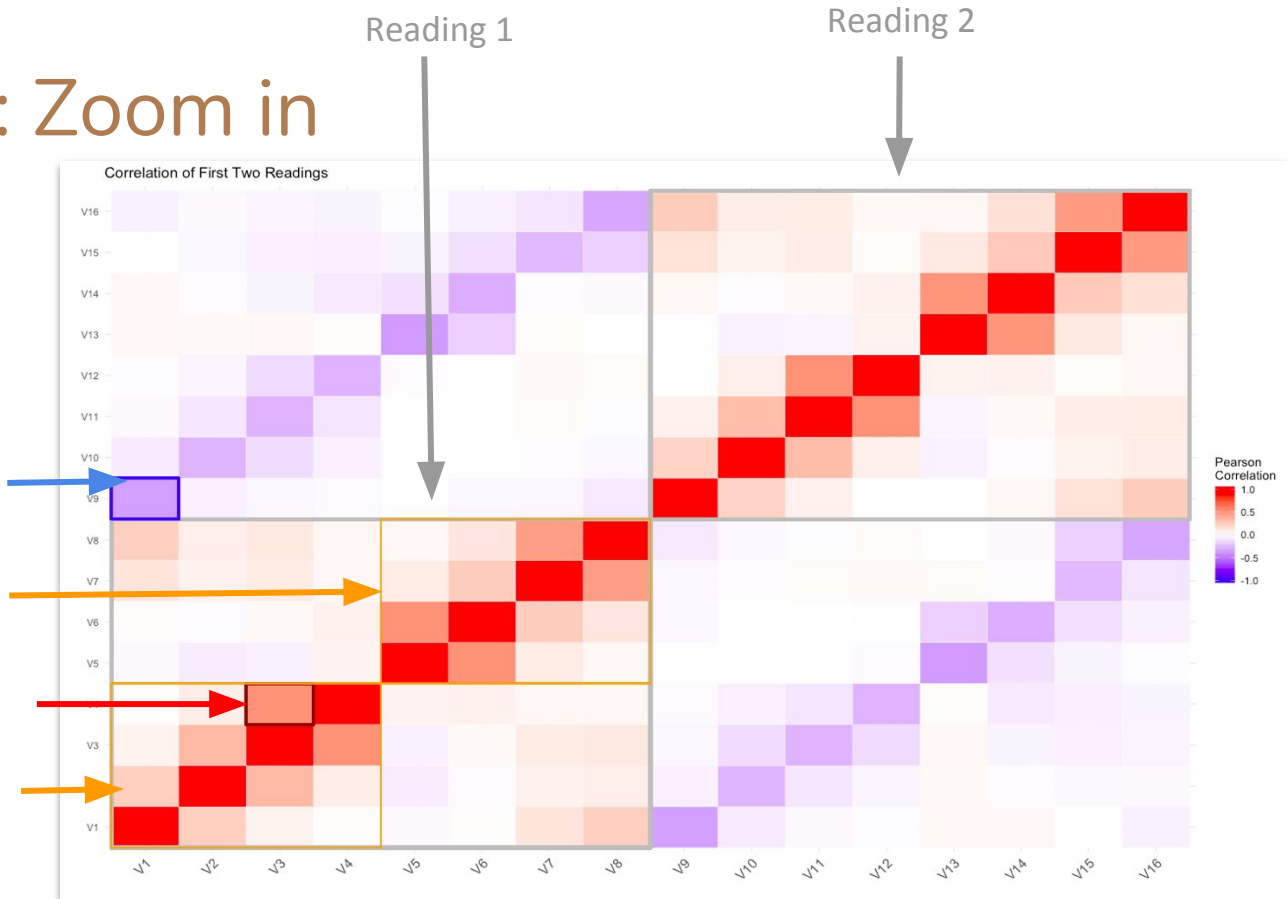
Correlation: Zoom in

Of the adjacent reading
the same sensor

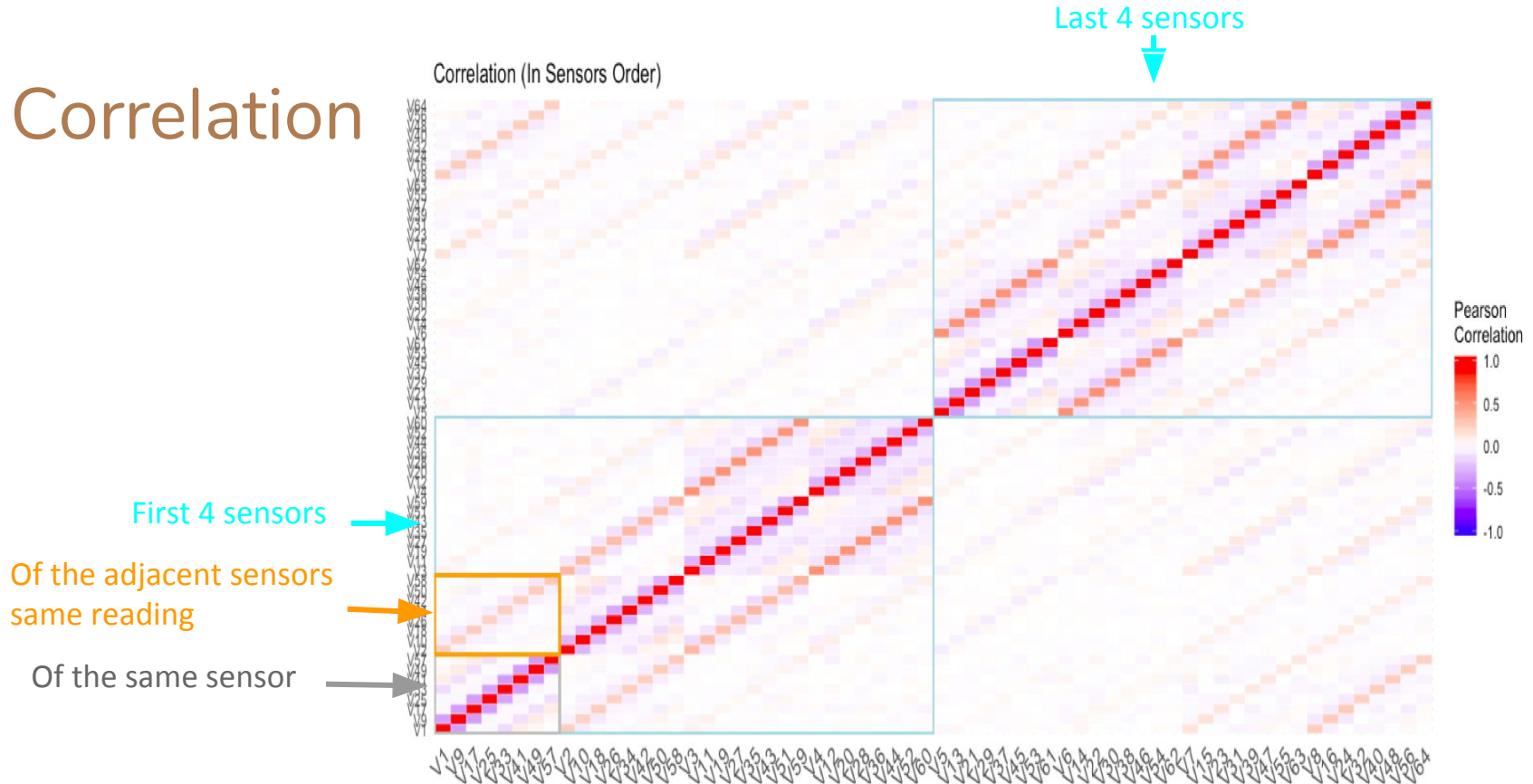
Last 4 sensors

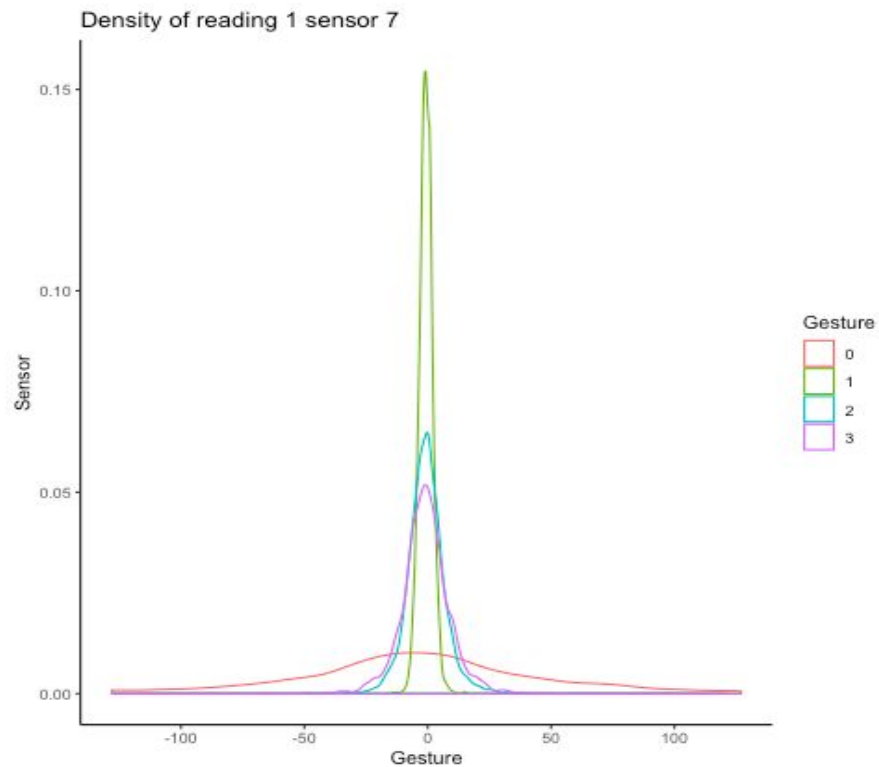
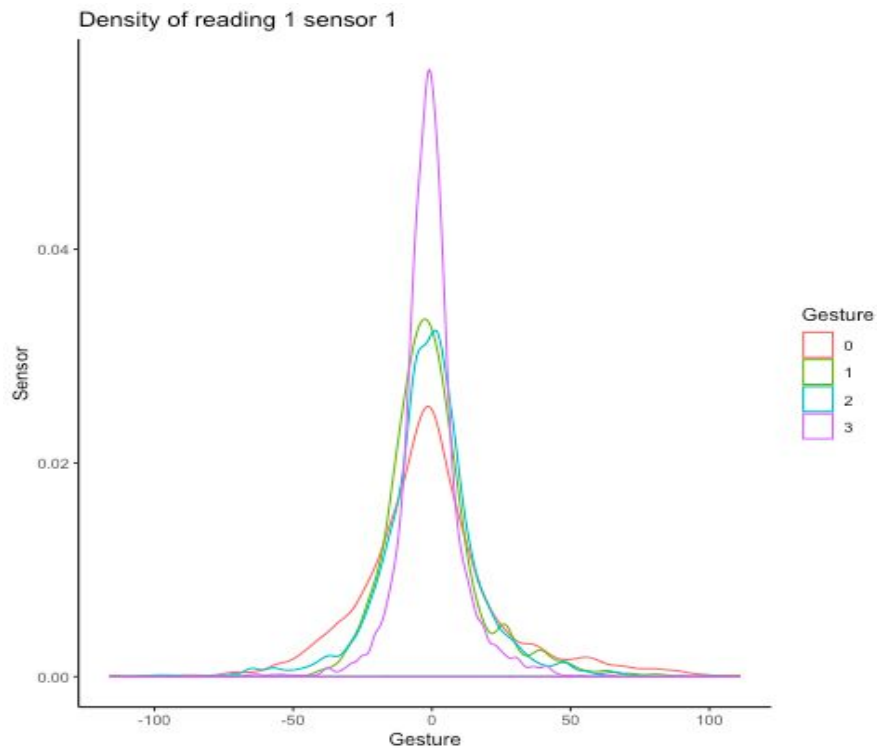
Of the adjacent sensors

First 4 sensors



Correlation



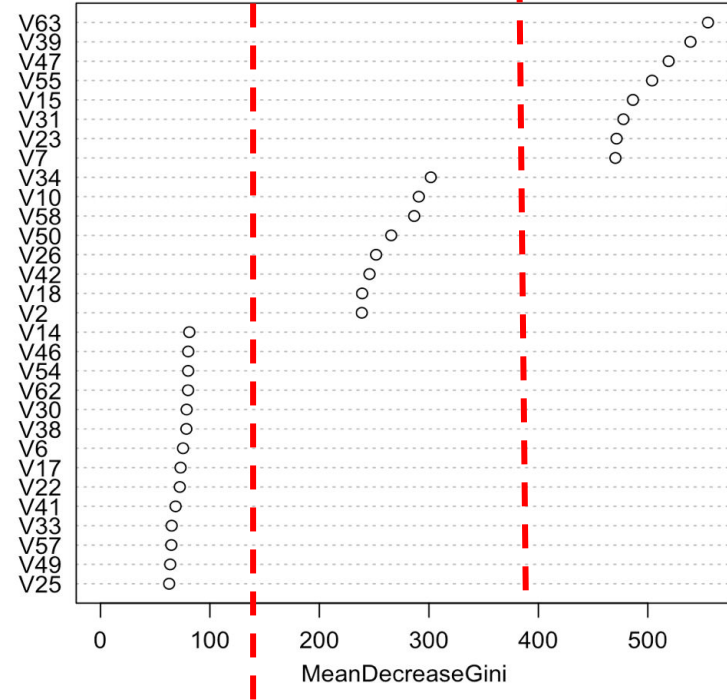
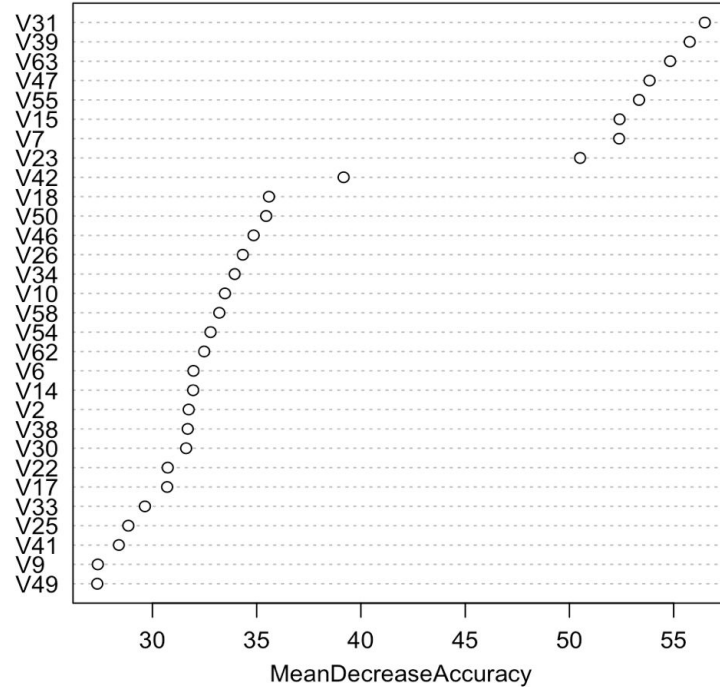


Some sensors are more important: check by preliminary model

Variable Selection

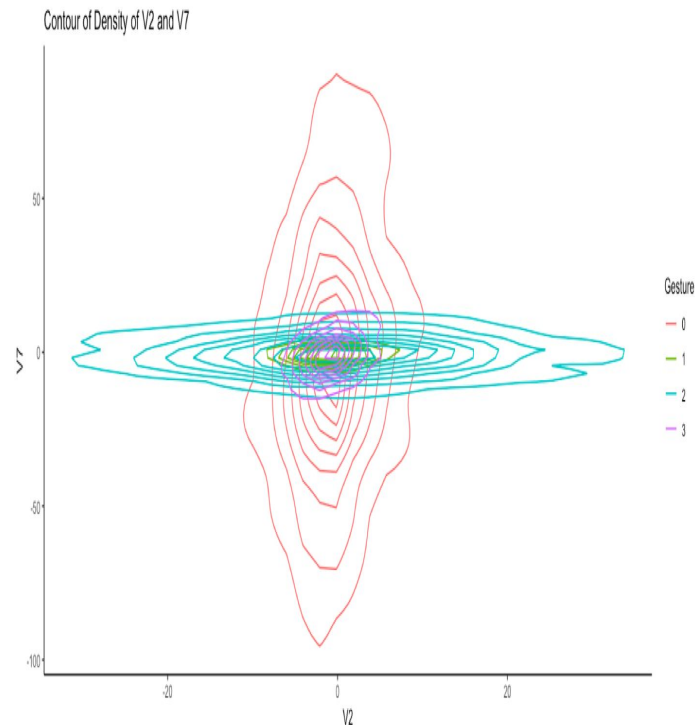
- Use random forest with $mtry=8$ to fit the whole dataset and check the importance of the variables, find that the last but one sensor in each reading is most important and the second sensor in each reading is next important.
- Use stepwise forward selection based on AIC/BIC values to choose the optimal logistic regression models respectively and find the important variables which are basically the same as the ones in random forest.
- Select the variables for the 7th and 2nd sensors, so 16 predictors.

Variable Selection



Model Selection: 16 predictors

- Training set : Testing set = 8 : 2
- Obviously non linear classification boundary
- LDA, Logistic Regression ?
 - Check: very bad performance 🙄
 - LDA: 34.85%
 - Multinomial Logistic Regression: 34.46%
 - Random Guess: 25%

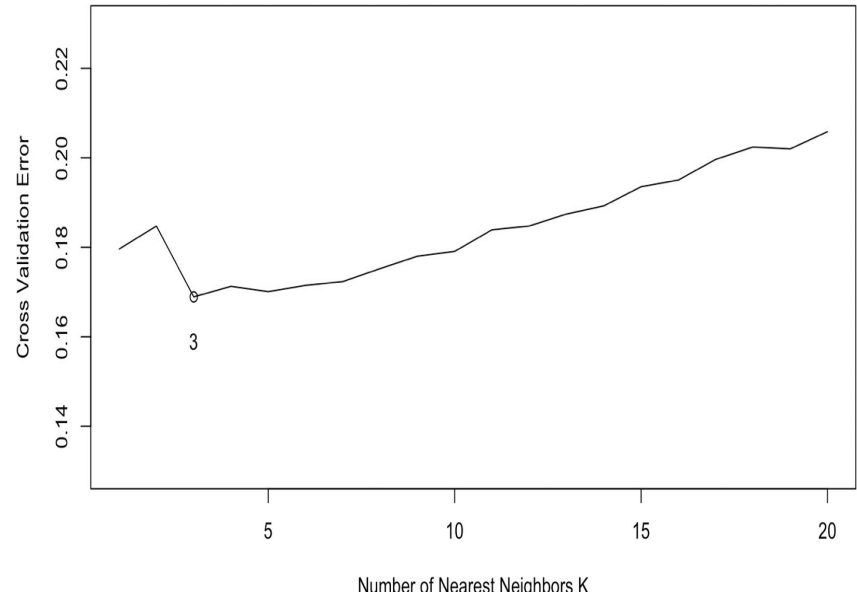


Model Selection: QDA

- QDA
 - 11678 observations: large enough
 - Approximately normally
 - Accuracy = 90.92% !!!

Model Selection: KNN

- KNN
 - High dimension?
 - 5-fold Cross Validation: K=3
 - Accuracy = 84.63%

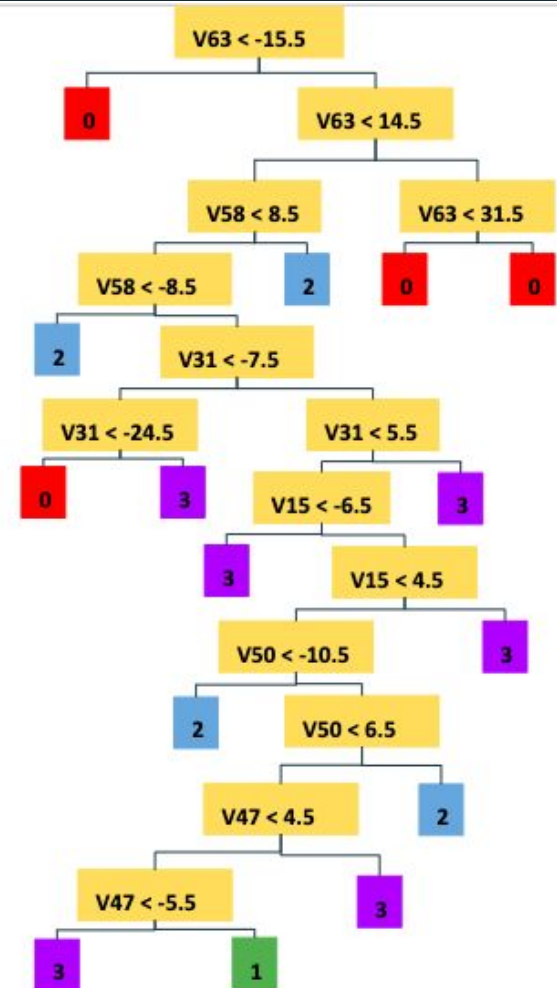


Model Selection: SVM

- SVM
 - radial kernel
 - Cross Validation: computational intensive
 - Instead, Validation Set
 - cost = 1, gamma = 0.5
 - Smallest validation error = 0.1734475
 - Accuracy = 83.39 %

Model Selection: Tree

- Classification Tree
 - Accuracy = 70.93%
 - Improvement: Random Forest



Model Selection: Random Forest

- Random Forest: mtry = 4
- Accuracy = 91.14%

Conclusion & Discussion

- The data here is balanced, but may not have “real structure”, to reduce dimension, use random forest to do variable selection.
- The motion of muscle 2 & 7 are almost decisive for these 4 gestures.
- Non linear boundary: Logistic Regression & LDA have poor performance.
- Random Forest has the best performance.
- Future: more sensors, more gestures, more complicated models

Prediction Accuracy of several methods

Method	Prediction Accuracy
Logistic Regression	0.3446062
LDA	0.3484589
QDA	0.9092466
KNN	0.8463185
SVM (radial)	0.8339041
Classification Tree	0.7093322
Random Forest	0.9113870