Final Paper STAT 602

Joseph McDonald

5/6/2020

The problem at hand is to answer the question of whether it is possible to build a reasonably accurate classifier that predicts three things simultaneously: 1. The writer of a short note (40 writers) 2. Which of 6 phrases is written 3. Whether the line is written in cursive or print using the kinematic features recorded by MovAlyzer. To begin the analysis we will do a brief investigation of structure of the data to look for any straightforward interactions or signals. We will also use principle component analysis to explore representations of higher order interactions or signals in the data. Next we will train several classifiers using LDA, KNN, MclustDA, and Random Forest algorithms. Finally, we will use the findings of these classifiers to select the best approach for building a reasonbly accurate classifier for all three independent variables. It is important to note that because the goal is proving whether a solution to this problem exists, most fine tuning will be left for future endeavors as we will prioritize breadth of analysis over depth.

We begin our analysis by examining the relationship of each dependent variable and each of the three independent variables. It should be noted briefly that the MovAlyzer data is recorded such that each row corresponds to one segment of one letter on a line. For our purposes, however, in the interest of computational efficiency we collapse this data to the unique trial level using mean vectors so that each row corresponds to the unique combinations of group, subject, condition, trial. For the most part, there is little discernable seperation between the classes for any combination of dependent and independent variables. The one exception, shown in **Figure 1**, is the relationship between RelativePendownDuration and Group. This makes sense intuitively as cursive writing entails far fewer lifts of the pen than print, and indicates a clear signal for classifying Group. There are some other variations and possible groupings present in some of the other relationships, such as in **Figure 2**, however, no clear signals or patterns are detectable.

Next we use principle component analysis to look for higher order signals and interactions within the data in reduced dimensions. Because all PCA components are orthogonal there is no correlation between the new covariates derived from PCA. Because of this, adding or removing components has a clear relation proportion of the variance that a model can explain. Due to the high variance within the covariates we build our PCA model using the correlation matrix, which is effectivley the same as scaling prior to conducting PCA. The Plot of the variance explained by the first ten components of our PCA model is shown in **Figure 3**. An examination of some of the plots of the first 10 components, such as **Figure 4**, shows signs of clear signals between cursive and print writing. Further, the Pairs plot for Group in **Figure 5** seems to indicate that component 2 is capturing the difference between print and cursive writing. Attempting similar graphical analysis with the independent variable Condition (6 classes) using one verus the rest methodologies does not yield any meaningful insights. There appears to be some potential clustering but it is not clear. Finally, we do not pursue this method for the Subject variable as there are two many classes to reasonably represent the principle components graphically, staying within the scope of this question.

We now move on to building up classifiers to determine if there is a signal in the data worth pursuing. We begin by constructing a very basic LDA model with 10 principle components to establish a baseline. Using cross-validation to obtain training accuracy, we first build up marginal classifiers to predict each independent variable seperately. We find that this model returns accuracy of 0.879 for Group, 0.385 for Subject, and 0.303 for Condition. The results from Group are not surprising as the high accuracy is supported by the graphical indications of clear signals within the data. Despite the lower cross-validation accuracies for Subject and Condition the fact that both are much higher than the probabilities of random selection (0.025 and 0.16 respectively) suggests that there is some signal resonating from the data. We next build up a single model by factorializing the three independent variables into a single independent variable with 480 classes. This is a naive place to start because we only have 3 replicates per class at this level, however, in attempt to establish a baseline it will suffice. We find that the cross-validated accuracy (0.281) is worse than any of the marginal accuracy measures but is similar to Condition. We will use 0.281 as our baseline accuracy for prediction all three independent variables.

In order to try to circumvent the issue of lack of replicates, we also build up a model which joins Group and Condition and predicts Subject marginally. While this method reduces the number of classes to 12 and 40 respectively, it fails to perform even half as well as the baseline. Additionally, we build up a classifier that combines the the three marginal predictions, but this also fails to perform as well as the baseline. With a baseline established, we next run aforementioned cross-validated models through a loop to determine the optimal number of principal components for each model. This step provides significant improvement to all

models, in particlar the joint class model which reached an accuracy of 0.586 with 21 components. In general, most of the models trained are optimized with about 20 out of 23 principle components explaining roughly 90% of the variance. Moving on from LDA we train similar set of KNN models. Using cross-validation for simple fine tuning we follow similar methodology as with the LDA models to determine the optimal number of principle components to use for each model. The KNN models trained tend to be optimized with fewer components (17-18) than the LDA models were. Within this model group both the Group and Condition marginal predictors outperformed the LDA models but the Subject classifier lagged behind substantially. While KNN models are known for their convenience when working with multiple classes, the high number of classes in the Subject variable combined with the low number of replicates likely resulted in the decreased performance. We will also build up a class of models using MclustDA. Because we have not spent much time with feature selection or dimension reduction these models are computationally expensive. The performance of the marginal predictors was approximately on par with the optimized LDA models, but the joint classification model proved computationally prohibitive within the scope of this question. The marginal models were built using 10-fold cross-validation.

Finally, we attempt to fit a class of Random Forest models for our data using 10-fold cross-validation. This results in substantial improvement across all marginal classifiers, in particular for Condition. This is good news as Condition has been the variable that has proven the most difficult to classify thus far. That being said, the accuracy of the joint classifier, while better than the baseline, is much lower than the LDA accuracy. Since random forests, by their nature, eliminate most correlation in the data, using principle components to train the models is a bit naive. Therefore, we also train a group of random forest classifiers using the mean vectors of the training data. For both sets of models we use cross-validation to select the best value for n.var which dicatates how many variables are randomly sampled for each random tree.

Now that we have several classes of models trained we evaluate their relative performances in the table below in order to select the best method of making joint predictions. Seeing that Random Forests without using principle components produces the best accuracy for each marginal classifier we build up a model that combines the three marginal predictions into one. We will not pursue other combinations of marginal and joint classifiers due their extremely poor performance while establishing our baseline. We find that our new model has a cross-validated accuracy of 0.430. While this is far better than our baseline it is not as accurate as the LDA model built with 21 principle components and a joint independent variable with 480 classes. Therefore, we select the joint LDA model as our best model and estimate that the testing accuracy will be approximatlely 0.586. (Predictions can be made by running the predictions script after changing the value of the csv variable at the top of the script to the path for the file of unlabeled data).

```
##           X       LDA       KNN  MclustDA      RFpc        RF
## 1     Group 0.9243056 0.9368056 0.9236111 0.9354167 0.9458333
## 2   Subject 0.7083333 0.6243056 0.6680556 0.7368056 0.7736111
## 3 Condition 0.4111111 0.4840278 0.4430556 0.5541667 0.5618056
## 4     Joint 0.5861111 0.3437500        NA 0.4277778 0.4895833
```

Through this analysis we have shown that there does appear to be some signal in the data that would allow a the construction of a reasonably accurate classifier for all three independent variables. This is supported both by graphical evidence and the accuracies achieved by our various classes of models. In particular, the LDA model seems to be picking up some signal or interaction between the independent variables deep within the principle components which is driving the joint classification accuracy to outperform the marginal classification accuracy for Condition by such a wide margin. Our analysis suggests that there is a signal on which to predict all three classes. The next best steps would be to fine tune the random forest models and to see if the high marginal accuracies can be leveraged into better joint predictors. Additionally, further dimensionality reduction and working to maximize the model degrees of freedom could open up MclustDA as another viable path for further detailed analysis.

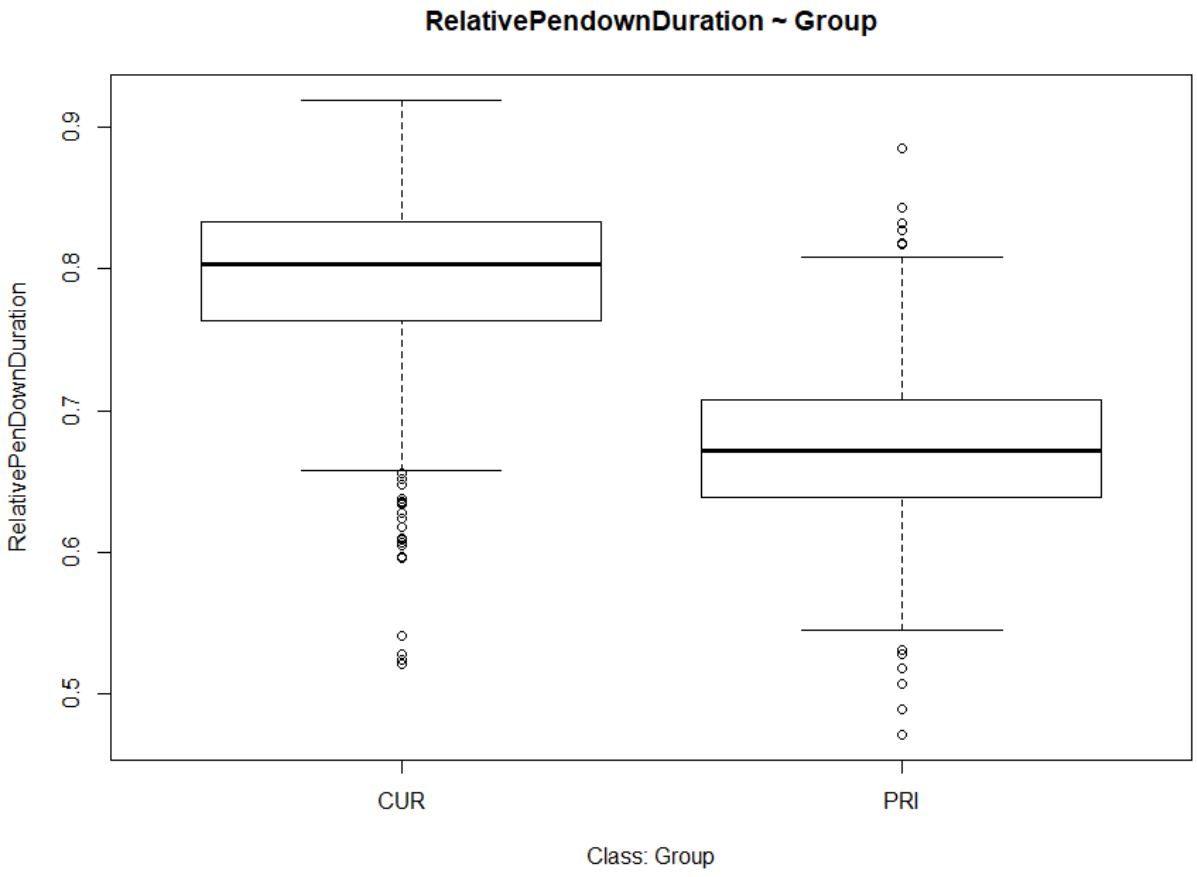

Figure 1: RelativePenDownDuration is the only single variable to show clear seperation within Group

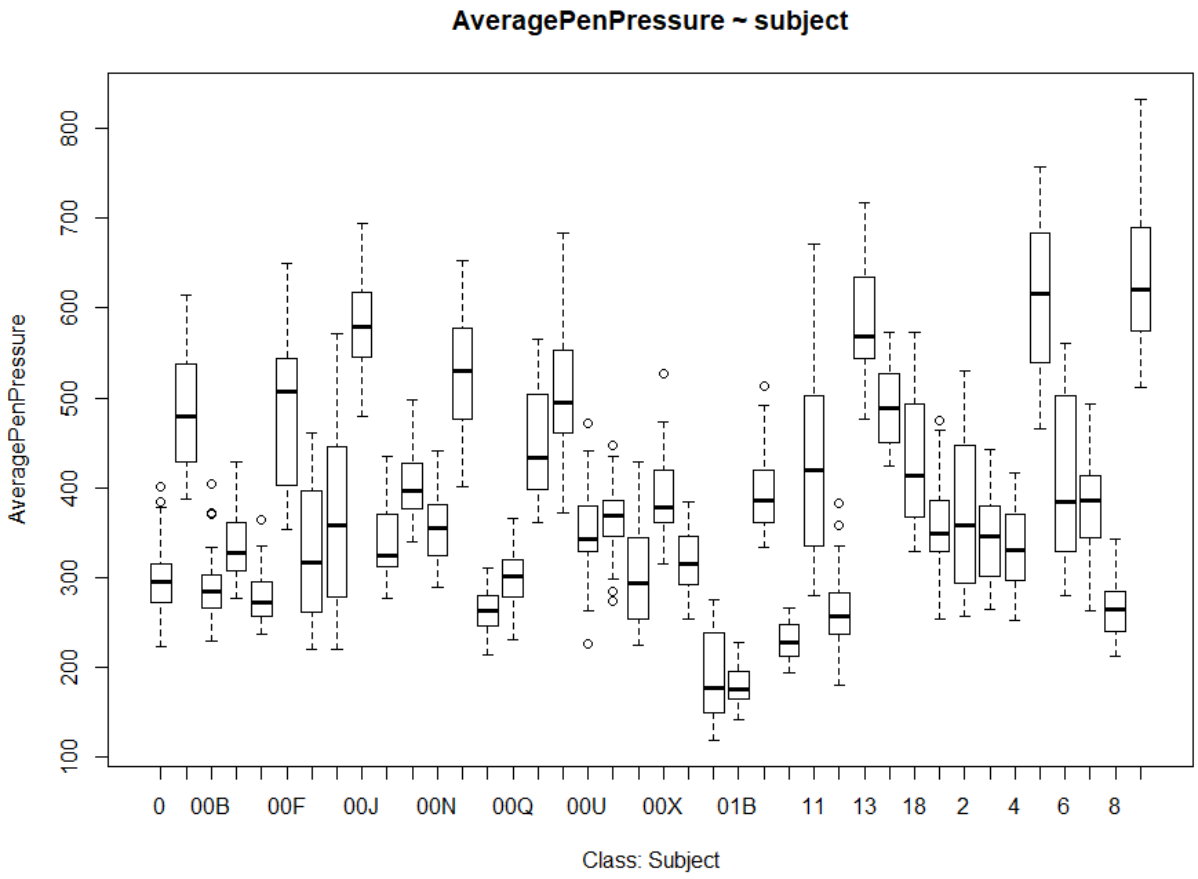

Figure 2: Other variables showed some variations or possible groupings but no discernable signals

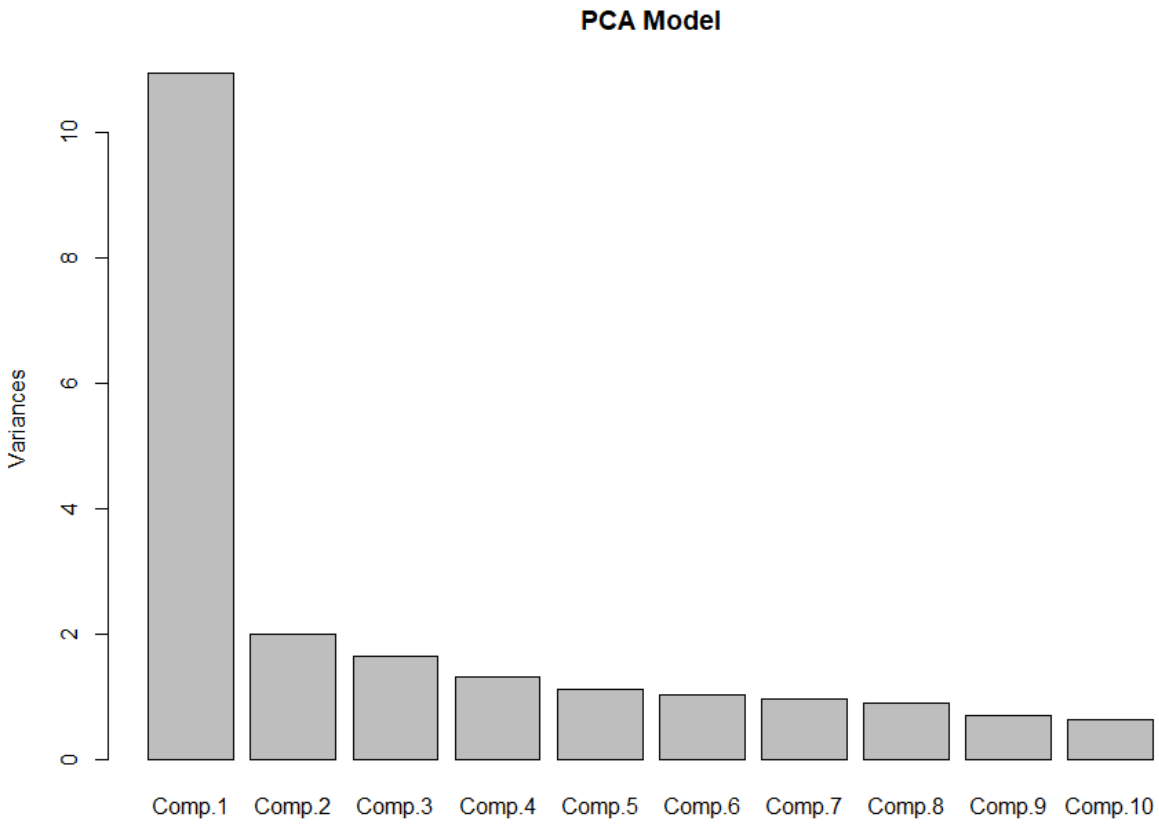

Figure 3: Plot of variances explained by principle components

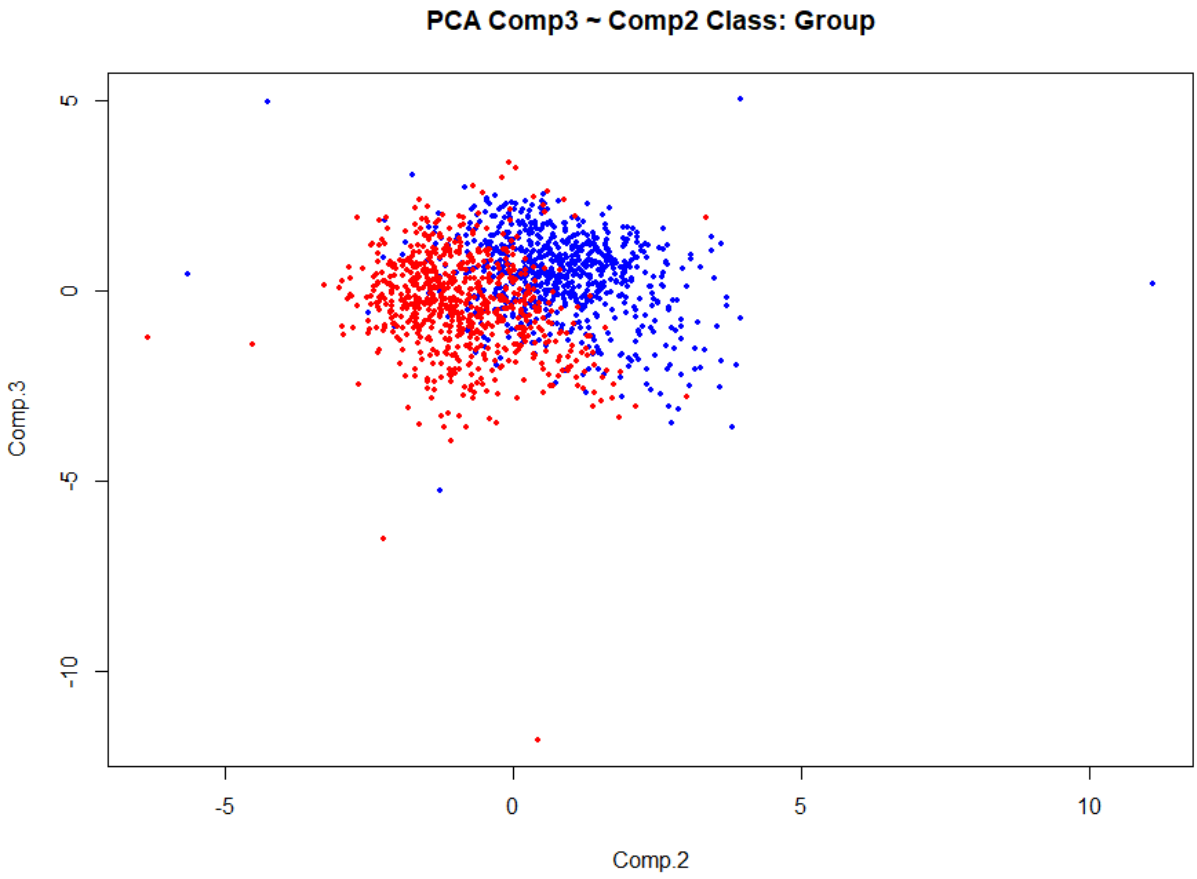

Figure 4: PCA confirms presence of clear signal for Group

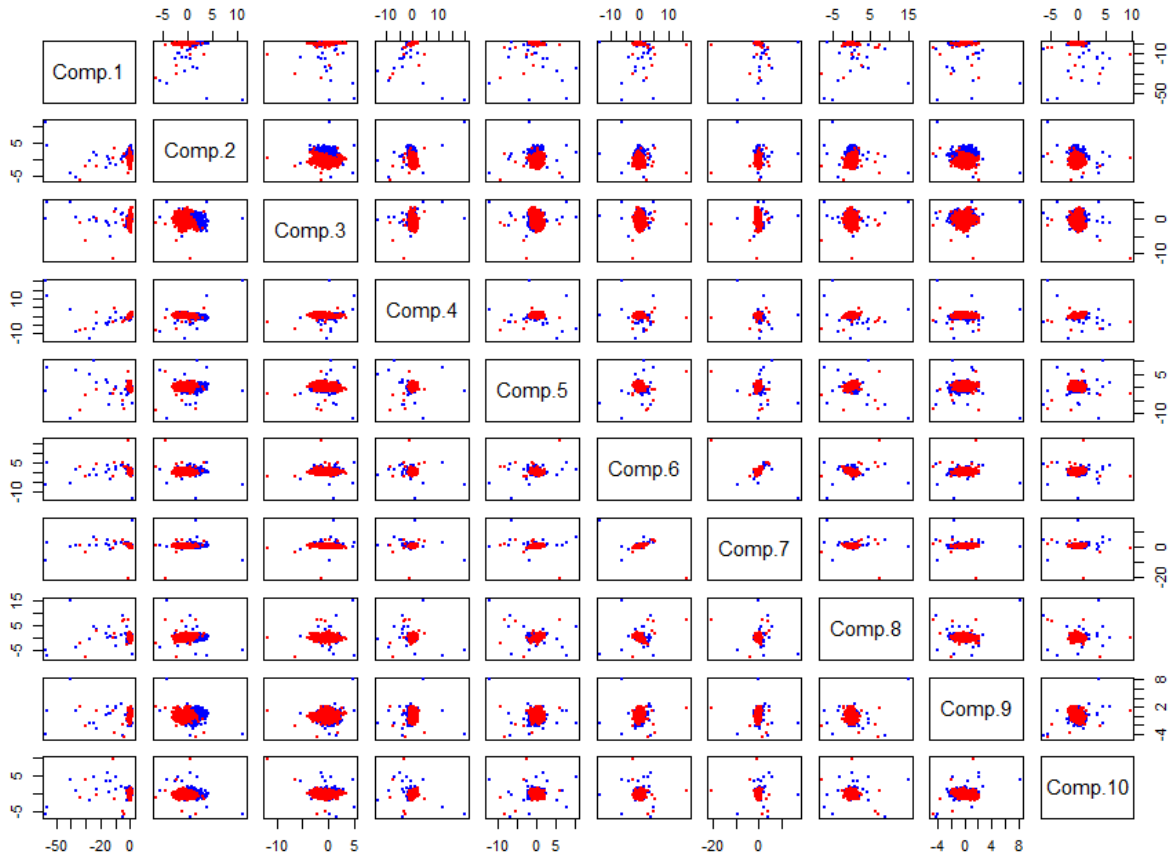

Figure 5: A pairs plot of PCA for Group shows that the second component seems to be capturing the difference between cursive and print writing