hw2

Part 1

Question 1

```r
samples <- sample(1:nrow(conf),
                  nrow(conf)*0.8,
                  replace = FALSE)
train <- conf[samples, ]
test <- conf[-samples, ]
```

Question 2

We can see that the bigger estimator is the Ecludian distance of the inferred ideal point, following by the qualification. The first one is negative meaning that dimminish the probability of voting for that particular candidate. In the case of the qualification is in the other way around. Finally the varuables if the president is string at the moment of voting and the if the share the party afiliation with the president we see that both increase the probability of the senator voting for that particular candidate.

*** Summary of Logit Results ***

```
##
## Call:
## glm(formula = vote ~ EuclDist2 + qual + strngprs + sameprty,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1279   0.1048   0.2201   0.4253   2.0781
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7702     0.2023  -3.808  0.00014 ***
## EuclDist2    -4.3520     0.3070 -14.175  < 2e-16 ***
## qual          3.9360     0.2457  16.020  < 2e-16 ***
## strngprs      1.0987     0.1376   7.984 1.42e-15 ***
## sameprty      1.4452     0.1641   8.808  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2361.8  on 3046  degrees of freedom
## Residual deviance: 1502.8  on 3042  degrees of freedom
## AIC: 1512.8
##
## Number of Fisher Scoring iterations: 6
```

If we estimate the confusion matrix we get that the accuracy in the test sample 91.4% (44+643/(44+12+63+643)) of the case correctly. In other hand, we can see that the precision is 97.33%, impliying that the model correctly assign the labels.

Confusion Matrix

```
## [1] 0.9146982
```

Accuracy in clasification of Logit

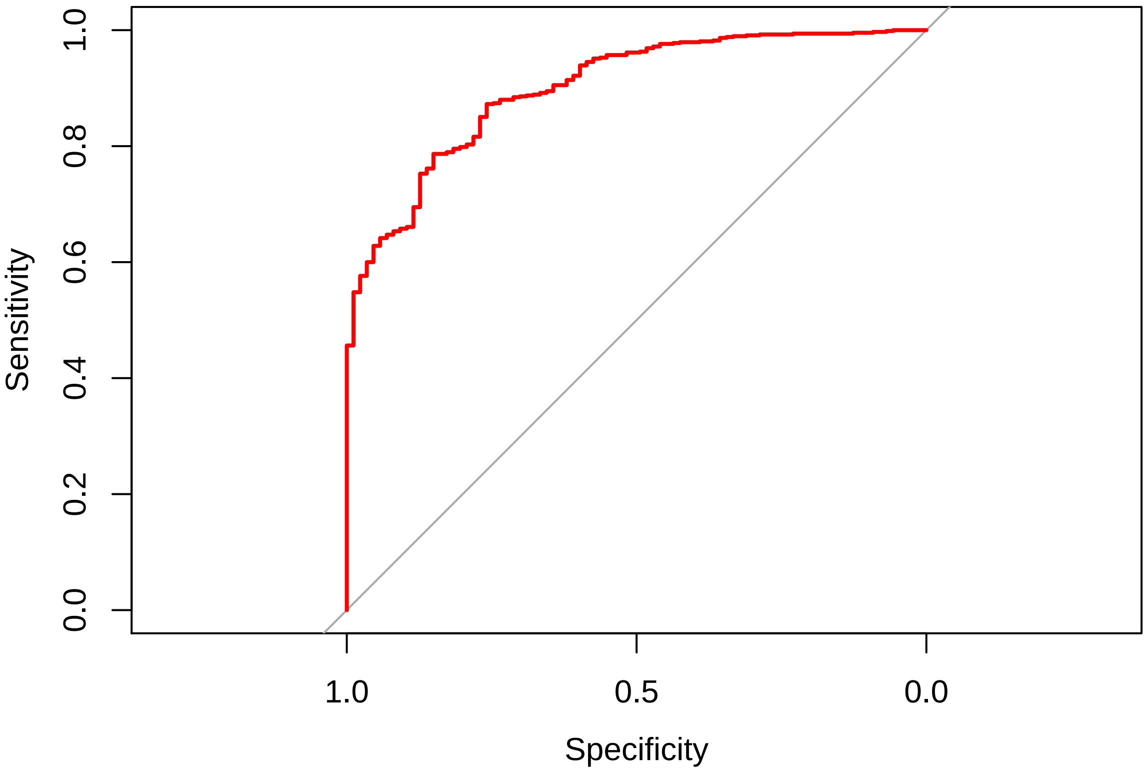
```
##            vote
## logit.pred1   0   1
##           0  40  18
##           1  47 657
```

Finally, in the case of the ROC graph we see that the model assign the labels correctly because the curve moves to the upper right corner of the graph.

*** ROC Plot***

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Question 3

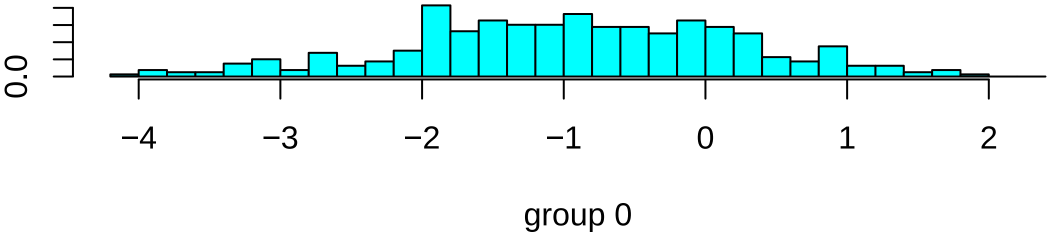
When we fit the LDA model into the data, we find that the results are not different to the parameters estimated to the logit in sign, but the relative intensity between them is different. For example, the qualifications and the euclidian distance to the ideal point in relative terms in the logit are quite similar, instead in the LDA the ecuaclidian distance is much bigger. Nevertheless, the rank as absolute value of the estimator is the same as in the Logit estimation.

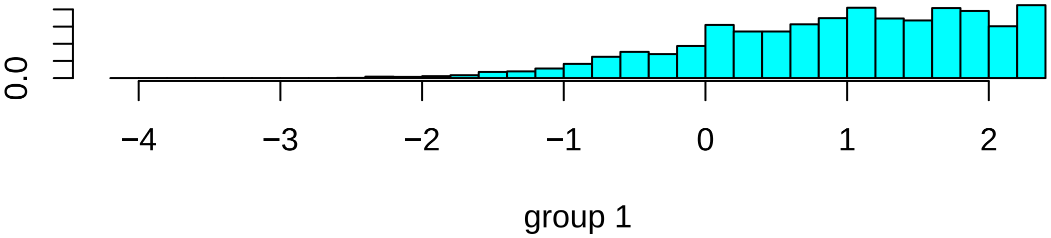
```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Call:
## lda(vote ~ EuclDist2 + qual + strngprs + sameprty, data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.1306203 0.8693797
##
## Group means:
##   EuclDist2      qual  strngprs  sameprty
## 0 0.3954099 0.5708920 0.3291457 0.1708543
## 1 0.1501610 0.8093318 0.6202341 0.6130615
##
## Coefficients of linear discriminants:
##                  LD1
## EuclDist2 -3.3609207
## qual       2.5862138
## strngprs   0.5475428
## sameprty   0.6131494
```

*** Graph group distribution ***

We can see the distribution of the groups and there is some difference between the both groups meaning that the estimator manege to differenciate both groups. The distribution are centered in different values and we observe the over the value of zero there is sustantial change in the histograms.





When we construct the confusion matrix we can see that overall precision and accuracy are 95.07% and 90.68% respectivly. Meaning that the overall assignation of the labels is fairly correct.

Confusion Matrix LDA

```
##   class posterior.0 posterior.1        LD1
## 1     0   0.9604128  0.03958716 -3.4080793
## 2     1   0.1606596  0.83934038 -0.8206684
## 3     1   0.1486640  0.85133598 -0.7716209
## 4     0   0.8951993  0.10480068 -2.8502886
## 5     1   0.0949921  0.90500790 -0.4996209
```

```
##    vote
##       0   1
##   0  47  31
##   1  40 644
```

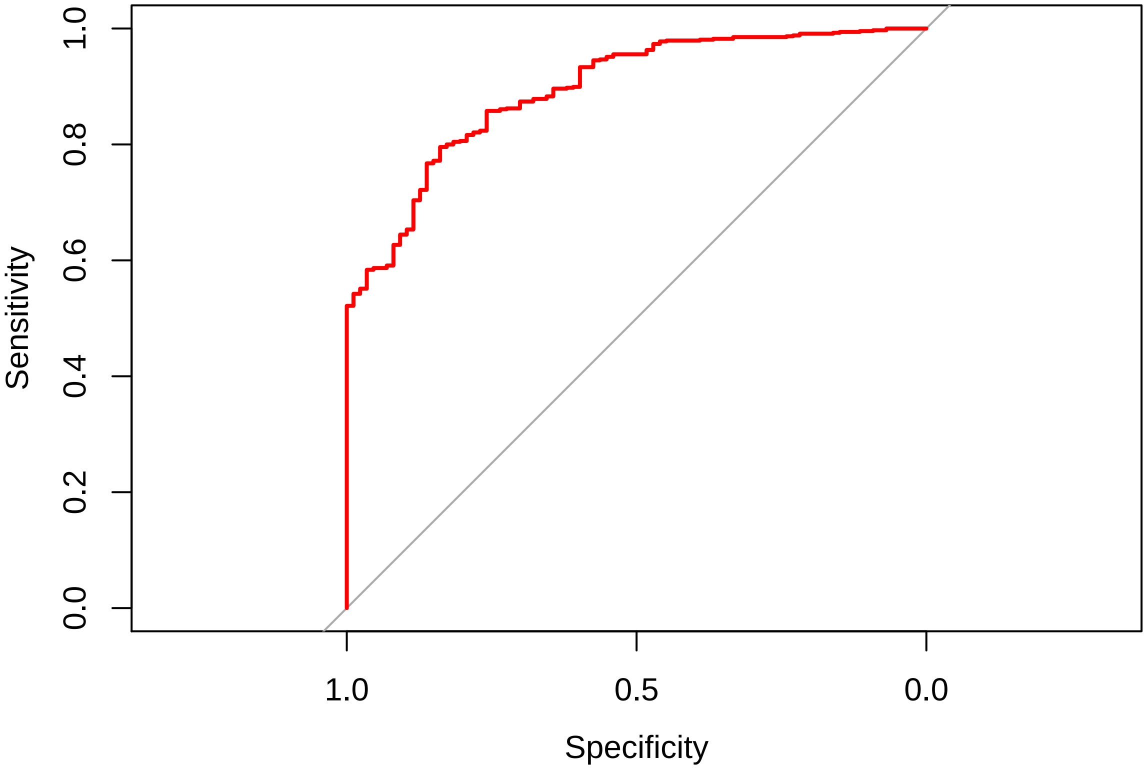
Accuracy in clasification of LDA

```
## [1] 0.9068241
```

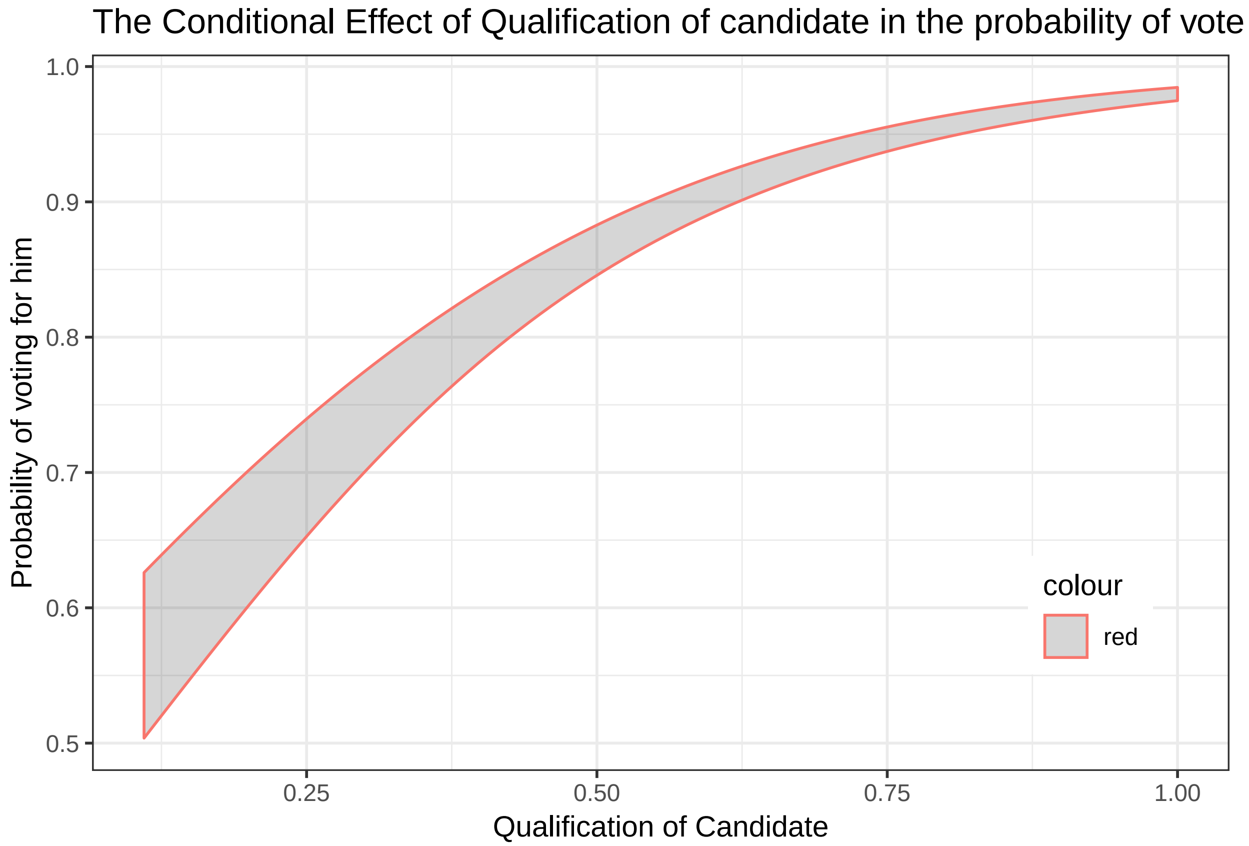
Finally, if we plot the ROC curve for the LDA and we can see that prediction is fearly good, because is in the upper left corner of the graph.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



Question 4



We can see as the qualification increase the probability of vote also increase. Nevertheless, is important to see that even for low levels of qualification the base probability is also high (60%). This implies that probably other factore are relevant to the analysis but a highly qualified judge will be always be voted in favor.

Question 5

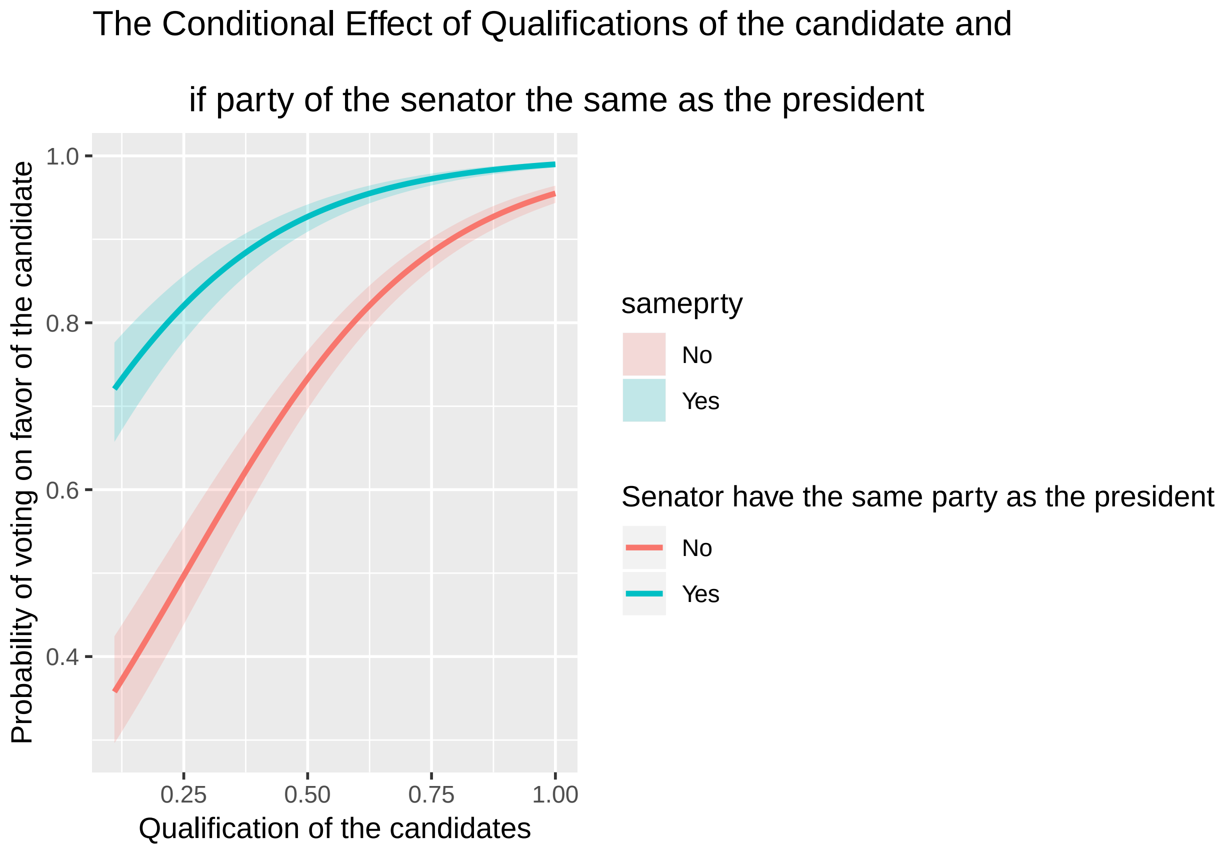
We can see that both models predict fairly the same results, actually we can show that in test sample, both models predict 97.37% of cases with the same label. Also we can see that both roc curves are quite similar in shape and the confusion matrix are behahve quiate similar in accuracy and precision. This imply that both models predict very well in the test set. In terms of which model predicts better depends which indicator we take into consideration to make the decision, if we are interested in an overall prediction is better to think in accuracy(i.e we are interested in the negative and positive cases) in the other hand if we are interesting only in the postive cases we may prefer to use precision. In this case the logit model have a better fit in any metrics that the LDA, have better precision(95% vs 97%) and better accuracy(90% vs 91%).

Average prediction of both model predicting the same for the same observation in the test set

```
## [1] 0.9737533
```

In conclusion, the main difference as we said before is the relative intensitivity between the variables more than the overall sign of the effect of them.

Question 6



Part 2

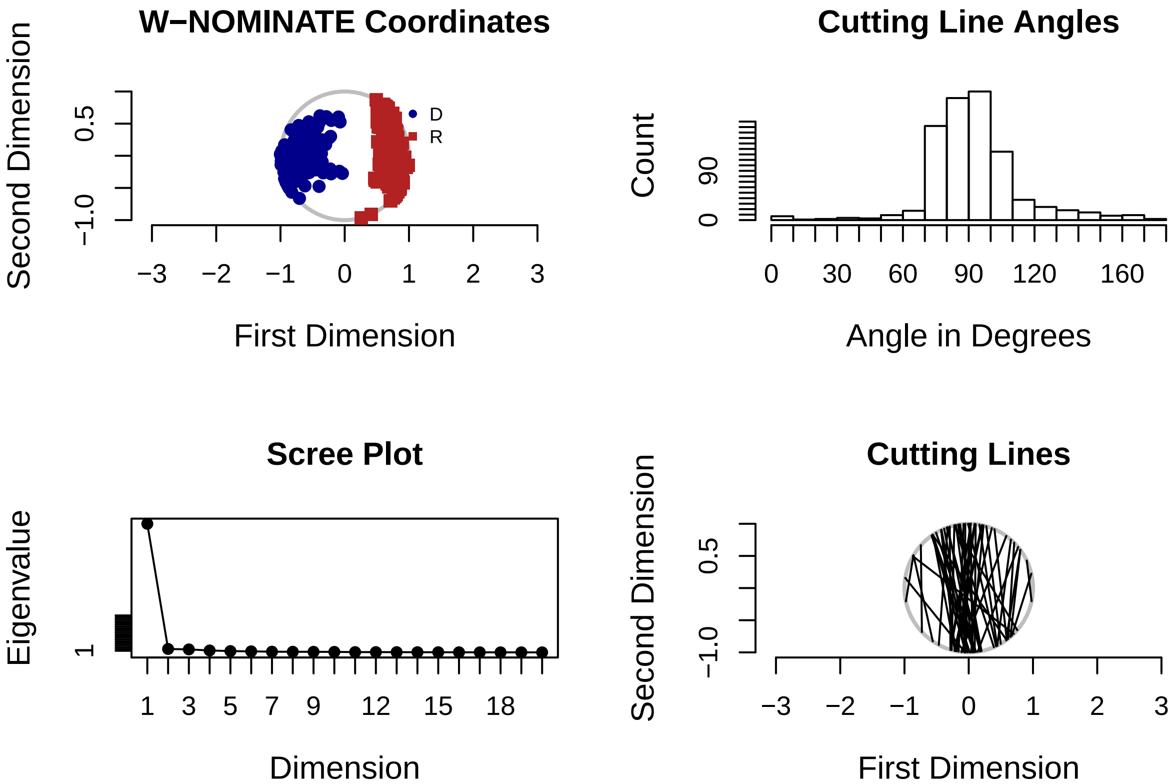
Question 1

```
## Attempting to read file in Keith Poole/Howard Rosenthal (KH) format.
## Attempting to create roll call object
## 113 th_ House _ Roll _ Call _ Data
## 445 legislators and 1202 roll calls
## Frequency counts for vote types:
## rollCallMatrix
##      0      1      6      7      9
##  14576 295753 202943    290  21328

##
## Preparing to run W-NOMINATE...
##
##  Checking data...
##
##      ... 1 of 445 total members dropped.
##
##      Votes dropped:
##      ... 181 of 1202 total votes dropped.
##
```

```
##  Running W-NOMINATE...
##
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##
##
## W-NOMINATE estimation completed successfully.
## W-NOMINATE took 175.08 seconds to execute.
```

```
##
## Preparing to run W-NOMINATE...
##
##  Checking data...
##
##      ... 1 of 445 total members dropped.
##
##      Votes dropped:
##      ... 181 of 1202 total votes dropped.
##
##  Running W-NOMINATE...
##
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##
##
## W-NOMINATE estimation completed successfully.
## W-NOMINATE took 462.195 seconds to execute.
```



```
## NULL
```

Question 2

Discuss the dimensionality of the space. You can present and inspect fit via the aggregate proportion reduction in errors (APRE), the geometric mean prediction (GMP) rate, scree plots, or any other diagnostic tool (visual or numeric) to inspect the overall fit of the algorithm.

For viewing the dimmensionality of the space to fit the data, we can see the eigen values of the fitting of the data.Examining the eigen values we see that the first four are over one. Meaning that maybe we can see that we can explain this data with four dimmension. Nervertheless, we can see that the first value is 69, meaning that most of the variance is explianed by the first dimension, then the second value is 1.97, the third is 1.73 and fourth one is 1.16. Impliying that it's not obvious that the other three dimensions have some explinatory power that is relevant even if the eiganvalue is over one.

Eigenvalues

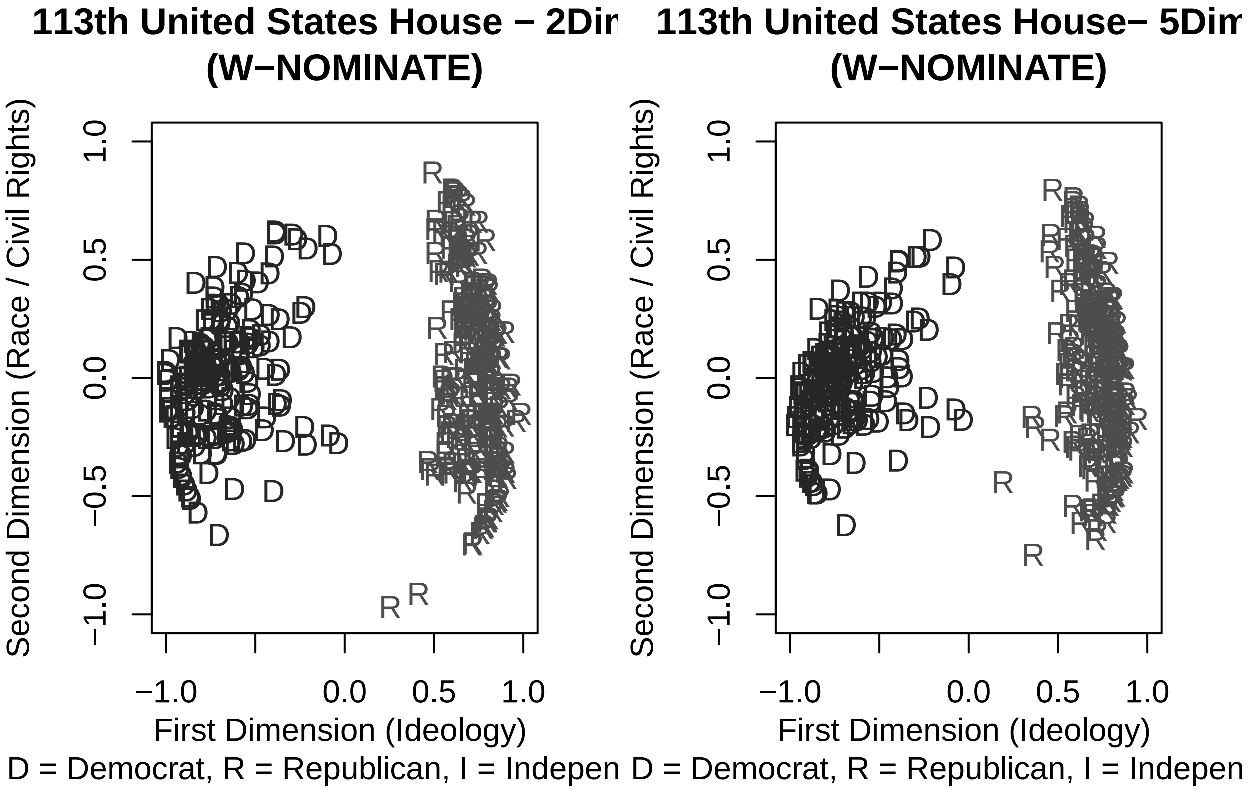
```
## [1] 69.2931290  1.9777323  1.7397527  1.1627728  0.8384978  0.6917516
```

For countinuing the analysis we are going to analyze the APRE and GMP rates, particullarilly we are going to analyze the difference between the difference dimentions that are calculated for doing this we are going to fit the algorithm for 2 and 5 dimensions.

The results show us that as we incorporate new dimmensions the delta in improvement in the APRE and GMP decrease showing that as new dimension is incoported less variance is explained. This correlate with what we see in the eigenvalues where the first one is very big and the following three are very close to one. Impliying that the improvement in the overall prediction for incorporating additional dimensions are not big enough to maybe incorporate them.

Also if we plot the results in the first two dimension for both fitted W-Nominate estimator. We see that five dimension plot is more compress in the one of five dimension.

Graph Five dimension vs two dimension



Results 2-Dims

```
##
##
## SUMMARY OF W-NOMINATE OBJECT
## ----------------------------
##
## Number of Legislators:     444 (1 legislators deleted)
## Number of Votes:   1021 (181 votes deleted)
## Number of Dimensions:      2
## Predicted Yeas:        212927 of 225718 (94.3%) predictions correct
## Predicted Nays:        185010 of 199413 (92.8%) predictions correct
## Correct Classifiction:     92.79% 93.6%
## APRE:              0.817 0.837
## GMP:           0.84 0.857
##
##
## The first 10 legislator estimates are:
```

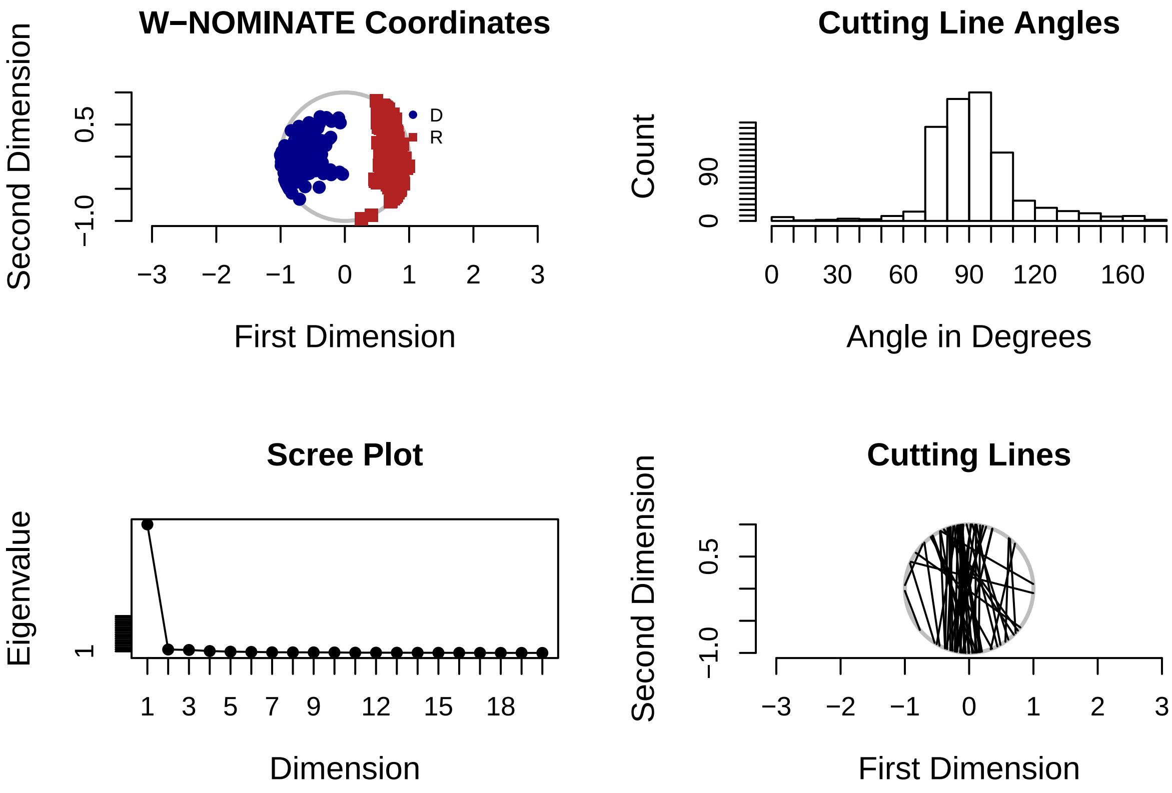
```
##                   coord1D coord2D
## OBAMA (D USA)      -0.936   0.171
## BONNER (R AL-1)     0.642   0.556
## BYRNE (R AL-1)      0.811   0.205
## ROBY (R AL-2)       0.636   0.772
## ROGERS (R AL-3)     0.724   0.393
## ADERHOLT (R AL-4)   0.678   0.735
## BROOKS (R AL-5)     0.792  -0.007
## BACHUS (R AL-6)     0.632   0.541
## SEWELL (D AL-7)    -0.560   0.024
## YOUNG (R AK-1)      0.565  -0.311
```

Results 5-Dims

```
##
##
## SUMMARY OF W-NOMINATE OBJECT
## ----------------------------
##
## Number of Legislators:     444 (1 legislators deleted)
## Number of Votes:   1021 (181 votes deleted)
## Number of Dimensions:      5
## Predicted Yeas:        214794 of 225718 (95.2%) predictions correct
## Predicted Nays:        187377 of 199413 (94%) predictions correct
## Correct Classifiction:     92.8% 93.46% 94.16% 94.33% 94.6%
## APRE:              0.817 0.834 0.851 0.856 0.863
## GMP:           0.84 0.857 0.867 0.871 0.875
##
##
## The first 10 legislator estimates are:
```

```
##                   coord1D coord2D coord3D coord4D coord5D
## OBAMA (D USA)      -0.904   0.056   0.061  -0.397  -0.135
## BONNER (R AL-1)     0.653   0.583  -0.332  -0.092   0.339
## BYRNE (R AL-1)      0.813   0.171   0.164   0.070   0.118
## ROBY (R AL-2)        0.622   0.706  -0.184  -0.127   0.257
## ROGERS (R AL-3)      0.716   0.392  -0.082   0.143   0.553
## ADERHOLT (R AL-4)    0.650   0.660  -0.192  -0.100   0.308
## BROOKS (R AL-5)      0.796  -0.073   0.448   0.005   0.290
## BACHUS (R AL-6)      0.635   0.508  -0.123  -0.190   0.212
## SEWELL (D AL-7)     -0.565   0.155  -0.452  -0.062   0.411
## YOUNG (R AK-1)       0.535  -0.160  -0.756  -0.301   0.165
```

Plots 5-D



```
## NULL
```

Question 3

The major problem of this methodology if we compare to a classification technique is that in the W-Nominate techniques we assume an specific utility function, instead in a clasification technique we assume normally a linear form that can be flexible enough to incorporate another non-linearity. This have especific consequences in how we treat extreme values, meaning that as one of the assumptions of W-N models is how we construct distances relative to a pivot point that create the extreme value that we choose.

Another issue, is that this technique have is that we don't know which dimensions explain the result what we are seeing in the different plots where we can build. Instead in clasification thecniques we already know which dimensions are, but at the same time we don't know if we are missing dimensions.