PCA EFA: Parkinson’s

Jose Tamez

2023-01-27

Table of Contents

# SALIVA-Parkinson’s PCA-EFA-GDSTM

### Loading the libraries

library("FRESA.CAD")  
library(psych)  
  
library(readxl)  
op <- par(no.readonly = TRUE)  
pander::panderOptions('digits', 3)  
pander::panderOptions('table.split.table', 400)  
pander::panderOptions('keep.trailing.zeros',TRUE)

## The Data

pd\_speech\_features <- as.data.frame(read\_excel("~/GitHub/FCA/Data/pd\_speech\_features.xlsx",sheet = "pd\_speech\_features", range = "A2:ACB758"))

### The Average of the Three Repetitions

Each subject had three repeated observations. Here I’ll use the average of the three experiments per subject.

rep1Parkison <- subset(pd\_speech\_features,RID==1)  
rownames(rep1Parkison) <- rep1Parkison$id  
rep1Parkison$id <- NULL  
rep1Parkison$RID <- NULL  
rep1Parkison[,1:ncol(rep1Parkison)] <- sapply(rep1Parkison,as.numeric)  
  
rep2Parkison <- subset(pd\_speech\_features,RID==2)  
rownames(rep2Parkison) <- rep2Parkison$id  
rep2Parkison$id <- NULL  
rep2Parkison$RID <- NULL  
rep2Parkison[,1:ncol(rep2Parkison)] <- sapply(rep2Parkison,as.numeric)  
  
rep3Parkison <- subset(pd\_speech\_features,RID==3)  
rownames(rep3Parkison) <- rep3Parkison$id  
rep3Parkison$id <- NULL  
rep3Parkison$RID <- NULL  
rep3Parkison[,1:ncol(rep3Parkison)] <- sapply(rep3Parkison,as.numeric)  
  
whof <- !(colnames(rep1Parkison) %in% c("gender","class"));  
avgParkison <- rep1Parkison;  
avgParkison[,whof] <- (rep1Parkison[,whof] + rep2Parkison[,whof] + rep3Parkison[,whof])/3  
## I apply the log transform to the data  
##avgParkison[,whof] <- signedlog(avgParkison[,whof])  
#avgParkison[,whof] <- FRESAScale(avgParkison[,whof],method="OrderLogit")$scaledData  
#pander::pander(table(avgParkison$class))

#### Standarize the names for the reporting

dataframe <- avgParkison  
outcome <- "class"  
  
trainFraction <- 0.5  
rhoThreshold <- 0.5  
TopVariables <- 5  
  
trainSample <- sample(nrow(dataframe),nrow(dataframe)\*trainFraction)  
  
trainDataFrame <- dataframe[trainSample,]  
testDataFrame <- dataframe[-trainSample,]

pander::pander(c(rows=nrow(dataframe),col=ncol(dataframe)-1))

| rows | col |
| --- | --- |
| 252 | 753 |

pander::pander(table(dataframe[,outcome]))

| 0 | 1 |
| --- | --- |
| 64 | 188 |

varlist <- colnames(dataframe)  
varlist <- varlist[varlist != outcome]  
varlist <- as.data.frame(cbind(name=varlist,desc=varlist))

## Univariate

univariate\_columns <- c("caseMean","caseStd","controlMean","controlStd","controlKSP","ROCAUC","wilcox.Zvalue")  
univar <- uniRankVar(varlist,  
 paste(outcome,"~1"),  
 outcome,  
 testDataFrame,  
 rankingTest = "AUC")

100 : std\_MFCC\_1st\_coef 200 : app\_entropy\_shannon\_10\_coef 300 : app\_LT\_entropy\_log\_9\_coef 400 : tqwt\_entropy\_log\_dec\_7 500 : tqwt\_TKEO\_std\_dec\_35  
600 : tqwt\_stdValue\_dec\_27 700 : tqwt\_skewnessValue\_dec\_19

## Decorrelation with GDSTM Blind

DEdataframe <- GDSTMDecorrelation(trainDataFrame,thr=rhoThreshold)  
varlistDe <- colnames(DEdataframe)[colnames(DEdataframe) != outcome];  
varlistDe <- as.data.frame(cbind(name=varlistDe,desc=varlistDe))  
  
pander::pander(head(getLatentCoefficients(DEdataframe)))

* **La\_PPE**:

| * PPE | * Ea | * Ed\_5\_coef | * det\_entropy\_shannon\_10\_coef |
| --- | --- | --- | --- |
| * 1 | * -0.957 | * 0.545 | * 2.75e-08 |

* **La\_DFA**:

| * DFA | * tqwt\_energy\_dec\_16 |
| --- | --- |
| * 1 | * 0.618 |

* **La\_RPDE**:

| * RPDE | * apq3Shimmer | * app\_LT\_entropy\_log\_8\_coef |
| --- | --- | --- |
| * 1 | * -3.86 | * 0.0344 |

* **La\_numPulses**:

| * numPulses | * stdDevPeriodPulses | * app\_entropy\_shannon\_4\_coef | * app\_entropy\_shannon\_8\_coef | * app\_LT\_entropy\_log\_8\_coef |
| --- | --- | --- | --- | --- |
| * 1 | * 42563 | * 4.31e-06 | * -3.94e-07 | * -53.2 |

* **La\_numPeriodsPulses**:

| * numPulses | * numPeriodsPulses | * det\_entropy\_shannon\_10\_coef |
| --- | --- | --- |
| * -1 | * 1 | * -1.26e-07 |

* **La\_meanPeriodPulses**:

| * meanPeriodPulses | * stdDevPeriodPulses | * app\_LT\_entropy\_log\_8\_coef |
| --- | --- | --- |
| * 1 | * -0.834 | * 0.000977 |

predTestDe <- predictDecorrelate(DEdataframe,testDataFrame)  
  
univarDe <- uniRankVar(varlistDe,  
 paste(outcome,"~1"),  
 outcome,  
 predTestDe,  
 rankingTest = "AUC")

100 : La\_std\_MFCC\_1st\_coef 200 : La\_app\_entropy\_shannon\_10\_coef 300 : La\_app\_LT\_entropy\_log\_9\_coef 400 : La\_tqwt\_entropy\_log\_dec\_7 500 : La\_tqwt\_TKEO\_std\_dec\_35  
600 : La\_tqwt\_stdValue\_dec\_27 700 : tqwt\_skewnessValue\_dec\_19

## Decorrelation with GDSTM Driven

DriDEdataframe <- GDSTMDecorrelation(trainDataFrame,Outcome=outcome,thr=rhoThreshold)  
varlistDe <- colnames(DriDEdataframe)[colnames(DriDEdataframe) != outcome];  
varlistDe <- as.data.frame(cbind(name=varlistDe,desc=varlistDe))  
  
pander::pander(head(getLatentCoefficients(DriDEdataframe)))

* **La\_PPE**:

| * PPE | * Ed\_4\_coef | * Ed\_5\_coef | * Ed\_6\_coef | * det\_entropy\_log\_5\_coef | * app\_entropy\_shannon\_8\_coef | * app\_det\_TKEO\_mean\_3\_coef | * app\_det\_TKEO\_mean\_4\_coef | * Ed2\_4\_coef | * tqwt\_energy\_dec\_11 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| * 1 | * -40.9 | * 14 | * 6.27 | * 0.000233 | * -5.48e-10 | * 7.13e-05 | * -5.32e-05 | * 700 | * -4.57 |

* **La\_DFA**:

| * DFA | * VFER\_NSR\_SEO | * tqwt\_energy\_dec\_25 |
| --- | --- | --- |
| * 1 | * 0.355 | * -0.396 |

* **La\_RPDE**:

| * RPDE | * ppq5Jitter |
| --- | --- |
| * 1 | * -55.1 |

* **La\_numPulses**:

| * numPulses | * det\_entropy\_log\_3\_coef | * det\_entropy\_log\_5\_coef | * det\_LT\_entropy\_log\_3\_coef | * app\_LT\_TKEO\_mean\_3\_coef |
| --- | --- | --- | --- | --- |
| * 1 | * -3.31 | * 0.705 | * 3.06 | * -28.7 |

* **La\_numPeriodsPulses**:

| * numPulses | * numPeriodsPulses | * ppq5Jitter | * tqwt\_energy\_dec\_11 |
| --- | --- | --- | --- |
| * -1 | * 1 | * 423 | * 80.1 |

* **La\_meanPeriodPulses**:

| * numPulses | * meanPeriodPulses |
| --- | --- |
| * 1.6e-05 | * 1 |

predTestDri <- predictDecorrelate(DriDEdataframe,testDataFrame)  
  
univarDeDri <- uniRankVar(varlistDe,  
 paste(outcome,"~1"),  
 outcome,  
 predTestDri,  
 rankingTest = "AUC")

100 : La\_std\_MFCC\_1st\_coef 200 : La\_app\_entropy\_shannon\_10\_coef 300 : La\_app\_LT\_entropy\_log\_9\_coef 400 : La\_tqwt\_entropy\_log\_dec\_7 500 : La\_tqwt\_TKEO\_std\_dec\_35  
600 : La\_tqwt\_stdValue\_dec\_27 700 : tqwt\_skewnessValue\_dec\_19

## PCA Analysis

### PCA   
noclassData <- trainDataFrame[,colnames(trainDataFrame) != outcome]  
noclassTestData <- testDataFrame[,colnames(testDataFrame) != outcome]  
pc <- principal(noclassData,TopVariables,rotate="varimax") #principal components  
pander::pander(head(pc$loadings))

|  | RC3 | RC1 | RC4 | RC2 | RC5 |
| --- | --- | --- | --- | --- | --- |
| **gender** | -0.6190 | 0.1187 | 0.14235 | 0.0667 | -0.124 |
| **PPE** | -0.0178 | -0.5263 | 0.12687 | -0.3380 | -0.247 |
| **DFA** | -0.3424 | -0.0512 | -0.39557 | 0.2410 | 0.342 |
| **RPDE** | -0.4345 | 0.2859 | -0.10439 | 0.4484 | 0.117 |
| **numPulses** | 0.8612 | 0.0969 | -0.00301 | -0.1970 | -0.114 |
| **numPeriodsPulses** | 0.8605 | 0.0896 | -0.00204 | -0.1984 | -0.115 |

PCA\_Train <- as.data.frame(cbind(predict(pc,noclassData),trainDataFrame[,outcome]))  
colnames(PCA\_Train)[TopVariables+1] <- outcome  
  
PCA\_Predicted <- as.data.frame(cbind(predict(pc,noclassTestData),testDataFrame[,outcome]))  
colnames(PCA\_Predicted)[TopVariables+1] <- outcome  
varlistPCA <- colnames(PCA\_Predicted)[colnames(PCA\_Predicted) != "class"];  
varlistPCA <- as.data.frame(cbind(name=varlistPCA,desc=varlistPCA))  
  
univarPCA <- uniRankVar(varlistPCA,  
 paste(outcome,"~1"),  
 outcome,  
 PCA\_Predicted)

## EFA

uls <- fa(noclassData,TopVariables,rotate="varimax") #unweighted least squares is minres   
pander::pander(head(uls$weights))

|  | MR3 | MR1 | MR4 | MR2 | MR5 |
| --- | --- | --- | --- | --- | --- |
| **gender** | -0.6152 | 0.1174 | 0.14178 | 0.0662 | -0.121 |
| **PPE** | -0.0174 | -0.5234 | 0.12532 | -0.3383 | -0.247 |
| **DFA** | -0.3419 | -0.0509 | -0.39150 | 0.2425 | 0.336 |
| **RPDE** | -0.4329 | 0.2841 | -0.10203 | 0.4476 | 0.117 |
| **numPulses** | 0.8605 | 0.0974 | -0.00434 | -0.1974 | -0.115 |
| **numPeriodsPulses** | 0.8599 | 0.0901 | -0.00336 | -0.1988 | -0.115 |

EFA\_Train <- as.data.frame(cbind(predict(uls,noclassData),trainDataFrame[,outcome]))  
colnames(EFA\_Train)[TopVariables+1] <- outcome  
EFA\_Predicted <- as.data.frame(cbind(predict(uls,noclassTestData),testDataFrame[,outcome]))  
colnames(EFA\_Predicted)[TopVariables+1] <- outcome  
varlistEFA <- colnames(EFA\_Predicted)[colnames(EFA\_Predicted) != "class"];  
varlistEFA <- as.data.frame(cbind(name=varlistEFA,desc=varlistEFA))  
  
univarEFA <- uniRankVar(varlistEFA,  
 paste(outcome,"~1"),  
 outcome,  
 EFA\_Predicted)

### The tables

pander::pander(univar$orderframe[1:TopVariables,univariate\_columns])

|  | caseMean | caseStd | controlMean | controlStd | controlKSP | ROCAUC | wilcox.Zvalue |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **std\_delta\_log\_energy** | 0.0403 | 0.02153 | 0.0226 | 0.01414 | 0.299 | 0.818 | 5.05 |
| **std\_delta\_delta\_log\_energy** | 0.0149 | 0.00737 | 0.0083 | 0.00433 | 0.315 | 0.807 | 4.87 |
| **tqwt\_maxValue\_dec\_12** | 0.0924 | 0.07546 | 0.1854 | 0.10206 | 0.826 | 0.792 | 4.62 |
| **tqwt\_minValue\_dec\_12** | -0.0923 | 0.07576 | -0.1844 | 0.10150 | 0.817 | 0.786 | 4.52 |
| **tqwt\_maxValue\_dec\_11** | 0.0646 | 0.04572 | 0.1266 | 0.07109 | 0.554 | 0.781 | 4.43 |

pander::pander(univarDe$orderframe[1:TopVariables,univariate\_columns])

|  | caseMean | caseStd | controlMean | controlStd | controlKSP | ROCAUC | wilcox.Zvalue |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **La\_tqwt\_entropy\_log\_dec\_33** | -3.21e+02 | 1.25e+02 | -1.07e+02 | 2.83e+02 | 0.6866 | 0.812 | 4.94 |
| **La\_std\_delta\_log\_energy** | 1.58e-02 | 1.03e-02 | 7.52e-03 | 6.47e-03 | 0.4436 | 0.794 | 4.65 |
| **La\_std\_6th\_delta** | 1.42e-02 | 7.21e-03 | 5.55e-03 | 8.41e-03 | 0.9252 | 0.787 | 4.53 |
| **tqwt\_entropy\_shannon\_dec\_36** | 3.96e+00 | 9.81e+00 | 9.75e+00 | 1.10e+01 | 0.0427 | 0.770 | 4.25 |
| **La\_std\_2nd\_delta** | 7.52e-03 | 8.61e-03 | -5.91e-04 | 8.34e-03 | 0.5680 | 0.769 | 4.22 |

pander::pander(univarDeDri$orderframe[1:TopVariables,univariate\_columns])

|  | caseMean | caseStd | controlMean | controlStd | controlKSP | ROCAUC | wilcox.Zvalue |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **La\_tqwt\_entropy\_log\_dec\_33** | -3.21e+02 | 1.25e+02 | -1.07e+02 | 2.83e+02 | 0.687 | 0.812 | 4.94 |
| **La\_std\_6th\_delta** | 1.42e-02 | 7.21e-03 | 5.55e-03 | 8.41e-03 | 0.925 | 0.787 | 4.53 |
| **tqwt\_entropy\_log\_dec\_12** | -1.68e+05 | 4.04e+04 | -1.32e+05 | 2.83e+04 | 0.494 | 0.763 | 4.13 |
| **La\_tqwt\_entropy\_shannon\_dec\_16** | -4.57e+01 | 1.92e+02 | 1.08e+02 | 1.83e+02 | 0.253 | 0.755 | 3.99 |
| **std\_9th\_delta\_delta** | 1.88e-02 | 4.54e-03 | 1.51e-02 | 2.70e-03 | 0.788 | 0.752 | 3.93 |

pander::pander(univarPCA$orderframe[1:TopVariables,univariate\_columns])

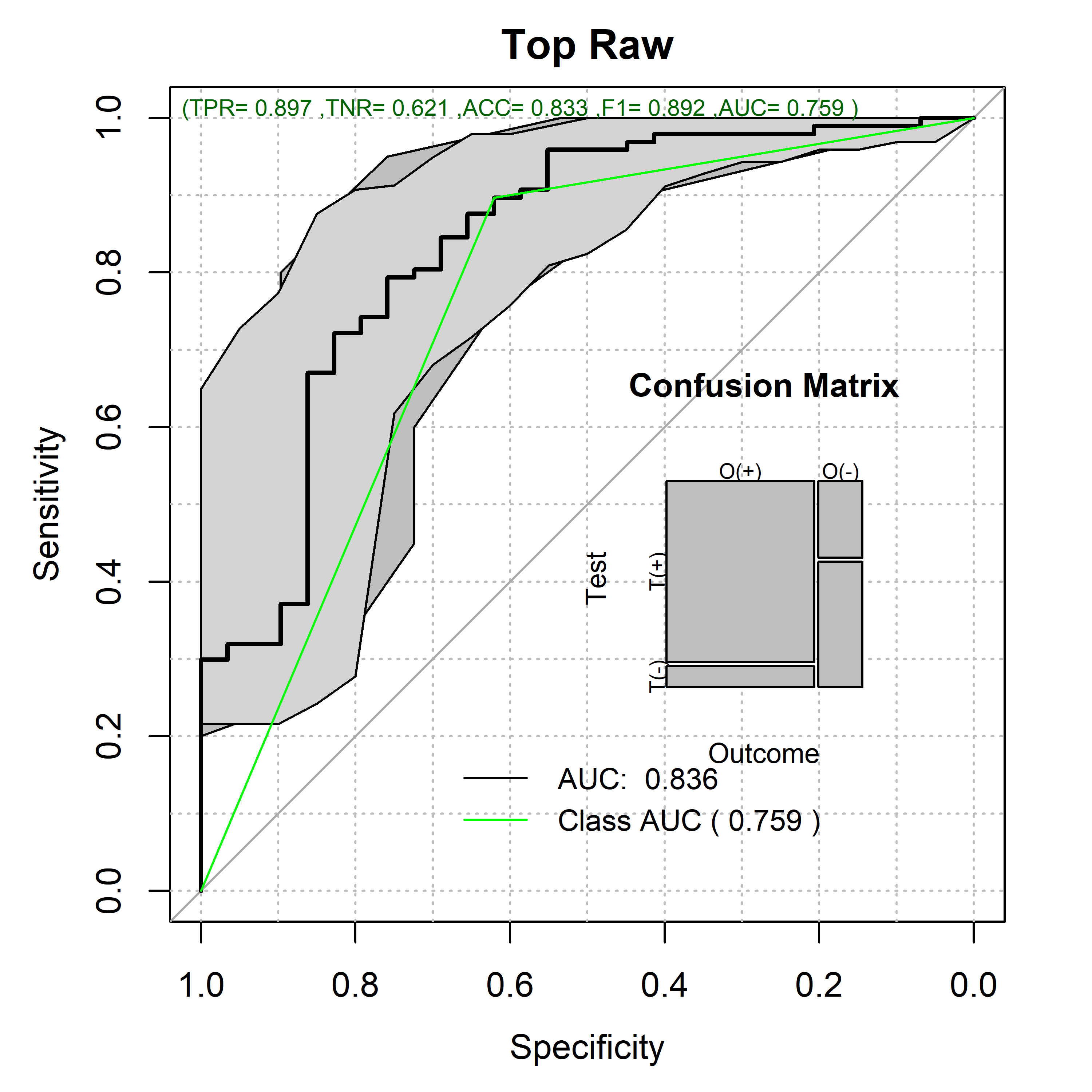
|  | caseMean | caseStd | controlMean | controlStd | controlKSP | ROCAUC | wilcox.Zvalue |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RC2** | 12.43 | 63.4 | -41.6 | 36.1 | 0.0909 | 0.784 | 4.49 |
| **RC4** | -13.33 | 73.9 | 44.6 | 44.2 | 0.3577 | 0.749 | 3.90 |
| **RC3** | -11.71 | 79.8 | 39.2 | 111.0 | 0.8258 | 0.640 | 2.01 |
| **RC5** | -5.34 | 45.5 | 17.9 | 37.3 | 0.3872 | 0.658 | 2.33 |
| **RC1** | -3.75 | 68.1 | 12.5 | 69.9 | 0.1100 | 0.625 | 1.74 |

pander::pander(univarEFA$orderframe[1:TopVariables,univariate\_columns])

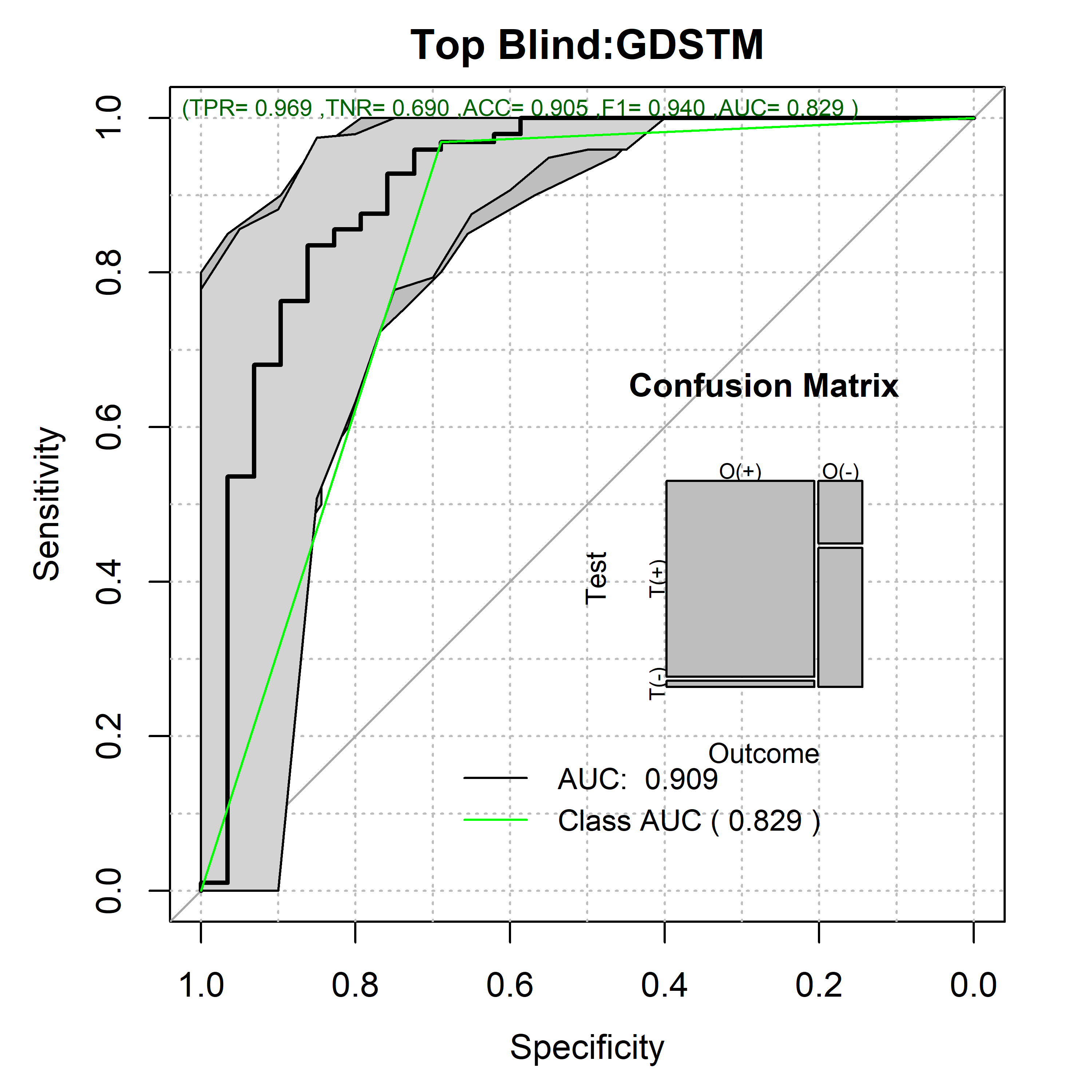
|  | caseMean | caseStd | controlMean | controlStd | controlKSP | ROCAUC | wilcox.Zvalue |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MR2** | 12.46 | 63.5 | -41.7 | 36.1 | 0.102 | 0.785 | 4.49 |
| **MR4** | -13.21 | 73.4 | 44.2 | 44.0 | 0.365 | 0.749 | 3.89 |
| **MR3** | -11.71 | 79.7 | 39.2 | 110.8 | 0.824 | 0.640 | 2.00 |
| **MR5** | -5.39 | 45.5 | 18.0 | 37.3 | 0.376 | 0.662 | 2.39 |
| **MR1** | -3.75 | 68.0 | 12.5 | 69.6 | 0.111 | 0.626 | 1.74 |

### Model of top five variables

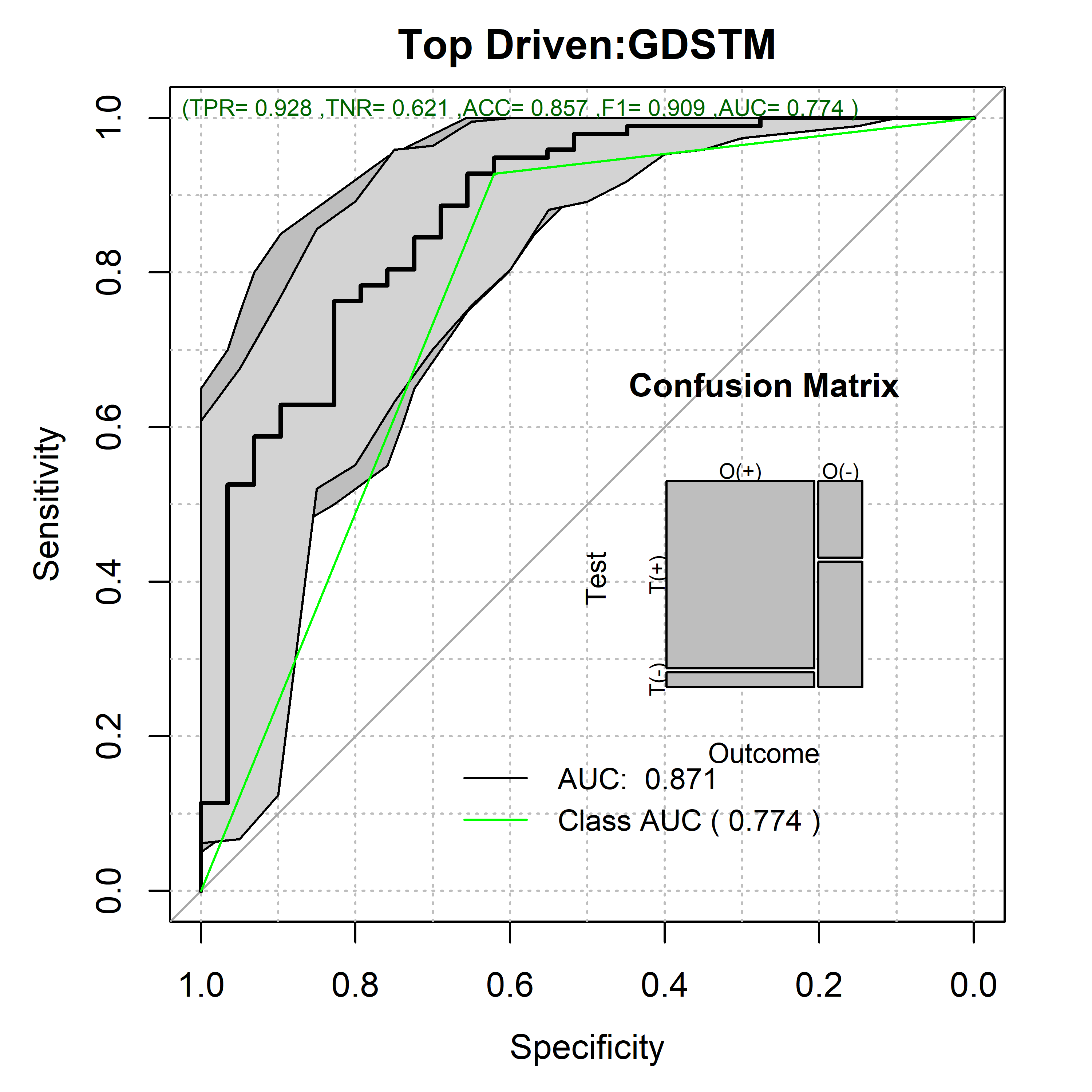
lmRAW <- glm(paste(outcome,"~."),  
 trainDataFrame[,c(outcome,rownames(univar$orderframe[1:TopVariables,]))],  
 family="binomial")  
pr <- predictionStats\_binary(cbind(testDataFrame[,outcome],predict(lmRAW,testDataFrame)),"Top Raw",cex=0.95)

Top Raw 

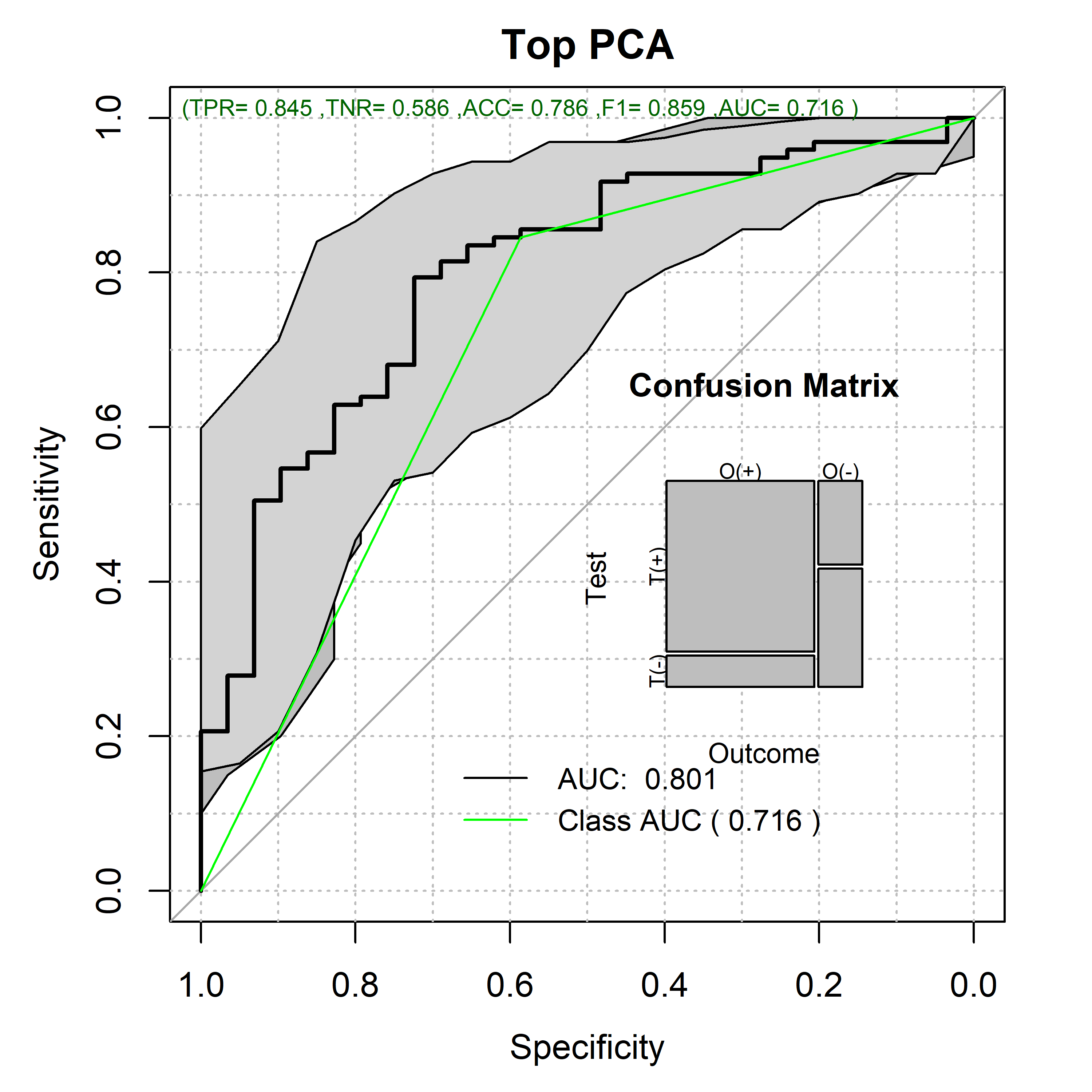
lmDe <- glm(paste(outcome,"~."),  
 DEdataframe[,c(outcome,rownames(univarDe$orderframe[1:TopVariables,]))],  
 family="binomial")  
pr <- predictionStats\_binary(cbind(predTestDe[,outcome],predict(lmDe,predTestDe)),"Top Blind:GDSTM",cex=0.95)

Top Blind:GDSTM 

lmDri <- glm(paste(outcome,"~."),  
 DriDEdataframe[,c(outcome,rownames(univarDeDri$orderframe[1:TopVariables,]))],  
 family="binomial")  
pr <- predictionStats\_binary(cbind(predTestDe[,outcome],predict(lmDri,predTestDri)),"Top Driven:GDSTM",cex=0.95)

Top Driven:GDSTM 

lmPCA <- glm(paste(outcome,"~."),  
 PCA\_Train,  
 family="binomial")  
pr <- predictionStats\_binary(cbind(PCA\_Predicted[,outcome],predict(lmPCA,PCA\_Predicted)),"Top PCA",cex=0.95)

Top PCA 

lmEFA <- glm(paste(outcome,"~."),  
 EFA\_Train,  
 family="binomial")  
pr <- predictionStats\_binary(cbind(EFA\_Predicted[,outcome],predict(lmEFA,EFA\_Predicted)),"Top EFA",cex=0.95)

Top EFA 