# QUESTION 1B

3 Principal Components were needed to capture atleast 95% of the data (0.9772 to be precise)

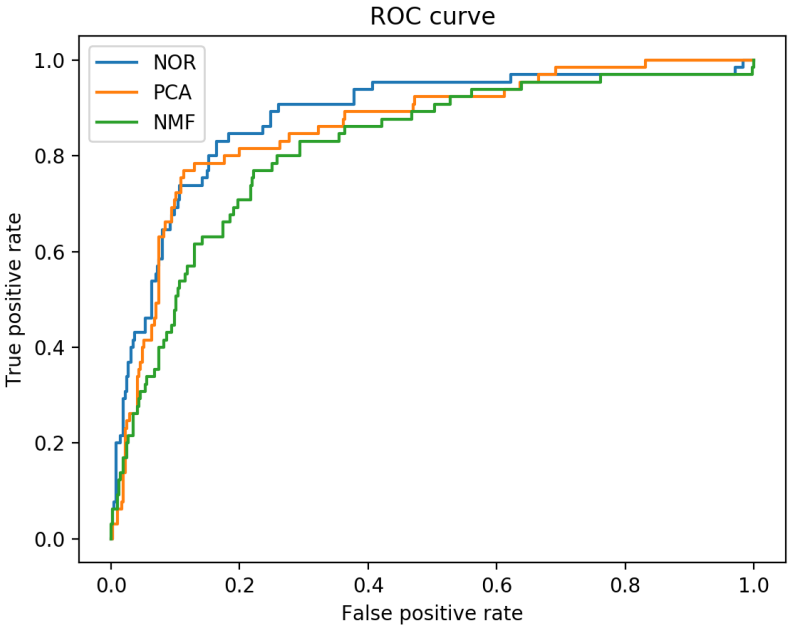
These components were:

- Component 1: **fixed acidity, citric acid, pH**
- Component 2: **residual sugar, free sulfur dioxide, total sulfur dioxide**
- Component 3: **volatile acidity, free sulfur dioxide, alcohol**

# QUESTION 1C

I picked **R = 10** by minimizing reconstruction error

# QUESTION 1D




From this ROC curve, we can make the conclusion that the normalized data performs far better than the NMF data since the NOR curve is above or equal to the NMF curve at every single FPR-TPR ratio. PCA performs worse than both NMF and NOR when the FPR and TPR are close to 0 however performs better than NOR and NMF when FPR and TPR are close to 1.

# QUESTION 2B

Refer to file: **performance_rf.csv**

# QUESTION 2C

OOB Error: 0.0741733690795353

Error on test set: 0.1166666666666667

**Error is higher on the test set than OOB which is expected**

# QUESTION 3A

Refer to file: **performance_xgboost.csv**

# QUESTION 3B

XGBoost Error: 0.08125000000000004

Random Forest Error: 0.1166666666666667

**XGBoost Performs better than Random Forest using misclassification error as a metric**