Project SETI

Classification of the Unknown *Squiggle* Time Series

Frank Fan, Kenny Smith, Jason Wang
Advised by Professor Jeffrey Ullman

Two Goals

Squiggle v.s. Non-Squiggle

1) Build a real-time classifier to distinguish between squiggle and non-squiggle

Intra-Squiggle

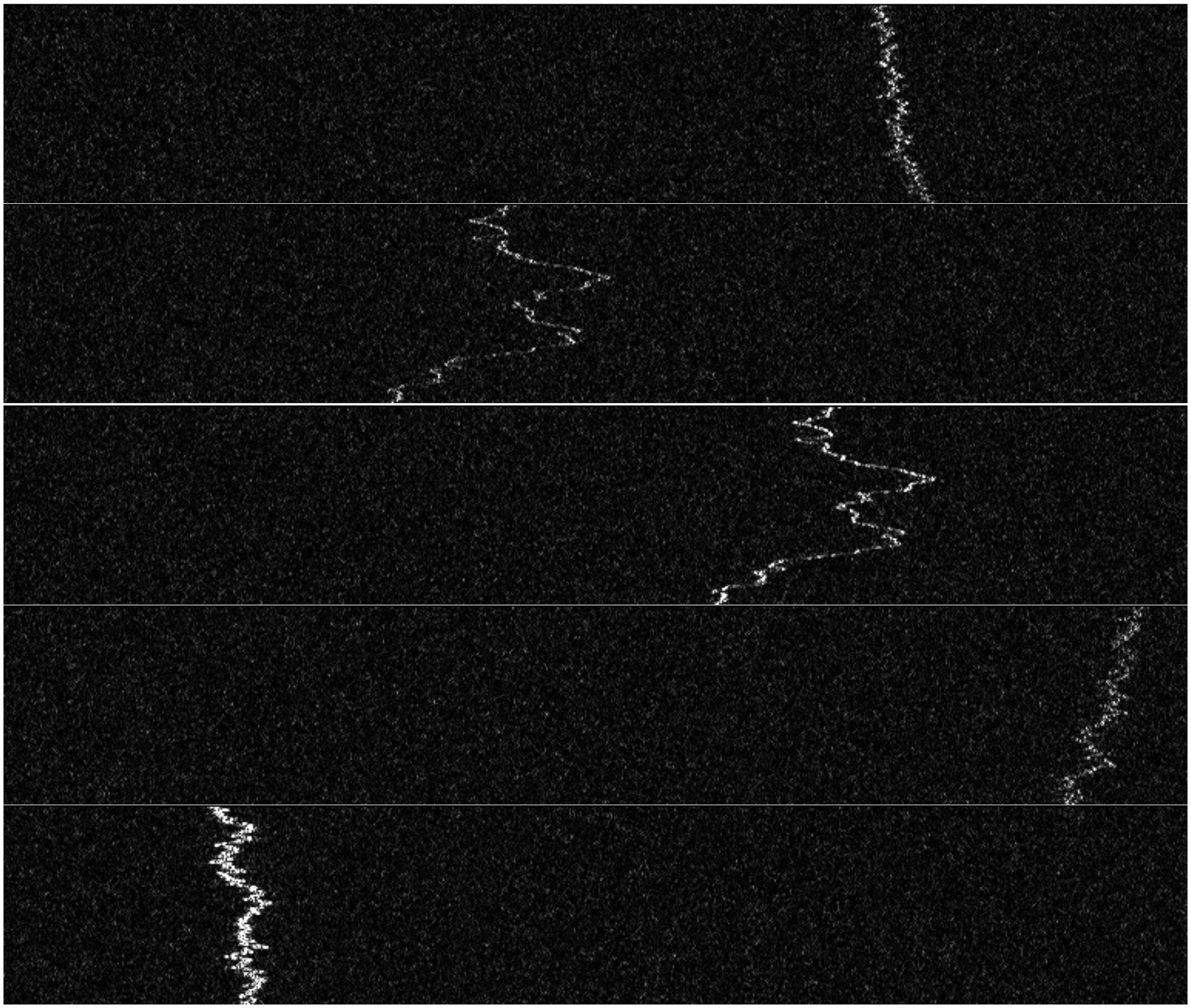
1) Identify squiggle subgroups
2) Pinpoint key characteristics of each subgroup
3) Build a multi-class classifier to stratify new squiggles into subgroups*

Squiggle

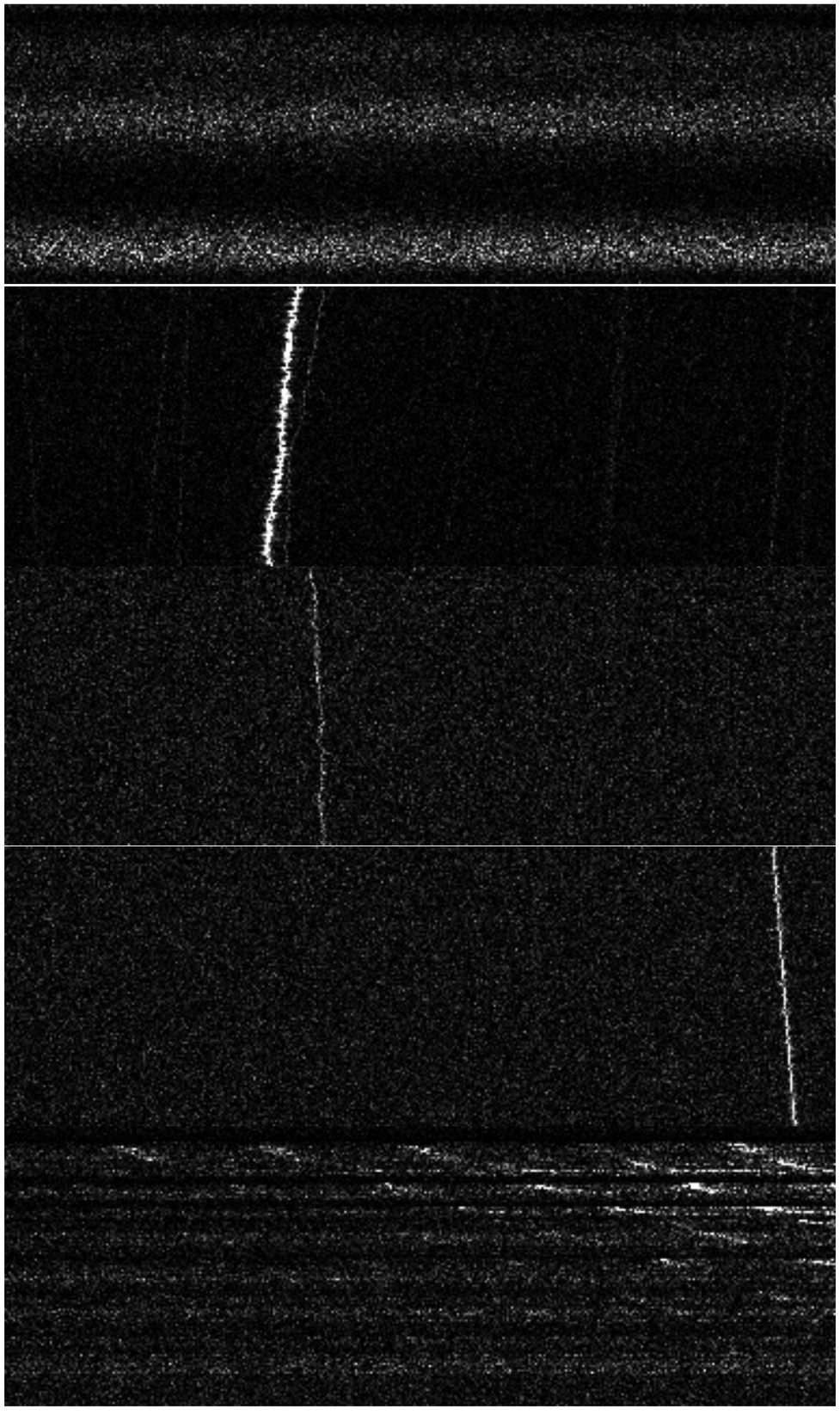
v.s.

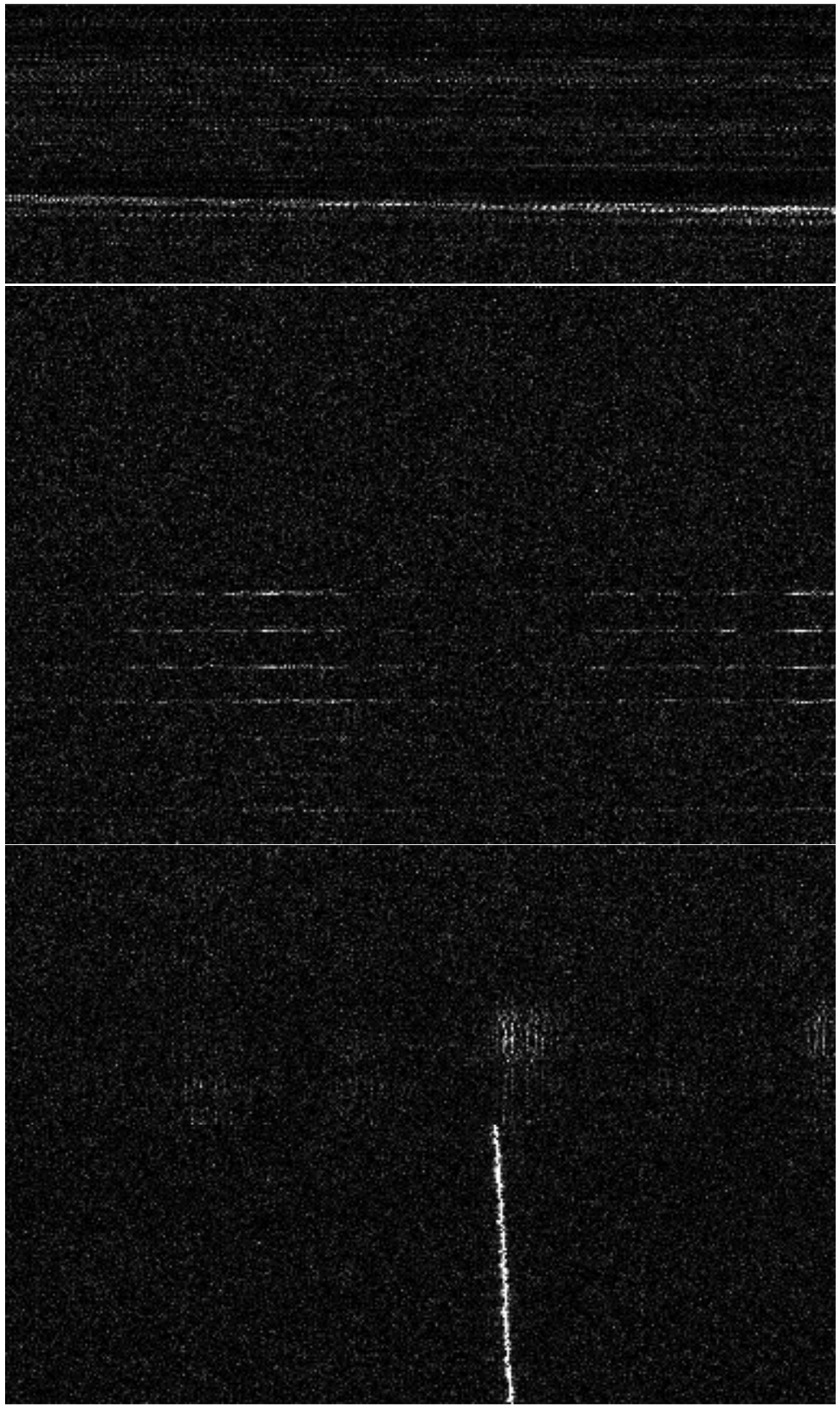
Non-squiggle

Squiggle



Non-Squiggle

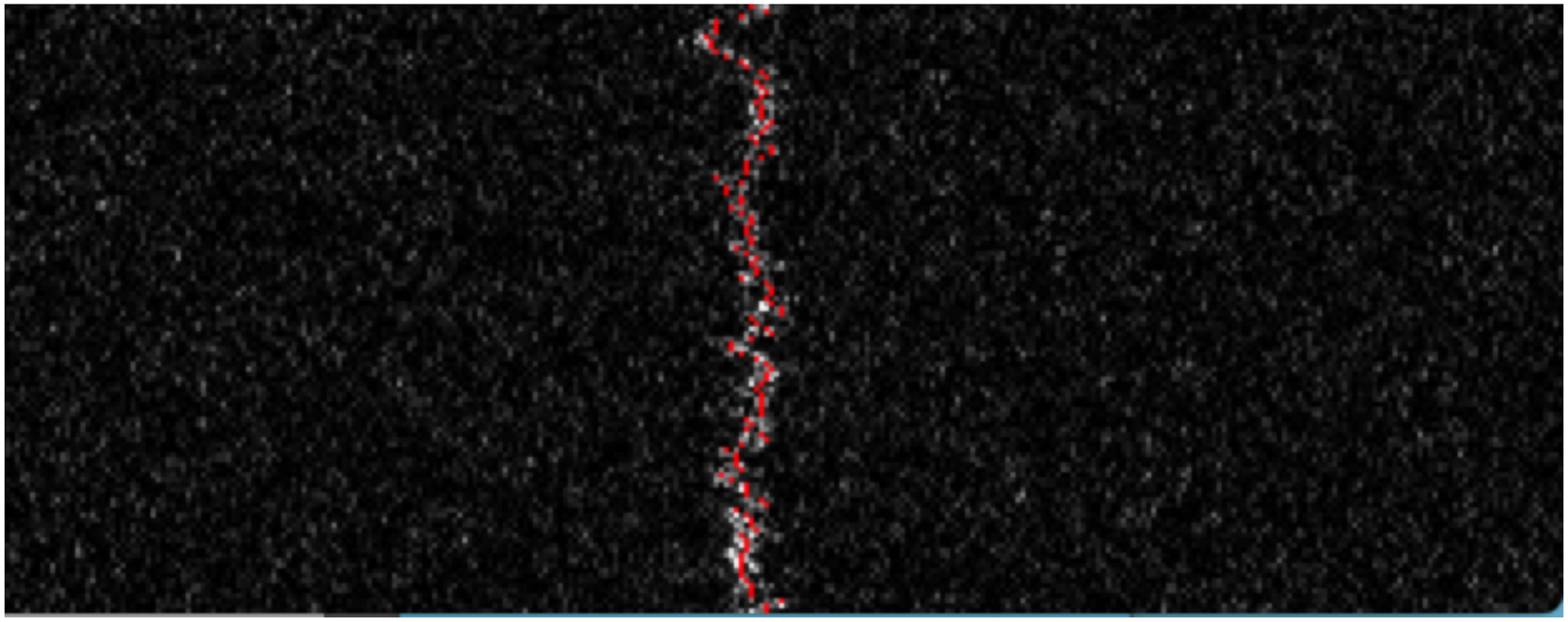




Insight 1: Discretization

Dynamic Prog. Algorithm

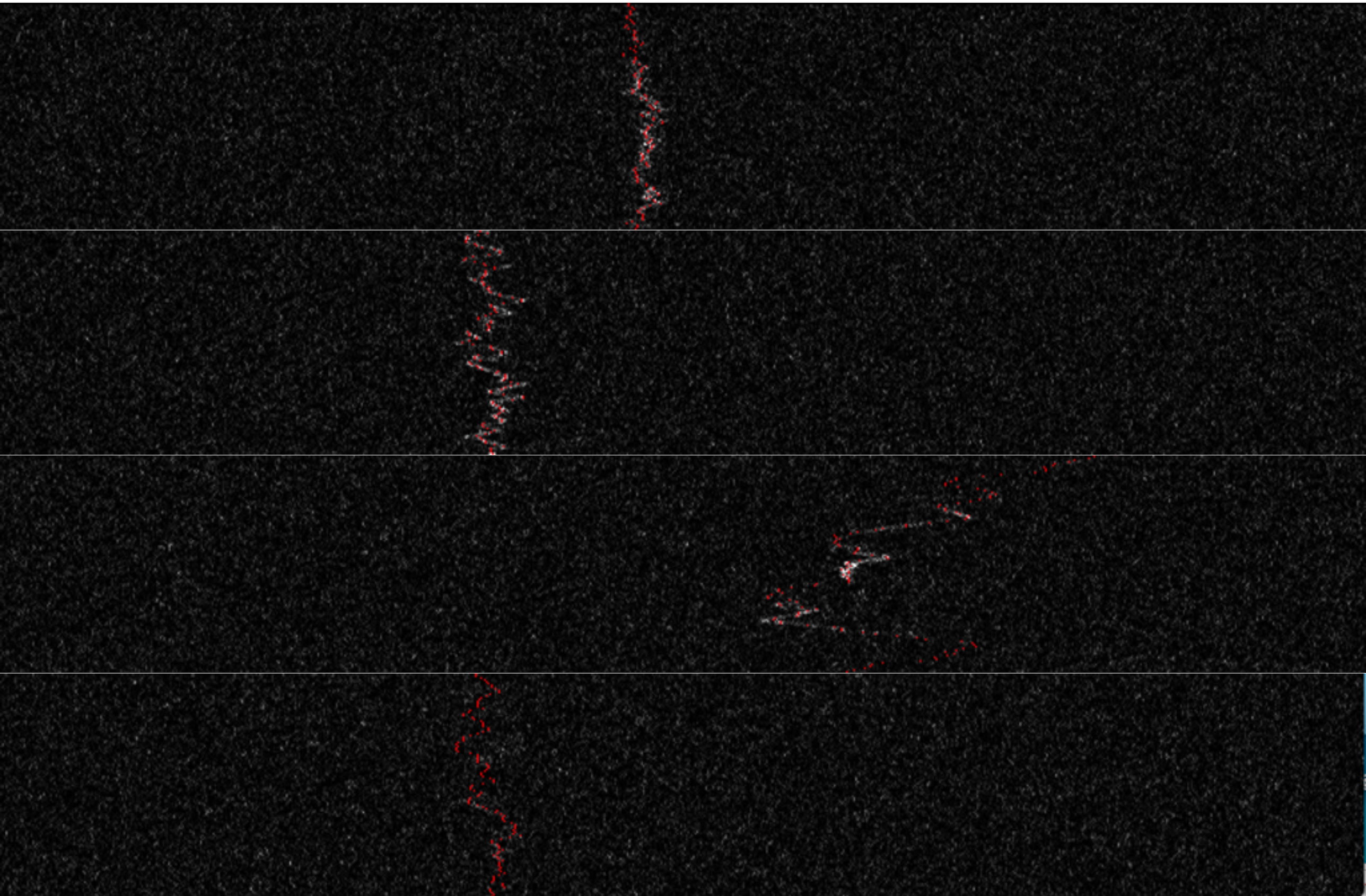
Spectrogram → Time Series



Choose path to minimize

$L(\alpha,\beta) = -\alpha*(\text{Intensity}) - \beta*(\text{Neighboring Intensities}) + (1-\alpha-\beta)(\text{Deviation})^2$

Discretization



Feature Overview

- 63 Discrete Fourier Transform samples
- Variance of raw time series
- Loss from DP algorithm
- AR, MA, Innovation, & Intercept parameters from ARIMA(1,1,1) fitted model
- Hurst exponent*
- Average signal thickness
- MSE from fitted linear regression

72 Total Features

Dataset

Initial

7657 unknown (squiggle and non-squiggle mixed)
833 squiggles (hand-curated)

Following initial clustering + hand-curation:
7607 non-squiggles
883 squiggles

8490 spectrograms total

Methodology

1. Split 8490 samples in **90% Training**, **10% Test**

2. Using **90% training:**
Conduct 10-fold cross validation on training set to tune model parameters

3. Using **10% test:**
Measure performance using validation set error

Baseline Model

Logistic regression
129 raw time series data points

Validation test error

Model	ACC	AUC
Unregularized	**0.875**	**0.504**
Lasso (L1)	0.875	0.500
Ridge (L2)	0.875	0.500

Intermediate Model

Logistic regression

63 discrete fourier transform samples

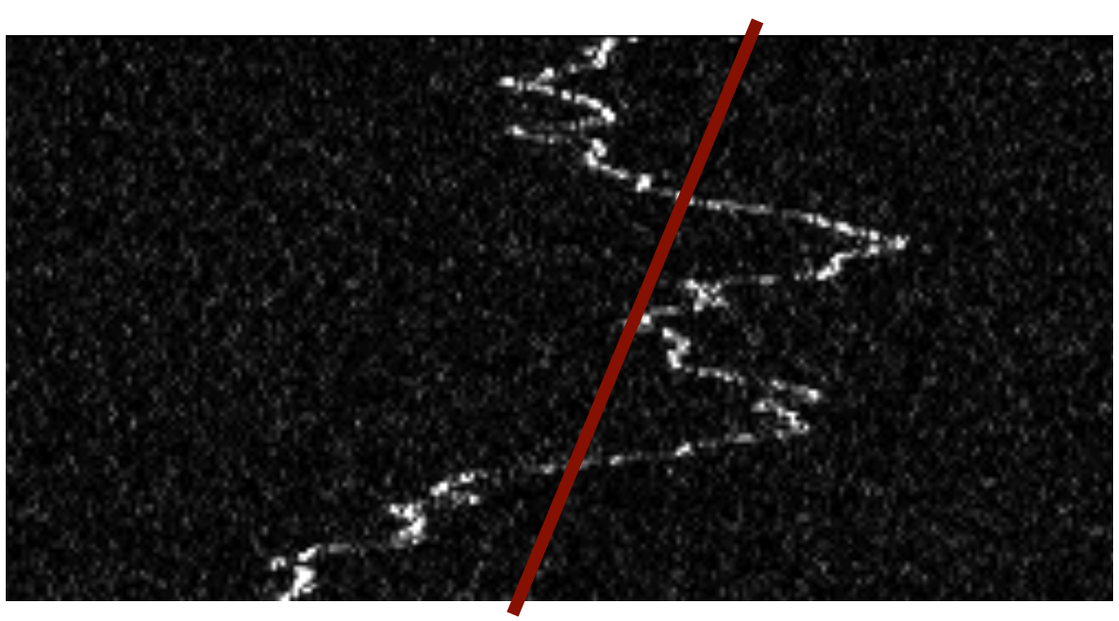
Validation test error

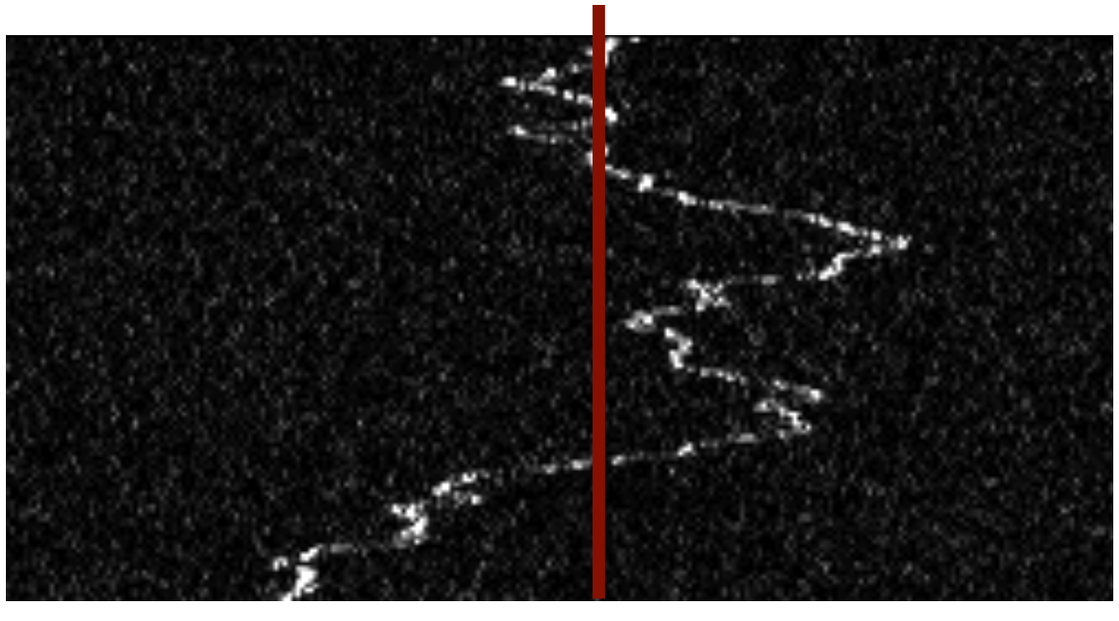
Model	ACC	AUC
Unregularized	0.955	**0.967**
Lasso (L1)	0.953	0.963
Ridge (L2)	**0.959**	0.962

Insight 2:
Re-visit Raw Time Series

Introduce 4 new features:

- Loss from DP algorithm
- Average signal thickness
- Modulation: MSE from fitted linear regression
- Variance of raw time series





Insight 3:
Time Series Parameters

Introduce 4 more new features from ARIMA(1,1,1)

- Intercept
- Autoregression
- Moving Average
- Innovation

ARIMA(1,1,1) Model

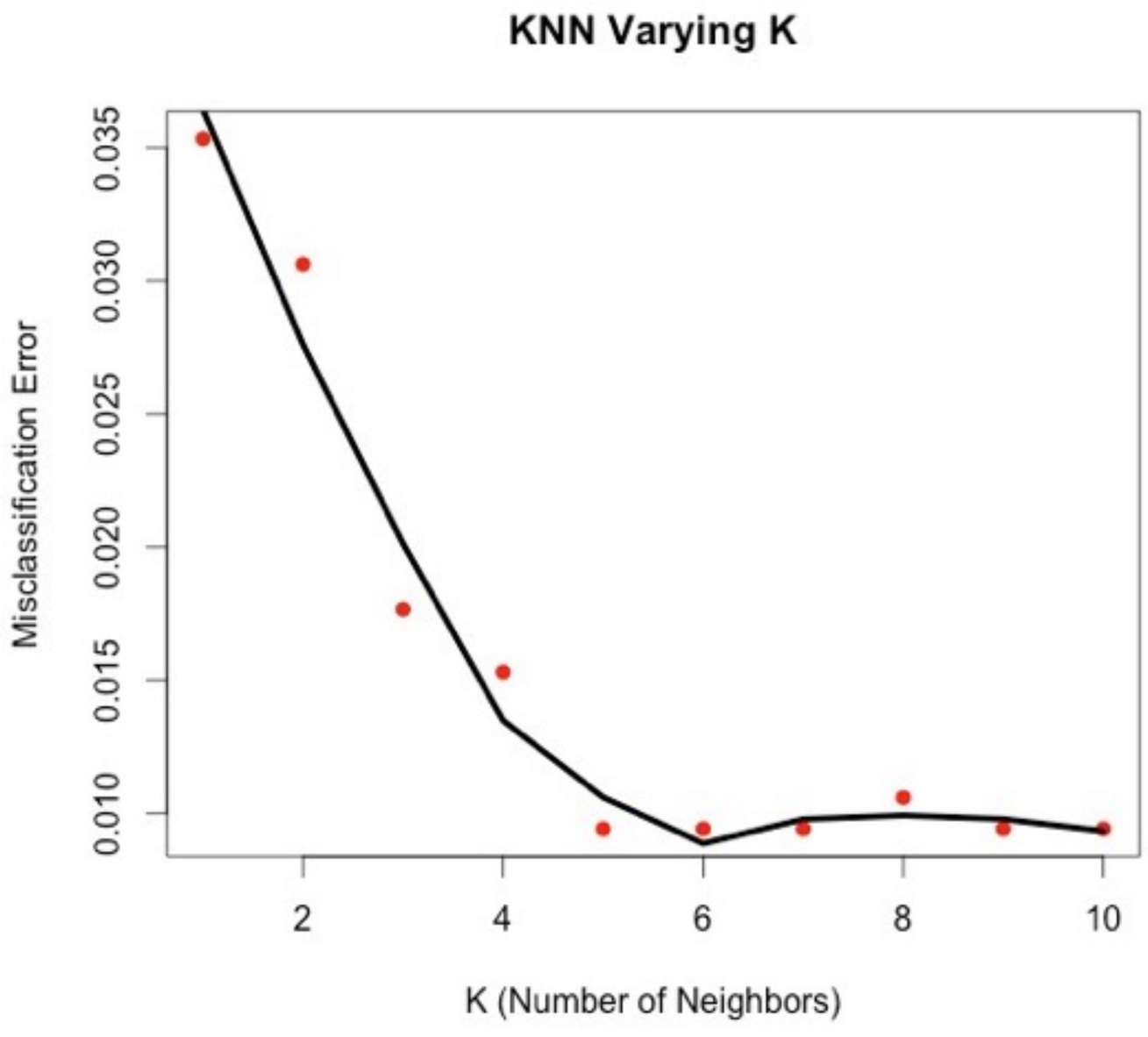
$$(1 - \phi B)\nabla x_t = (1 + \theta B)w_t$$

$$x_t = (1 + \phi)x_{t-1} - \phi x_{t-2} + w_t + \theta w_{t-1}$$

Full Model

K-Nearest Neighbors

Full **72** feature set



Validation test error

Model	ACC
K = 4	0.987

Full Model

Logistic Family

Full **72** feature set

Validation test error

Model	ACC	AUC
Unregularized	0.986	0.981
Lasso (L1)	0.987	0.986
Ridge (L2)	**0.988**	**0.991**

Full Model

Support Vector Machines

Full **72** feature set

Validation test error

Kernel	ACC	AUC
Linear	**0.989**	0.979
Radial	0.988	0.981
Polynomial	0.982	**0.987**
Sigmoid	0.954	0.983

Full Model

Tree-Based Methods

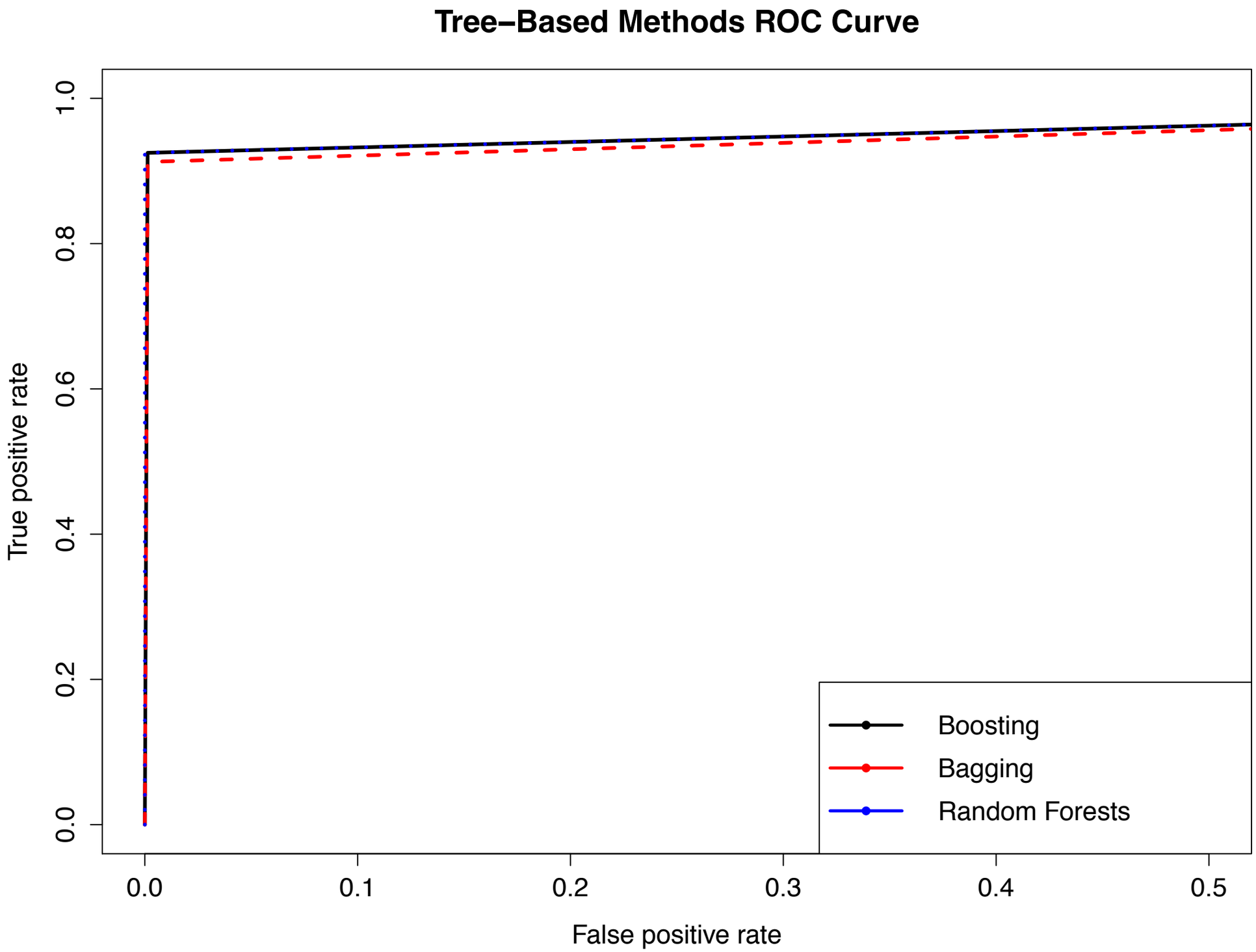
Full **72** feature set

Validation test error

Model	ACC	AUC
Boosting	**0.992**	0.962
Bagging	0.992	0.956
Random Forests	0.991	**0.963**

Full Model

Tree-Based Methods

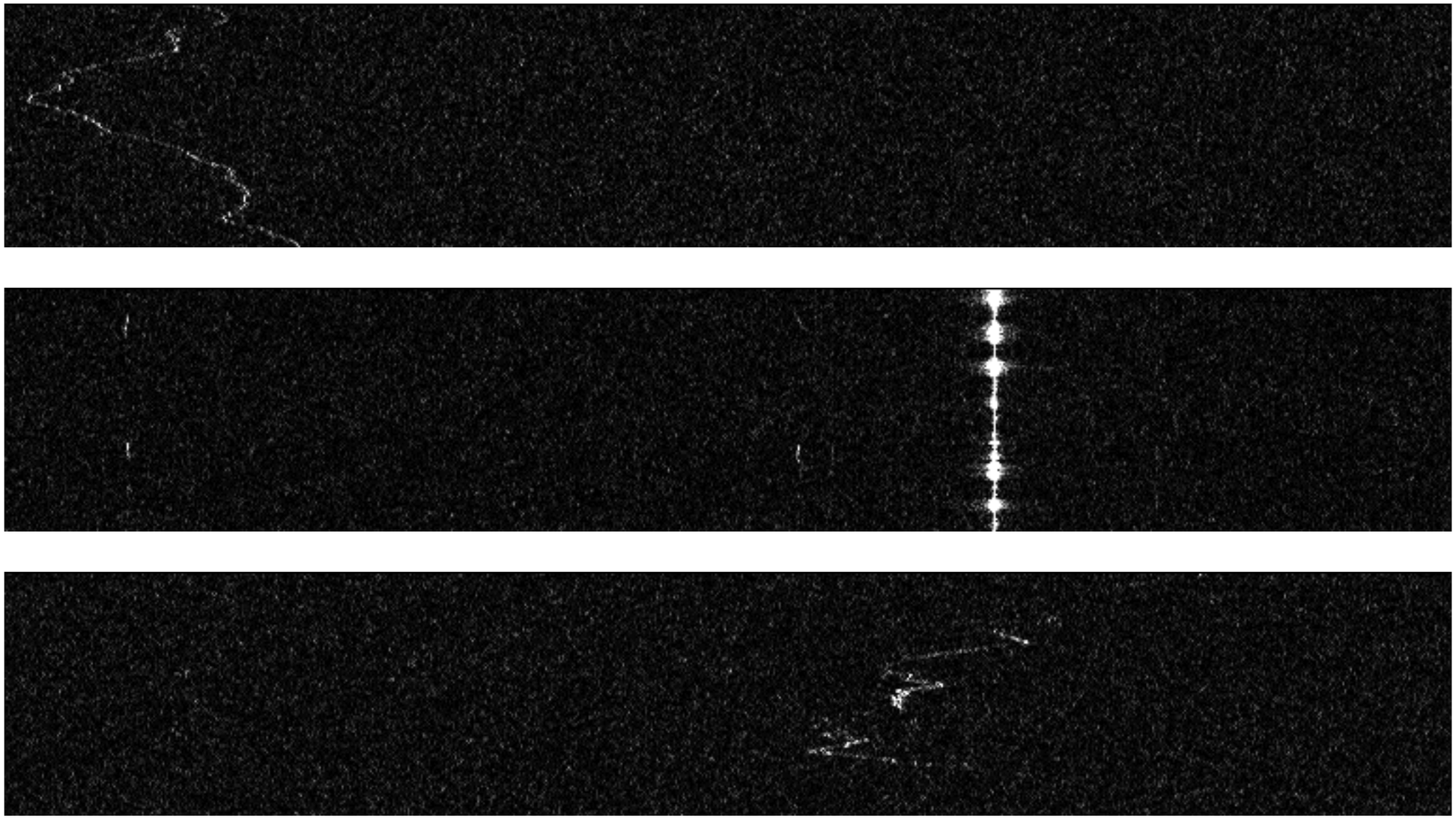


Full Model
Significant Features

Significant Feature: Yes/No?

Model	Logistic (Lasso)	Logistic (Unreg. + Ridge)	SVM	Tree-Based
DP Loss	X	X		
Width	X	X		X
Innovation	X	X	X	X
Modulation	X	X	X	X
MA		X	X	
AR		X	X	
DFT		X		

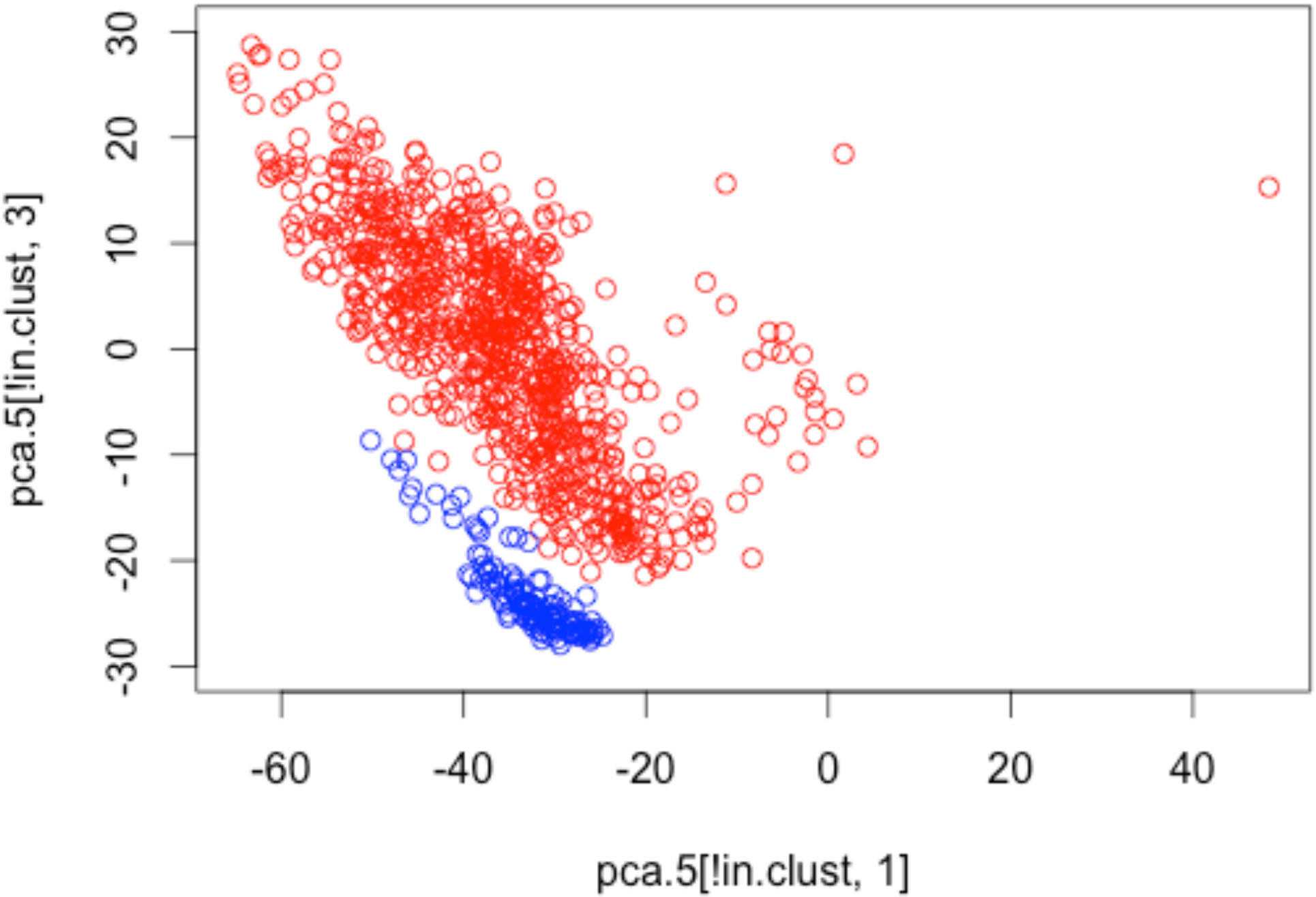
Failed Classifications

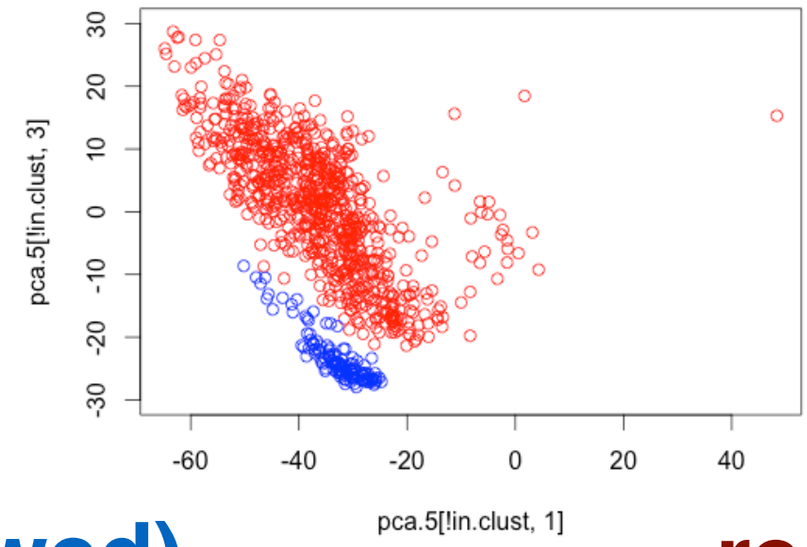


Intra-Squiggle Exploration

Principal Components Analysis

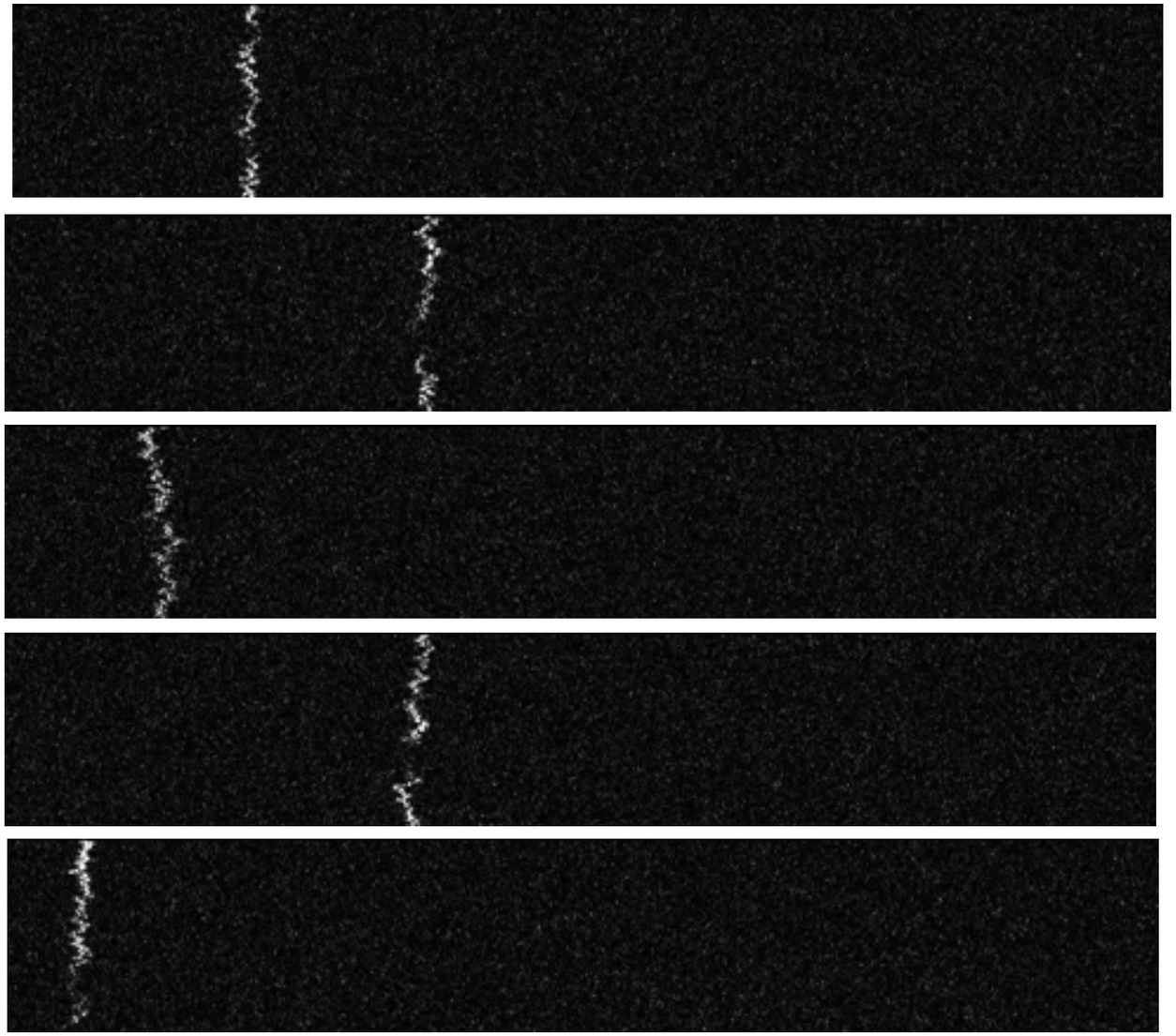
PC 1 v.s. PC 3

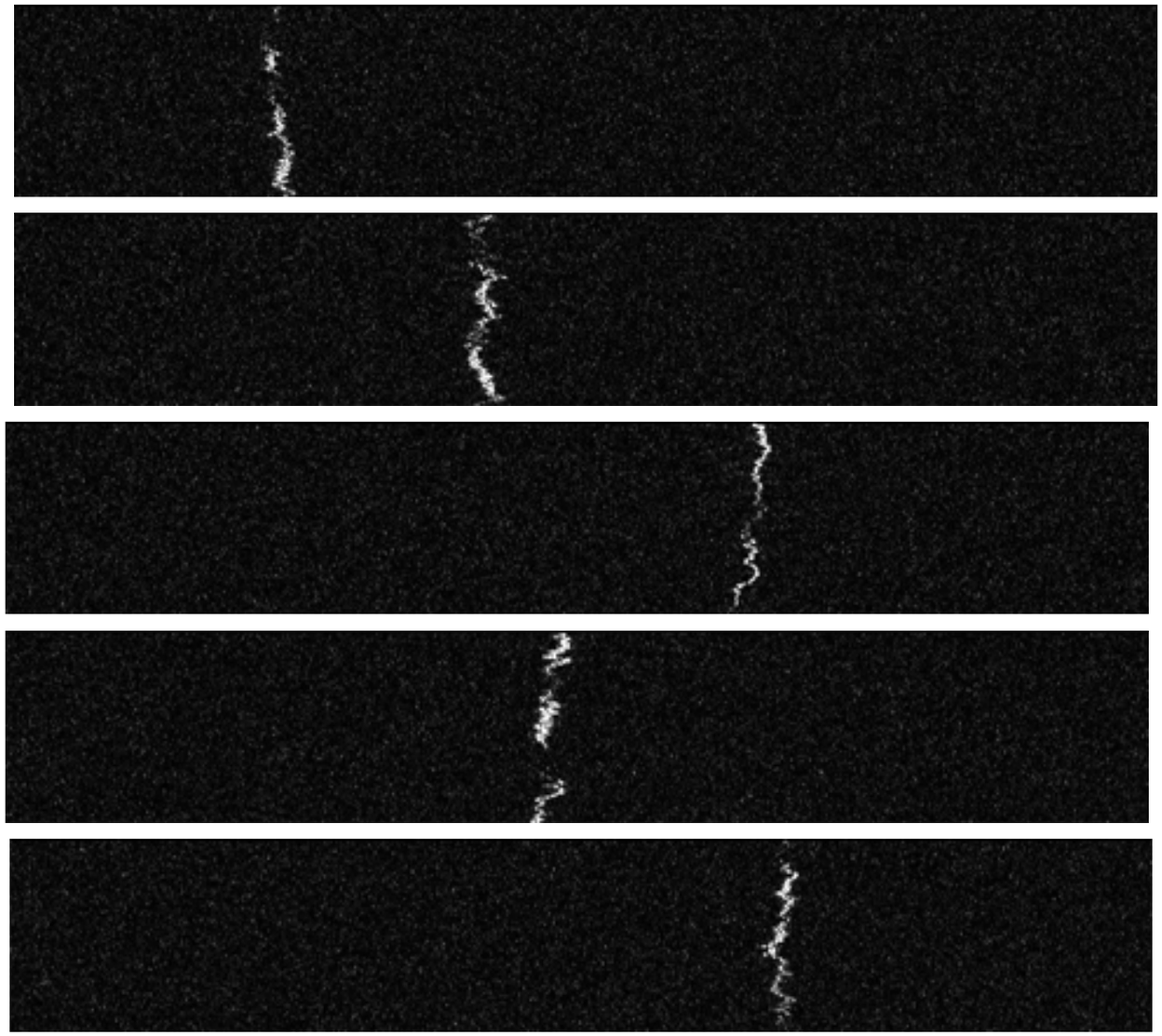




blue (AR-skewed)

red (MA-skewed)

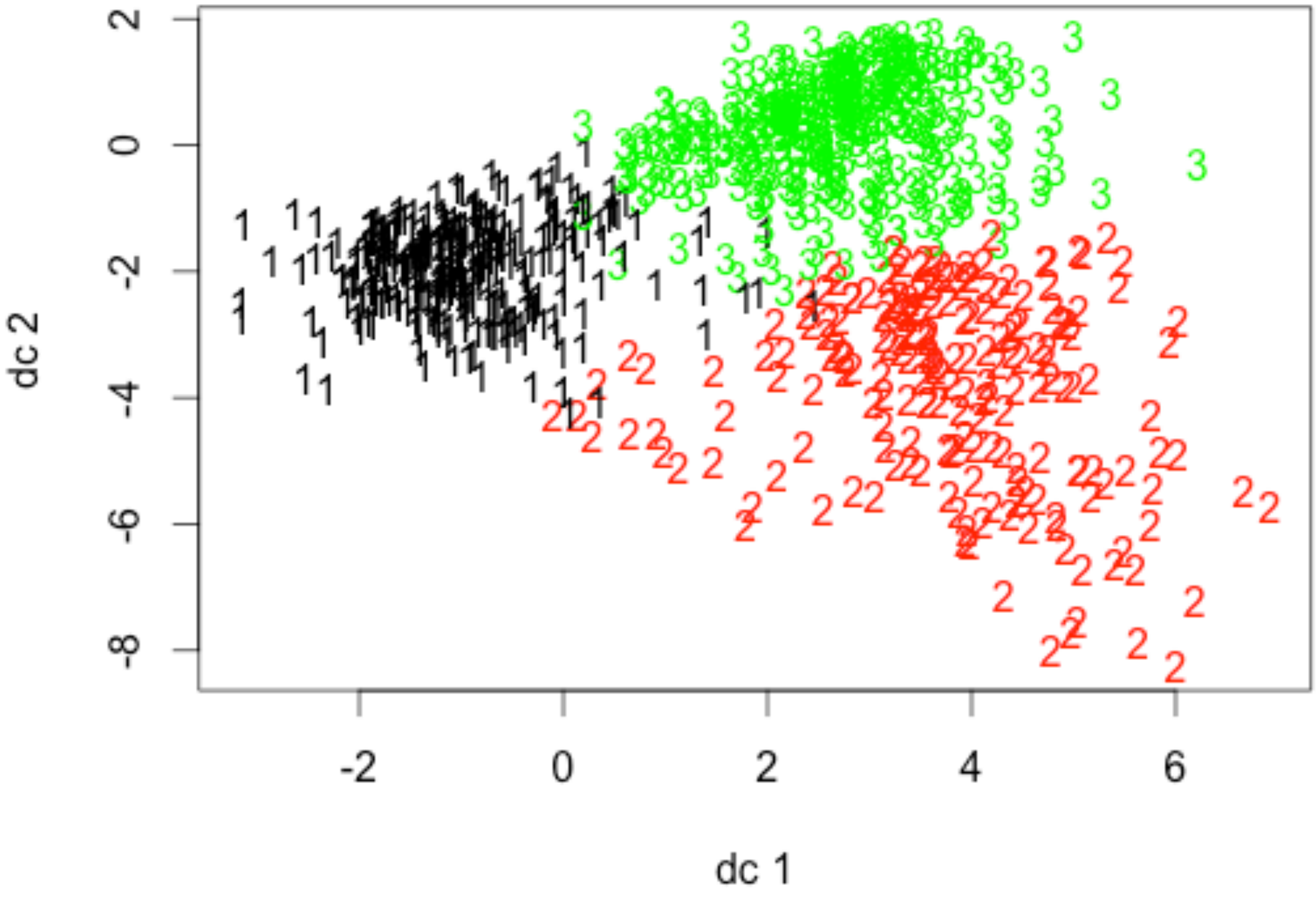


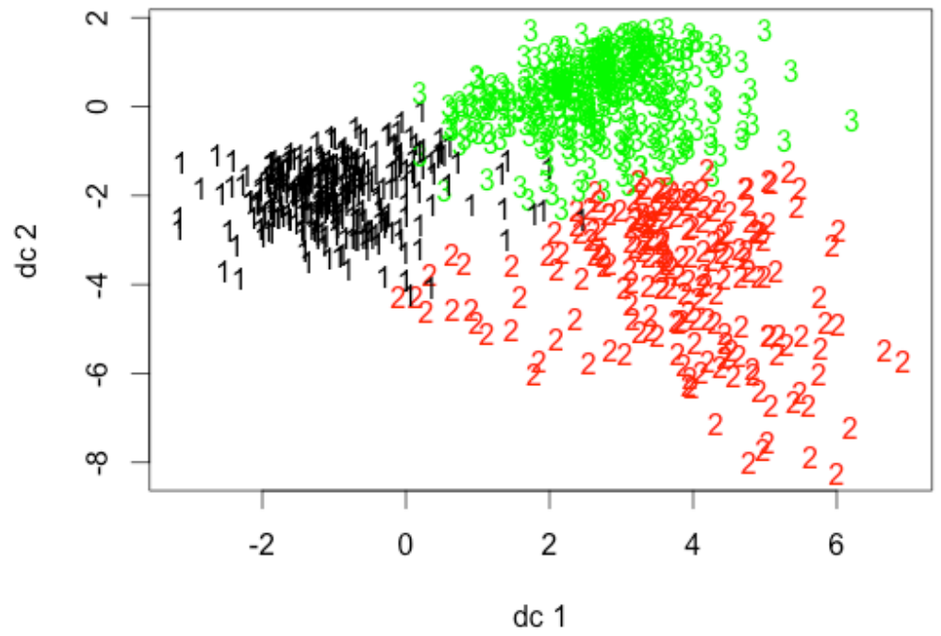


0.1353110

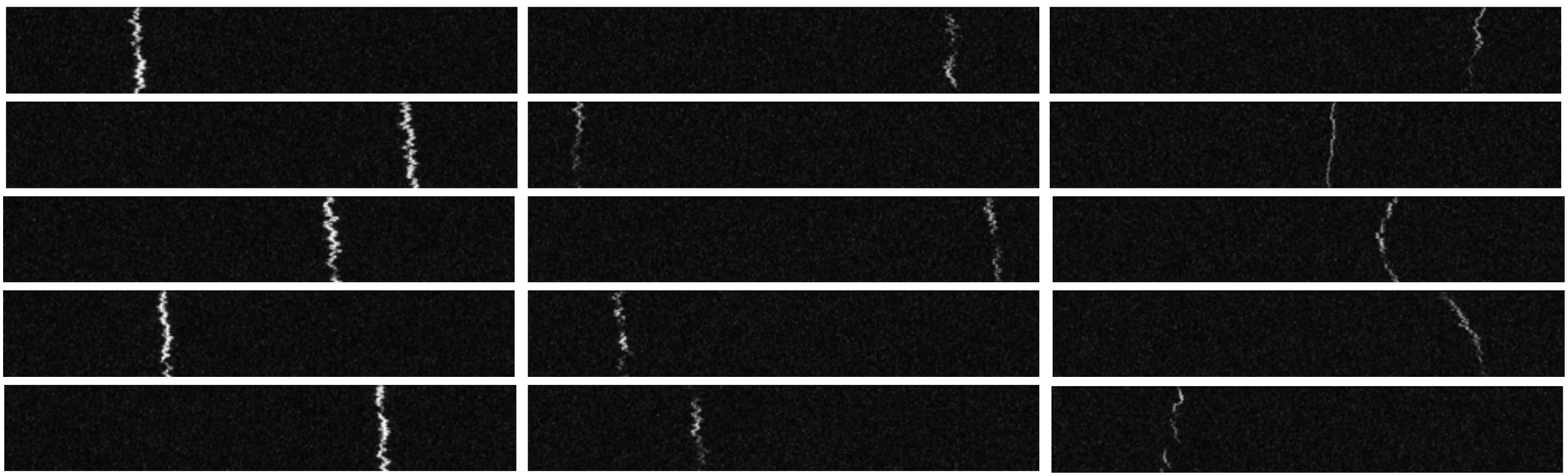
0.5392916

K-Means
Euclidean



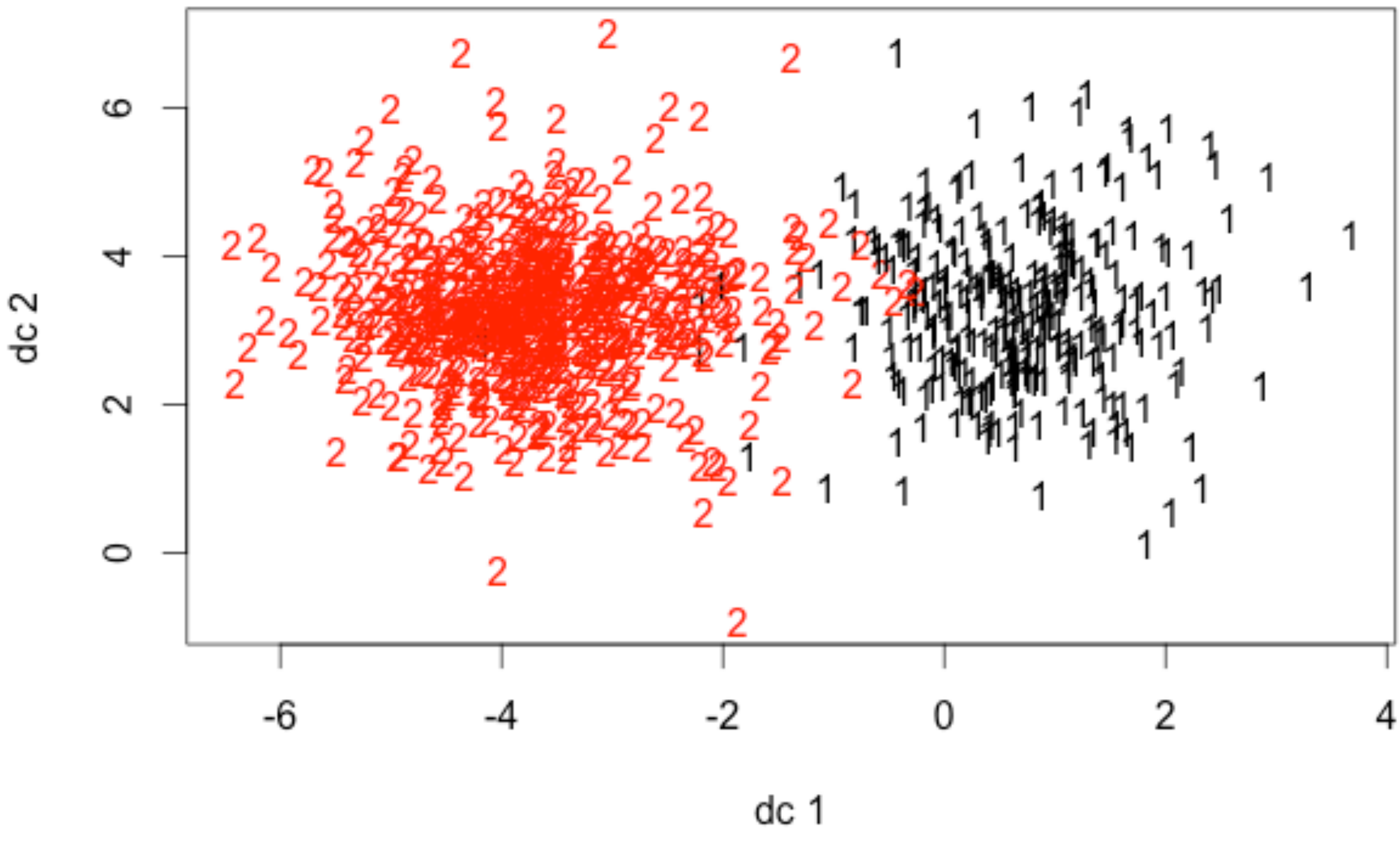


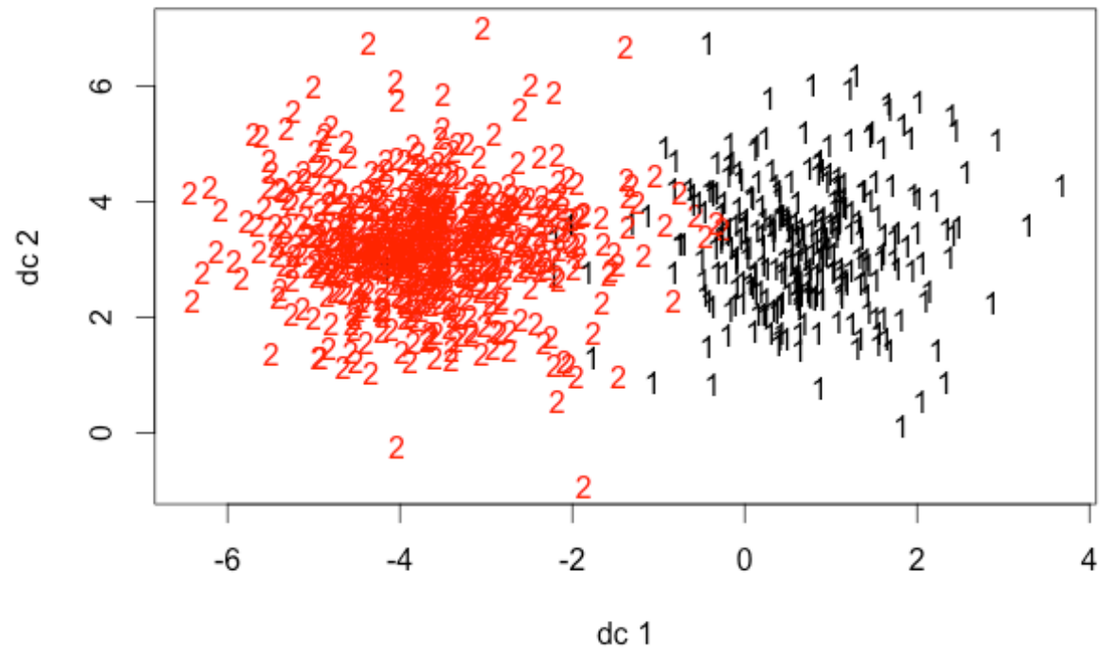
1 **2** **3**



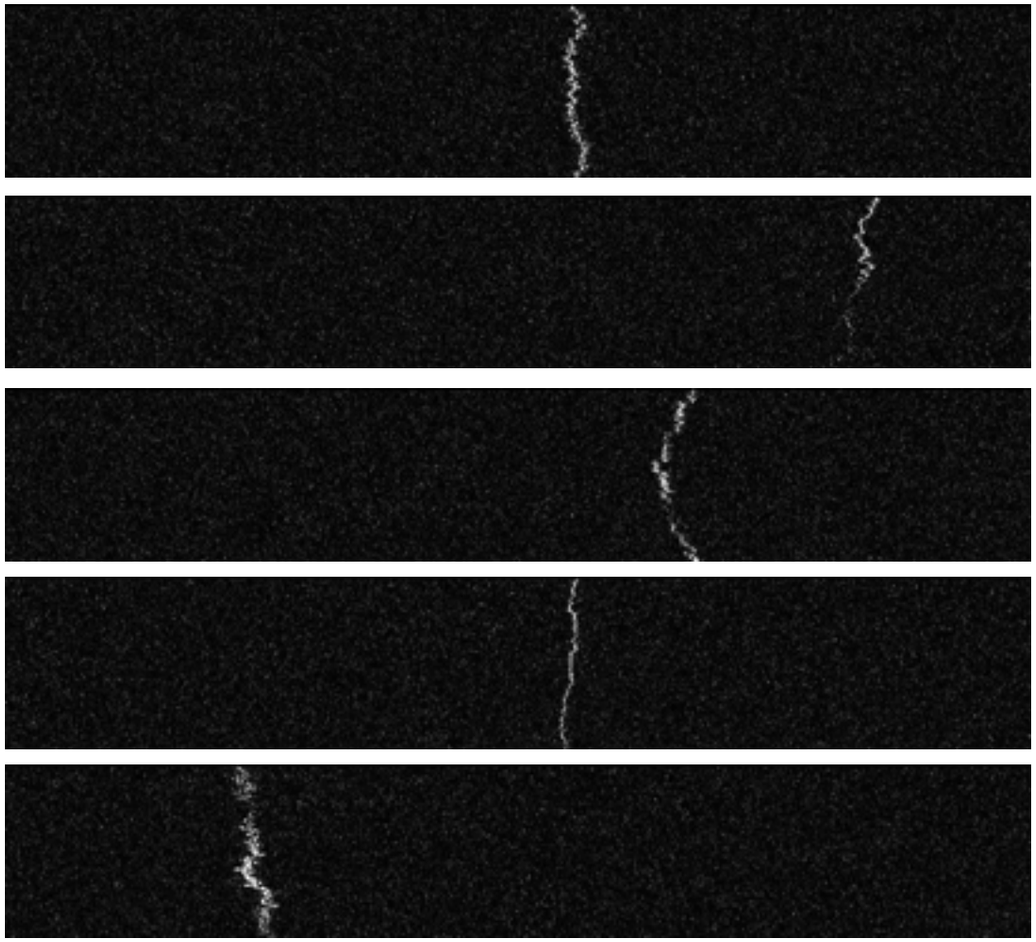
0.2302370 **0.4378288** **0.3623165**

Hierarchical
Ward D1: Euclidean



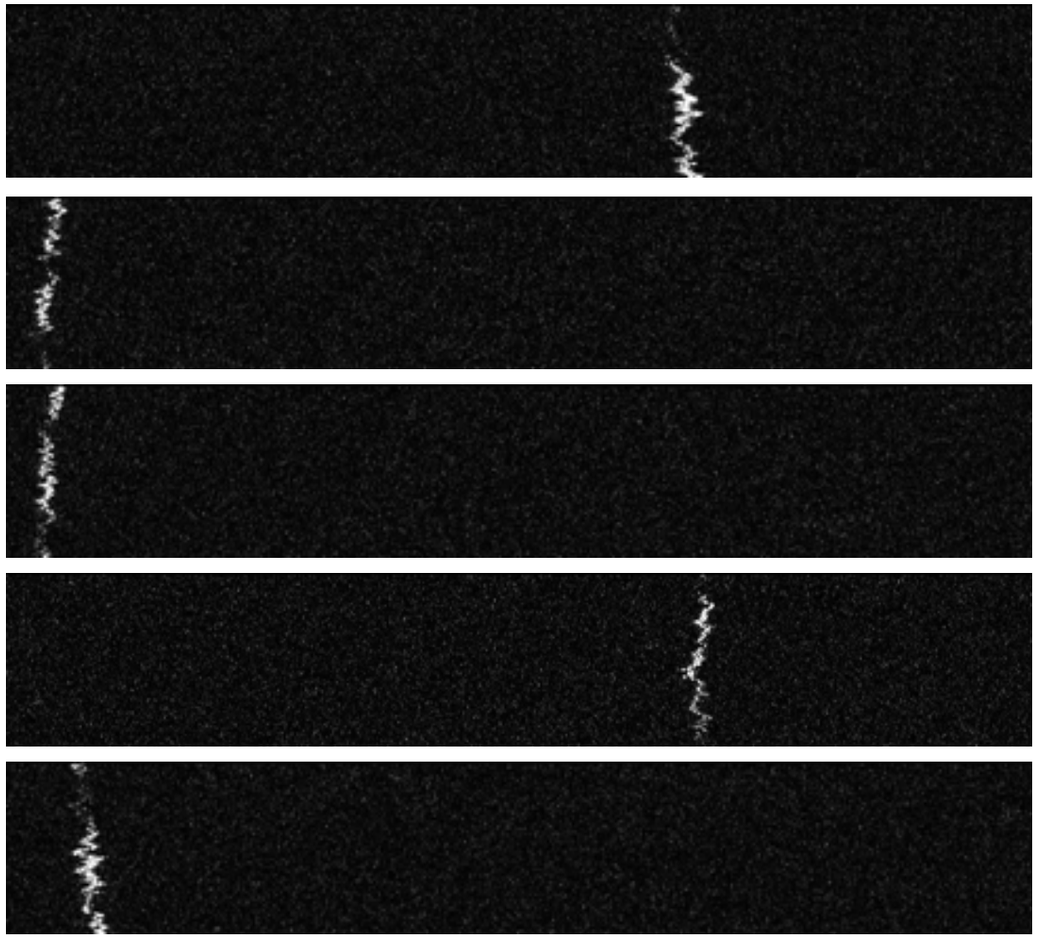


1



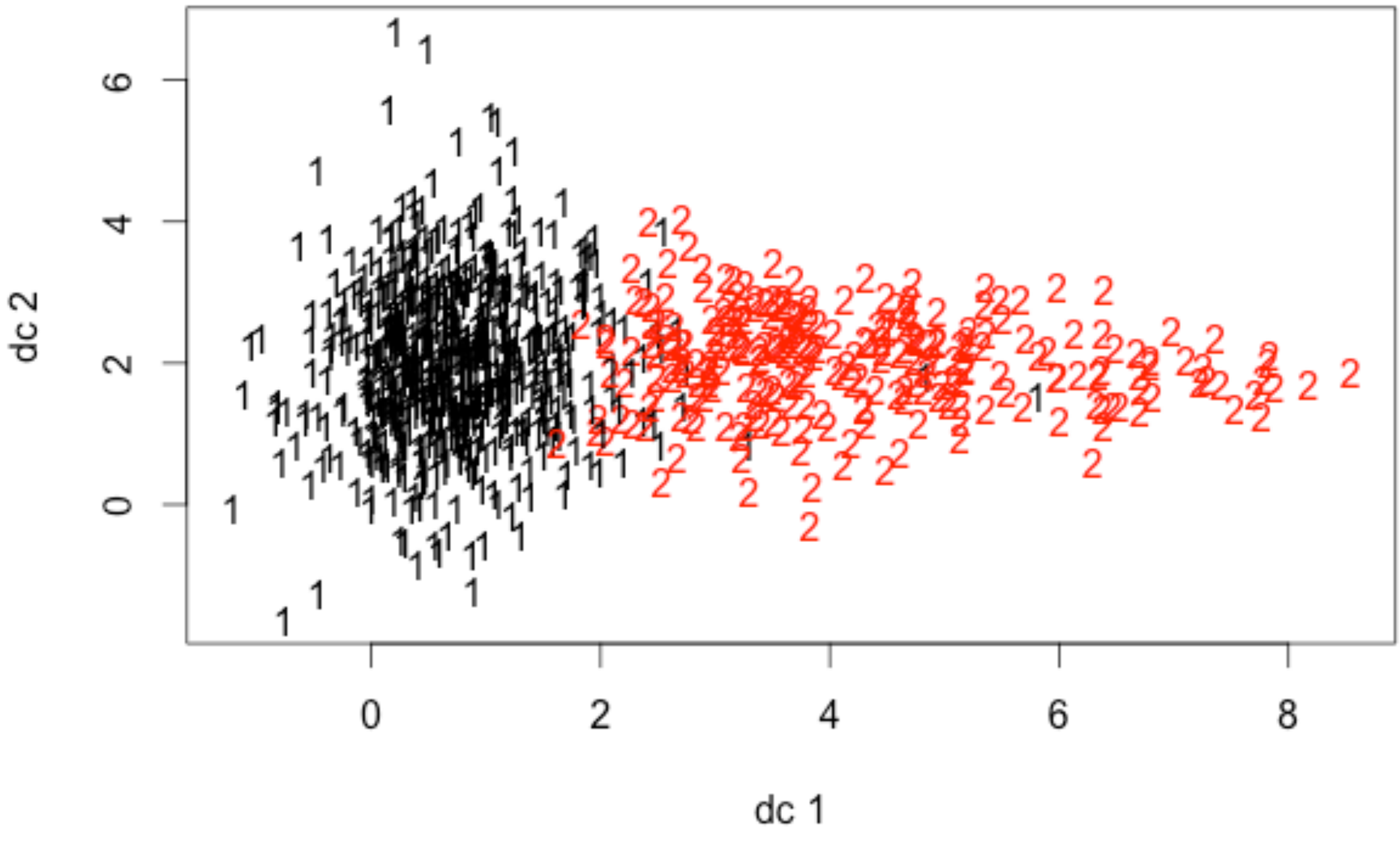
0.2568928

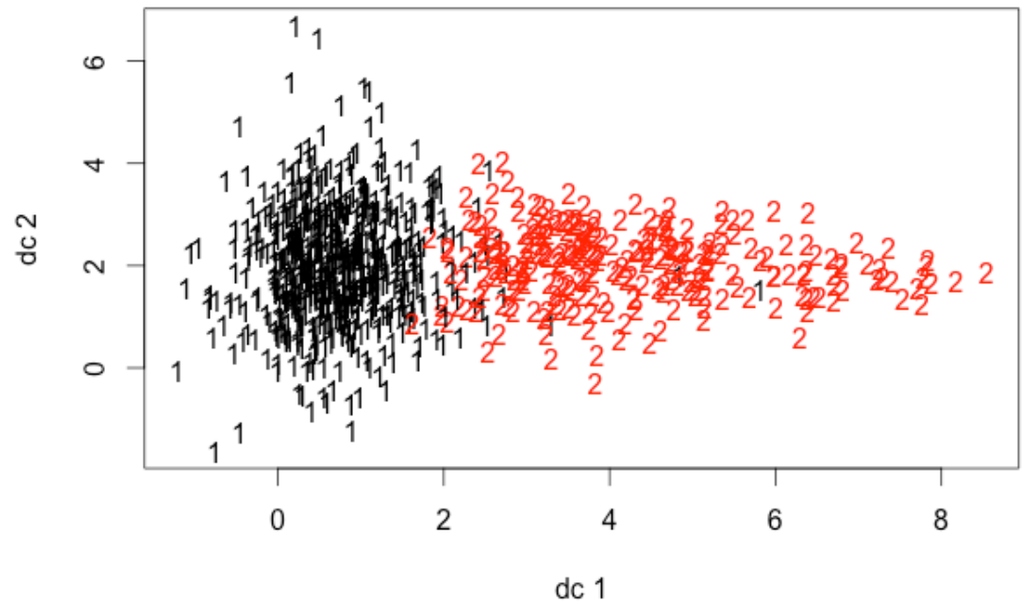
2



0.2811006

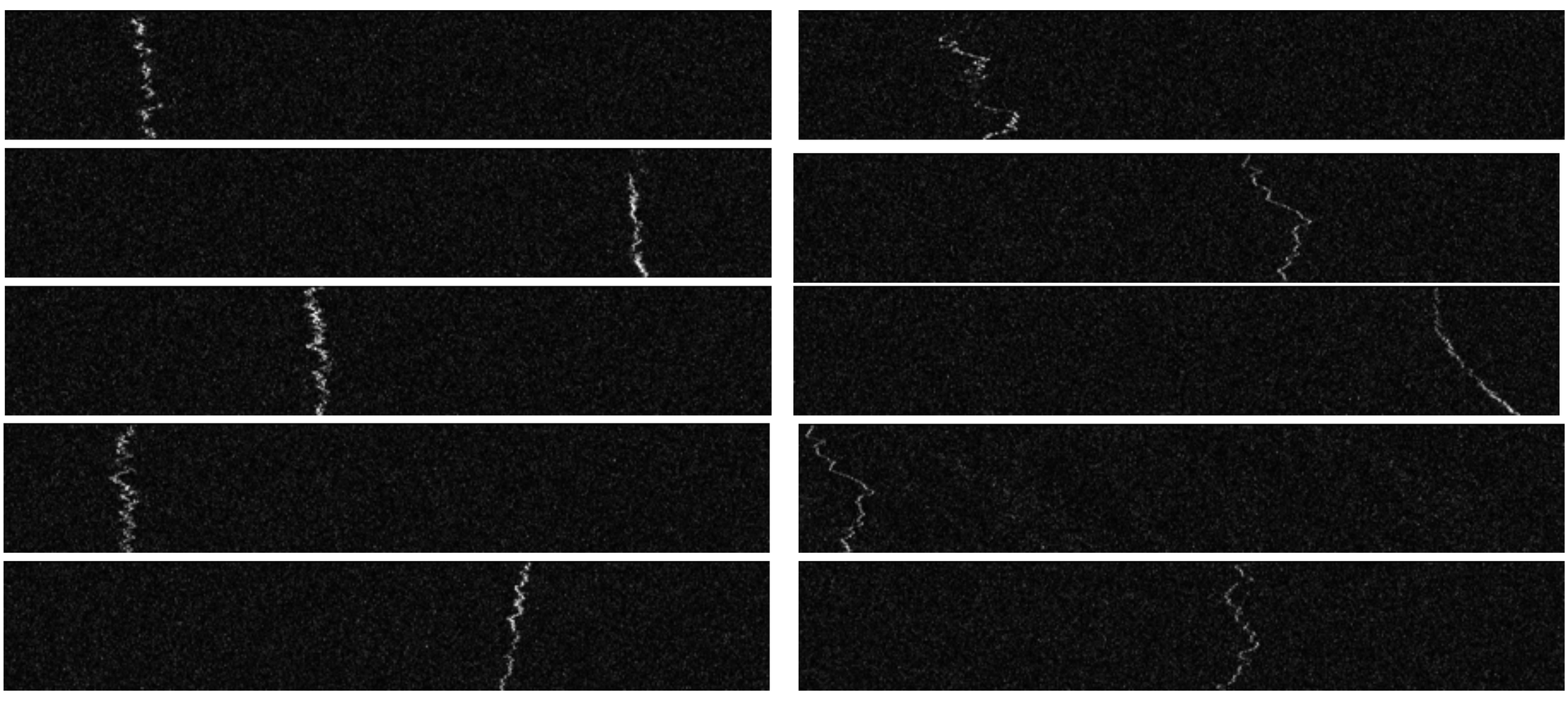
Hierarchical
Ward D2: Euclidean





1

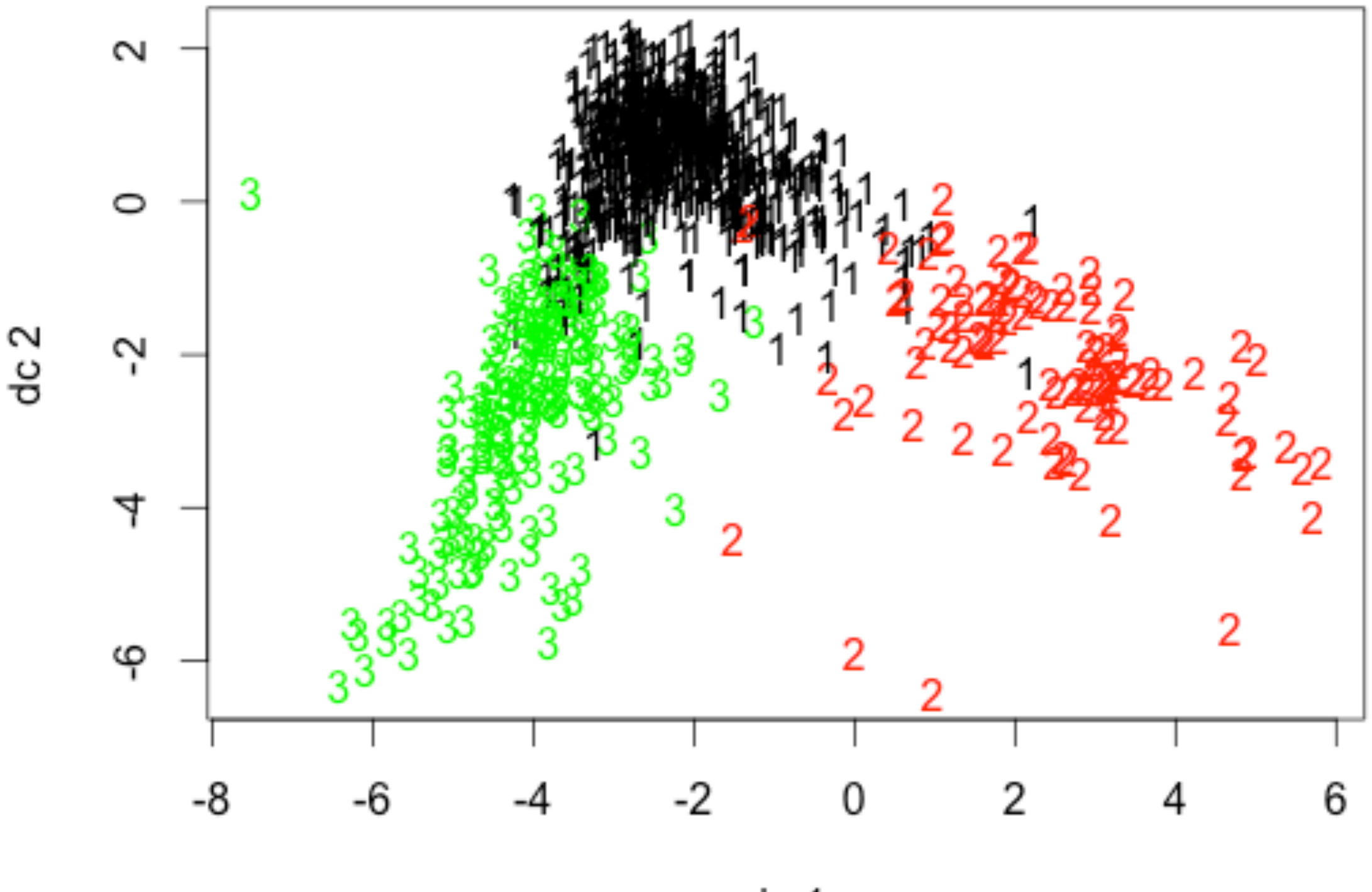
2

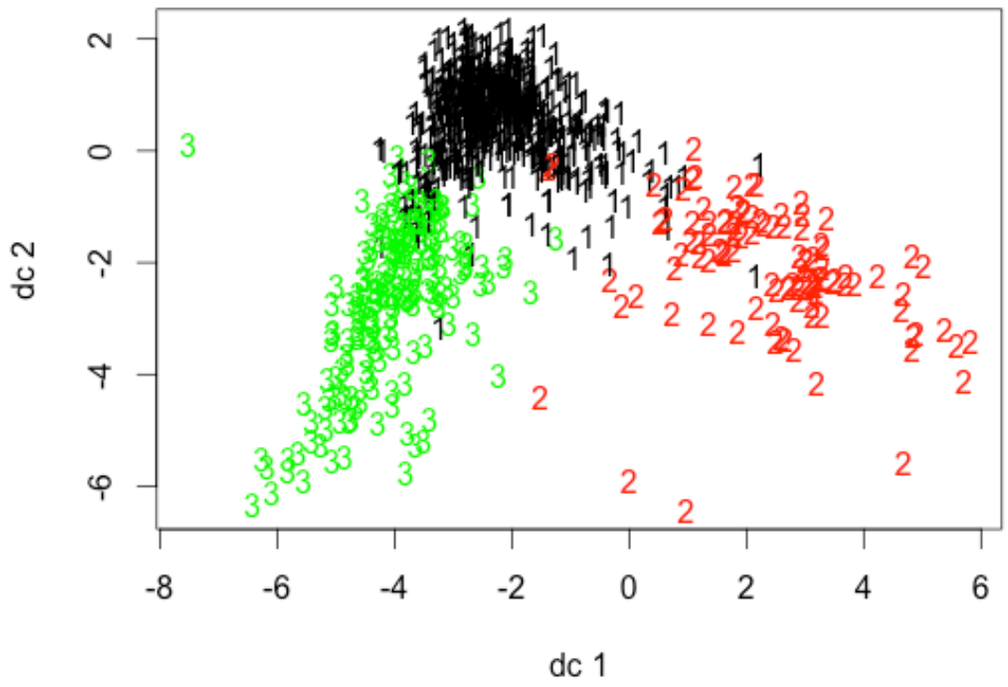


0.2528106

0.3820813

Hierarchical
Ward D2: Manhattan

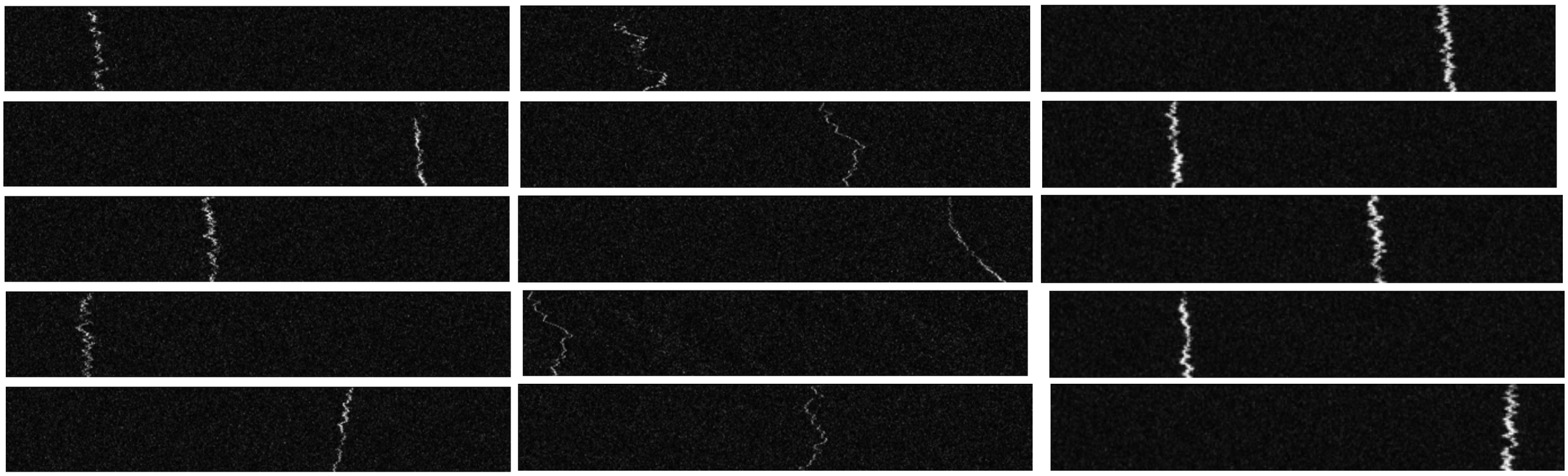




1

2

3

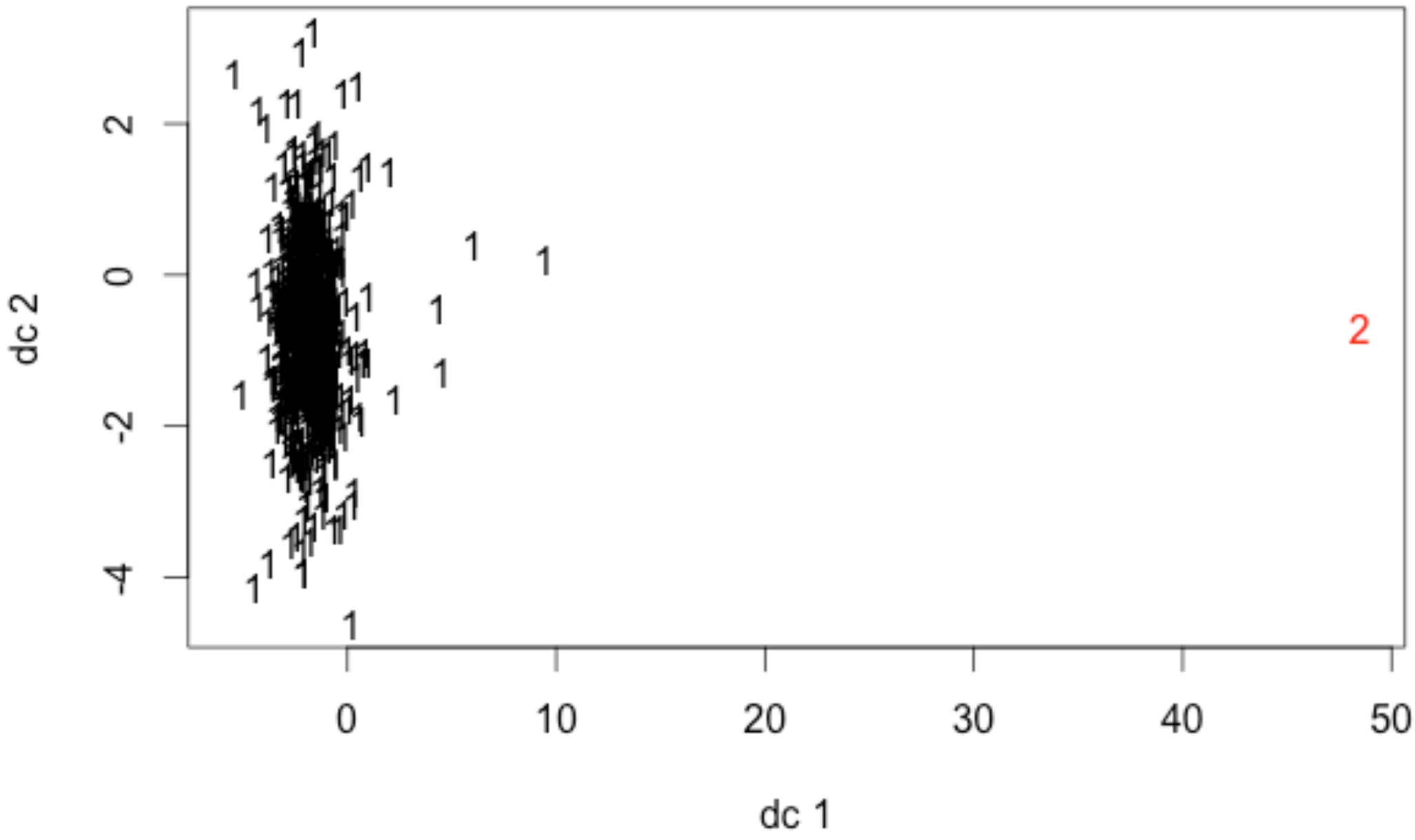


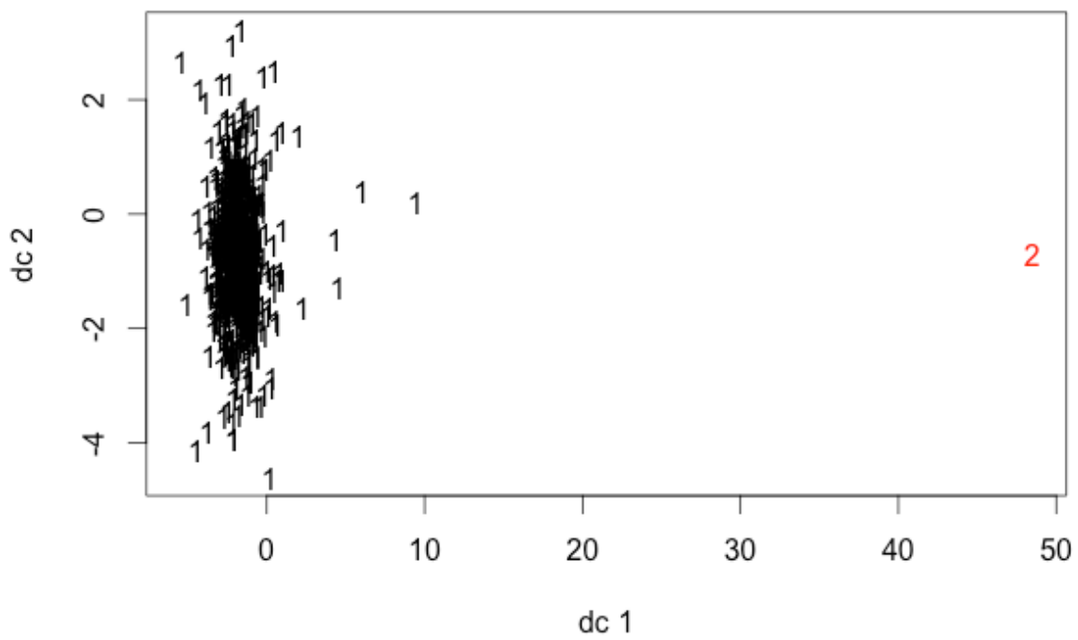
0.12713811

0.05949398

0.41644788

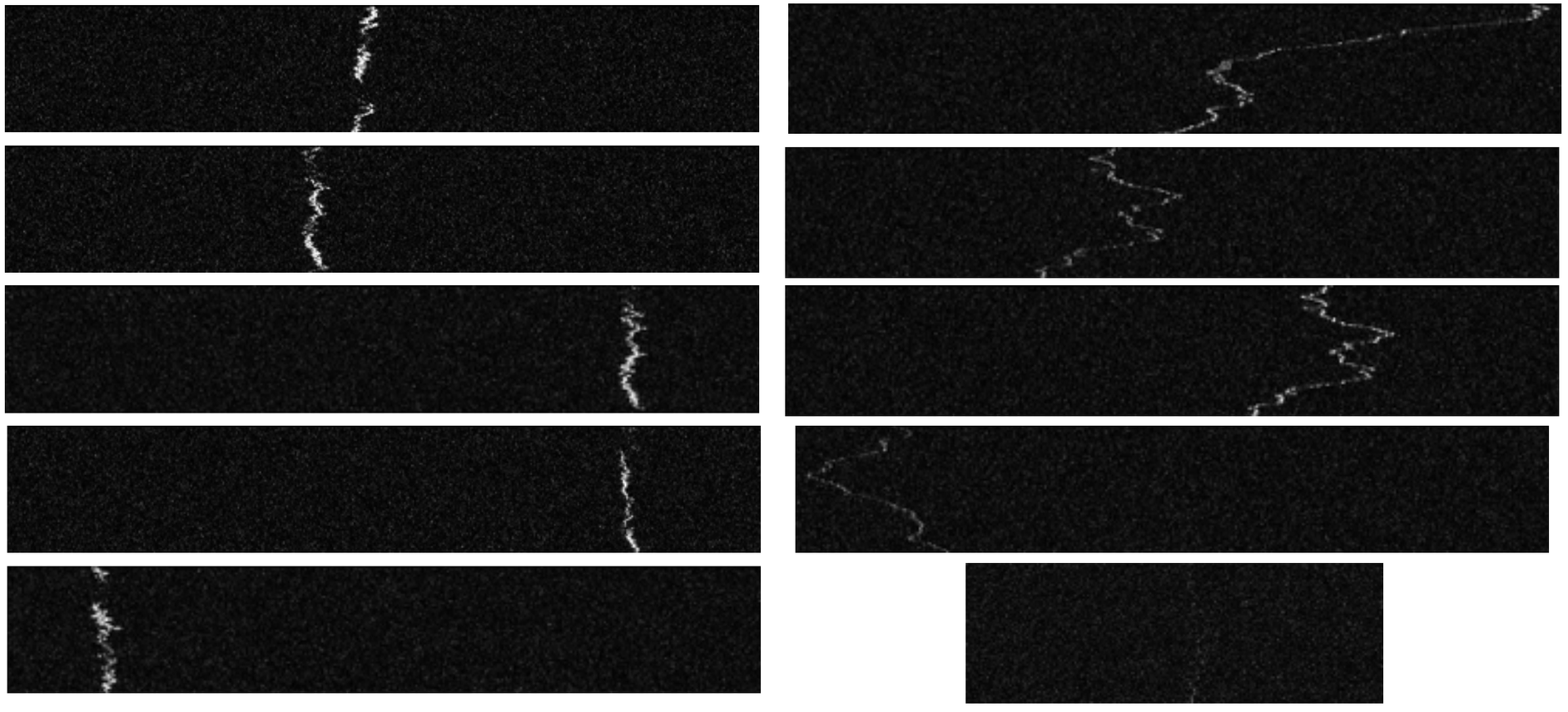
Divisive
DIANA: Euclidean





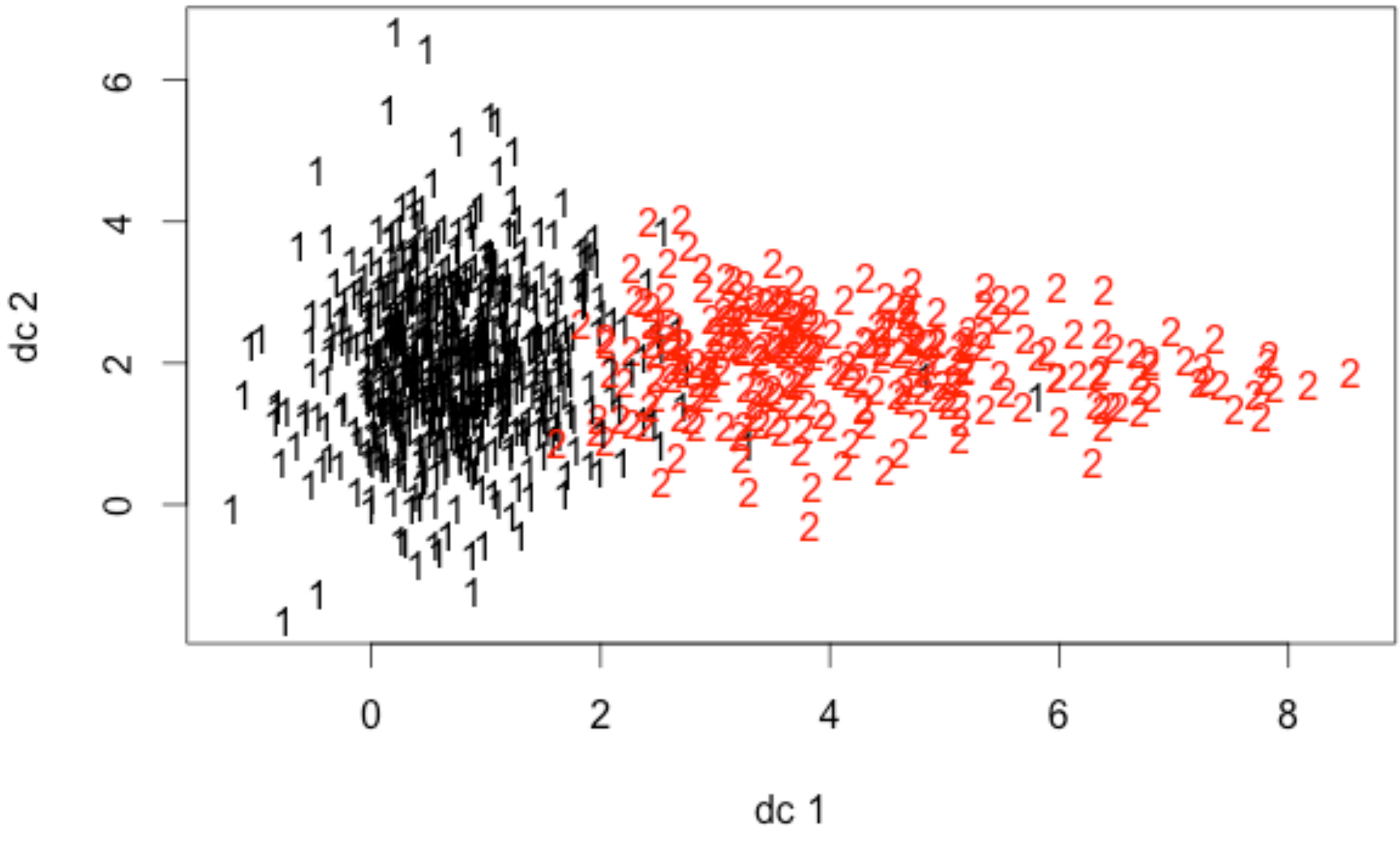
1

Near 2

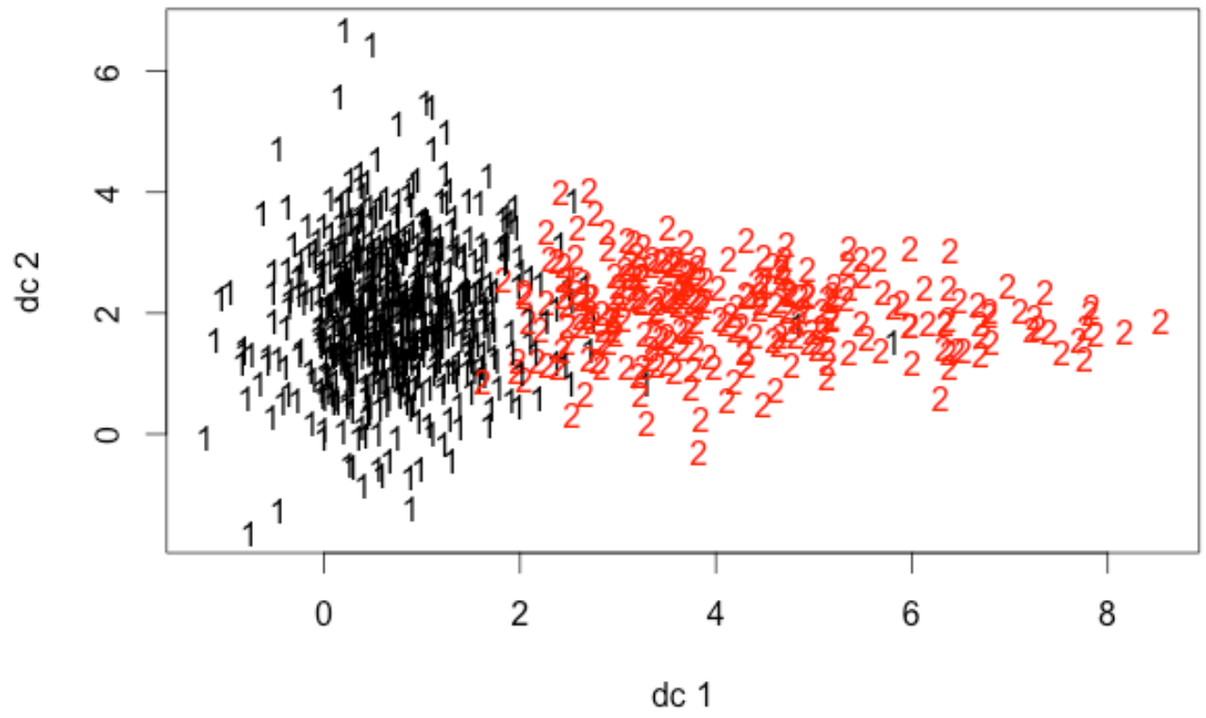


Revisit

Hierarchical: Ward D2



Insights
Chi-Square Test



Characteristic	p-value
August	5.75E-03
4 AM - 8 AM	2.31E-06
8 AM - 12 PM	5.79E-06
12 PM - 4 PM	7.32E-16
L-Polarization	1.24E-03
R-Polarization	3.14E-05