TOPOLOGICAL DATA ANALYSIS FOR CLOUD CLASSIFICATION

TYPES OF CLOUDS

Cirrus

Cumulus

Altocumulus

DATASET – CCSN DATABASE

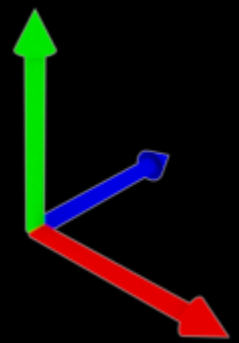
- Images of various clouds sorted by their type
- Unfortunately some images contain other objects such as buildings and trees.
- Solution – manually crop the images

TRANSFORMING IMAGES INTO GREYSCALE MATRIX

First Method – use blue channel

Save the number corresponding to the intensity of the blue light in the pixel and disregard the red and green channel.

Second Method – use PCA on all the pixels of the image



Find the linear combination of channels in which the pixels of the image differ the most, i.e. the first component given by PCA.

TRANSFORMING IMAGES INTO GREYSCALE MATRIX

Original image

Blue channel

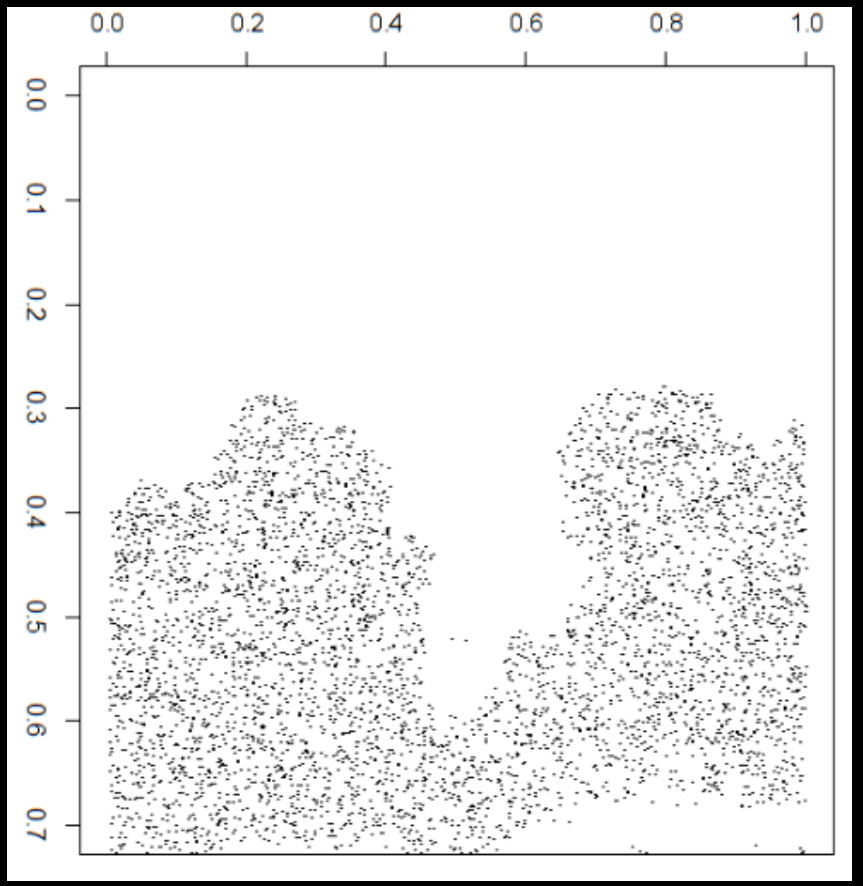
PCA

FINAL PREPROCESSING – TURNING BACKGROUND PLAIN BLACK

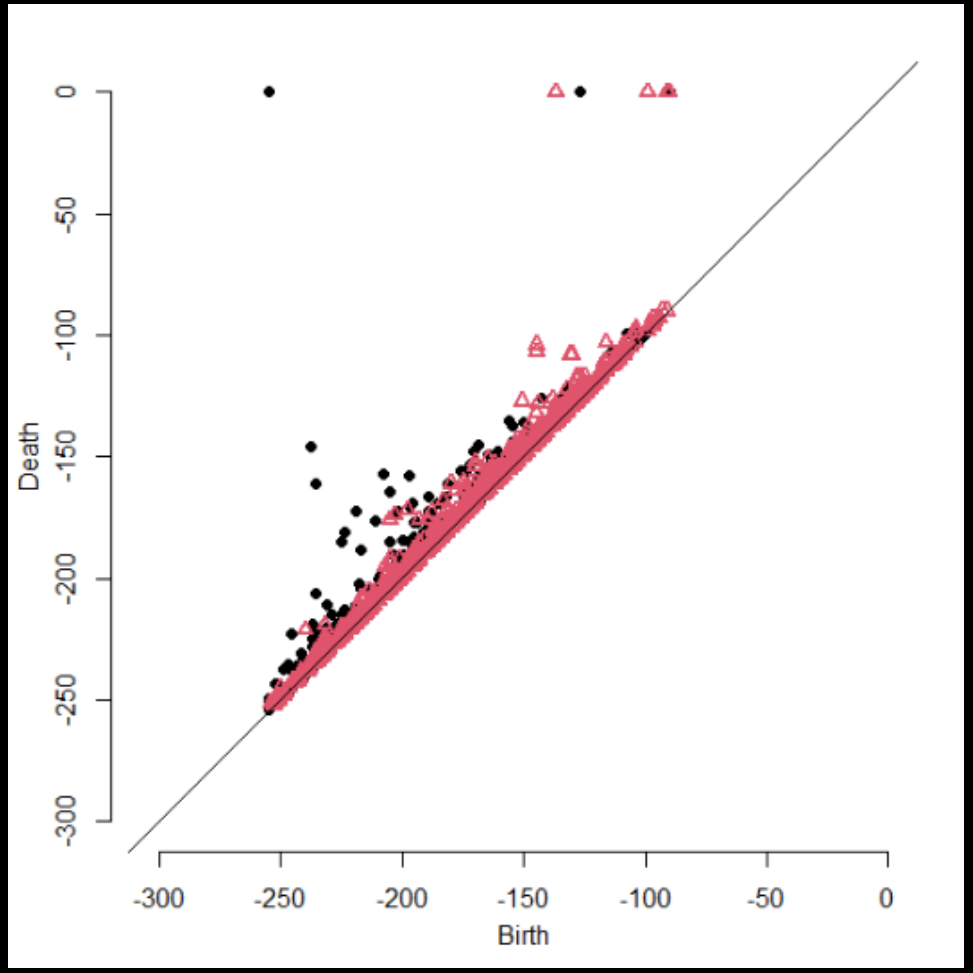
Procedure:

- Compute the mean greyscale value for all the pixels in the image.
- Set all the pixels with the value lower than the mean to 0.

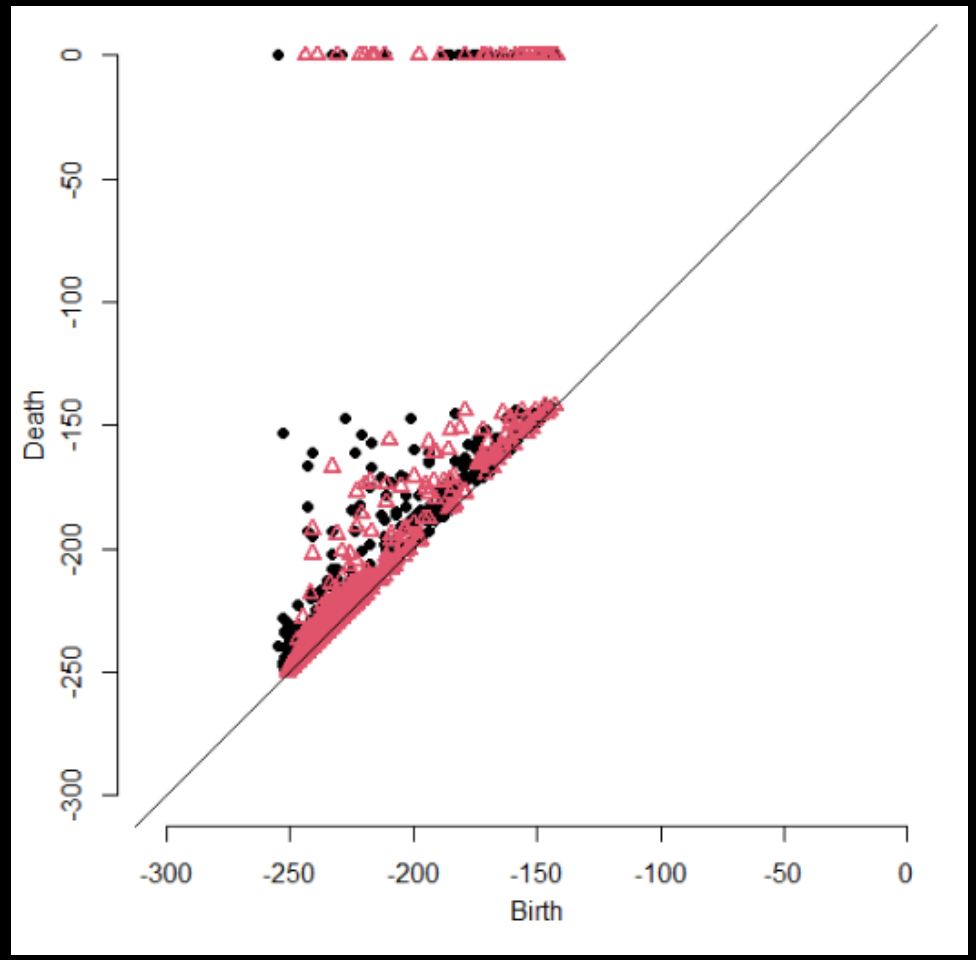
FILTERED COMPLEX



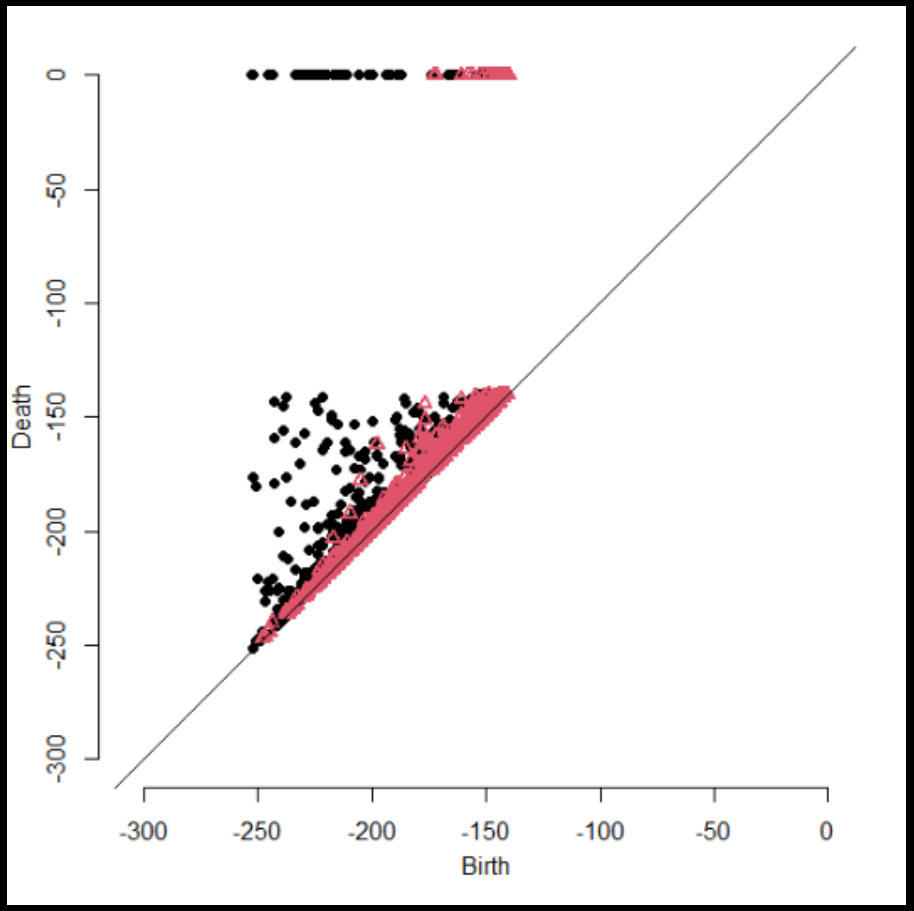
PERSISTENCE DIAGRAMS



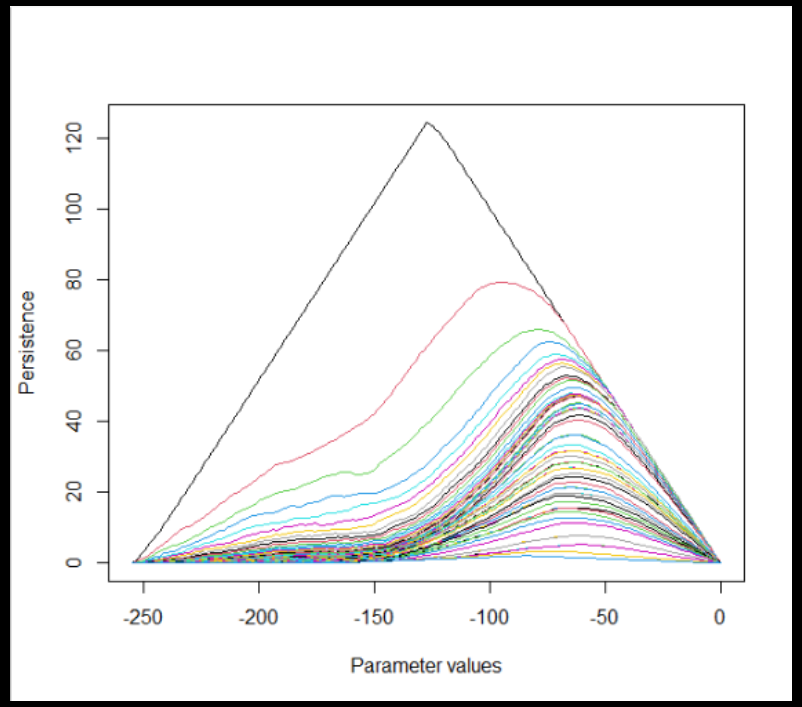
PERSISTENCE DIAGRAMS



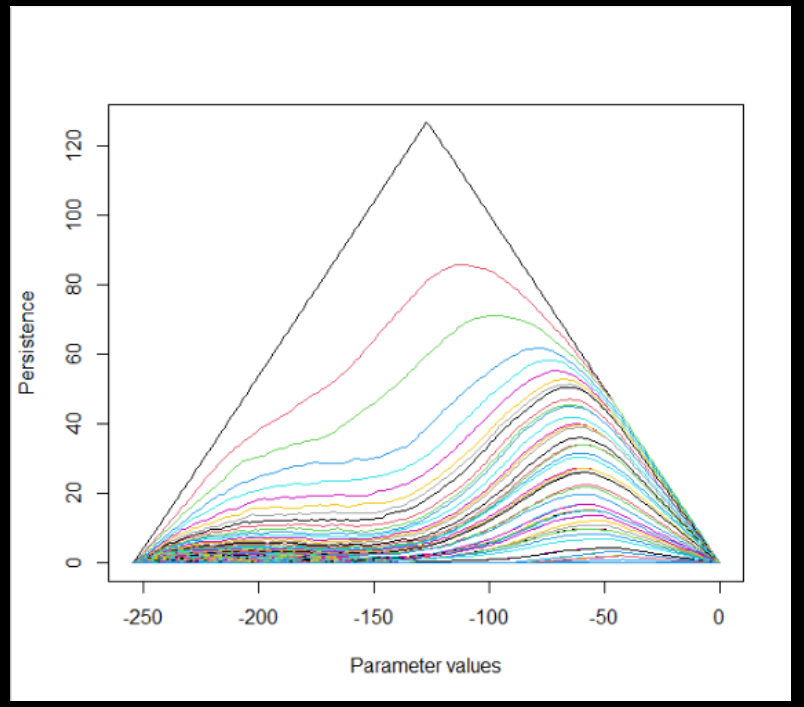
PERSISTENCE DIAGRAMS



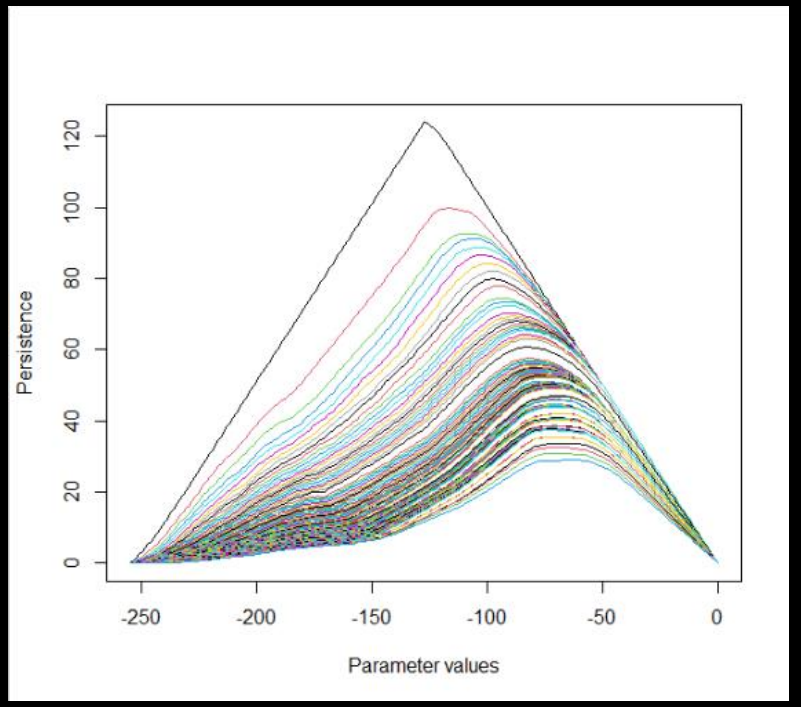
PERSISTENCE LANDSCAPES – HOMOLOGY IN DEGREE 0



Cirrus

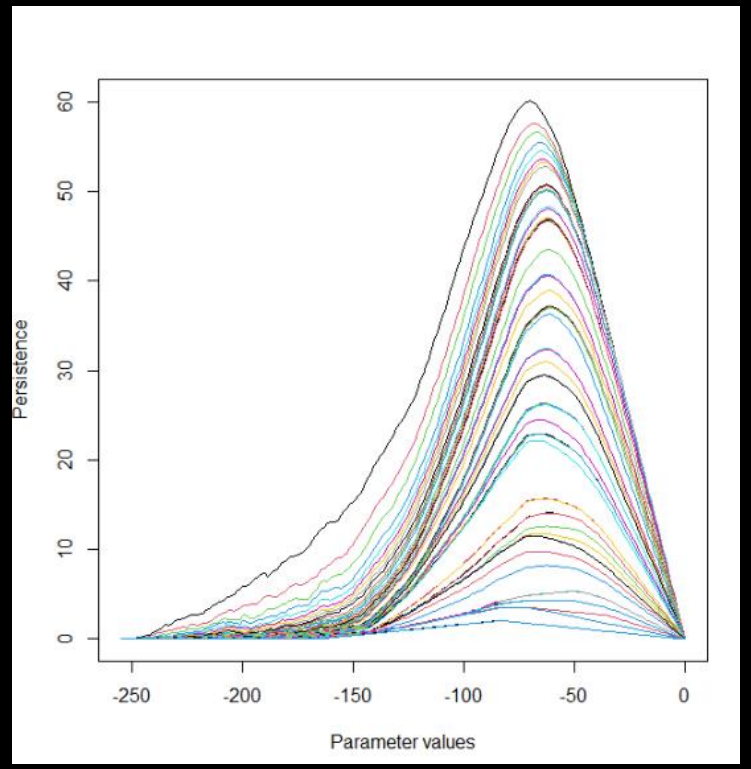


Cumulus

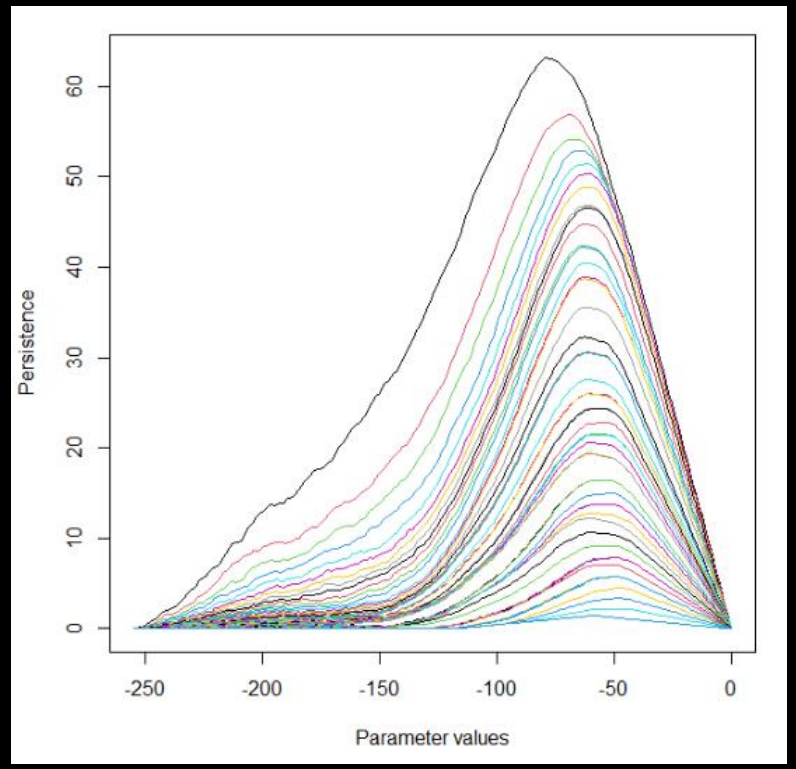


Altocumulus

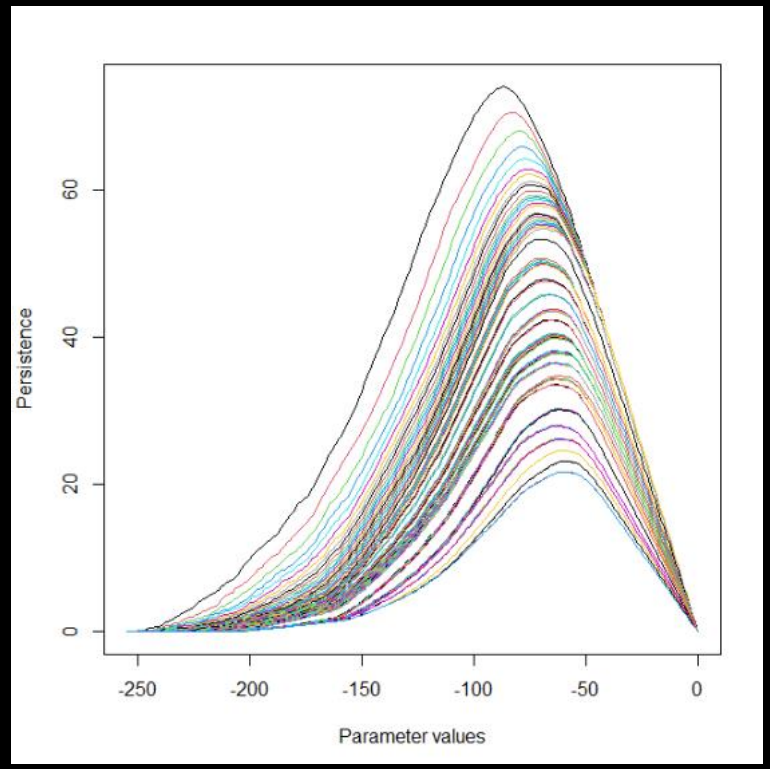
PERSISTENCE LANDSCAPES – HOMOLOGY IN DEGREE 1



Cirrus



Cumulus

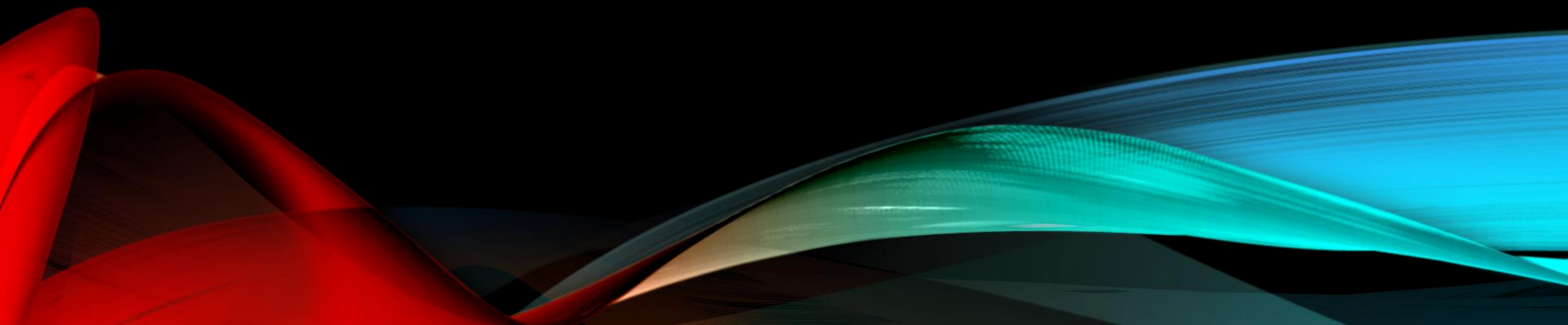


Altocumulus

PERMUTATION TEST

	Cirrus vs. Cumulus	Cirrus vs. Atlocumulus	Cumulus vs. Altocumulus
Homology degree 0	0.002	0	0
Homology degree 1	0.028	0	0

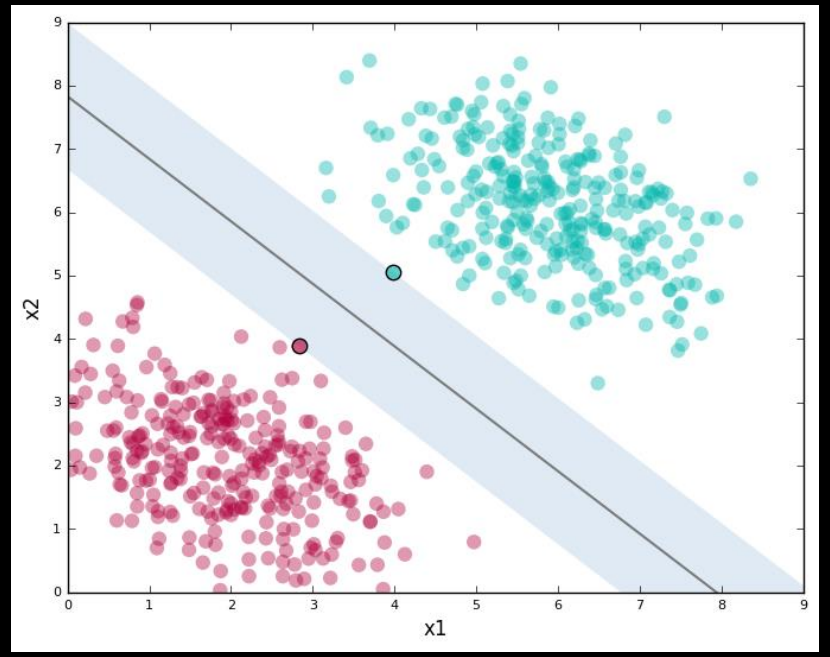
CLASSIFICATION ALGORITHMS



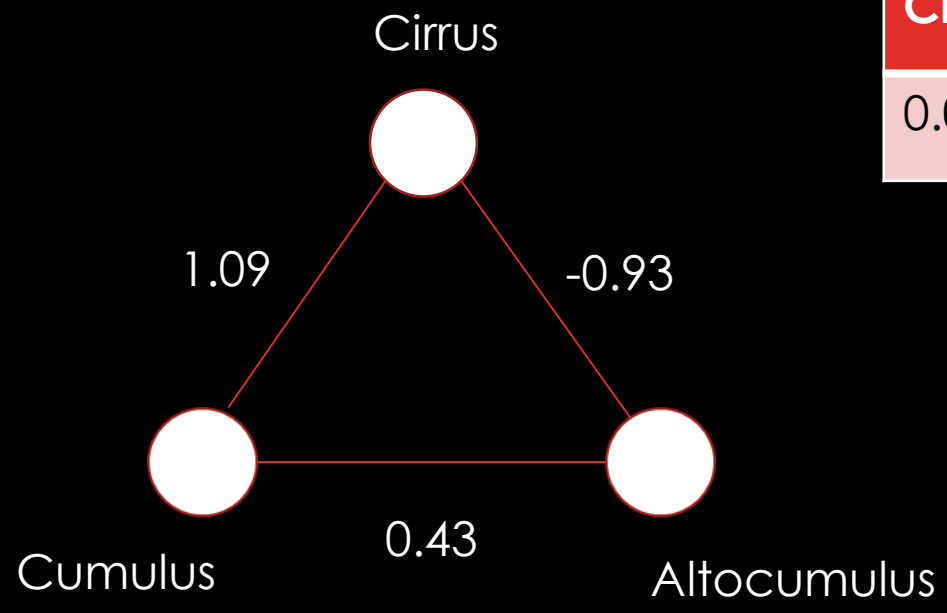
LINEAR SVM FOR PERSISTENCE LANDSCAPES OF HOMOLOGY IN DEGREE 0 AND 1

Standard procedure, but there is a problem.

SVM is a **binary classifier** and we have 3 categories here!



SOLUTION - VOTING SCHEME



Cirrus	Cumulus	Altocumulus
0.0	1.09	0.43 + 0.93 = 1.36

Look at classifiers for each pair of classes. Record the **decision value**.

LINEAR SVM - RESULTS

- Evaluation method: 10-fold cross validation applied several times.
- Error rate for homology in degree 0:
- Error rate for homology in degree 1:

ENSEMBLE METHOD

- Combine both methods (homology in degree 0 and 1) by taking a **weighted sum** of the **decision values** given by each of them.
- Goes along the lines of **Bayesian model averaging**.

Pros:

- Higher stability
- Lower variance
- Potentially the biases can cancel each other out.

Cons:

- Lower interpretability of the model

ENSEMBLE METHOD

In our case:

v_0, v_1 - decision values obtained from, resp. homology in degree 0 and 1.

v – decision value of the ensemble method.

where

$$v = 0.55 \cdot v_0 + 0.45 \cdot v_1$$

ENSEMBLE METHOD

In our case:

v_0, v_1 - decision values obtained from, resp. homology in degree 0 and 1.

v – decision value of the ensemble method.

where

$$v = 0.55 \cdot v_0 + 0.45 \cdot v_1$$

Error rates

SVM for homology in degree 0: 0.34

SVM for homology in degree 1: 0.39

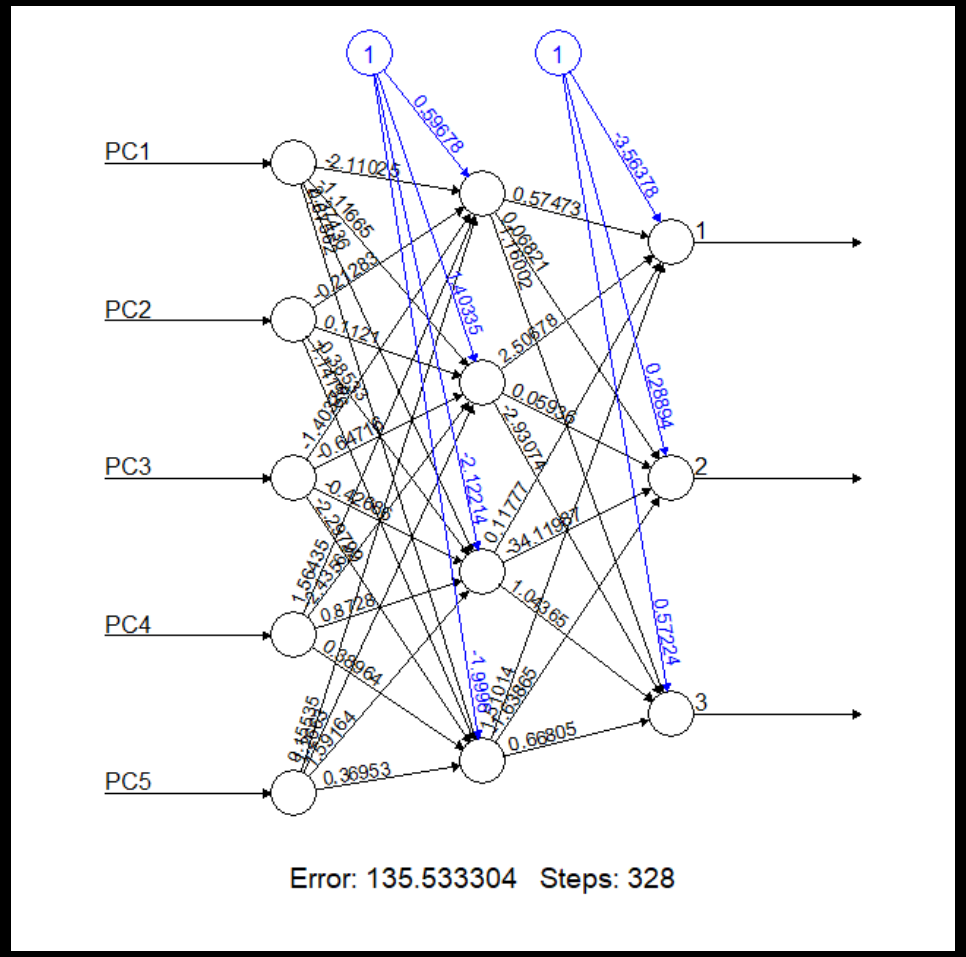
Ensemble method: 0.28

ARTIFICIAL NEURAL NETWORK

Error function: **cross-entropy**

Optimization algorithm: **resilient backpropagation** with weight backtracking.

Input: first 5 components from PCA on persistence landscapes for homology in degree 0



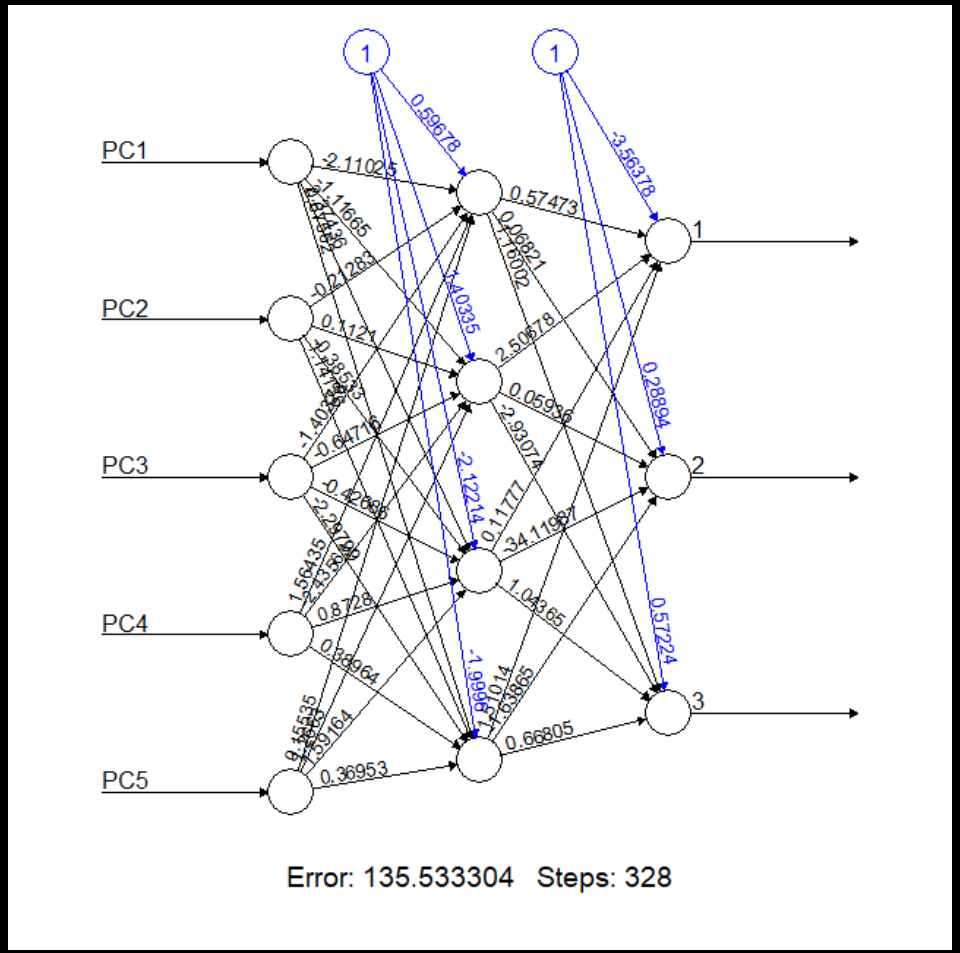
ARTIFICIAL NEURAL NETWORK

Error function: **cross-entropy**

Optimization algorithm: **resilient backpropagation** with weight backtracking.

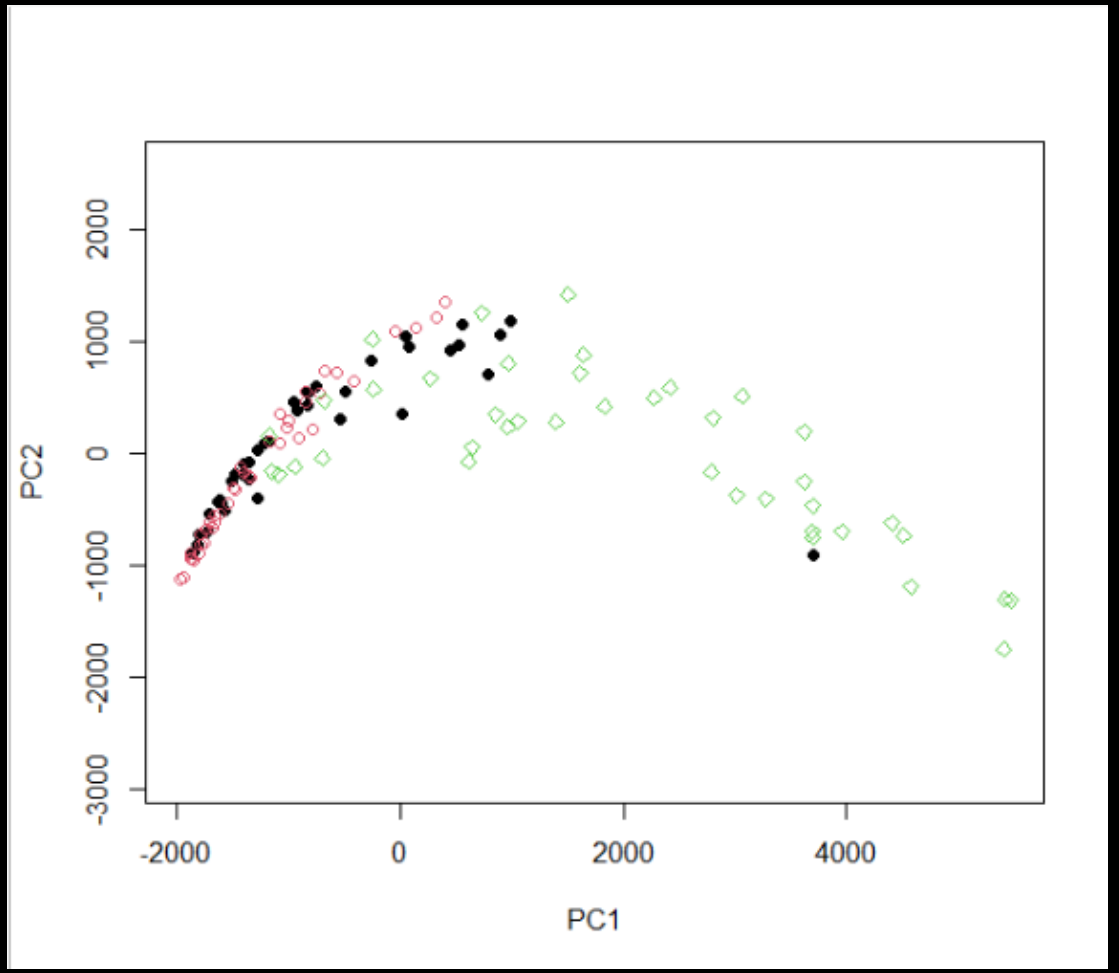
Input: first 5 components from PCA on persistence landscapes for homology in degree 0

Error rate: 0.38



PCA SHORTCOMINGS

- PCA does not take into account the labels of the data points.
- A lot of spread and scatter of the data but not necessarily good separation of the data.
- The “centers of mass” of the classes are very close to each other compared to the spread of the data points.
- Consequence: impossible to create good separation curves.



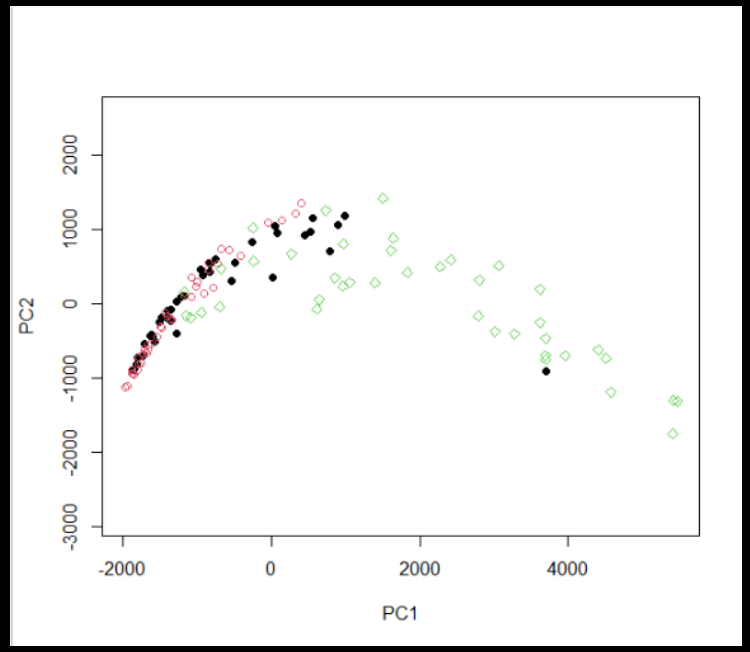
CLASSIFICATION ALGORITHMS SUMMARY

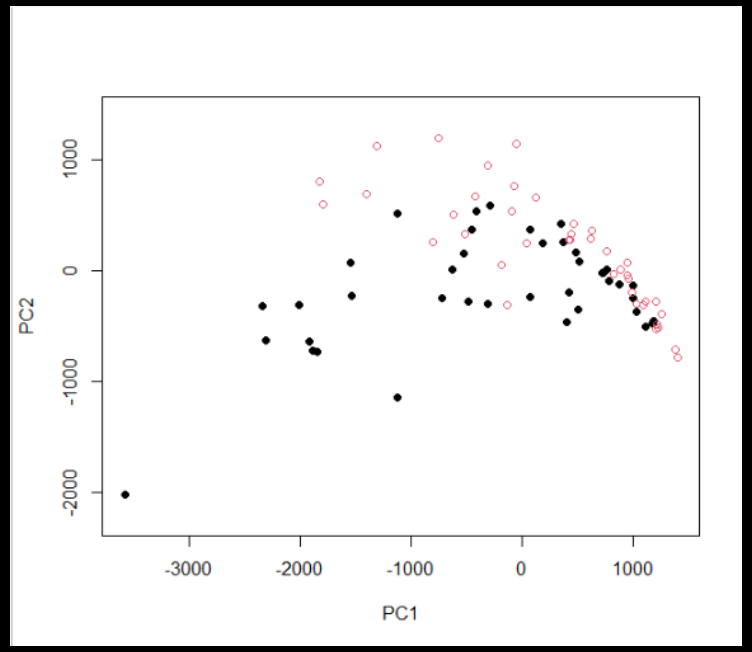
Algorithm	Linear SVM for homology in degree 0	Linear SVM for homology in degree 1	Ensemble method with both SVMs	Artificial Neural Network
Error rate	0.34	0.39	0.28	0.38

FURTHER INVESTIGATION – NEURAL NETWORK

Build 3 binary classifiers based on neural networks each using PCA on its two classes of data. Then use the voting scheme

- Better separation of the classes when we only have two of them

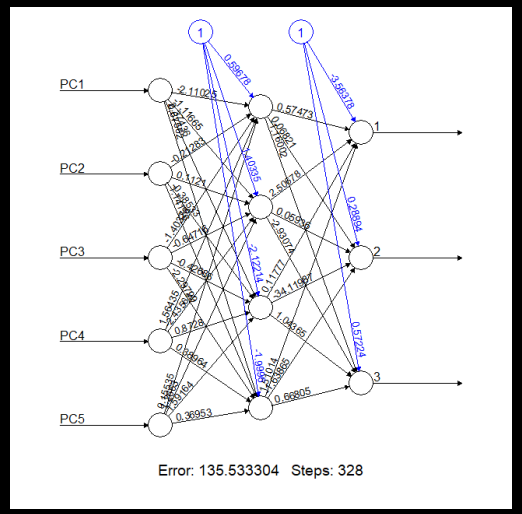




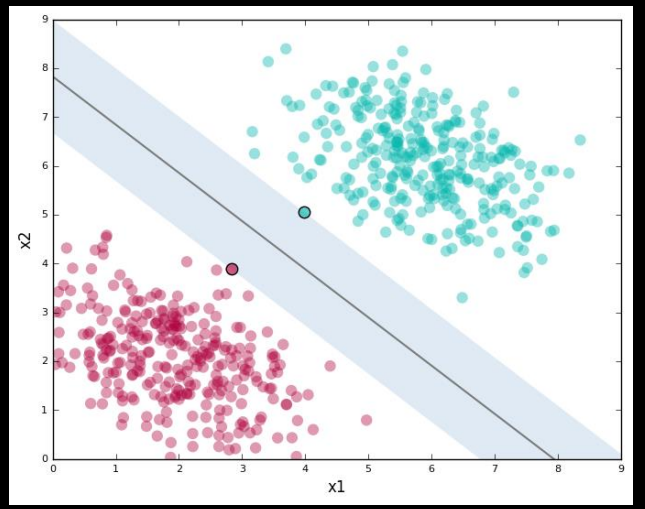
FURTHER INVESTIGATION – NEURAL NETWORK

Incorporate the neural network into the ensemble method.

Neural network is a very different model from a linear SVM. Their predictions might have low covariance and they can have opposite biases – perfect scenario for the ensemble method.

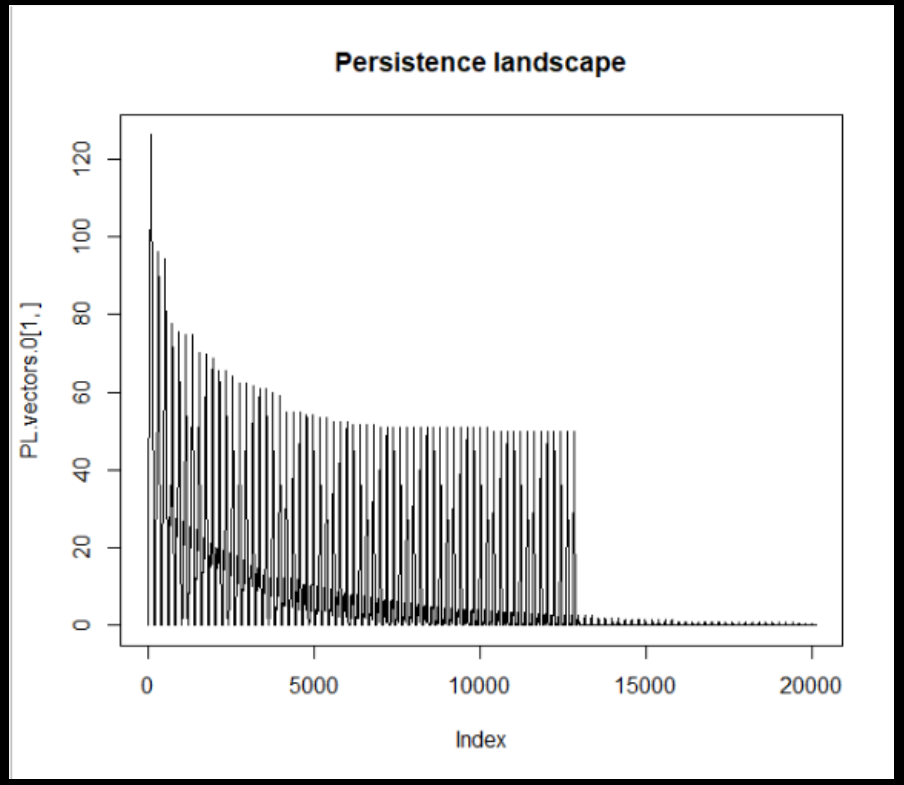


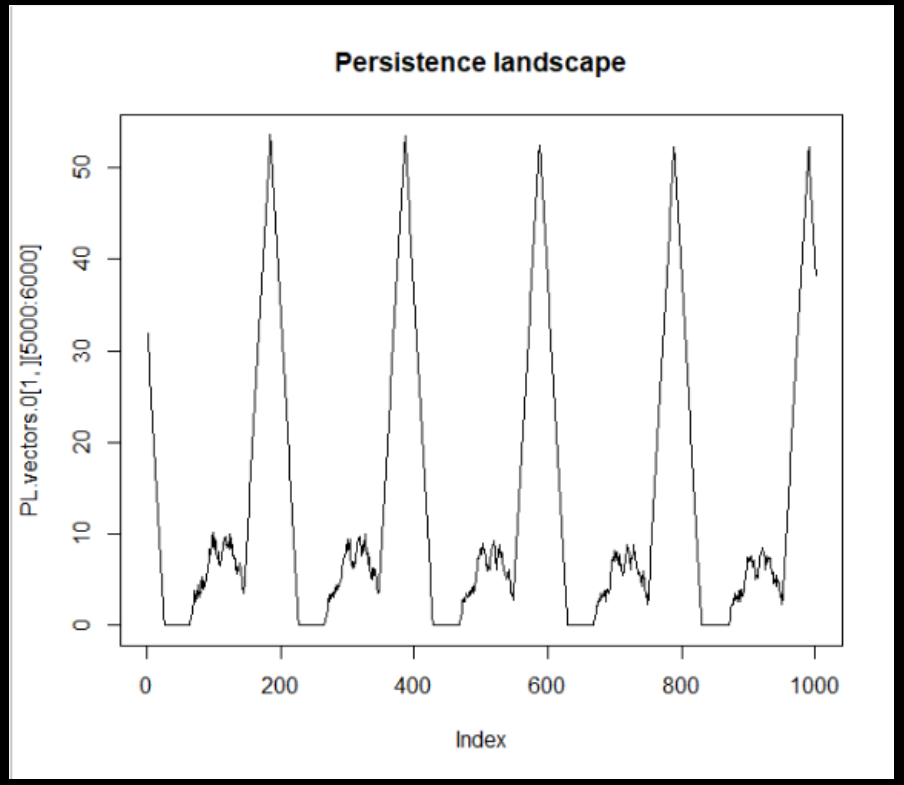
+



FURTHER INVESTIGATION – SIGNAL PROCESSING

Treat persistence landscapes as a time signal.

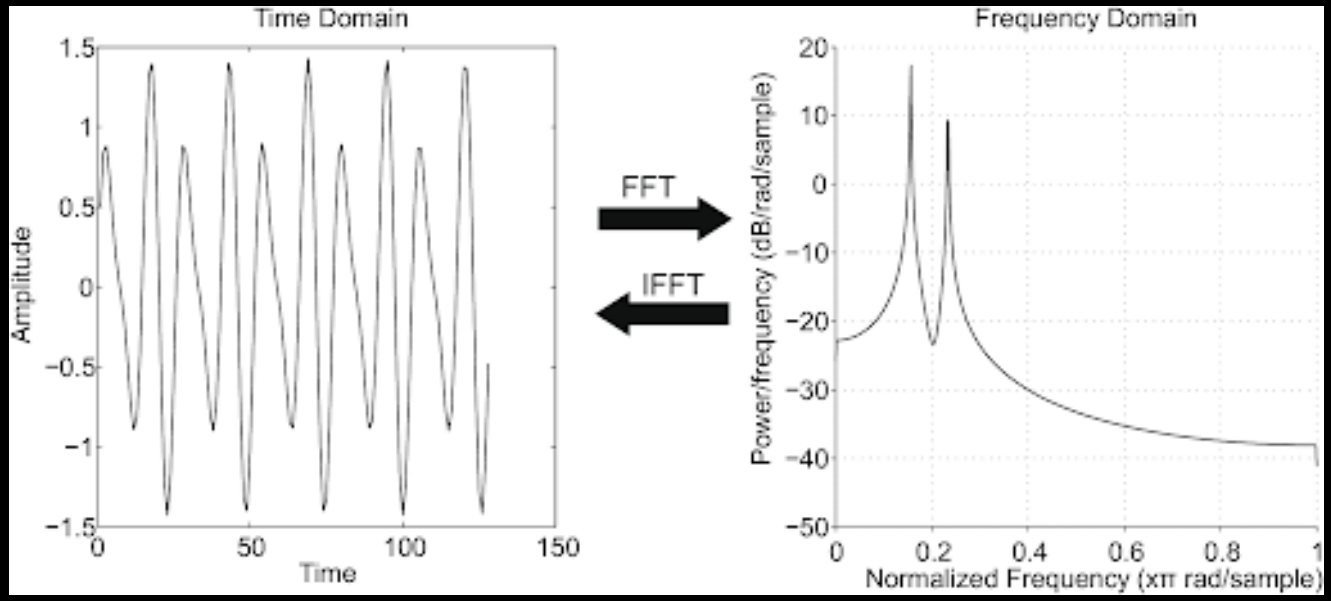




FURTHER INVESTIGATION – SIGNAL PROCESSING

Apply Fourier transform to persistence landscape

We might be able to get rid of thousands of highly correlated covariates which do not contribute almost any useful information.



CONCLUSION

- Topological Data Analysis is certainly capable of producing classifiers for the cloud types which work with reasonable accuracy.
- The biggest problem seems to be reducing the dimensionality of the persistence landscape in order to successfully apply more sophisticated algorithms such as artificial neural networks.

REFERENCES

- CCNS database

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CADDPD

- Wikipedia:

https://en.wikipedia.org/wiki/Fourier_transform

https://en.wikipedia.org/wiki/Ensemble_learning

- Fragoso, Tiago & Bertoli, Wesley & Louzada, Francisco. (2018). Bayesian Model Averaging: A Systematic Review and Conceptual Classification. International Statistical Review. 86. 1-28. 10.1111/insr.12243.

THANK YOU!