This week, I read the article in depth and sorted out the main logic of the code. And I started to write the code.

The problem I meet: How to use the data?

The amount of data is huge, should I train with some of the data, train better results, and then go test the whole?

Or should I train with the whole data?

Should I choose from all h5 files (nearly 1T), or a particular h5 file (20GB) ?

Literature Review

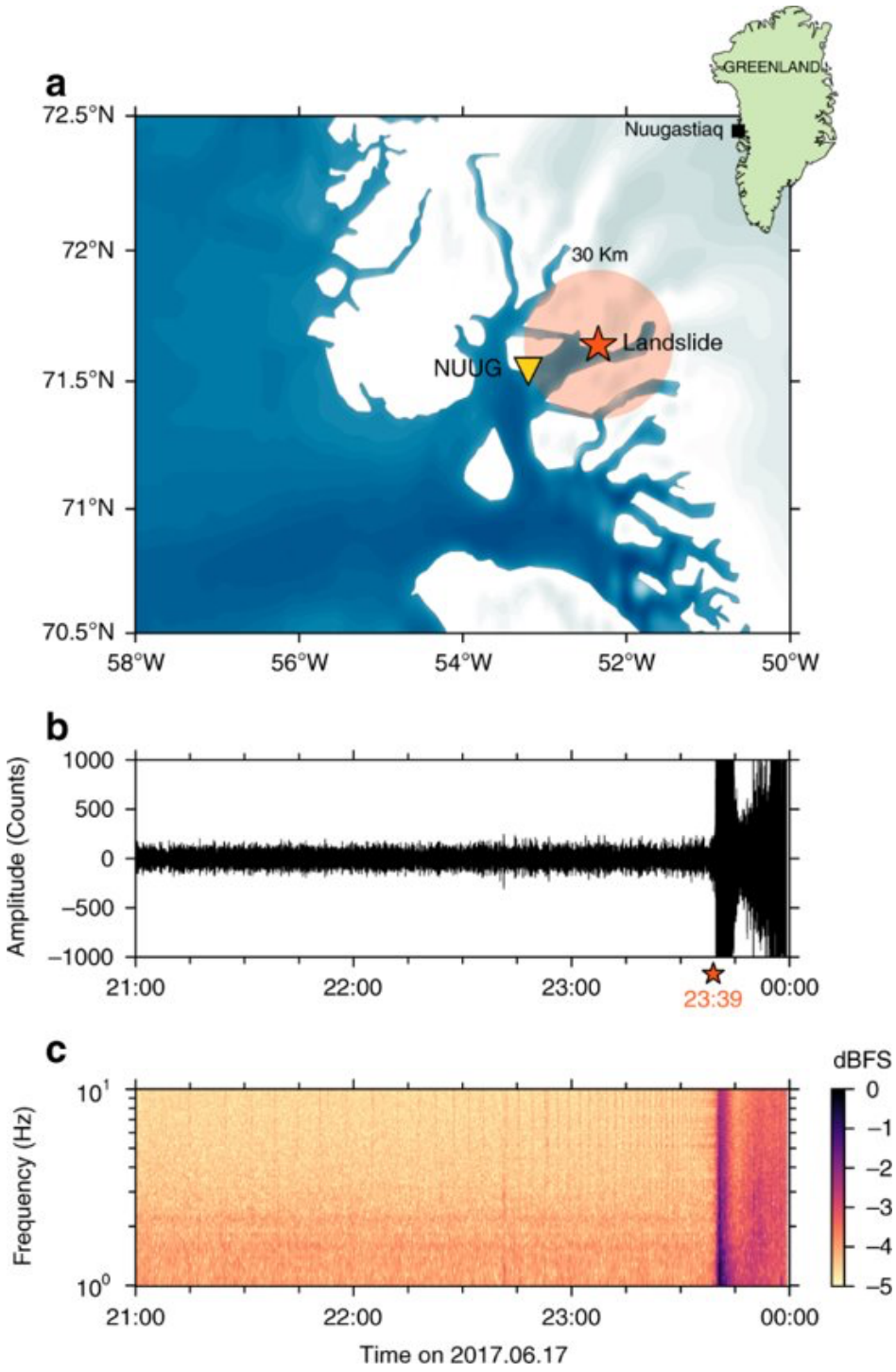
Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning

Background

Seismic signal detection has always been a concern, but with the increase in data volume, there are many challenges

- The amount of data increases, and the original method model cannot be applied to large data
- The original are supervised models, which need to label the data, and unsupervised learning methods are needed
- more small problems will be encountered in practice. For example, the seismic signal is very small and the noise is very large.

Example： Seismic records of the 2017 Nuugaatsiaq landslide, not directly visible and weak amplitude



Three-component template matching

The fundamental principle of the template-matching strategy is that seismic events exhibit certain similarities in waveforms, especially those within the same seismic zone.

- Choose one or more templates, which are usually the waveforms of known seismic events.
- Cross-correlate these templates with continuous seismic data.
- If the similarity between a segment of the continuous data and the template exceeds a preset threshold, it is considered that a seismic event similar to the template has been found.

Disadvantages

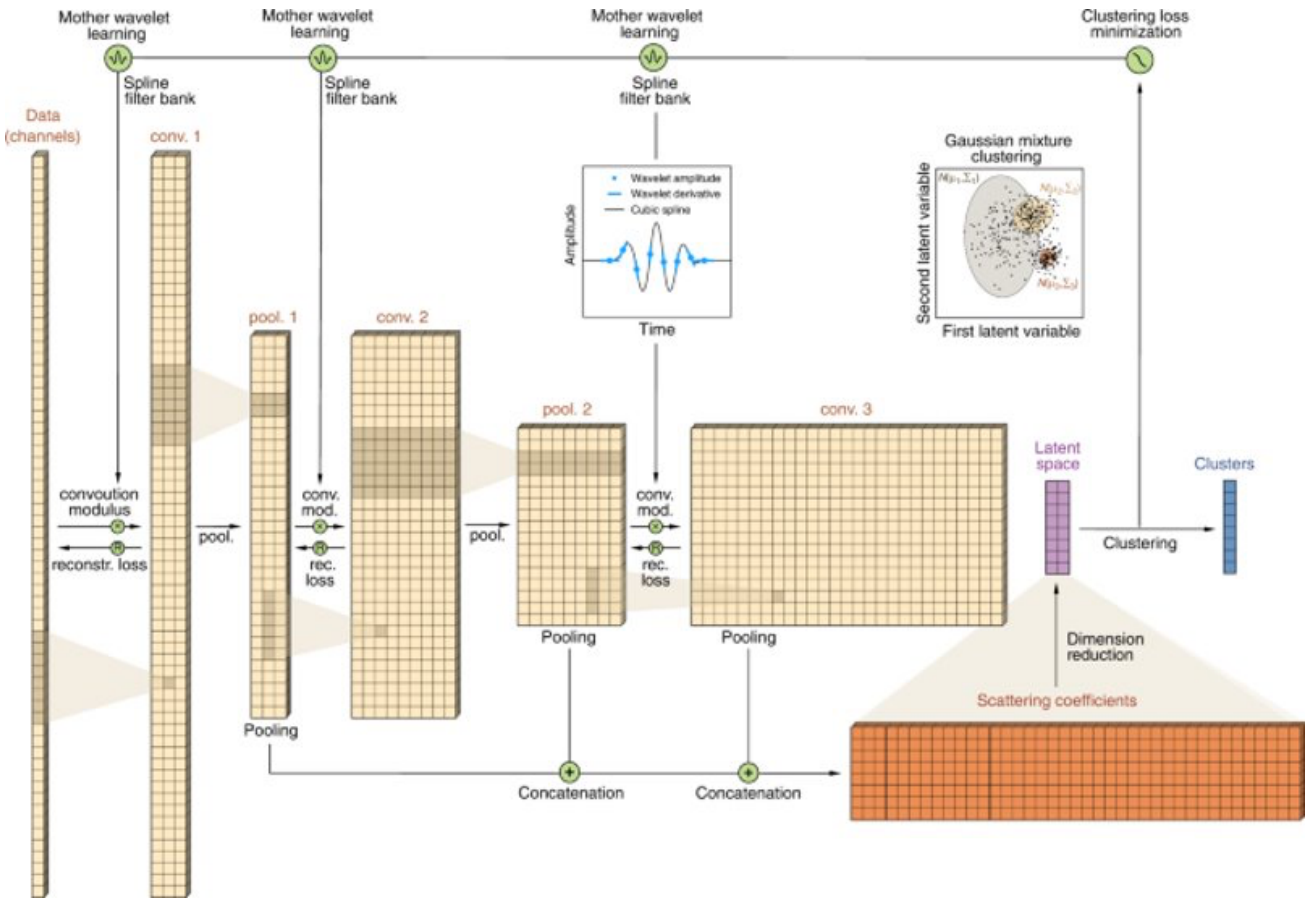
- Dependence on template quality: The template-matching strategy highly relies on the quality of the template (including the duration, frequency band, etc.). If the template selection is inappropriate, it may lead to inaccurate detection results.
- Large computational load: Template matching requires cross-correlation calculations for every possible seismic event. If the amount of data is large, the computational load can be very large, which may pose challenges in practical applications.

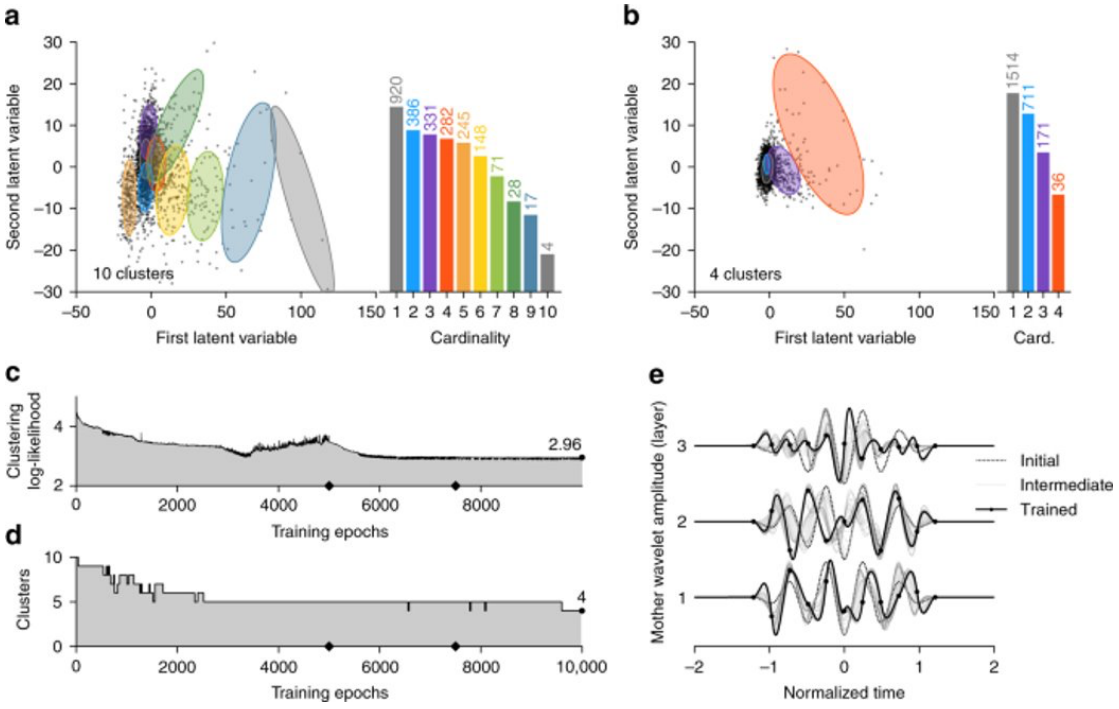
Our aim is not to surpass technologies like template matching, but to provide the first preliminary statistical result that can simplify further detailed analysis, such as the selection of templates for template matching detection.

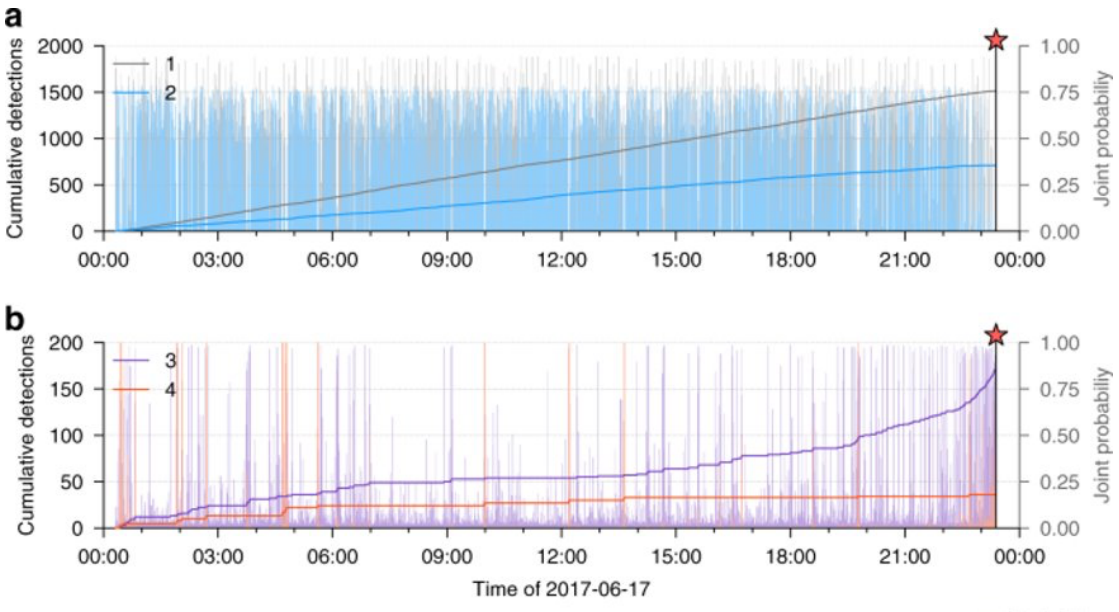
Methods

Deep scattering network for automatic feature extraction

Gaussian mixture model for clustering




a
Second latent variable
First latent variable
10 clusters
Cardinality
920
386
331
282
245
148
71
28
17
4
b
Second latent variable
First latent variable
4 clusters
Card.
1514
711
171
36
c
Clustering log-likelihood
2.96
Training epochs
d
Clusters
4
Training epochs
e
Mother wavelet amplitude (layer)
Normalized time
Initial
Intermediate
Trained


a
Cumulative detections
1
2
Joint probability
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
00:00
b
Cumulative detections
3
4
Joint probability
Time of 2017-06-17