MANCHESTER 1824

The University of Manchester

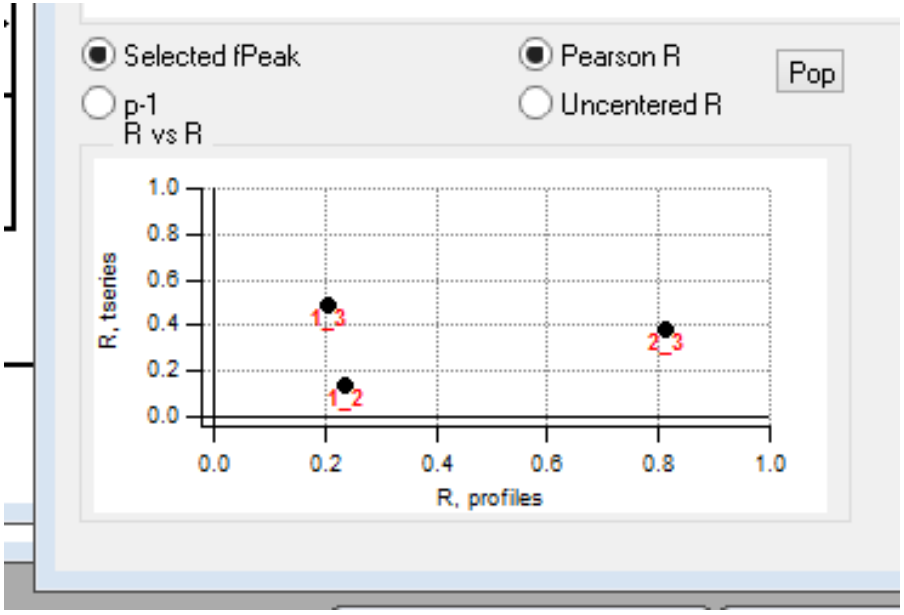
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centred vs Uncentred *r*

James Allan

Motivation

- Frequently need to compare mass spectra, e.g. PMF output with library spectra
- Need a metric that is a good discriminator



Generic unweighted closeness metric

- Data vectors *A* and *B* of length *n*
- Generate closeness metric *S*

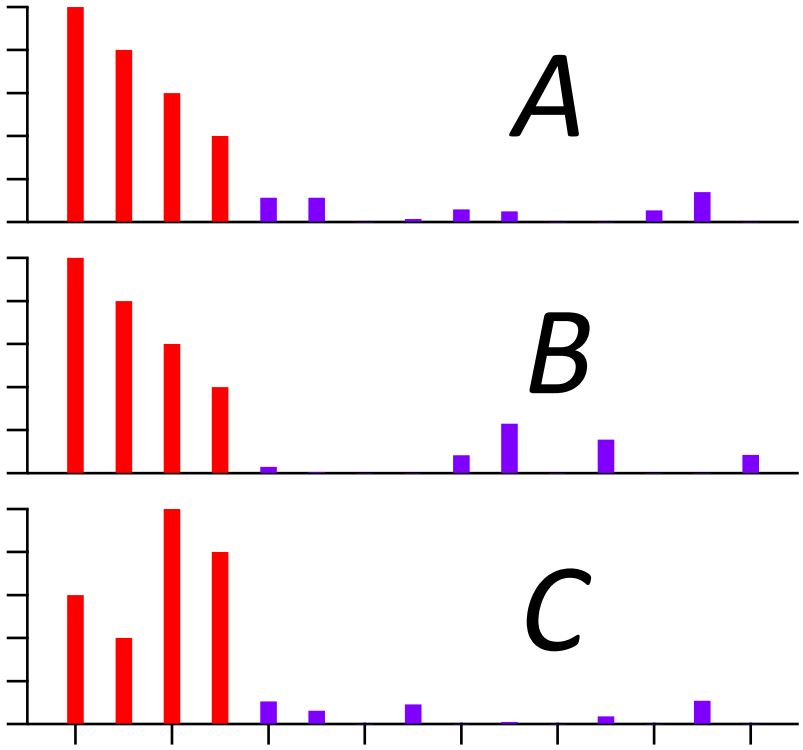
$$S(A,B) = \frac{\sum_{i=1}^{i=n}(A_i - A')(B_i - B')}{\sqrt{\sum_{i=1}^{i=n}(A_i - A')^2 \sum_{i=1}^{i=n}(B_i - B')^2}}$$

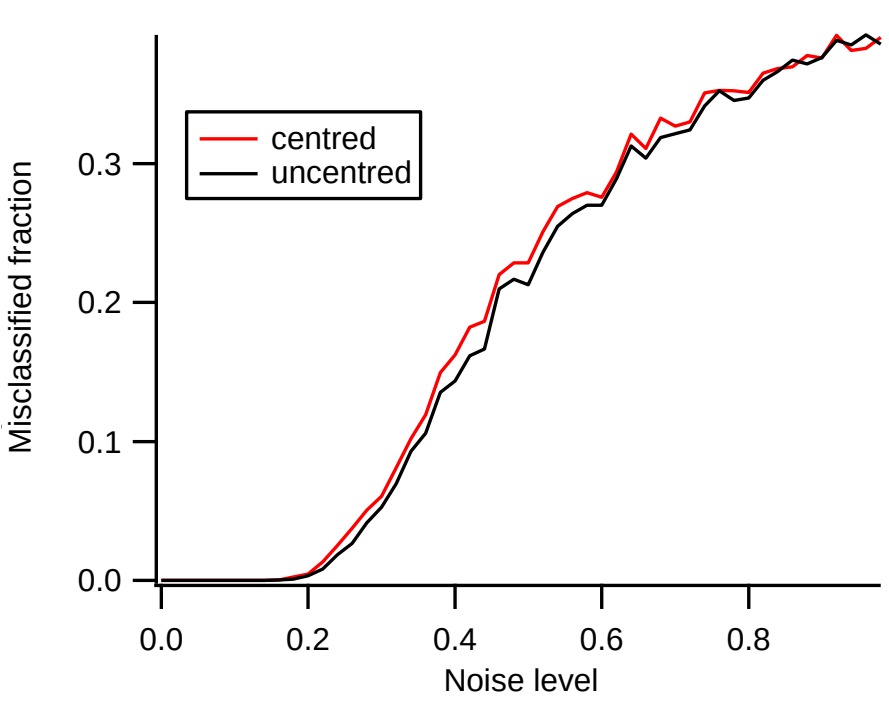
- For *r*: $A' = \bar{A} \ = n^{-1}$ (if normalised)
- For uncentred *r*: $A' = 0$
  - AKA normalised dot product:

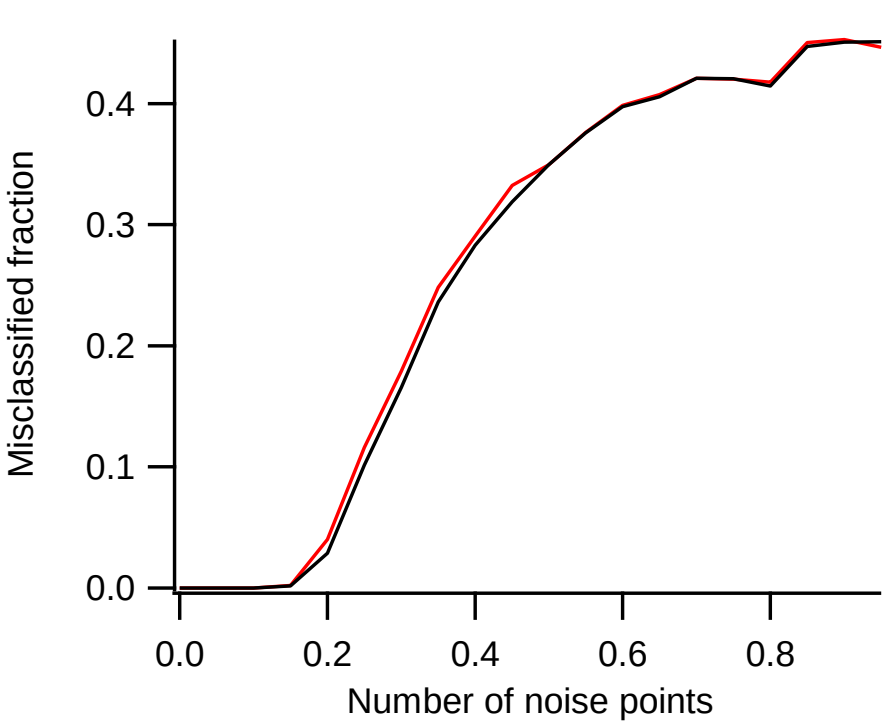
$$S(A,B) = \frac{A \cdot B}{\sqrt{(A \cdot A)(B \cdot B)}}$$

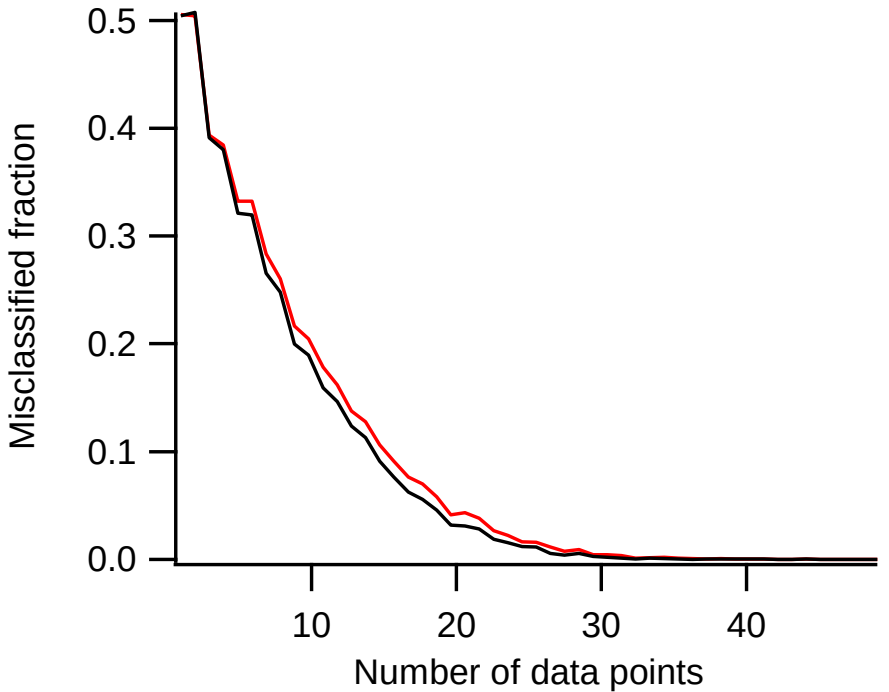
Testing Metrics

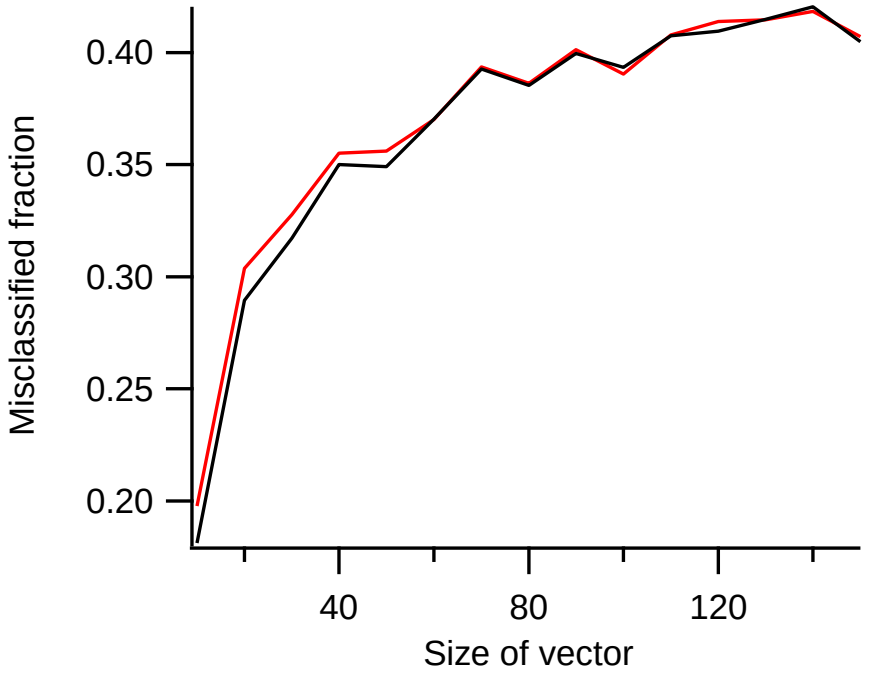
- Generate three vectors of synthetic data
  - Part ‘real’, part ‘noise’
- *S*(*A*,*B*) should always be greater than *S*(*A*,*C*)
- Repeat multiple times with different random data, record fraction of misclassified results




centred
uncentred
Misclassified fraction
0.0
0.1
0.2
0.3
0.0
0.2
0.4
0.6
0.8
Noise level


Misclassified fraction
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
Number of noise points


Misclassified fraction
0.0
0.1
0.2
0.3
0.4
0.5
10
20
30
40
Number of data points


Misclassified fraction
0.20
0.25
0.30
0.35
0.40
40
80
120
Size of vector

Conclusions

- Uncentred *r* is more mathematically sound and performs better than centred *r* for comparing mass spectra
  - But not by much

- Centred *r* remains better for time series as the mean of a time series is meaningful


知
coa
asian feelgoodfood
知
asian feelgoodfood
patagonia