Exploring unsupervised classification of AMS spectra

Mikko Äijälä et al.,
University of Helsinki
(mikko.aijala@helsinki.fi)

16th AMS users’ meeting
Milan, September 11-13th 2015
Our motivation & research questions

• The SMEAR II station in Hyytiälä, Finland often experiences air pollution plumes (transported locally or regionally)

- What are the most common types of pollution and their sources?
- How to derive reference spectra for the pollution types?
- How much is the variability within a pollution type?
Mapping the pollution episodes:

- ~150 [organic] pollution plumes found in 3x1 month data sets
  - (EUCAARI intensives ’08–’09)
Extracting the pollution spectra with PMF

- Resulted in 100 extracted pollution spectra
- Bonus: solution selection based on timetraces avoids PMF rotational ambiguity 😊
Classifying pollution mass spectra (1-28/100)
Classifying pollution mass spectra (29-56/100)
Classifying pollution mass spectra (56-84/100)
Alternative approach: unsupervised classification

• **Exploratory cluster analysis** to classify pollution samples
  • Tested with simple ‘k-means’ clustering

+ Mathematical solution -> **less subjectivity** involved
+ Can easily handle **large amounts of samples**
  • actually improves performance!

- **No [a priori] weighting** [for individual samples] in standard ’k-means’
- ”Hard” clustering -> **intermediate cases** are forced to one cluster only
  • (Unless using e.g. fuzzy k-means or gaussian methods)
Classification for two clusters
Cluster centroids: average mass spectra
Effect of mass scaling $S \times \text{a.m.u.}^{1.5}$

- Edited; figures saved for publication (Äijälä et al., 2015-16),
- I can send them privately if needed, -MÄ

Mass scaling allows k-means to find higher cluster number solutions
With optimised settings:
8 distinct pollution types found:

Sawmill (monoterpene SOA)

BBOA (wood burning, aged)

A-SV-OOA (semi-volatile
/semi-oxid. anthropogenic mix)

HOA (traffic, fresh)

COA (cooking, fresh, local)

Amine containing pollution aerosols? (3 types):

• Edited; figures saved for publication (Äijälä et al., 2015-16),
• I can send them privately if needed, -MÄ
Take-home messages:

- Clustering is a simple, objective method for MS classification, and can provide more quantitative and less subjective results than traditional "expert opinion" classification.

- With optimised data pre-processing (via e.g. "mass scaling") we can automatically classify aerosol samples within the "traditional AMS aerosol classifications" framework.

- High a.m.u. range (m/z > 45 Th) signals seem to contain a more unique fingerprint for aerosol classification than low am.u’s (m/z < 45 Th).

Open question:

- Should we use up-weighting of high a.m.u. signals also when calculating spectral similarity (e.g. correlation) values and running PMF?

Äijälä et al., 2015 (in prep)
Things to consider:

• Needs a good **metric for spectra (dis)similarity**
  • E.g. correlation, (dot product) cosine, squared Euclidean distance...

• **Modified algorithms** (k-medoids, fuzzy k-means, weighted k-means)?

• Data pre-processing (e.g. intensity or mass scaling)?

• [Approximate] **number of clusters** expected

• Metric for clustering **solution quality**
  • E.g. Silhouette, Davies-Bouldin, Calinski-Harabasz, Gap criterion...
On distance (dissimilarity) metrics for MS

Some options for dissimilarity metrics:
Pre-processing can improve classification

• Intensity scaling
  • Scale variable intensities (signals S at each m/z) by a root function (e.g. $S^{1/2}$)
  • Reduces the dominance of the highest signals on the solution
    ➢ Actually degraded solution quality for normalized AMS MS data... 😞

• Mass scaling
  • Scale variable intensities by their mass in a.m.u. (m/z), (e.g. $S \times (m/z)^2$)
  • High a.m.u signals become relatively more important
    ➢ Markedly improves solution quality and enables more detailed classification😊