How to quantify similarity of aerosol mass spectra?

M. Äijälä[1], H. Junninen[1], T. Petäjä[1], M. Kulmala[1], D. Worsnop[1] and M. Ehn[1]

Department of Physics, University of Helsinki, Helsinki, 00560, Finland

Keywords: mass spectrometry, AMS, similarity metric.

Presenting author email: mikko.aijala@helsinki.fi

Introduction

In analysis of mass spectrometric data it is often necessary to evaluate and quantify how similar the mass spectra obtained from two samples are. This applies to *e.g.* algorithm based classification and identification of mass spectra. Often Pearson's product-moment correlation "*r*" is used to describe mass spectral similarity, without giving the matter further consideration. In this work we wish to highlight the importance of suitable similarity (or conversely dissimilarity) metric selection, and aim to optimise such a metric in an example case study involving classification of 70 eV electron ionisation (EI) aerosol mass spectra.

Methods

We studied an example set of 81 mass spectra, obtained with an Aerosol Mass Spectrometer (AMS; Jayne *et al.*, 2000). The samples each contained a "fingerprint" mass spectrum of an individual air pollution episode measured at the SMEAR II station in southern Finland, and deconvolved from ambient measurements by applying factor analysis to separate the pollution mass spectra from the background aerosol spectra.

We applied k-means++ clustering to reproduce the already well known divisions to aerosol general chemotypes such as low-volatile, semi-volatile, hydrocarbon-like organic aerosol types (LV-OOA, SV-OOA, HOA). By optimising the dissimilarity metric and weighting needed in the classification process, we also obtained valuable information on the effects of using various metrics and weights.

For this work we tested the aptness of four dissimilarity metrics (Pearson correlation, dot-product cosine, squared Euclidean "distance" and Manhattan distance for describing the (dis)similarity between the aerosol mass spectral samples. We additionally probed the effects of variable (here mass-to-charge ratios, '*m/z'*) weighting as well as weighting based on signal intensity.

Results and conclusions

We conclude there are indeed differences between the metrics' performances. Both 'dot-product cosine' and 'Pearson correlation' were found to produce very similar, robust classification results, with 'squared Euclidean' dissimilarty also providing satisfactory results. Based on our tests 'Manhattan distance' is not to be recommended for aerosol mass spectra similar to ours, as it does significantly worse in representing the similarities between aerosol types. This clearly leads to problems in finding the mass spectral structures corresponding to the aerosol chemotypes.

We additionally explored the effects of applying different weight distribution between mass spectral variables (*m/z*), as is commonly done and advocated for in many mass spectrometric applications outside of aerosol sciences (Stein & Scott, 1994). Specifically we applied 1) 'mass scaling'

$$\text{weight (i)}_{\text{(mass)}} = \text{m/z (i)}^{\text{s\_m}} \qquad \text{(Eq. 1)}$$

where i are our mass spectral variables, and 's_m' is a mass scaling factor ranging from zero to three in our tests., and 2) 'intensity scaling'

$$\text{weight (i)}_{\text{(int.)}} = \text{signal (i)}^{1/\text{s\_i}} \qquad \text{(Eq. 2)}$$

where 's_i' is the scaling coefficient for signal intensity.

Comparing to uniform weight distribution we conclude mass weighting using s_m of 1 to 2 enhances the aerosol chemotype classification while signal weighting appears detrimental to it with any s_i > 1.

Based on this (albeit limited) study, we would like to encourage metric and variable weight distribution optimization in connection to any data analytical tasks involving aerosol mass spectra classification or algorithm-based identification. We find that Pearson correlation seems a suitable metric for identification and classification of 70 eV EI aerosol mass spectra, although theoretical considerations would seem to favour dot-product cosine metric instead. The two seem to produce almost identical results in our tests. We also recommend exploring 'mass scaling' as a basis of weight distribution among variables, as in our case it does markedly enhance aerosol classification to aerosol chemotypes.

We thank the ERC (Grant 638703-COALA)

Jayne, J.T., *et al.* (2000). Development of an Aerosol Mass Spectrometer for Size and Composition. Analysis of Submicron Particles, *Aerosol Science and Technology*, 33, 49-70,

Stein, S.E., & Scott, D.R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9), 859-866.