Improved Aerosol Apportionment by Bayesian Classifier

Thomas P. Rebotier, Stephen M. Toner and Kimberley A. Prather
Department of Chemistry/Biochemistry, University of California, San Diego, 92093, La Jolla, U.S.A.

INTRODUCTION

Apportioning ambient aerosol measured by ATOFMS is made by comparing their mass spectra with seeds obtained from source studies. Currently, matching depends on the dot product between the particle spectrum and the average spectrum of a particle class, the "seed" (Song et al., 1999). An aerosol matches a class when its dot product is larger than a given "vigilance factor" (typically 0.8).

This approach ignores cluster size: seeds obtained from 5 particles pull as much weight as seeds obtained from 2,000. Dot product matching also ignores differences in variability: Δ(m/z) is always larger for the peaks that have large m/z on average, so that the existence of smaller significant peaks is entirely swamped by noise on larger peaks. For example, Na_2Cl creates two peaks (at +81 and +83) which very dependable marks of fresh sea salt but their influence in matching is nothing compared to that of the noise from the large Na peak which occurs in fresh and aged sea salt, as well as in biomass.

These problems can be addressed by a "mixture of Gaussians" Bayesian Classifier (Duda and Heart, 1973), in which each spectral class is represented by a Gaussian distribution in 700 dimensions (350 positive and negative m/z). To match an unknown particle, its likelihood is computed by the Gaussian formula, then weighted by the prior probabilities (the expected number of particles in each cluster). The particle is apportioned to the class of highest posterior probability.

$$likelihood = \exp\left[-\frac{1}{2}(spectrum - seed)^t CovMatrix^{-1}(spectrum - seed)\right]$$

$$posterior = \frac{prior \cdot likelihood}{\sum_{allclusters} prior \cdot likelihood}$$

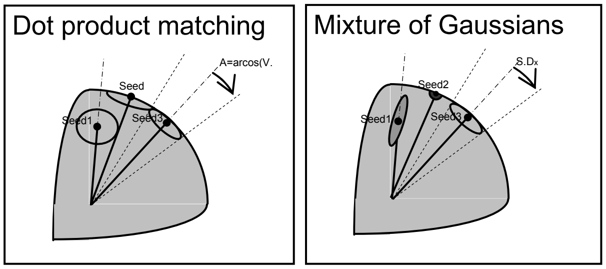

Figure 1. Comparing the distributions assumed by each matching method. The mixture of Gaussians model allows variations of shape and intensity.

METHODS

To compare the efficiency of both matching methods, they were given to apportion aerosols for which the results were already known. Aerosols collected by a fine/ultrafine ATOFMS in a dynanometer study of 28 Light Duty Vehicles (LDVs) (Sodeman et al., 2005) and 6 Heavy Duty Diesel Vehicles (HDDV) (Toner et al., 2005) were divided by size and origin into two training and two testing sets. One training/testing pair consisted of ultrafine particles (aerodynamic diameter Da < 100 nm) and the other of fine particles (Da between 100 nm and 300 nm). The testing sets used all particles collected from 6 of the LDVs and 1 of the HDDVs, and the training sets used only particles from the other vehicles. Thus, the seeds were computed entirely from the exhaust of different vehicles than the ones used for testing.

Three methods of apportionment were compared: dot product apportionment using seeds obtained from the ART-2A algorithm, Bayesian matching using the exact same seeds, and Bayesian matching using seeds obtained by a Hierarchical Clustering program. Seeds were obtained from training particle clusters by averaging their spectra. In the case of ART-2A, the clusters were obtained separately for LDV and HDDV particles, with a vigilance factor of 0.85. In the case of Hierarchical Clustering, spectra from both sources were mixed and gradually clustered into increasingly larger and fewer clusters, the process being stopped just before those became too heterogeneous.

No new program works without a few adjustments, and in applying Bayesian matching to aerosols, three are important to point out. **Adjusting the priors** may be necessary. When all classes come from the same source study, the priors are simply computed as the proportion of particles of each class in that study. When several studies are merged (for example a study on LDVs and one on HDDVs) a ratio of the effective sources has to be estimated. Within a study the priors frequently vary by a factor 1,000; therefore, when estimating the relative abundance of sources measured in one study compared to the other, a between-study ratio within a factor 2 of the actual abundance ratio is still a very good guess. **Simplifying the covariance matrix** is a computational necessity-- with a 700x700 spectrum, computations would be too expensive with a full covariance matrix. In this study, the matrix has been approximated by its diagonal, which in practice reduces the Gaussian distribution to the product of independent Gaussian distributions for each m/z. In future studies, it should be possible to use a sparse covariance matrix, with a few non-zero diagonal elements. **Cutting off the low likelihood matches** is a way of avoiding to apportion particles that come from sources not represented in the seed set. With the given priors, the probabilities sum to one, even though some particles may not belong to any of the offered classes. There are two ways to manage this; one can sort the resulting best posteriors and select only a given proportion (for example, match 80% or 99% only, these being the most certain matches); or one can put an arbitrary cutoff value on the (prior*likelihood) value. This study used an automatic matching of the most certain 99%.

RESULTS

Competing methods are compared on the basis of the percentage of testing set aerosol that were matched into the type they came from (LDVs into LDV classes, HDDVs into HDDV classes. Results, shown in Figure 1, clearly favor Bayesian matching, in particular for HDDV aerosols, even though the Bayesian classifier was forced to apportion 99% of the particles, whereas the dot product matching (operated at v.f. of 0.85) only recruited 85 to 96% of test particles. When matching only equal numbers, Bayesian correctness is consistently above 99%.

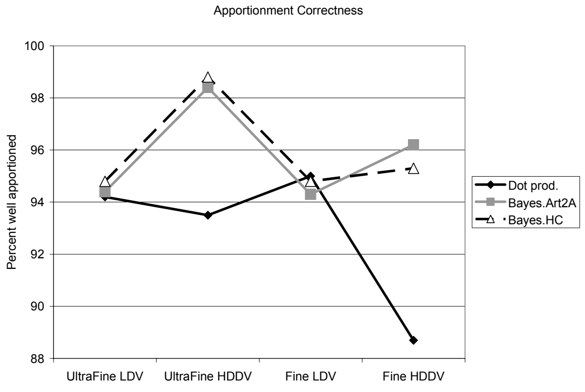

Figure 2. Percent of test particles well apportioned depending on size, source, and apportionment method.

Several reasons contribute the better performance of Bayesian matching. One is that seeds corresponding to very few training particles no longer recruit as many testing particles. Another is that small peaks that are consistently present in the training particles of a particular cluster are now influential in the matching. Figure 3 illustrates this by comparing the spectrum of a seed with the median spectrum of the particles it recruits with either dot product or Bayesian matching. The seed shown in Figure 3 has the largest common recruitment of both methods for fine LDV testing particles (that is, some other seeds recruit more particles with one method but not with the other). The positive spectra are shown; in the median spectra, the different shades of grey indicate the relative area reached respectively by 25% (very light grey), 50% (dark grey), and 75% (black) of the particles.

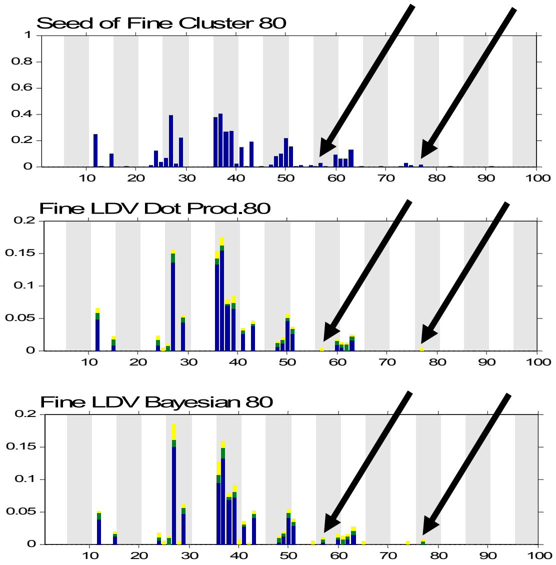

Figure 3. Bayesian matching respected small peaks +57 and +77, dot product matching did not.

CONCLUSIONS

The matching method makes a large difference to apportionment error. Bayesian matching takes into account the importance of a seed (number of particles it was averaged from) and the small marker peaks. Apportionment error is lower with Bayesian than with dot product matching.

Keywords: *Apportionment, Bayesian Classifier, Hierarchical Clustering, ART-2A*

REFERENCES

Duda, R. O., and Hart, P. E. (1973). *Pattern Classification*, Wiley Interscience.

Sodeman, D. A., Toner, S. M., and Prather, K. A. (2005).Determination of Single Particle Mass Spectral Signatures from Light Duty Vehicle Emissions, *Environmental Science & Technology*, in press.

Song, X.-H., Fergenson, D. P., Hopke, P. K., and Prather, K. A. (1999).Classification of Single Particles Analyzed by Atofms Using an Artificial Neural Network, Art-2a, *Anal. Chem*, *71* (4), 860-865.

Toner, S. M., Sodeman, D. A., and Prather, K. A. (2005).Single Particle Characterization of Ultrafine- and Fine-Mode Particles from Heavy Duty Diesel Vehicles Using Aerosol Time-of-Flight Mass Spectrometry, *in preparation*.