Spectral Library Clustering Using a Bayesian Information Criterion

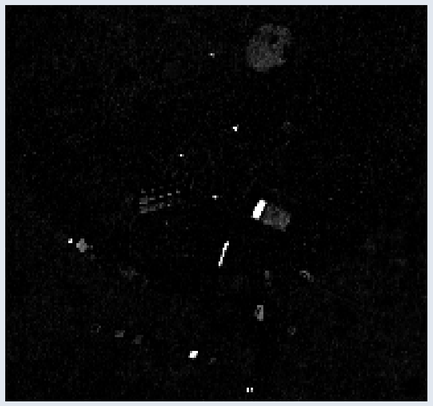
Jonathan Piper and John Duselis
Dstl Porton Down, Salisbury, Wilts., SP4 0J

Abstract
Unsupervised classification of spectral libraries provides groups of spectra which may be of use in material detection or identification processes for hyperspectral imagery. It can also be used to discover associations between materials' spectra that can help analysts to interpret spectral data. We use the assumption of a Gaussian Mixture Model, combined with a Bayesian Information Criterion (BIC) for determination of the number of mixture components, to model a spectral library and assign its members to clusters. This process provides a natural method to include spectra within multiple classes where this is appropriate. The process is demonstrated for a library of mineral spectra. Results are compared with the minerals' geological classifications and with the results of other published clustering processes.

Why do unsupervised clustering of spectral libraries?
Clusters of library spectra can assist algorithms for target detection, making it possible to detect materials that have not been measured in the laboratory but are related to others that have, as illustrated in Figure 1.

Clusters can be defined manually or with the aid of supervised clustering algorithms. However, this is labour-intensive; requires trained analysts; may be difficult in libraries with poor metadata; and limits the cluster analysis to groups that have already been identified by the analyst.

Our process is unsupervised and also allows for spectra to be allocated to more than one cluster – valuable for mechanical mixtures of different compounds, or spectra showing features from more than one ion, bond or functional group.



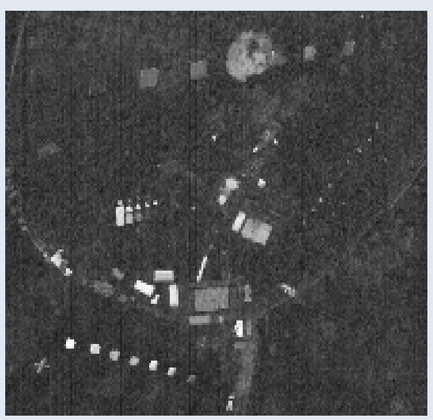


Figure 1 – left: aerial photograph showing targets deployed for a field trial, including coloured Perspex panels in the bottom-left of the image. **Top-right**: detection image of the same scene from a visible and short-wave infrared (SWIR) hyperspectral imager. Lighter tones indicate higher likelihood that a pixel contains the target as opposed to the background. The image was searched for a target made of red Perspex. Only one of the Perspex targets is detected. **Bottom-right**: Hyperspectral detection image searched for a subspace including many different colours of Perspex. Note that all the square targets in the bottom-left are detected. Crucially, this includes targets that were not represented in the library used to define the target subspace. The target subspace is an example of a cluster of spectra.

Technical approach
Our process employs several established techniques for unsupervised learning:

Spectral libraries typically record reflectance for hundreds or thousands of wavebands. To avoid solving a grossly under-determined problem we reduced the dimensionality of the spectral signals using principal components analysis. In our example, spectra with 420 wavebands were summarised using 9 principal components.

Expectation maximization is used to fit a Gaussian mixture model to the data. This process requires that the number of clusters is specified in advance, so it is performed for a range of numbers of clusters. During the process spectra may be associated with clusters in one of two ways: either they are assigned to the single cluster with the highest estimated probability (single membership), or they can be assigned to any cluster for which the probability exceeds a threshold (multiple membership).

The Bayesian information criterion (BIC) is evaluated for each model; the model with the smallest value is selected as the best fit for the data.

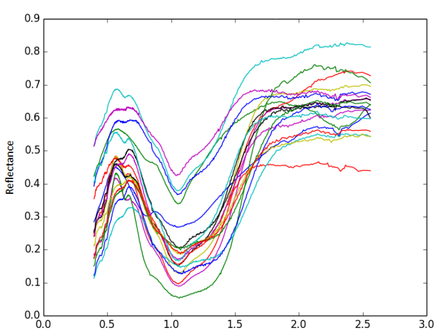
$$BIC = -2 \cdot \ln(M) + k \cdot \ln(n)$$

Here M is the maximised likelihood function for the model, n is the number of spectra in the library and k is the number of free parameters in the model.
The covariance matrices of each cluster may be constrained in various ways to improve estimation of the matrices' elements with the limited data available. We use four different approaches. In three the parameters are estimated individually for each cluster: "spherical", in which covariances are zero and all variances equal; "diagonal", in which covariances are zero but variances can be unequal; and "full", in which variances and covariances can take any value. We also employ the "tied" covariance matrix type: a full covariance matrix that is the same for each component. Since the BIC penalises models with more free parameters in order to avoid over-fitting, less constraint on the covariance matrix usually results in models with fewer clusters.

Performance assessment
The clustering process was tested using data from the U.S. Geological Survey Digital Spectral Library. This library is accompanied by metadata that can be used to generate a hierarchy of truth clusters. In our analysis the three tiers of the hierarchy are described as "Type", "Group" and "Mineral". For example, a spectrum for the rock chert is labelled as mineral – chert; group – silicate; type – tectosilicate.
Results from the algorithm are compared with the truth clusters using the Jaccard Index (JI):

$$JI = \frac{|truth\ cluster \cap test\ cluster|}{|truth\ cluster \cup test\ cluster|}$$

This concept can be extended to analysing the quality of the entire clustering result (as opposed to measuring the similarity of two clusters) by assigning an identity to each cluster produced by the algorithm (based on comparison with all truth classes) and counting numbers of true positives (TPs), false positives (FPs) and false negatives (FNs). Then $JI = \frac{TPs}{(TPs+FPs+FNs)}$
Results are shown in Table 1.



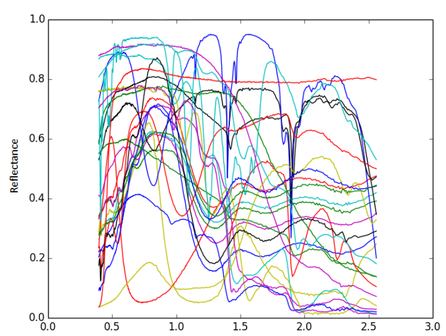


Figure 2 – example outputs of the clustering process. Both were generated assuming a diagonal cluster covariance matrix and disallowing multiple membership (so each spectrum is a member of only one cluster). The cluster on the left is a good match for the Olivine group (Jaccard Index of 0.85) but the cluster on the right cannot be clearly identified with any truth class.

Further work
Various adaptations of the clustering process could be tested: for example, use of different distance measures and different criteria for determining the optimal cluster number (in addition to the BIC). More detailed consideration should also be given to setting the probability threshold used when assigning spectra to multiple clusters.
Of equal importance to developing the clustering algorithm is developing methods for testing it. More work on analysing clustering quality is needed, as is a set of truth classes known to accurately reflect relationships between spectra. Figure 3 shows a subset of the members of one of the truth clusters used in this analysis cluster. It can be seen that they display considerable diversity.

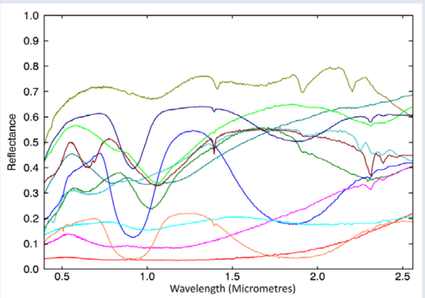


Figure 3 – a subset of spectra from the Pyroxene Group, one of the truth classes used in this analysis.

Covariance type and multiple membership	Type	Group	Mineral
Spherical, single	0.12	0.21	0.22
Diagonal, single	0.21	0.28	0.22
Full, single	0.25	0.25	0.19
Spherical, multiple	0.10	0.15	0.16
Diagonal, multiple	0.15	0.18	0.14
Full, multiple	0.20	0.18	0.13

Table 1 – Jaccard indices comparing clustering results for different algorithm settings (covaraince type and whether multiple membership is allowed) with different sets of truth clusters, corresponding to different levels of the truth cluster hierarchy. Results lie in the range 0..1, where 1 is a perfect match with the truth. Although the scores are reasonably poor, some good clusters were produced, as illustrated in Figure 2.

Conclusion
We have shown that a set of standard techniques for unsupervised learning, whilst producing some good results, were not well adapted to the specific task of clustering spectral libraries. More sophisticated methods are required to produce tools that would be of use to assist operational exploitation of HSI.

Acknowledgments
The authors would like to thank Professor Jim Davis of Ohio State University for his guidance in this investigation.