Using dimensionality reduction and clustering techniques to classify space plasma regimes

M. R. Bakrania[1], I. J. Rae[1,2], A. P. Walsh[3], D. Verscharen[1], A. W. Smith[1]

[1]Mullard Space Science Laboratory, UCL, UK
[2]Department of Mathematics, Physics and Electrical Engineering, University of Northumbria, UK
[3]European Space Astronomy Centre, Spain

Email: mayur.bakrania@ucl.ac.uk

1. Introduction

Particle populations in collisionless space plasma environments, such as the Earth's magnetotail, are traditionally characterised by the moments of their distribution functions. Distribution functions, however, provide the full picture of the state of each plasma environment, especially when non-thermal particle populations are present that are less easily characterised by a Maxwellian fit.

Distribution functions, unlike moments, are not easily classified by a small number of parameters. We therefore propose to apply dimensionality reduction and clustering methods to particle distributions in pitch angle and energy space as a new method to distinguish between the different plasma regions. With these novel methods, we robustly classify variations in particle populations to a high temporal and spatial resolution, allowing us to better identify the physical processes governing particle populations in near-Earth space.

2. Machine Learning Models

In unsupervised learning, algorithms discover the internal representations of the input data without requiring training on example output data. Dimensionality reduction is a specific type of unsupervised learning in which data in high-dimensional space is transformed to a meaningful representation in lower dimensional space. This transformation allows complex datasets to be characterised by analysis techniques with much more computational efficiency.

We use the autoencoder to compress the data by a factor of 10 from a high-dimensional representation. We subsequently apply the PCA algorithm to further compress the data to a three-dimensional representation. The PCA algorithm has the advantage of being a lot cheaper computationally than an autoencoder, however the algorithm only captures variations that emerge from linear relationships in the data. After compressing the data, we use the mean shift algorithm to inform us of how many populations are present in the data using this three-dimensional representation. And finally, we use an agglomerative clustering algorithm to assign each data-point to one of the populations.

3. Autoencoders

Autoencoders are type of neural networks. They learn compressed representations of data by using a bottleneck layer which maps the input data to a lower dimensional space, and then subsequently reconstructing the original input. By minimising the `reconstruction error', or `loss', the autoencoder retains the most important information in a representative compression and reconstruction of the data.

Each neuron computes the following sum:

$$y = \sum_i w_i x_i + b, \qquad (1)$$

where w_i denotes the weights of each neuron, x_i the input from the previous layer, and b the bias term. The autoencoder adjusts the weights and biases to minimise the reconstruction loss.

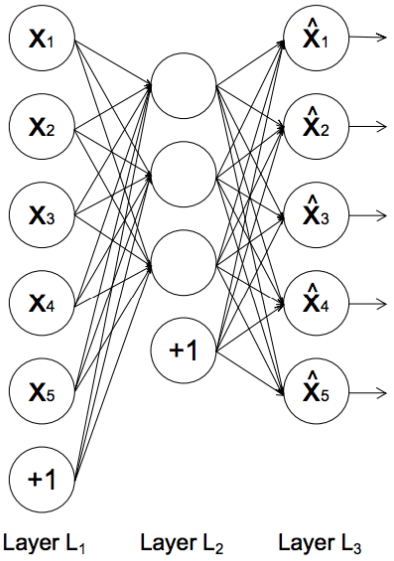


Fig.1: The architecture of an autoencoder [1]. Each circle represents a neuron corresponding to a data-point. Layer L_1 represents the input data, layer L_2 the encoded data in latent space, and layer L_3 the reconstructed data.

4. Principal Component Analysis

A PCA algorithm reduces the dimensionality of input data by transforming this data from a large number of correlated variables to a smaller number of uncorrelated variables, known as principal components. This is achieved by calculating the covariance matrix associated with the input data, and extracting the eigenvectors.

5. Mean Shift

The mean shift algorithm is a clustering technique that locates the maxima of a density function in a sample space. The algorithm does not require prior knowledge of the number of clusters.

6. Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering that uses a 'bottom-up' approach, whereby each data-point is first assigned a different cluster. Then pairs of similar clusters are merged until the specified number of clusters has been reached. During each recursive step, the agglomerative clustering algorithm combines clusters typically using Ward's criterion (Ward, 1963), which finds pairs of clusters that lead to the smallest increase in the total intra-cluster variance after merging. The increase is measured by a squared Euclidean distance metric:

$$d_{ij} = \left\| C_i - C_j \right\|^2 \qquad (2)$$

where C_i represents a cluster with index i.

7. Magnetotail Data

We use electron data from the magnetotail in order to test the effectiveness of our method. The magnetotail is traditionally divided into three different regions: the plasma sheet (PS), the plasma sheet boundary layer (PSBL), and the lobes. To ensure that we test our method on a large number of data from each of the magnetotail regions (>50 000 samples), we obtain Cluster-PEACE data from times when the C4 spacecraft has spent at least 1 hour in each region, according to Cluster-ECLAT dataset [2]. The dimensionality of each of our distribution samples is 312 (12 pitch angle bins x 26 energy bins).

[1] Sakurada, M. and Yairi, T. (2014). doi:10.1145/2689746.2689747.
[2] Boakes, P. D., Nakamura, R., Volwerk, M., and Milan, S. E. (2014). doi:10.1155/2014/684305.

8. Method

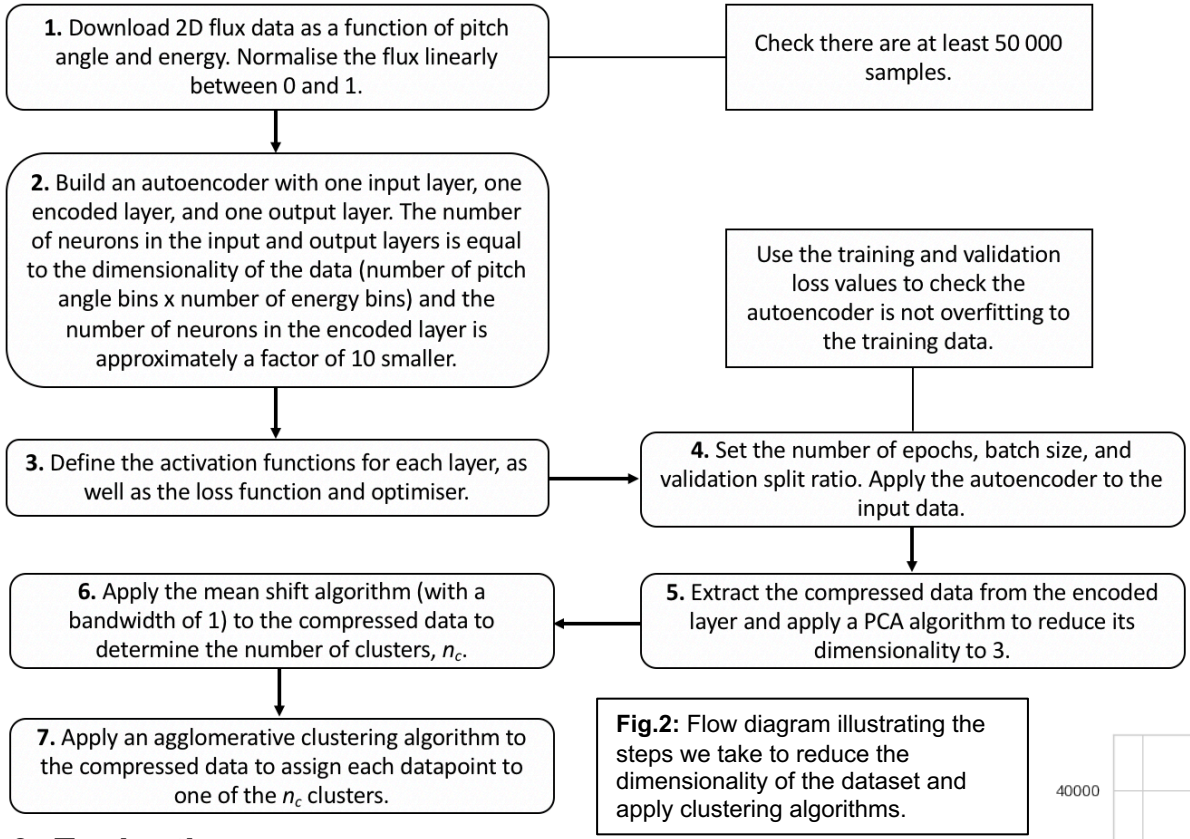


Fig.2: Flow diagram illustrating the steps we take to reduce the dimensionality of the dataset and apply clustering algorithms.

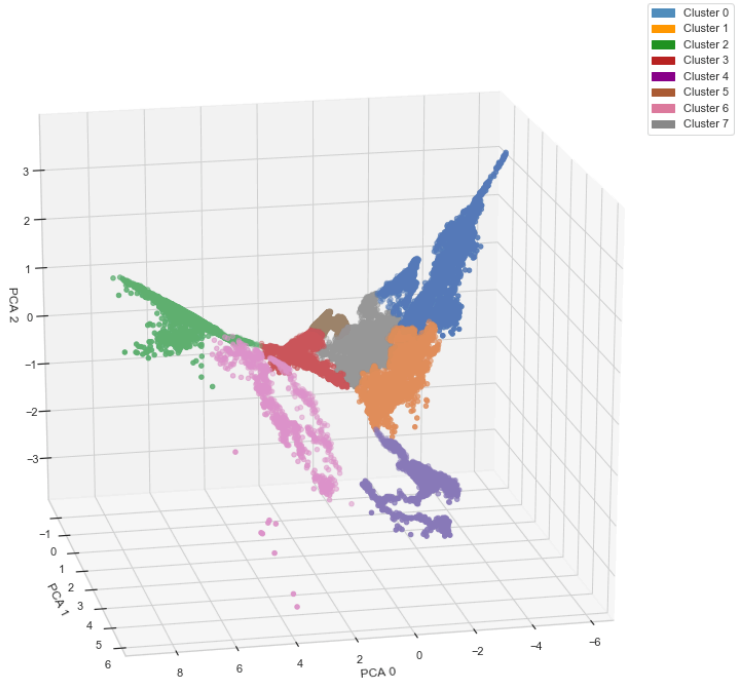


Fig.3: Three-dimensional representation of the magnetotail data after undergoing dimensionality reduction via an autoencoder and PCA algorithm.

9. Evaluation

Fig. 3 shows the result of applying the agglomerative clustering algorithm to the compressed magnetotail electron. The plot shows that the clustering algorithm is able to assign data-points of varying PCA values to the same cluster if they belong to the same complex non-spherical structure, e.g. clusters 0, 4, and 6. The clustering algorithm is able to form clear boundaries between clusters with adjacent PCA values, e.g. between clusters 0, 1, and 7, with no mixing of cluster labels on either side of the boundaries.

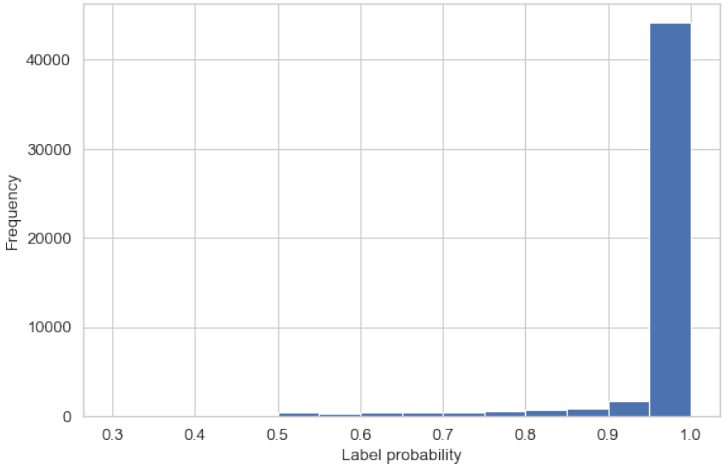
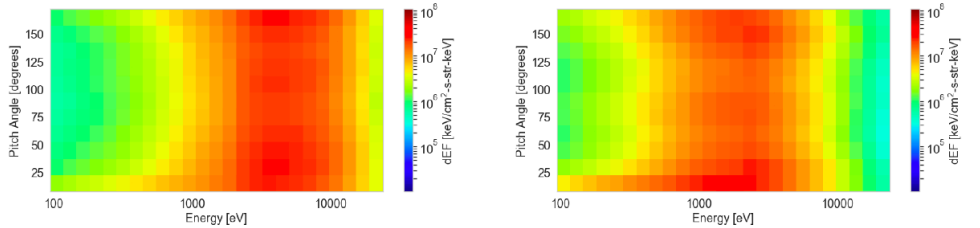
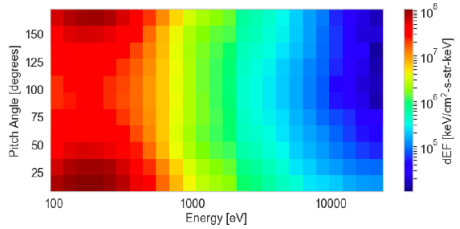


Fig.4: Histogram showing the probabilities, generated by GMMs, that the data-points belong to the cluster assigned to them.

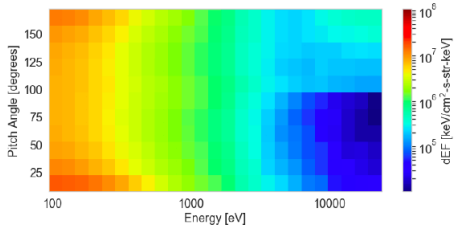


(5a) Cluster 0 - PS a

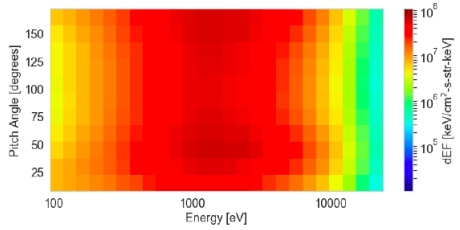
(5b) Cluster 1 - PS b



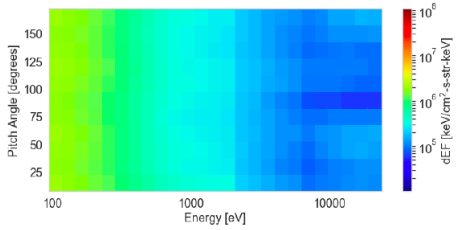
(5c) Cluster 2 - PS c



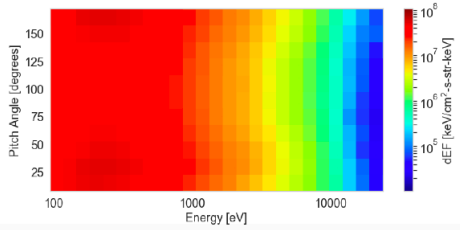
(5d) Cluster 3 - PSBL



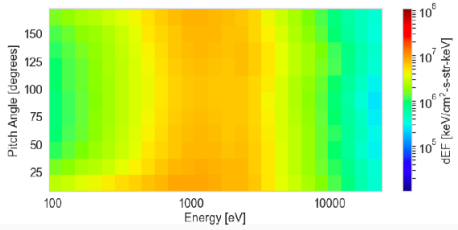
(5e) Cluster 4 - PS d



(5f) Cluster 5 - Lobes



(5g) Cluster 6 - PS e



(5h) Cluster 7 - PS f

Fig.5: Average electron differential energy flux distributions as a function of pitch angle and energy for each of the eight clusters. Each cluster is assigned a magnetotail region based on our interpretation of their plasma and magnetic field parameters.

Table 1: Contingency table comparing the agglomerative clustering (AC) labels of the magnetotail electron data to the original ECLAT labels (0 = PS, 1 = PSBL, and 2 = lobes).

AC labels	ECLAT labels		
	0	1	2
0	6549	0	0
1	3074	0	0
2	5092	0	0
3	1590	4188	0
4	2097	0	0
5	156	2228	15641
6	1029	0	0
7	7020	1057	0

9. Evaluation (continued)

We use Gaussian mixture models to find the probabilities of each data-point belonging to the cluster it has been assigned to. Fig. 4 shows that >92% of the data-points have an associated probability of over 0.9, and <1% of the data-points have a probability of <0.5. This indicates a high certainty in our clustering method.

Fig. 5 shows large differences in the average pitch angle/energy distributions. Each distribution differs by the: peak flux energy, peak flux value, or the pitch angle anisotropy. The lack of identical distributions shows mean shift has not overestimated the number of clusters.

10. Conclusion

In table 1, the majority of clustering labels are in agreement with the ECLAT regions. For AC labels 0, 1, 2, 4, and 6, which represent various populations within the plasma sheet, there is 100% agreement with the ECLAT label 0. By using this method to characterise pitch angle and energy distributions, instead of using the derived moments, we successfully distinguish between multiple populations within what has historically been considered as one region, due to the lack of variation in the plasma moments as well as the similarity in spatial location. In a follow up study, we will use the results from applying this method to link the occurrence of these populations to other high-resolution spacecraft measurements in different plasma regions, in order to understand the physical processes driving changes in the less abundant particle populations.

Acknowledgments

We thank the Cluster instruments teams (PEACE, FGM, CIS, EFW) for the data used in this study. Data can be obtained from the Cluster Science Archive (https://csa.esac.esa.int/csa-web/).