Analysis of Electron Distribution Functions using the Gaussian Mixture Model

Beniamino Sanò[1,2,3] and Nathan N. Maes[4] and David L. Newman[5] and Marty Goldman[5] and Francesco Valentini[2] and Giovanni Lapenta[4]

[1]University of Trento, Italy
[2]Università della Calabria, Italy
[3]ASI - Italian Space Agency, Italy
[4]Center for mathematical Plasma Astrophysics (CmPA), Department of Mathematics, KULeuven, University of Leuven, Belgium
[5]University of Colorado, USA

Key Points:

- The Gaussian Mixture Model (GMM) is assessed in its ability to identify multiple Gaussian and kappa-distributed beams depending on their relative average means and standard deviations.
- The Gaussian Mixture Model (GMM) can be used to define the complexity of a velocity distribution function based on the optimal number of beams determined by information theory criteria.
- The Gaussian Mixture Model (GMM) is applied to burst intervals of Magnetospheric Multiscale (MMS) mission for electrons in the dayside and nightside with different counts levels and noise to signal ratios.

Corresponding author: Beniamino Sanò, beniamino.sano@unitn.it

Abstract

Velocity distribution functions (VDFs) measured by the Magnetospheric Multiscale (MMS) mission are complex 3D datasets that can be represented as a superposition of multiple beams (M. V. Goldman et al., 2020). A recent work (Dupuis et al., 2020) proposed the use of the Gaussian Mixture Model (GMM). Here we investigate the approach by considering first synthetic distributions made by artificially creating beams of either Maxwellian distributions or kappa distributions with varying power law index. By varying the inter-beam average difference and the beam standard deviation we evaluate the ability of the GMM in recognizing correctly the beam. We then apply the method systematically to MMS data in the tail and in the dayside. In this case, the data need preparation before being processed by the GMM to account for the specifics of the instrument and in particular the lack of data at low energy and to account for the noise in the counts. The conclusion of the analysis is that the GMM is capable of detecting the presence of multiple beams when their distinction is significant. The GMM can define reliably the complexity of a measured data-set in terms of the number of optimal beams provided by information theory criteria. Visual inspection confirms this automatic definition of complexity.

Plain Language Summary

This work investigates regions of interest in electrons distribution functions from Magnetospheric Multiscale (MMS) mission, using an unsupervised machine learning technique called Gaussian Mixture Model (GMM). First we tested the ability of the GMM to identify multiple Gaussian and kappa-distributed beams on synthetic distributions, and then we analysed real data from MMS. The data is downloaded and preprocessed through AIDApy, a Python package for the analysis of spacecraft data from heliospheric missions. A Gaussian mixture model search through the particles and identify the presence of different subpopulations whithin an overall population. The optimal number of subpopulations is determined by a model selection technique, and the presence of certain distributions can be utilized to find magnetic reconnection regions.

1 Introduction

The study of the Earth's magnetosphere and its complex system of electromagnetic interactions is a key goal in understanding the fundamental physics of space. Magnetic

reconnection and plasma turbulence are both closely interrelated fundamental processes in the dynamics of the magnetosphere (Biskamp, 2000). Magnetic reconnection is a process during which the magnetic field energy is converted into kinetic energy, thermal energy, and particle acceleration energy. Reconnection occurs in small-scale electron diffusion regions within a current sheet (Lapenta et al., 2016). As the field lines flow into the region, they reconnect at the X-point. The reconnected field has a strong magnetic tension, which pulls the reconnected field away from the X-point, expelling the plasma coupled to it as bi-directional outflow jets (Li et al., 2021). Plasma turbulence is the result of multi-scale nonlinear interactions and instabilities of large-scale fluid motions. Collisionless space plasmas are often in a turbulent non-equilibrium state, characterized by strong fluctuations of field and plasma parameters (Scott, 2021). Turbulence and reconnection research is focused on how magnetic reconnection occurs in a turbulent system and how the dynamics of turbulence and reconnection interact (Yokoi & Hoshino, 2011).

To study this relation, several spacecraft have been sent into space in recent years. Cluster mission observed for the first time in-situ magnetic reconnection in turbulent plasma (Retinò et al., 2007). NASA's Magnetospheric Multiscale (MMS) mission has the goal of observing at an unprecedented rate traces of magnetic reconnection in Earth's magnetosphere (Burch et al., 2016). Fast Plasma Investigation (FPI) instrument measures incoming particles through a filter which selects certain particle speeds and directions; then a 3D picture of the ion plasma is produced every 150 milliseconds, while for electron plasma FPI captures a picture every 30 milliseconds. Because of these frame rates, MMS measures more than 100 GB of data every day. However, due to limitations of the probes, a continuous overwriting of data takes place and a large part of it is lost irreversibly: in fact, approximately 4 GB of data per day are transmitted to Earth.(Baker et al., 2016) At first the task of looking at the raw data and selecting the interesting ones was done by so-called scientists in the loop, who would observe the data by eye and select which ones to store. Nowadays, due to the size of the measurements, this kind of filtering is no longer possible nor desirable. An automatized procedure is necessary for a preliminary analysis of the data in order to choose which ones to select and send to Earth. Nevertheless, researchers are able to understand all type of information and interpret them critically by simultaneously using a combination of optimization, model learning, planning, prediction, and diagnostic analysis. This is challenging for many automated systems: as a result, artificial intelligence has become the perfect candidate for this task

thanks to the ability to recognize patterns and extract information from data by simulating human learning. Following this paradigm shift, the European Commission (EC)'s Horizon 2020 project started the Artificial Intelligence Data Analysis (AIDA) project, which not only aims to automatize the pre-processing of space data, but also to introduce modern data assimilation, statistical methods and machine learning (ML) to heliophysics data processing: *Aidapy*, an high level Python package for the analysis of spacecraft data from heliospheric missions has been developed as a result.

A new EC project has now followed up AIDA, the project Automatics in SpAce exPloration (ASAP) to study the deployment of ML tools onboard space missions, using the type of processors that can resist the hostile environment of space.

In this work we used Aidapy along with unsupervised ML clustering techniques to characterize particle velocity distributions. The goal of the analysis is to differentiate between simple and more complex regions within the velocity distribution functions measured by MMS: in particular, complex shaped electron distributions, thus represented with a greater number of clusters, have been shown to be good indicators for magnetic reconnection and turbulence (Shuster et al., 2014; Hoshino et al., 2001). The necessity of using unsupervised ML techniques arises from the fact that supervised ML needs huge databases of input features and labeled outputs to work correctly: this kind of database of distributions would be very problematic and time-consuming for researchers to build. Nevertheless, unsupervised learning extracts patterns and information from untagged data, thus being much more efficient and suitable for the task (Chollet, 2017). As the literature demonstrates, ML gives good results when applied to data from simulations (Dupuis et al., 2020): however, such data is less noisy and more smooth than real distributions from MMS. A long pre-processing is therefore necessary, where it is critical to deal with problems such as optimization and missing data in order to smooth the clustering.

In addition to the already mentioned Aidapy, the platform PySPEDAS (Grimes et al., 2022) and other packages for data analysis, ML, numerical, and visualization libraries in *Python* help to facilitate rapid development and deployment of ML algorithms. Thanks to its simple, user-friendly nature, Python has become the most popular language to build, train and test neural networks, and a fundamental tool for developing ML solutions with high iteration velocity.

2 GMM Approach to Synthetic and Observational Data

We motivate the choice of unsupervised techniques in the classification of the velocity distribution functions. The main difference between supervised and unsupervised ML is the usage of labelled train data in supervised learning. The model learns the relation between the labelled inputs and outputs and applies this knowledge to the unseen data. The scientist in the loop would thus identify some features in the data so the algorithm can train on these chosen features. This however imposes some bias into the learning algorithm because the features are based on the existing knowledge of the scientist. Unsupervised methods on the other hand impose no such bias and are preferable to extract features or clusters from the existing data.

In addition to the imposed bias of supervised ML; we also prefer unsupervised due to the relatively small size of the downloaded VDF datasets.

2.0.1 The Gaussian Mixture Model

A Gaussian mixture is defined as function comprised of several Gaussians (usually the same amount as clusters in the data). Each of these Gaussians has the usual parameters of mean μ and covariance Σ, defining the centre and width of the Gaussian respectively (see figure 1). The final parameter is called the mixing probability π, and will de-

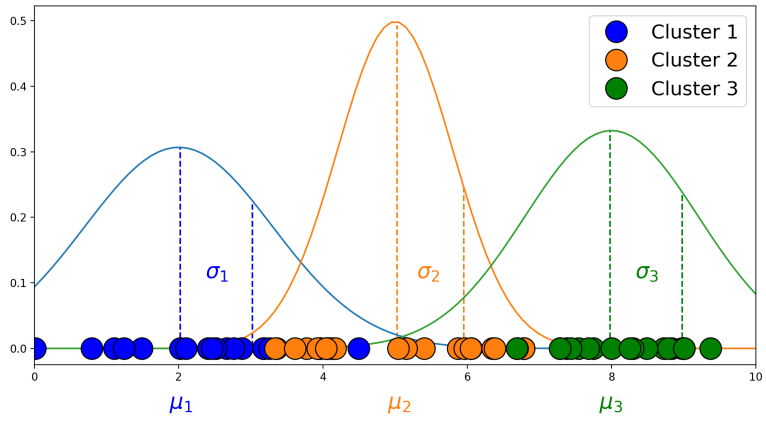

Figure 1: Three Gaussians depicted above their respective clusters with means and covariances shown.

fine how big the Gaussian function will be (Bouguila & Fan, 2020)(Moitra, 2018). The

Gaussian density function is given by:

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)). \tag{1}$$

Here the $\mathbf{x}$ represents the data points. The mixture model therefore states that all the data points are generated from a mixture of such distributions:

$$P(x) = \sum_{i=1}^{k} P_i \mathcal{N}(\mathbf{x}|\mu,\Sigma), \tag{2}$$

with k the amount of clusters and P_i is the weight of the respective Gaussian. The parameters of the distributions are fitted with the expectation-maximization method (EM), which starts of with randomly chosen values for the parameters and, after calculating the probability that the data points were generated by these Gaussians, changes them iteratively. This process maximizes the likelihood that the data were generated from the Gaussians with the chosen parameters.

The GMM is one of the primary models we will be interested in when studying the velocity distributions observed by the MMS mission. The amount and complexity of the clustering will be a good indicator of interesting regions in space where magnetic reconnection or other magnetic action might be present.

2.1 Clustering Statistics

We need a way to analyse the appropriate amount of clusters k that best describes the dataset. One way to evaluate this would be to manually go over certain particle generations and pick out the visible clusters. This way a training dataset of 'truth values' can be created and used to train a neural network to find the optimal amount of clusters. Because this is a tedious and largely subjective task, we will prefer to work with unsupervised training algorithms, as mentioned before, which do not require a predefined cluster assignment. This does not mean that we will not be evaluating the performance of the different algorithms visually at all, because this is of course still the best clustering tool at our disposal (Pedregosa et al., 2011).

2.1.1 Bayesian Information Criterion

The Bayesian Information Criterion or BIC is in simple terms an information criterion that tries to balance the correctness of the fit of the model and the complexity of the model. If the complexity would not be taken into account, we run the risk of over-

fitting. On the other hand the correctness of the fit should be seen as a way to counter underfitting. The criterion is defined as follows:

$$BIC = kln(n) - 2ln(\hat{L}). \tag{3}$$

The two parts that are balanced are: k, the number of parameters estimated by the model and n, the number of data points that express the complexity of the model, while $ln(\hat{L})$ is the maximum log-likelihood that computes the goodness of the fit.

This criterion is often used to evaluate the number of clusters in Gaussian mixture models and is also a build-in in the scikit-learn GMM-module.

An often used alternative is the Akaike Information Criterion or AIC. This is given by:

$$AIC = 2k - 2ln(\hat{L}). \tag{4}$$

The difference is immediately clear: it does not involve a logarithmic term in the complexity measure (BIC penalizes the number of parameters in the model to a greater extent). While BIC assumes that the true model is in the candidate set and we simply want to find it, AIC only tries to find the model that most adequately describes the dataset (Konishi & Kitagawa, 2007).

2.1.2 Silhouette Coefficient

The Silhouette Coefficient uses the means intra-cluster distance (a) and the mean nearest-cluster distance (b) to calculate a measure of the goodness of the fit. This is a simple heuristic given by:

$$Silhouette - coefficient = \frac{b-a}{max(a,b)}. \tag{5}$$

As we can tell by the formula, the clustering is better for higher values of the coefficient. This means that points in a cluster lie closer to each other than to points from another cluster (b larger than a). As this heuristic can not assess the goodness of a fit of only one cluster (unlike BIC and AIC that can) we should be careful when applying it to decide whether there are two or one clusters present in the experimental data. It will however still give a good insight in the complexity of the distributions.

2.1.3 Calinsky-Harabasz Index

Another often used metric for evaluating clustering algorithms that uses the ratio of the sum of between-cluster dispersion and of within-cluster dispersion matrices (Pedregosa et al., 2011). This score is also easy to interpret: the higher the index, the more dense and well separated the clustering is, and thus the better. This score also fails to evaluate single cluster allocations like the silhouette coefficient.

3 GMM Applied to Synthetic Distributions

3.1 Single Gaussian

Let us first asses the performance of the GMM (from the sklearn library in python) on synthetic distributions. To do this we look at the most trivial example, namely a number of particles generated from one single Gaussian distribution. To select the optimal number of clusters we will always use the BIC-score since it is much faster and more precise than the previously mentioned silhouette- and CH-score. This information criterion always selects a one-component model and therefore the GMM can always successfully reconstruct the mean speed and temperatures of the generated Gaussian distribution.

3.2 Mixture of Gaussians

The performance of the GMM on multiple Gaussian clusters is dependant on two main factors; the variances of the different clusters and the separation between the clusters. When generating clusters from distributions with the same variance, it is of course optimal to select the 'tied' covariance type inside the GMM. This means that all components share the same general covariance matrix. When dealing with a more general case where particles can be drawn from Gaussian distributions with different variances, the 'full' covariance type will be optimal. To illustrate this, figure 2 shows the BIC-scores for 1-10 clusters using both covariance types on a dataset of 4 clusters generated from Gaussian distributions with different variances. We can see that the 'full' covariance type leads to the correct prediction of the number of components.

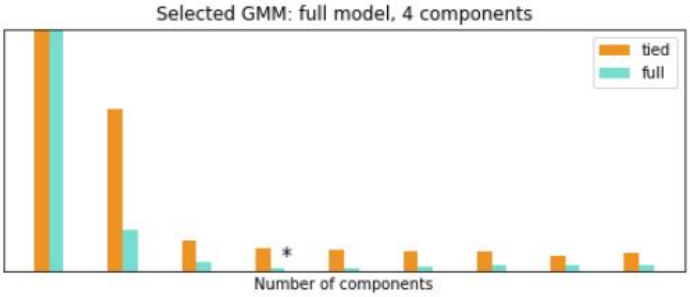

Figure 2: BIC-scores for two different covariance types for an instance where the clusters have different variances.

As expected, the GMM succeeds in identifying the optimal number of components when dealing with all-Gaussian clusters. We can also show that the model is capable of accurately predicting the means and variances of the clusters. The accuracy of these predictions do however go down when dealing with clusters that are not well separated. To illustrate this, we will plot the normalized error of the predicted means and variances for varying distance between the clusters. Lets look at an instance of 4 clusters.

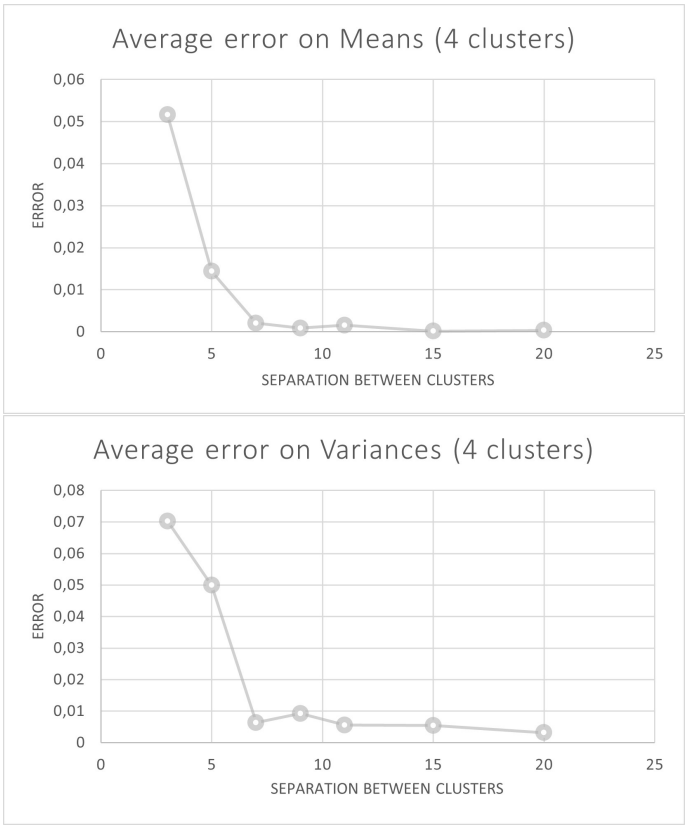

Figure 3: Error on the mean and variance predictions of the GMM on 4 Gaussian clusters for decreasing distance between the respective clusters. The distance on the horizontal axis is the maximum distance between the three velocity components of the means of different clusters (in units of 10^6 m/s). On the vertical axis, the error on the total velocity is displayed, averaged over the 4 clusters.

As expected, the Gaussian mixture model works excellent for recreating the means and variances of a number of Gaussian sampled clusters. It is only at the point where the clusters almost completely coincide that the accuracy of the predictions start to decline rapidly.

3.3 Single Kappa

The previous results look rather promising however, as we know the clusters observable in the electron VDF will not be completely Gaussian in nature. These clusters are characterised by non Maxwellian suprathermal tails. These tails decrease as a power law of the velocity (Pierrard & Lazar, 2010). As in most literature about space plasmas,

we will fit these distributions by the Kappa distribution (Pierrard & Meyer-Vernet, 2017; Maksimovic et al., 1997; Kim et al., 2015). The spectral index of these distributions determines the slope of the distribution. In the limit where $\kappa \to \infty$, the distribution simplifies to a Maxwellian (see fig. 4).

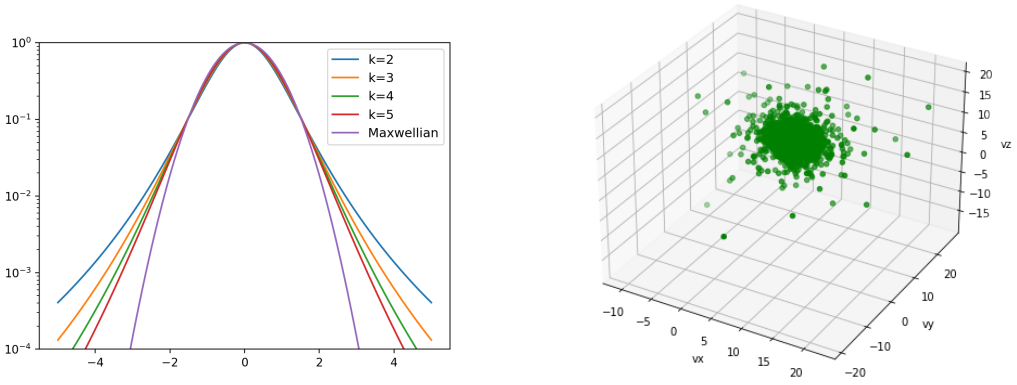

Figure 4: (Left) The Kappa distribution function for several values of the spectral index. (Right) Particles randomly generated from a Kappa distribution with $\kappa = 2$ (in units of 10^6 m/s).

It is also important to note that the value of the spectral index must be chosen as such that it is not too close to the critical value $\kappa_c = 1.5$. At this value the distribution function collapses (Pierrard & Lazar, 2010). Following observations and satellite data, kappa distributions with a spectral index $2 < \kappa < 6$ seem to be a good fit (Shohaib et al., 2022) and thus satisfies this requirement.

We will now see if the GMM is still capable of identifying the correct number of clusters and their means/variances if these clusters are not sampled from a Gaussian distribution, but from a kappa distribution. Let us start with one single kappa distributed cluster. In the previous section, when generating Gaussian clusters, we selected $\kappa = 200$. We now bring down this value until the GMM does no longer makes the right prediction. Below $\kappa = 6$, the GMM predicts a significantly larger amount of clusters due to the outliers generated from the kappa distribution. Since we are interested in kappa values below this, we need to improve the performance. This can be done by increasing the convergence threshold of the EM-algorithm that the GMM uses. This way the algorithm goes to fewer iterations and is less likely to overestimate the number of actual clusters.

Using this method, we get down to $\kappa = 2.6$ before the model starts overfitting the number of clusters. This process also inherently speeds up the computation since we need fewer iterations. Remarkably, as we will see in the next section, this overfitting does not occur when more than one cluster is present and we can get our spectral index as low as $\kappa = 2$. This way we can correctly predict all distributions that model the velocity in space plasmas.

We conclude that the GMM can correctly reconstruct the mean speeds and temperatures of a single kappa distributed cluster for $\kappa = 2.6$ and up.

3.4 Mixture of Kappa's

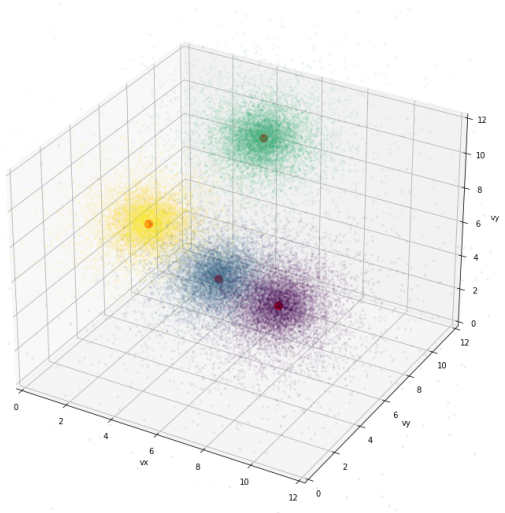

Figure 5: Predicted clusters by BIC-score after GMM. Means of clusters are highlighted in red (graph in units of 10^6 m/s).

In complete analogy with the mixture of Gaussian clusters, we look at the GMM performance when the dataset would consist of a mix of different kappa distributed clusters. We do however now have an extra parameter to take into account, namely: Do all clusters have the same kappa value or not? Contrary to the single-cluster case, the GMM model is able to correctly predict the number of components in a multi-cluster dataset, even when they are generated by distributions with a spectral index between 2 and 2.6. The primary deciding factor of the precision of these predictions will thus once again be the separation between the different clusters. We will analyse the results for an instance

where all clusters are generated by a full $\kappa = 2$ distribution, and another instance where the spectral indices can vary randomly between $\kappa = 2$ and $\kappa = 6$.

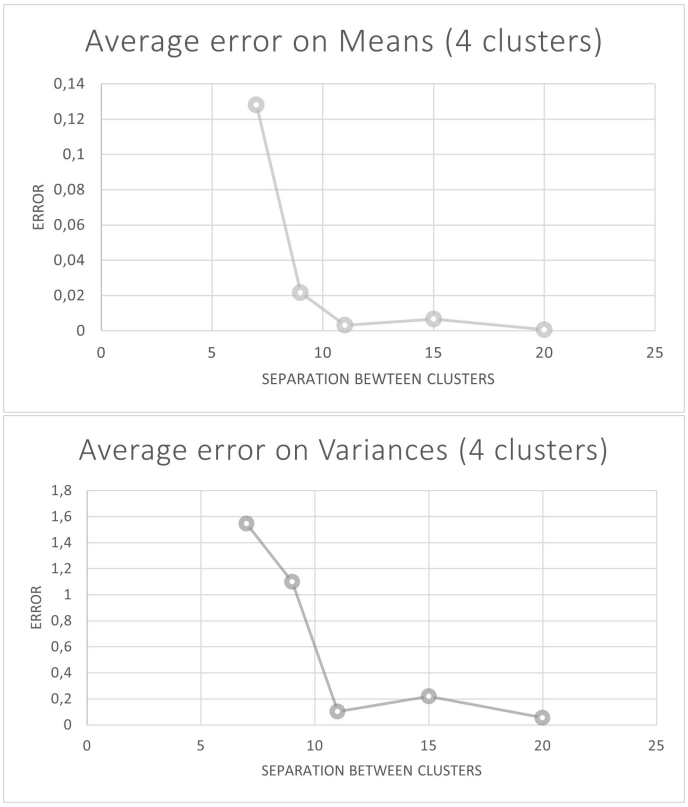

Figure 6: Error on the mean and variance predictions of the GMM on 4 Kappa-distributed clusters (all same κ) for decreasing distance between the respective clusters. The distance on the horizontal axis is the maximum distance between the three velocity components of the means of different clusters (in units of 10^6 m/s). On the vertical axis, the error on the total velocity is displayed, averaged over the 4 clusters.

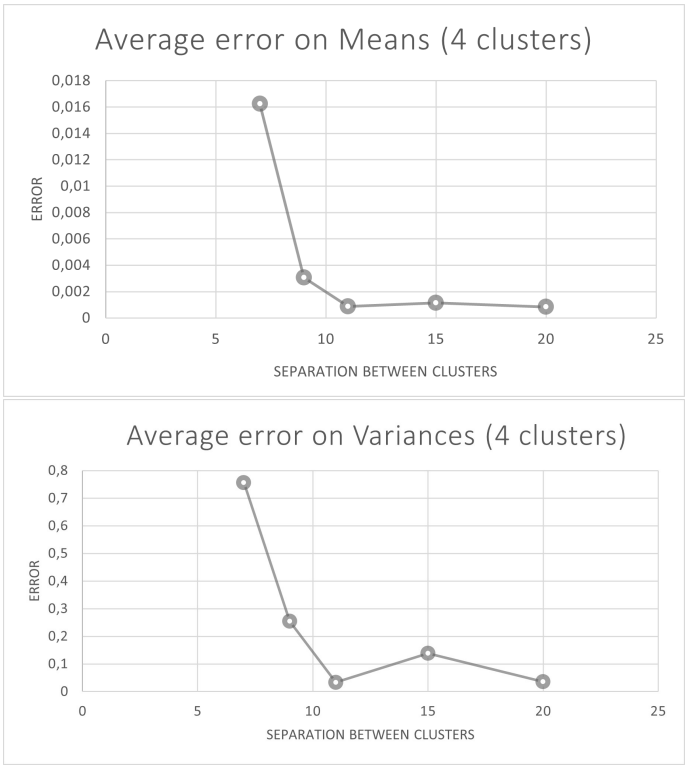

Figure 7: Error on the mean and variance predictions of the GMM on 4 Kappa-distributed clusters ($2 < \kappa < 6$) for decreasing distance between the respective clusters. The distance on the horizontal axis is the maximum distance between the three velocity components of the means of different clusters (in units of 10^6 m/s). On the vertical axis, the error on the total velocity is displayed, averaged over the 4 clusters.

From these results we conclude that the performance of the GMM is not really affected by clusters with a non-equal spectral index. In previous examples we always ran the GMM for 1-10 components. If we pick a higher number for the maximum amount of components, we risk overfitting some simpler cases with few components by assigning a really high number of clusters. This also reduces the run time of the program when we will run it for more than 100 electron distributions. The model is still capable of representing the complexity of the distribution, which is in essence what we are after. This means that instances with over 10 components will almost always be assigned 10 components for the optimal solution, and not some smaller number.

3.5 Spectral Clustering

Another clustering technique that can be considered in the context of electron VDFs is spectral clustering. This method performs a dimensionality reduction based on the spectrum of the similarity matrix (Bonaccorso, 2017)(Pedregosa et al., 2011). When spectral clustering is applied to a Gaussian/Kappa distributed dataset like before, we see that the algorithm is indeed able to correctly classify most clusters using the Silhouette Coefficient and Calinsky-Harabasz Index as clustering statistics (BIC and AIC are not available for this clustering algorithm). However, using this algorithm comes with a big jump in time complexity as well as the added run time that results from the change in clustering statistics. Nevertheless, spectral clustering could become relevant when using the 'SpectralClustering' function of *scikit-learn*. This function allows the user to use parallelization, which will split the work across different CPU cores and decrease the execution time (Brownlee, 2020). Several efforts were made to to get the time complexity down this way to make it comparable to GMM, but this does not yet yield the desired results. These attempts included using the cores of our local machines as well as utilizing several cores of the tier-2 Genius cluster of the VSC (Flemish Supercomputer Center). As for now, the achieved time complexity is not yet one to rival our GMM results.

4 Data and Methods

For the purpose of testing the capability of the GMM to define the complexity of a measured dataset, we utilized data from NASA's MMS Mission. Fast Plasma Investigation (FPI) instrument measures incoming particles through a filter which selects certain particle speeds and directions; then a 3D picture of the ion plasma is produced every 150 milliseconds, while for electron plasma FPI captures a picture every 30 milliseconds (Pollock et al., 2016). *Aidapy*, an high level Python package for the analysis of spacecraft data from heliospheric missions developed by ESA, is used to download data from FPI and other on-board instruments.

Unlike simulation data, when working with real data one has to deal with specifics of instruments. In the case of FPI, low energy particle counting is perturbed by the electrical charge of the probe: this leaves a gap in the center of measured VDFs. Before processing the data with GMM, the Python package Scikit-learn is used to perform linear interpolation to fill the gap in the VDFs. Particles are then generated from VDFs for

input into the GMM: to avoid noisy distributions with few particles and ones which are too demanding in terms of numerical resources, a number of particles of 40,000 is chosen by authors' experience. The information criterion chosen to select the number of optimal beams of the mixture is BIC, as it is to be found preferable to AIC in working with a large number of real data.

Data were selected from two distinct intervals in which magnetic reconnection signatures were found. In the literature, different types of reconnection signatures have been identified (M. Goldman et al., 2016) . Several of these signatures have been observed in the events analyzed in this work. The first event is from December 8, 2015, when the reconnecting dayside magnetopause was crossed by MMS probes. The second one is from July 3, 2017 and it was observed in the magnetotail.

5 GMM applied to MMS data in the dayside magnetopause

During the event on December 8, MMS spacecraft was at first in the magnetosheath, but the magnetopause moved outward causing the spacecraft to move to the magnetopause (Burch & Phan, 2016). The crossing of a reconnecting magnetopause is recognized at 11:20:42-11:20:45 UT, because of the behaviour of several physical quantities: intense current density, strong electric field and high speed electron outflows. Four seconds of data observations from MMS3 were processed with the GMM technique, from 11:20:41 to 11:20:45. Every 0.03 s an electron VDF is measured by the spacecraft, for a total of 133 VDFs analyzed within the interval. For each VDF the BIC information criterion assignes a score based on how accurate the fit is. The fit is then repeated with a different number of clusters. The maximum number of possible clusters is an input for the GMM: from the authors'experience, 10 is a reasonable compromise between result accuracy and computation time.

The bottom plot of figure 8 shows the normalized results of the GMM analysis on the VDFs. High values of the score (white lines) shows a difficulty in the fit of the distribution function, which may be associated with a complex VDF. In fact, the information criterion tends to assign smaller scores to the best fits, which correspond to more Maxwellian VDFs. As the plot shows, during the reconnecting magnetopause crossing (between 11:20:43 and 11:20:44), the VDFs tend to become more complex. For a visual verification, three VDFs taken at different times are shown in figures 9, 10 and 11. The

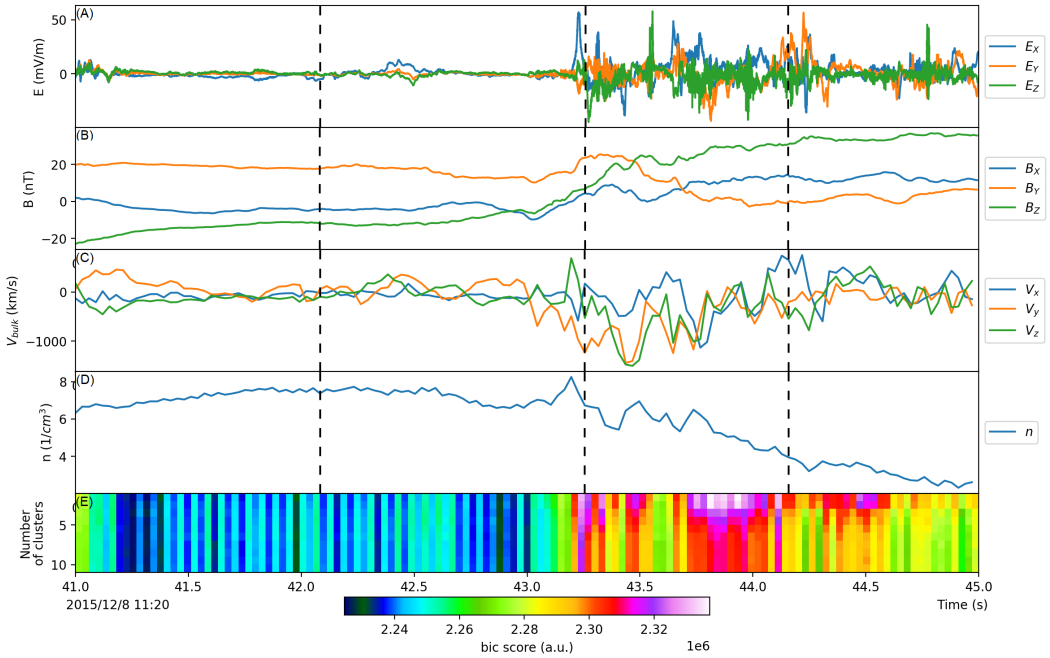

Figure 8: Data from 8 December 2015. Vectors are expressed in Geocentric Solar Ecliptic System (GSE), with X-axis pointing from the Earth towards the suns, its Y-axis chosen to be in the ecliptic plane pointing towards dusk and Z-axis parallel to the ecliptic pole. GSE components of magnetic and electric field are shown. Electric and magnetic field are shown in panels (A) and (B), along with electron bulk velocity (C) and density (D). Panel (E) shows the GMM results. Vertical dashed lines represent the times when VDFs are measured.

first one is taken at 11:20:42:08, when the crossing of the reconnecting region had not yet happened and the BIC score is near its minimum. The second and third one, which appear visually complex and show strong asintropy, are taken after the reconnection event, where the BIC score indicates a worst fit. This capability of the GMM to automatically recognize regions where VDFs present non-Maxwellian features is of crucial importance for the detection of reconnection regions within the plasma.

Figure 12 illustrates the Gaussians parameters found by the GMM applied to one of the distributions measured before the reconnection. The number of components provided as input to the algorithm is three. The ellipses evidence the mean and the vari-

ance of each Gaussian of the mixture, and the weight represents the probability that a random particle belongs to one of the three components.

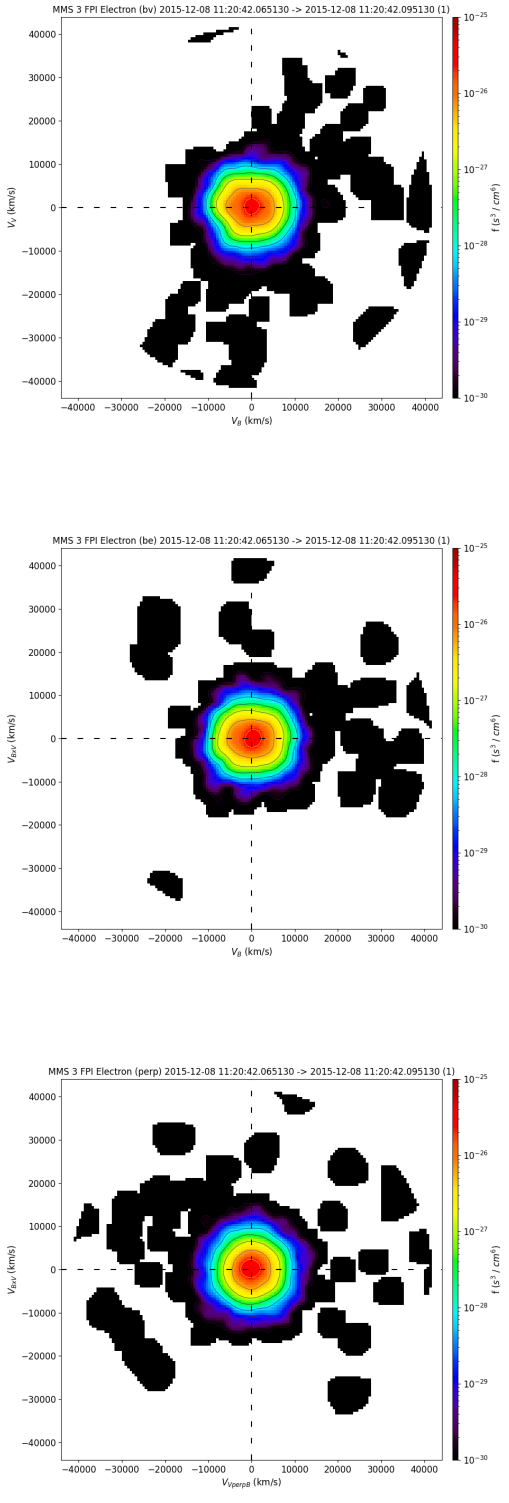

Figure 9: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

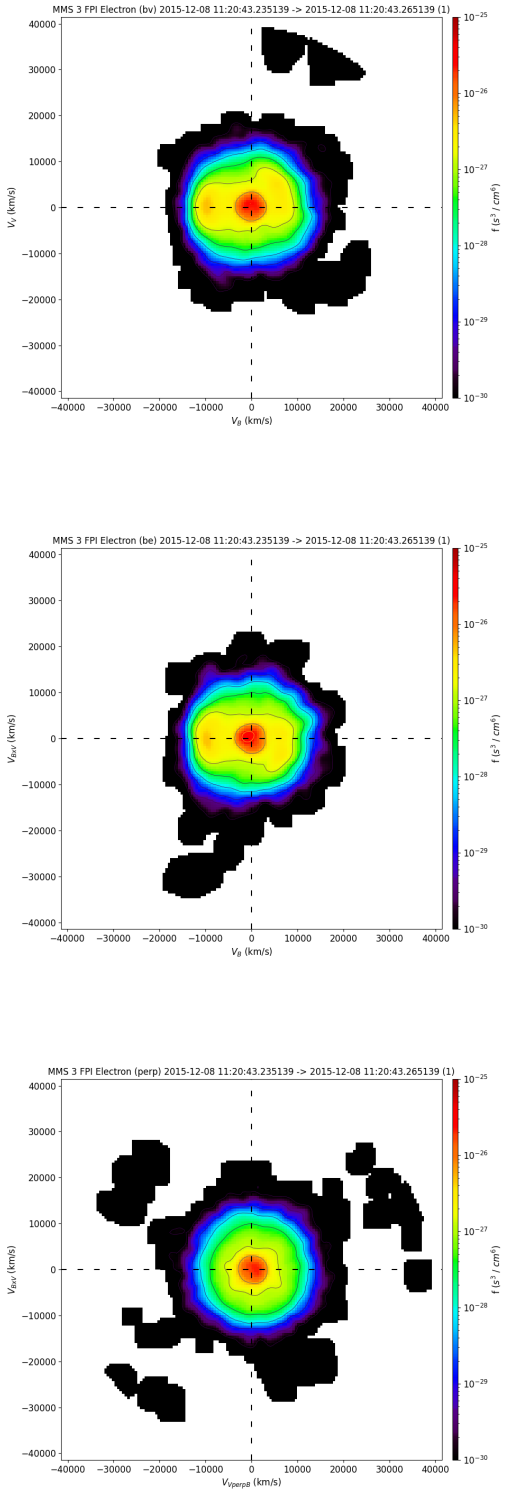

Figure 10: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

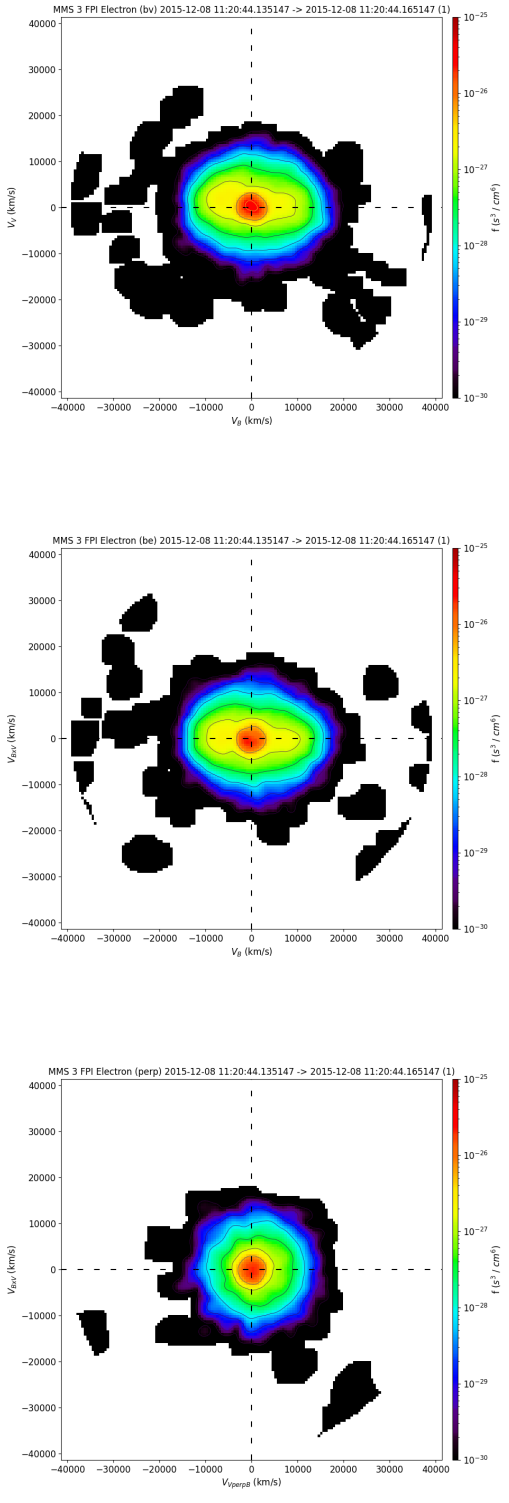

Figure 11: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

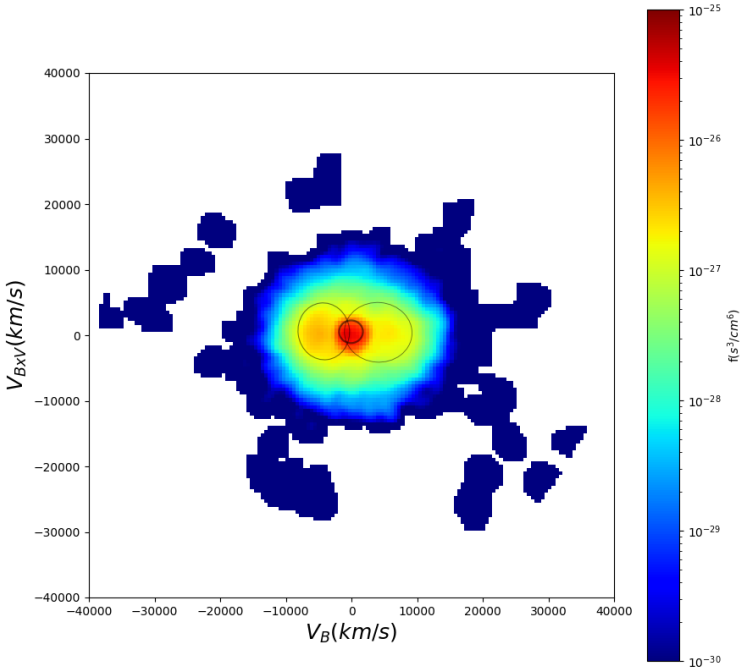

Figure 12: VDF cut from the dayside magnetopause taken at 11:20:44.18, before the reconnection. $\mathbf{V_B}$ is parallel to the magnetic field and $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$. The black ellipses show the different Gaussians of the mixtures based on their mean and variance. The transparency of the ellipses is determined by the weight of each Gaussian.

6 GMM applied to MMS data in the magnetotail

During the event of 3 July 2017 MMS3 spacecraft was in the magnetotail, and observed another reconnection event (Burch et al., 2019). Again, four seconds of data were analyzed, from from 05:26:48 to 05:26:52. Reconnection X-line is observed near 05:25:50:72.

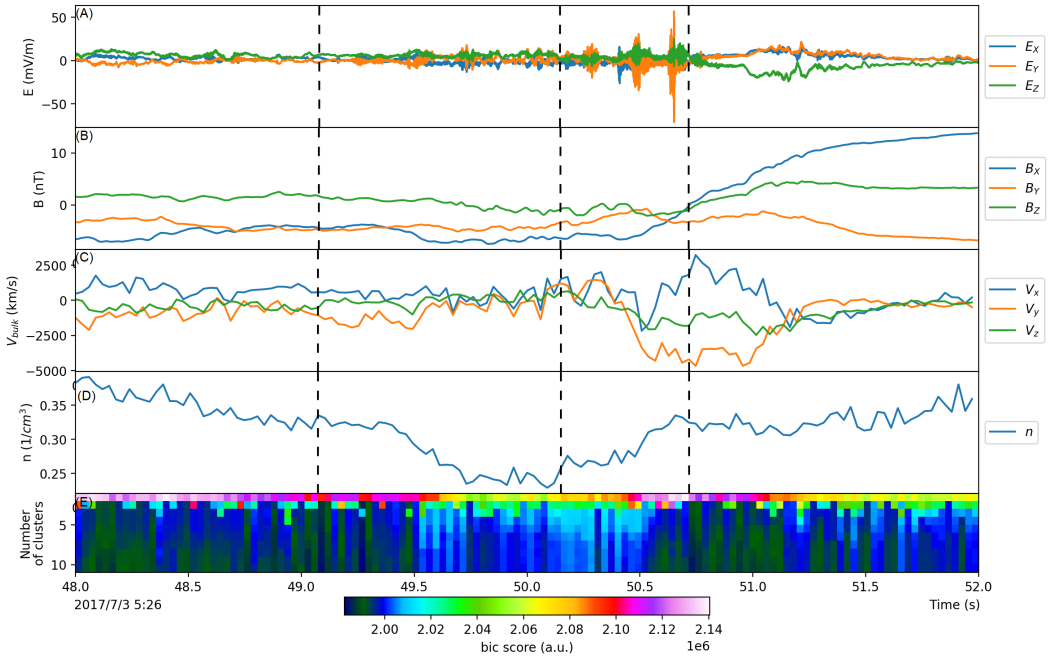

Figure 13: Data from 3 July 2017. Vectors are expressed in Geocentric Solar Ecliptic System (GSE), with X-axis pointing from the Earth towards the suns, its Y-axis chosen to be in the ecliptic plane pointing towards dusk and Z-axis parallel to the ecliptic pole. GSE components of magnetic and electric field are shown. Electric and magnetic field are shown in panels (A) and (B), along with electron bulk velocity (C) and density (D). Panel (E) shows the GMM results. Vertical dashed lines represent the times when VDFs are measured.

As shown in figure 13, until 05:25:49:50 VDFs appear Maxwellian. More complex VDFs are recognized for nearly a second, until they return simpler. The simplification of the VDFs occurs after the time when reconnection is observed, near 05:25:50:72. The GMM again succeeds in detecting the reconnection region through the clustering result. Three VDFs are shown in figures 14, 15 and 16 for visual verification. As expected from GMM results, the two VDFs with smaller BIC scores show Maxwellian features, while

the VDF from 5:26:50:15, before the reconnection event, appears strongly complex with large peaks in the $\mathbf{V_B}$ direction. The algorithm was again able to automatically recognize the most complex distributions within the time interval with great accuracy.

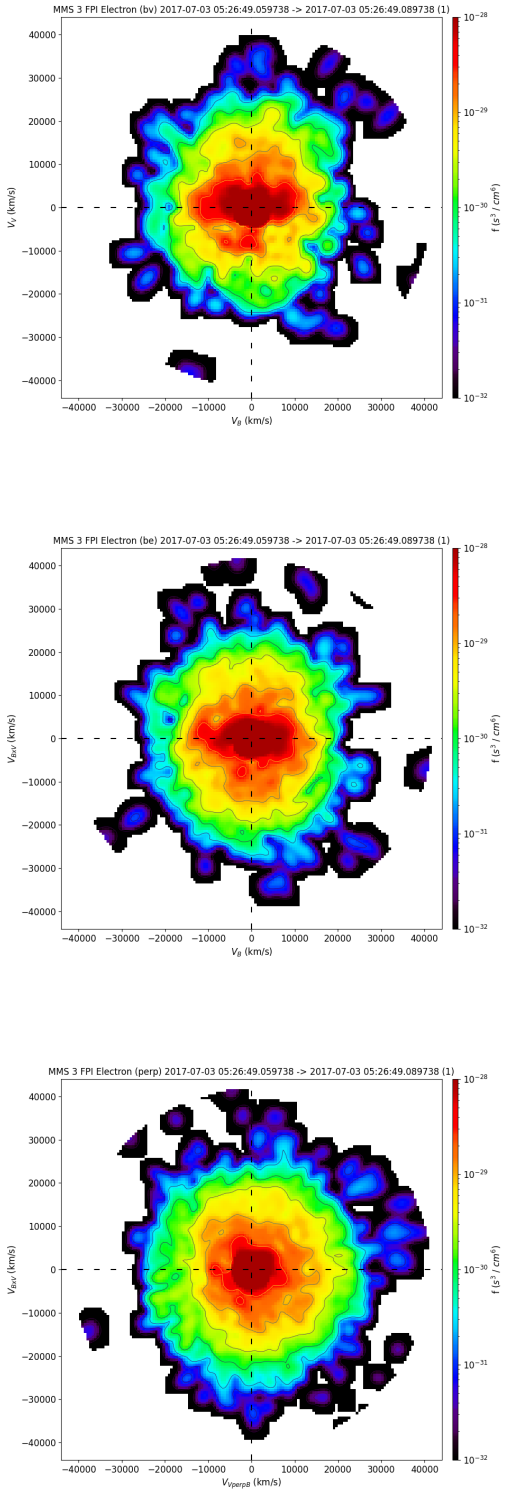

Figure 14: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

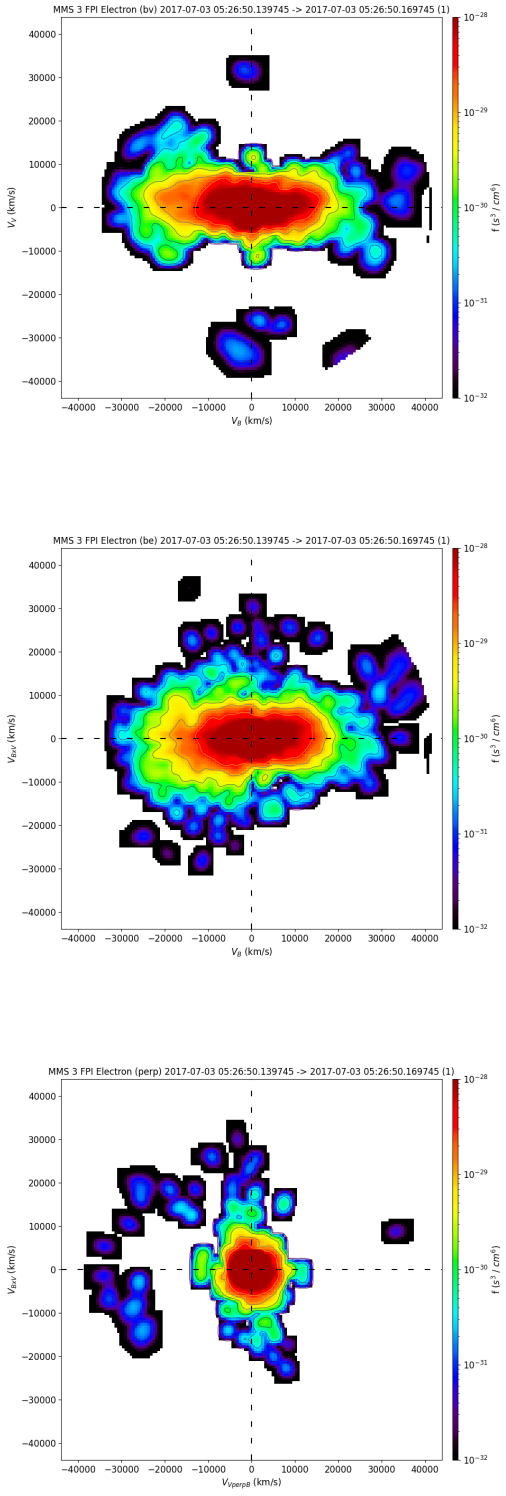

Figure 15: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

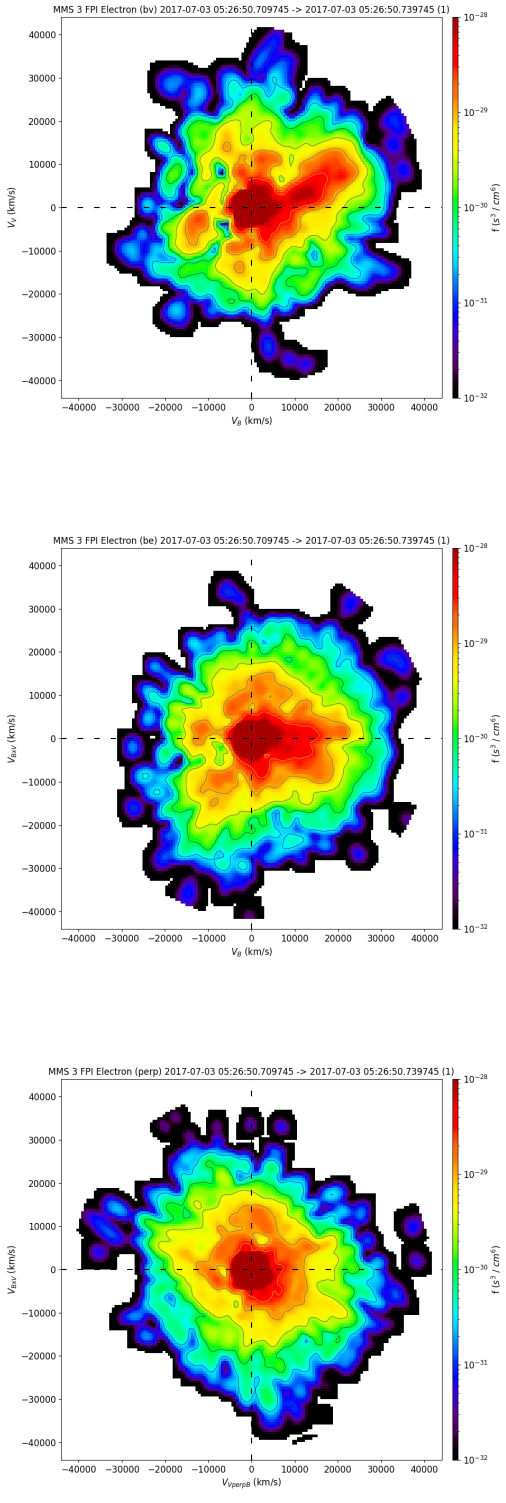

Figure 16: Three cuts of the VDF taken at 11:20:42:08, before the reconnection event. $\mathbf{V_B}$ is parallel to the magnetic field, $\mathbf{V_V}$ is in the direction of the bulk velocity, $\mathbf{V_{BxV}}$ is in the direction of $\mathbf{B} \times \mathbf{V}$ and $\mathbf{V_{VperpB}}$ is the bulk velocity projected onto the plane normal to $\mathbf{B}$.

7 Conclusions

In this paper we have investigated the effectiveness of the Gaussian mixture model in 1) recognizing the complexity of the distributions (i.e. the number of components) and 2) recreating the means (mean velocity) and variances (temperatures) of the clusters in electron VDF's.

The tests that were run on synthetic distributions proved the ability of the algorithm to accurately predict the complexity of kappa distributed clusters (with the maximum number of components set tot 10 as to prevent overfitting) as well as reconstructing the means and variances of these clusters. From the numerical results, it is obvious that the GMM algorithm still performs the better on the Gaussian distributed clusters as expected. Nevertheless, for clusters whose means are well separated (i.e. more than 10^7 m/s apart), the results for the kappa distributed particles are comparable to the Gaussians. We also remark that it is not important if the clusters have the same spectral index ($\kappa = 2$), or if the spectral index is different for each cluster ($2 < \kappa < 6$).

With regard to the analysis of real data, we applied GMM to MMS data in order to automatically identify magnetic reconnection sites through the complexity of the velocity distribution functions.

For the analysis, we selected time intervals from articles where magnetic reconnection events were identified from observations of particular variations in electric and magnetic fields, current density, and particles behavior. After preprocessing the data, which includes filling in the gap in the VDFs at low energies, we analyzed the particles generated by distribution with the GMM. We utilized the Bayesian Information Criterion to choose the best fitting amount of clusters within the distributions. The model has shown that it is able to capture the variation in complexity of the functions, arriving through this to automatically locate reconnection sites with good accuracy. In addition to this, a visual test showed that the BIC scores can accurately indicate the most complex distributions that show strong non-Maxwellian features.

In recent years the task of looking at the raw data from the spacecrafts and selecting the interesting ones was done by eye by scientists. Due to limitations of the probes, a continuous overwriting of data takes place and a large part of it is lost. Future goals include further improving the unsupervised ML techniques so that them can be used to analyze the data and collect the most interesting ones without having to lose a lot of im-

portant information. Just as with synthetic data, we favor GMM with Bayesian information criterion thanks to its efficiency and accuracy.

Acknowledgments

This project has received funding from the *ERC Advanced Grant TerraVirtuale* of GL (grant agreement No. 1101095310) and from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101082633 (ASAP). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Further funding was provided by the KULeuven Bijzonder OnderzoeksFonds (BOF) under the C1 project TRACESpace and from the NASA HSR Grant 80NSSC21K1689. FV acknowledges the support of the PRIN 2022 project "The ULtimate fate of TuRbulence from space to laboratory plAsmas (ULTRA)" (2022KL38BK), funded by the Italian Ministry of University and Research. This paper and related research have been conducted during and with the support of the Italian national inter-university PhD programme in Space Science and Technology.

Computing has been provided by the Flemish Supercomputing Center (VSC).

Data Availability Statement

The source code for data preprocessing and GMM analysis is available at `https://github.com/NathanNoaMaes/VDFClusteringMMS`. MMS data are available at `https://lasp.colorado.edu/mms/sdc/public/data/`. Data analysis was performed using Aidapy available at `https://gitlab.com/aidaspace/aidapy`, PySPEDAS available at `https://github.com/spedas/pyspedas` and scikit-learn available at `https://github.com/scikit-learn/scikit-learn`.

References

Baker, D. N., Riesberg, L., Pankratz, C. K., Panneton, R. S., Giles, B. L., & Wilder, R. E., F. D.and Ergun. (2016, mar). Magnetospheric multiscale instrument suite operations and data system. *Space Science Reviews*, *199*. Retrieved from `https://doi.org/10.1007/s11214-014-0128-5` doi: 10.1007/s11214-014-0128-5

Biskamp, D. (2000). *Magnetic reconnection in plasmas*. Cambridge University Press. doi: 10.1017/CBO9780511599958

Bonaccorso, G. (2017). *Machine learning algorithms.* Packt Publishing Ltd, 2017.

Bouguila, N., & Fan, W. (2020). *Mixture models and applications.* doi: 10.1007/978-3-030-23876-6

Brownlee, J. (2020, 09). Multi-core machine learning in python with scikit-learn. `https://machinelearningmastery.com/multi-core-machine-learning-in-python/`.

Burch, J. L., Dokgo, K., Hwang, K. J., Torbert, R. B., Graham, D. B., & Webster, e. a., J. M. (2019). High-frequency wave generation in magnetotail reconnection: Linear dispersion analysis. *Geophysical Research Letters*, *46*, 4089–4097. doi: https://doi.org/10.1029/2019GL082471

Burch, J. L., Moore, T. E., Torbert, R. B., & Giles, B. L. (2016). Magnetospheric multiscale overview and science objectives. *Space Science Reviews*, *199*. doi: https://doi.org/10.1007/s11214-015-0164-9

Burch, J. L., & Phan, T. D. (2016). Magnetic reconnection at the dayside magnetopause: Advances with mms. *Geophysical Research Letters*, *43*, 8327–8338. doi: https://doi.org/10.1002/2016GL069787

Chollet, F. (2017). *Deep learning with python.* Manning.

Dupuis, R., Goldman, M. V., Newman, D. L., Amaya, J., & Lapenta, G. (2020, jan). Characterizing magnetic reconnection regions using gaussian mixture models on particle velocity distributions. *The Astrophysical Journal*, *889*(1), 22. Retrieved from `https://doi.org/10.3847%2F1538-4357%2Fab5524` doi: 10.3847/1538-4357/ab5524

Goldman, M., Newman, D., & Lapenta, G. (2016). What can we learn about magnetotail reconnection from 2d pic harris-sheet simulations? *Space Sci Rev*, *199*, 651–688. doi: https://doi.org/10.1007/s11214-015-0154-y

Goldman, M. V., Newman, D. L., Eastwood, J. P., & Lapenta, G. (2020). Multibeam energy moments of multibeam particle velocity distributions. *Journal of Geophysical Research: Space Physics*, *125*(12), e2020JA028340. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JA028340` (e2020JA028340 2020JA028340) doi: https://doi.org/10.1029/2020JA028340

Grimes, E. W., Harter, B., Hatzigeorgiu, N., Drozdov, A., Lewis, J. W., Angelopoulos, V., ... Le Contel, O. (2022). The space physics environment data

analysis system in python. *Frontiers in Astronomy and Space Sciences*, *9*. Retrieved from https://www.frontiersin.org/articles/10.3389/fspas.2022.1020815 doi: 10.3389/fspas.2022.1020815

Hoshino, M., Hiraide, K., & Mukai, T. (2001, 06). Strong electron heating and non-maxwellian behavior in magnetic reconnection. *Earth Planets Space*, *53*, 627-634. doi: 10.1186/BF03353282

Kim, S., Yoon, P., Choe, G., & Wang, L. (2015, 06). Asymptotic theory of solar wind electrons. *The Astrophysical Journal*, *806*. doi: 10.1088/0004-637X/806/1/32

Konishi, S., & Kitagawa, G. (2007). *Information criteria and statistical modeling.* doi: 10.1007/978-0-387-71887-3

Lapenta, G., Goldman, M., Newman, D., & Markidis, S. (2016, May). Where should mms look for electron diffusion regions? *Journal of Physics: Conference Series*, *719*, 012011. Retrieved from http://dx.doi.org/10.1088/1742-6596/719/1/012011 doi: 10.1088/1742-6596/719/1/012011

Li, T. C., Liu, Y.-H., & Qi, Y. (2021, mar). Identification of active magnetic reconnection using magnetic flux transport in plasma turbulence. *The Astrophysical Journal Letters*, *909*(2), L28. Retrieved from https://doi.org/10.3847%2F2041-8213%2Fabea0b doi: 10.3847/2041-8213/abea0b

Maksimovic, M., Pierrard, V., & Riley, P. (1997). Ulysses electron distributions fitted with Kappa functions. *Geophysical Research Letters*, *24*, 1151-1154. Retrieved from https://hal.archives-ouvertes.fr/hal-03801373 doi: 10.1029/97GL00992

Moitra, A. (2018). *Algorithmic aspects of machine learning.* Cambridge University Press. doi: 10.1017/9781316882177

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pierrard, V., & Lazar, M. (2010, 03). Kappa distributions: Theory and applications in space plasmas. *Solar Physics*, *267*. doi: 10.1007/s11207-010-9640-2

Pierrard, V., & Meyer-Vernet, N. (2017). Chapter 11 - electron distributions in space plasmas. In G. Livadiotis (Ed.), *Kappa distributions* (p. 465-479). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/

pii/B9780128046388000115 doi: https://doi.org/10.1016/B978-0-12-804638-8.00011-5

Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., ... Zeuch, M. (2016). Fast Plasma Investigation for Magnetospheric Multiscale. *Space Science Reviews*.

Retinò, A., Sundkvist, D., Vaivads, A., Mozer, F., André, M., & Owen, C. J. (2007, April). In situ evidence of magnetic reconnection in turbulent plasma. *Nature Physics*, *3*(4), 236-238. doi: 10.1038/nphys574

Scott, B. (2021). *Turbulence and instabilities in magnetised plasmas, volume 1*. IOP Publishing. Retrieved from https://dx.doi.org/10.1088/978-0-7503-2504-2 doi: 10.1088/978-0-7503-2504-2

Shohaib, M., Masood, W., Siddiq, M., Alyousef, H., & El-Tantawy, S. (2022, 04). Formation of electrostatic solitary and periodic waves in dusty plasmas in the light of voyager 1 and 2 spacecraft and freja satellite observations. *Journal of Low Frequency Noise, Vibration and Active Control*, *41*, 146134842210913. doi: 10.1177/14613484221091340

Shuster, J. R., Chen, L.-J., Daughton, W. S., Lee, L. C., Lee, K. H., Bessho, N., ... Argall, M. R. (2014). Highly structured electron anisotropy in collisionless reconnection exhausts. *Geophysical Research Letters*, *41*(15), 5389-5395. doi: https://doi.org/10.1002/2014GL060608

Yokoi, N., & Hoshino, M. (2011, October). Flow-turbulence interaction in magnetic reconnection. *Physics of Plasmas*, *18*(11). Retrieved from http://dx.doi.org/10.1063/1.3641968 doi: 10.1063/1.3641968