Adaptive PCA-Based Outlier Detection for Multi-Feature Time Series in Space Missions*

Jonah Ekelund[1], Savvas Raptis[2], Vicki Toy-Edens[2], Wenli Mo[2], Drew L. Turner[2], Ian J. Cohen[2], Stefano Markidis[1]

[1] *KTH Royal Institute of Technology, Computer Science Dept., Stockholm, Sweden*
[2] *Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA*

Abstract. Analyzing multi-featured time series data is critical for space missions making efficient event detection, potentially onboard, essential for automatic analysis. However, limited onboard computational resources and data downlink constraints necessitate robust methods for identifying regions of interest in real time. This work presents an adaptive outlier detection algorithm based on the reconstruction error of Principal Component Analysis (PCA) for feature reduction, designed explicitly for space mission applications. The algorithm adapts dynamically to evolving data distributions by using Incremental PCA, enabling deployment without a predefined model for all possible conditions. A pre-scaling process normalizes each feature's magnitude while preserving relative variance within feature types. We demonstrate the algorithm's effectiveness in detecting space plasma events, such as distinct space environments, dayside and nightside transients phenomena, and transition layers through NASA's MMS mission observations. Additionally, we apply the method to NASA's THEMIS data, successfully identifying a dayside transient using onboard-available measurements.

Keywords: Outlier detection, Incremental PCA, Space mission data, Online learning

1 Introduction

Space missions, such as NASA's Magnetospheric Multiscale (MMS) Mission [1] and Time History of Events and Macroscale Interactions during Substorms (THEMIS) [2], generate large volumes of multi-featured time series data from in-situ measurements. These data require efficient methods for identifying and prioritizing scientifically relevant events. More broadly, large-scale data collection is essential across various scientific domains, including Earth-based weather monitoring [3] and Hyper Spectral Imaging (HSI) from Earth observation (EO) satellites [4]. In space missions, limited onboard computational resources and

* This work is supported by the European Commission, with Automatics in Space Exploration (ASAP), project no. 101082633.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

restricted downlink capacity necessitate real-time event detection to prioritise the most valuable data for transmission [1].

A challenge in analyzing space mission data is detecting outlier data points. These outliers may indicate both instrument failures and important physical phenomena occurring in the near-earth space environment, such as plasma environment crossings, magnetic reconnection events, or transient events. While some outliers may result from sensor anomalies or noise, others may reveal structures or processes not accounted for by existing models [5]. Traditional anomaly detection methods often assume static data distributions, which is generally not true for space plasma environments: dynamic changes in solar wind conditions and magnetospheric interactions can lead to evolving data characteristics.

One additional challenge for in-situ automatic data analysis is that onboard computational resources are largely constrained. These devices can only support lightweight and adaptive algorithms capable of processing high-dimensional data streams in real time [6]. Many existing methods require extensive prior knowledge of the data or rely on complex models that are impractical for deployment on resource-limited spacecraft. An effective approach should balance computational efficiency with the ability to detect scientifically meaningful outliers without extensive pre-training on specific datasets.

In this work, we present an outlier detection algorithm tailored for space missions. The method leverages reconstruction error from Principal Component Analysis (PCA) for feature reduction and adapts dynamically to new data distributions using Incremental PCA. This enables real-time outlier detection onboard spacecraft. Our focus is on evaluating the algorithm's principle for use in detecting outliers and demonstrating the algorithm's effectiveness in identifying plasma events such as bow shock crossings in MMS data and foreshock bubbles in THEMIS observations. The main contributions of this work are the following:

- Introduce an adaptive outlier detection algorithm based on the PCA reconstruction error, which can be applied to streaming data. The algorithm works without a pre-training step and dynamically adapts to the dominant features in new data.
- Extend the algorithm to handle different feature types with different magnitudes by coupling the feature scaling to a group with features of the same type. This approach preserves the relative variance within each feature group.
- Demonstrate the algorithm's effectiveness in finding boundary crossings and other scientifically interesting events in MMS data using multiple features.
- Show that the algorithm can detect scientifically relevant events in data available onboard the spacecraft using the THEMIS mission, potentially supporting real-time identification.

2 Background

In this work, we focus on the use-case of finding outliers in multi-featured space plasma time-series data using MMS [1], launched in 2015 and THEMIS, launched in 2007 [2], as examples. MMS was launched to investigate magnetic reconnection

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

in the boundary regions of Earth's magnetosphere using unprecedented time scales. It consists of four spacecraft flying in formation through the dayside magnetopause and the nightside magnetotail regions [1].

The THEMIS mission was launched to investigate the trigger and evolution of substorms. The mission consists of five satellites, lining up to track particles along the magnetotail. While the primary mission goal was to perform measurements in the magnetotail, in the nightside region of Earth's magnetosphere, the spacecraft also obtained measurements from the dayside region [2]. In this work, we have used an interval of dayside data from THEMIS.

Both MMS and THEMIS have limited storage capabilities, constraining the amount of data that can be collected. Furthermore, the presence of limitations in the downlink capabilities means that the collected data of the highest possible value has to be prioritized. The collection limitation and prioritization were performed using pre-specified temporal region-of-interests (ROI) with corresponding locations in space, using onboard calculated indicator values and human-made selections. These selections are intended to prioritize the most valuable mission data, especially those that are significantly different from the general state of the regions and have a high likelihood of containing interesting events [1, 2].

As spacecraft orbit Earth, they transition through multiple regions, including the magnetosphere, where Earth's magnetic field shapes the plasma environment and areas beyond its protective influence, where the solar wind and the Sun's magnetic fields dominate. They also traverse key boundaries such as the bow shock, where the solar wind slows and deflects and the magnetopause, which separates Earth's magnetic domain from the solar wind. Classifying which region the measurements belong to can help direct scientists toward phenomena of more interest [7]. Determining the regions and finding the crossings between them can help to optimize the data collection onboard the spacecraft.

2.1 Outlier Detection in Streaming Data

Outlier detection deals with finding samples, or groups of samples, that differ from the general structure of the remaining samples. Traditionally, most outlier detection methods deal with static data where the entirety of the data range is known [8]. If the data is high-dimensional, feature reduction techniques, such as Principal Component Analysis (PCA) [9] or Autoencoders [10], can be utilized to simplify the problem and reduce the search space [5].

However, streams of data add to the complexity of finding outliers. The temporal variability of the data means that what is considered an outlier can shift with time. The nominal values of a feature and feature importance may also shift in time [5]. This can render an initial data model incorrect and any outlier detection algorithm needs to be able to adapt to these shifts in the data.

PCA is a linear feature reduction method that surfaces the top N features, the Principal Components (PC) containing the most variance (i.e. information), from the original feature space [11]. Once in the reduced feature space, the samples can be separated into groups using different clustering techniques [10, 12]. Then, outliers can be located based on the distance to these clusters' centroids or

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

the density. Another option for finding outliers is to transform the samples back to the original feature space and evaluate how much the reconstructed features (R_f) differ from the original features (F_f). This reconstruction error (E_f):

$$E_f = R_f - F_f \tag{1}$$

is the loss of information from the feature reduction. Samples where features have large reconstruction errors will deviate from the PCA model and can, therefore, be considered outliers compared to the other samples [13]. Incremental PCA is a variant of PCA that can be used to build the PCA model incrementally [11]. This can be necessary if the data is too large to load all the samples into memory at once, or for the use-case presented in the paper, not all the data is available when the initial model is created.

3 Methodology

In this work, we present a method for detecting outliers in streaming multi-feature data based on evaluating the reconstruction error stemming from feature reduction using Incremental PCA [11]. The reconstruction error E_i, for a sample i, is calculated as the Euclidean norm of the reconstruction errors E_f for all the features f. By looking for large and rapid changes in this error, we can find samples that deviate from the model and the previous samples.

3.1 Outlier Detection

The outlier detection algorithm[1] operates in three different modes, Initialization, Check and Calibrate, as shown in Fig. 1. In the initialization mode, (Fig. 1 - top), the initial model is built based on the first samples retrieved by the algorithm. Following the initialization mode is the check mode, which is the main operating mode (Fig. 1 - bottom right). Here, samples are retrieved and evaluated for outliers. If multiple sequential samples are labeled as outliers, then a calibration using these outlier samples is initiated. In the calibration mode, the PCA model is updated based on the latest outlier samples.

There are five parameters controlling the execution of the algorithm: the calibration buffer size (S_c), the number of components (N) used in the PCA, the mean buffer size (S_m), the threshold (λ) and the outlier limit (L_o). S_c controls the size of the calibration buffer (c_buffer). The calibration buffer is used to store samples to use for calibration. When this buffer is full, either in the initialization mode or the check mode, the PCA is calculated using the samples in this buffer. N controls how many components of the PCA are calculated and thereby the size of the reduced feature space. A lower value will, in general, mean a higher reconstruction error, but changing this parameter can also change what events are discovered by the algorithm. S_m controls the size of the mean buffer (m_buffer). The mean buffer is a fixed width circular buffer used to store the reconstruction error (E_i) of the latest S_m samples not labeled as outliers. This buffer is then used to calculate the current mean μ and standard deviation σ. The μ and σ

[1] `https://github.com/Jonah-E/multi-feature-outlier-detection`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

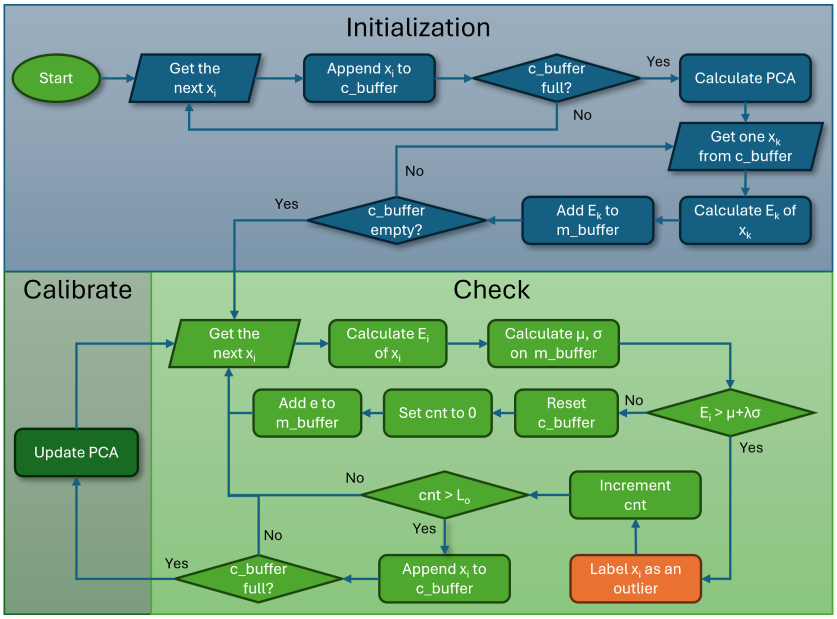

Fig. 1: Flowchart describing the working principles of the adaptive algorithm.

values are then used together with the threshold λ to calculate the maximum allowed reconstruction error, or the error threshold T_i, according to the equation

$$T_i = \mu + \lambda\sigma \tag{2}$$

A sample x_i with error E_i exceeding T_i will be labeled an outlier. The last parameter, L_o, controls how many samples x_i can be labeled as outliers before samples are added to the calibration buffer. Setting this to a value larger than zero will exclude the L_o first outlier samples from the calibration buffer.

Initialization mode: is the start of the algorithm. Here, the algorithm builds the initial model by retrieving the next sample x_i and then adding it to the calibration buffer. When the calibration buffer is full, the PCA is calculated using the samples in the calibration buffer. Following this, the reconstruction error (E_k) of the samples (x_k) in the calibration buffer is calculated and added to the mean buffer. Depending on the application of the algorithm, it could be beneficial to pre-calculate the PCAs on a curated dataset containing data from expected regions. Then, the initialization step could be skipped.

Check mode: The first step in the check mode is to retrieve the next sample x_i and calculate the E_i of this sample. Then μ and σ are calculated using the samples in the m_buffer and T_i can be calculated according to Equation 2. If the E_i is larger than T_i, then the sample is labeled as an outlier and a counter (cnt) is incremented. If this counter exceeds L_o, then the sample x_i is also added to the calibration buffer. If the sample is not an outlier, the calibration buffer is cleared, the counter is reset to 0 and the E_i is added to the mean buffer. Only samples not labeled as outliers are added to the mean buffer to prevent outlier samples from inflating the error threshold. The last step is to evaluate if the calibration buffer is full. If it is full, multiple samples in a row have been labeled as outliers and the PCA model is likely no longer correct. Then, a calibration is initiated.

The calibration mode: is where the PCA is updated using the samples from the calibration buffer, bottom left in Fig. 1, before the algorithm reenters check mode and the next sample is processed.

Table 1: MMS intervals with primary ROI, dayside data intervals are originating from Ref. [14] and nightside intervals from Ref. [15].

Nr	Data Interval	ROI	Day or Nightside
0	2017-12-14 16:00 → 2017-12-14 05:00	-	Dayside
1	2017-12-17 16:00 → 2017-12-17 22:00	17:52 → 17:54	Dayside
2	2018-01-12 00:50 → 2018-01-12 06:00	01:50 → 01:52	Dayside
3	2018-12-14 02:00 → 2018-12-14 07:20	04:21 → 04:22	Dayside
4	2018-12-10 04:00 → 2018-12-10 11:00	04:40 → 04:42 05:12 → 05:25 06:27 → 06:31	Dayside
5	2019-01-05 16:00 → 2019-01-05 19:00	17:38 → 17:41	Dayside
6	2021-01-12 00:00 → 2021-01-12 06:00	01:18 → 01:21	Dayside
7	2021-02-13 10:00 → 2021-02-13 18:00	11:05 → 11:06	Dayside
8	2022-05-02 18:00 → 2022-05-02 22:00	18:23 → 18:25	Dayside
9	2022-11-24 02:00 → 2022-11-24 09:00	04:16 → 04:18	Dayside
10	2023-01-16 06:00 → 2023-01-16 11:00	08:21 → 08:24	Dayside
11	2017-07-23 12:00 → 2017-07-23 18:00	16:55 → 16:56	Nightside
12	2021-08-14 16:00 → 2021-08-15 06:00	01:23 → 01:25	Nightside

3.2 Data and Pre-processing

MMS Data: The primary data used in this paper is from MMS. Specifically, the omni-directional ion spectrum and ion velocities in GSE coordinates from the Fast Plasma Investigation (FPI) instrument [16] and the magnetic field (hereafter referred to as the B-field) from fluxgate magnetometers (FGM), part of the FIELDS instrument suite [17], from the MMS-1 spacecraft. Table 1 lists eleven dayside intervals of MMS data together with some regions of interest. These intervals are either from MMS passing into or out of the Earth's magnetosphere from the Solar wind. Transitions from the solar wind to the magnetosheath are called *bow shock* crossings, while transitions from the magnetosheath to the magnetosphere are called *magnetopause* crossings. The data from the FGM is collected at a significantly higher frequency, ∼ 8Hz to ∼ 16Hz, than the FPI data, ∼ 0.2Hz, we therefore down-sampled to the sample frequency of FPI. The data used is level-2 data, which is post-processed on Earth and would not be available onboard the spacecraft.

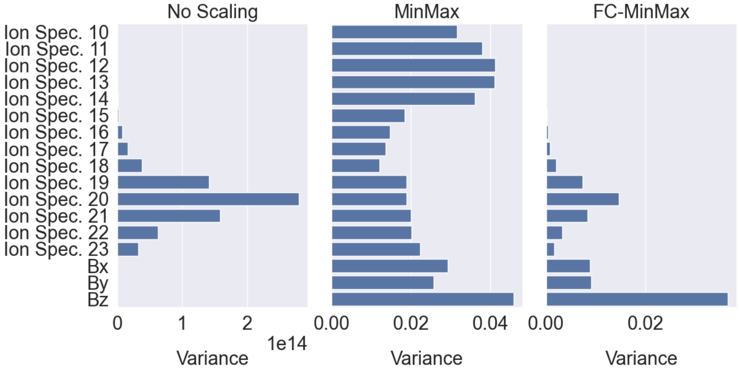

Fig. 2: Comparison of the variance of the different features for no scaling (Left), MinMax scaling (Middle) and FC-MinMax scaling (Right). The features are MMS omnidirectional ion spectrum, channel 10 to 23 and B-field from 2017-12-17, 20:00 to 21:50, while MMS is traversing the magnetosheath region.

The dayside ROI in Table 1 are taken from Raptis *et al.* [14] and are transient phenomena upstream of the bow shock in the ion foreshock region, which are known to energize particles and cause space weather effects [18]. Therefore, finding these, even in level-2 data, is important for further scientific research. In addition to these transient events, the data intervals also contain other phenomena, such as bow shock and magnetopause crossings, which are operationally critical to identify automatically.

We have used two nightside data intervals from MMS-1, listed in Table 1, to evaluate how the algorithm generalizes to different regions. These are intervals from the nightside magnetosphere of Earth, containing two events analyzed by Richard *et al.* [15]. These are fast plasma flows associated with geomagnetic disturbances and can, therefore, cause space weather effects [19, 20].

THEMIS spacecraft is used as an alternative source of space plasma data with different instruments compared to MMS. In particular, we use THEMIS-C probe measurements of the ion energy flux and ion velocity from the Electrostatic Analyzer (ESA), these are measurements with minimal processing, which are readily available for onboard algorithms [2, 21]. The interval investigated is from 2008-07-15 and contains a foreshock transient [22].

Multiple Features: The PCA method finds the axis with the highest variance; therefore, for PCA to be effective, all features must be normalized to the same range. When PCA is used in data with one type of feature, for example, the omnidirectional ion spectrum used in this paper, the values of all features can be expected to have the same magnitude. However, if multi-feature types are included, these new features can have values with magnitudes that differ significantly. Then, the feature types with the highest magnitudes will dominate the PCA. This can be seen in the leftmost plot of Fig. 2, where Ion spectrum channels 19 to 21 have the highest variance, while the B-field variance is not visible. A common way to solve this issue is to scale the features, for example,

DOI: 10.1007/978-3-031-97626-1_18

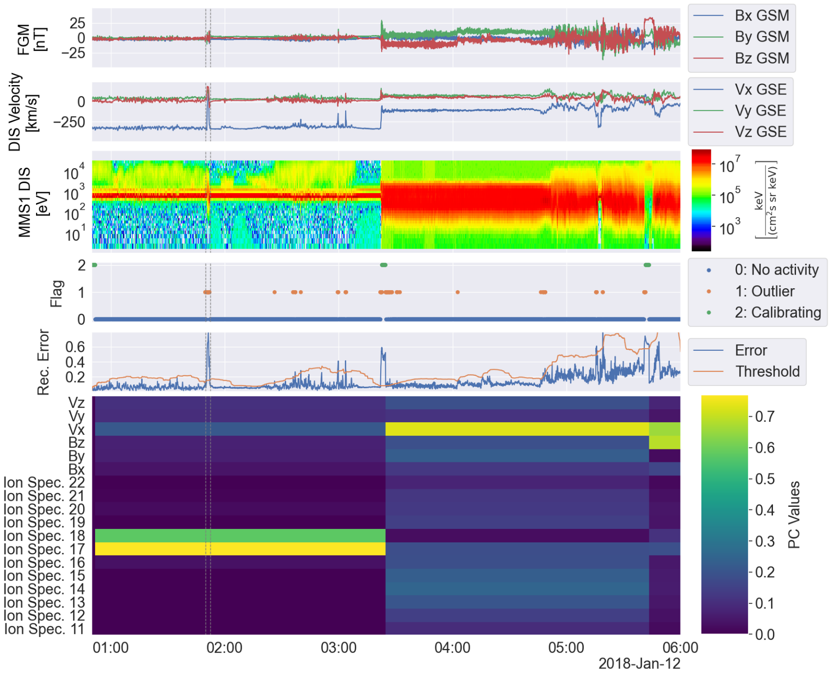

Fig. 3: Algorithm applied to multi-feature data from MMS dayside interval 2, the primary ROI is marked with dashed lines, $S_c = 25$, $S_m = 150$, $L_o = 10$ and $\lambda = 4$. The bottom plot is the principal component (PC) at each time; ion spectrum channels 0-10 and 23-31 have values close to zero and are not shown.

MinMax scaling, which scales each feature to be between zero and one. However, this destroys the relative variance within a feature type. This can be seen in the middle plot of Fig. 2, where the highest variance for the ion spectrum data has now moved to channels 12 and 13. To solve this issue, we introduce a coupling between features of the same type when scaling. This Feature Coupled MinMax (FC-MinMax) scaling scales the coupled features to the same min and max values, which are calculated over the group of features of the same type.

The variance after using the FC-MinMax scaling can be seen in the rightmost plot of Fig. 2. Here, the relative variance within each group, ion spectrum and B-field are retained while scaling the features to have the same magnitude. We can now see that the B-field, together with the ion spectrum, will affect the PCA. In our tests, we have used data interval 0 from Table 1 to calculate the scaling for each feature group when testing the data from the MMS dayside intervals. For other data, we have calculated the scaling based on the specific data.

4 Results

Fig. 3 shows the outlier detection algorithm performance when applied to a single time-series of MMS data. The features used are the ion omnidirectional

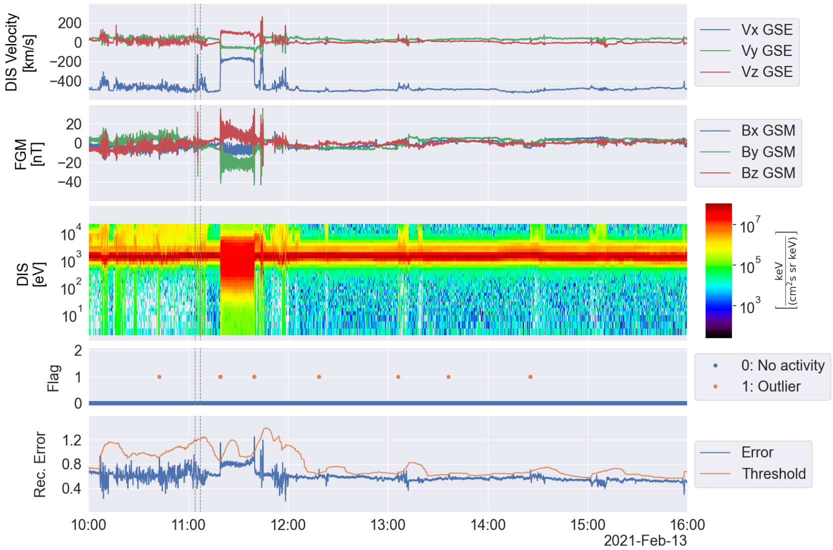

Fig. 4: Example of a ROI (dashed lines) the algorithm showed no detection.

spectrum, the B-field and the ion velocities. Each feature is scaled using the FC-MinMax scaling with the scaling factors calculated on data interval 0, see Table 1. The bottom three plots describe the functionality of the algorithm. Samples with the flag ‘outlier’ or ‘calibration’ are outlier samples. However, samples with the flag ‘calibration’ are used to update the PCA model.

The initialization stage can be seen furthest to the left in the plot; here, there are no outlier detections and the flag is set to ‘Calibrating’. Following this, we can see how the reconstruction error and detection threshold are updated as more samples are ingested. At 01:50, the algorithm detects ROI from Table 1. The next region of interest in the data is the bow shock crossing at 03:22, where the spacecraft enters the magnetosheath. The algorithm detects the crossing by the sudden increase in the reconstruction error. However, as the error does not decrease below the threshold of 35 samples ($L_o + S_c$), the entry triggers a calibration for the new region, which is shown by the samples labeled ‘Calibration’. After the calibration, the error decreases below the threshold. One more calibration is triggered at the short entry into the magnetosphere region at around 05:40 after a magnetopause crossing. Both of these crossings highlight different physical dynamics and their automatic finding can facilitate scientific research.

The bottom panel in Fig. 3 shows how each feature affects the principal component, inversely, this also shows which feature affects the reconstruction error the most. We can see there that in the solar wind region, the PCA model is dominated by the ion spectrum channels 17 and 18, while after entering the magnetosheath, multiple spectrum channels, together with the B-field and velocities, affect the model, where the Vx feature is now dominant.

Optimal parameters: As mentioned in 3.1, five parameters control the algorithm. We have performed a rudimentary parameter optimization on the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

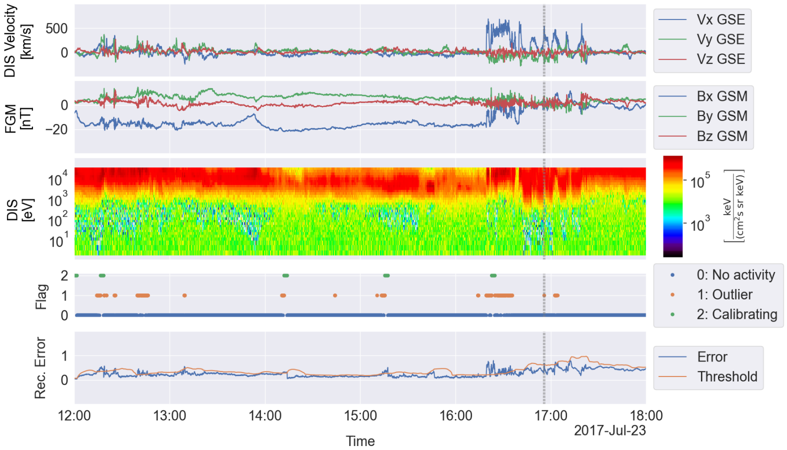

Fig. 5: The algorithm applied to MMS Nightside data interval 11 from Table 1.

set of dayside data intervals from Table 1, maximizing the detection of the listed ROIs. For this data, with the feature types, ion spectrum, B-field and ion velocities, we have found that setting the parameters as $S_c = 25$, $N = 2$, $S_m = 150$, $\lambda = 5$ and $L_o = 10$, balances finding the ROIs with false detections. With these parameters, the algorithm can find eleven out of the twelve dayside ROI specified in Table 1 when all the data intervals are fed in sequence to the algorithm without restarting it between intervals. However, the time shift between the intervals also introduces an artificial shift in the reconstruction error. When this shift is towards a larger error, the algorithm over-calibrates to the region at the start of the new interval, which can help it detect the primary ROIs. For a more realistic analysis, the mean buffer was reset to the first two samples when entering each new data interval. With these settings, most of the samples are marked as either 'No Activity' or 'Outlier', 98.7% and 1.1% respectively, only 0.2% are used to update the model.

The detection result for data interval 7 is plotted in Fig. 4. Here, the algorithm is missing the primary ROI. However, it does find the two crossings into and out of the magnetosheath at 11:19 and 11:39. The following quick crossing in and out of the magnetosheath, at 11:43 to 11:44, is, however, missed due to the sudden drop in reconstruction error from the first exit causing an increase in the standard deviation and thereby in the error threshold. If the ion spectrum is removed from the feature space, thereby only using the velocities and B-field, the marked ROI is found. All the crossings above are also found. However, multiple recalibrations are triggered at the entry to the magnetosheath.

MMS-1 Nightside Data: When applying the algorithm to the nightside data from Table 1, it can find the ROI at 2021-08-15 01:23 from the 12th data interval but not the ROI from the 11th interval. However, by increasing the number of PCAs to three and lowering the threshold to 3.5, the algorithm can find both regions. Here, the FC-MinMax scaling is calculated on the same data interval as the algorithm is applied. If scaling is instead calculated on dayside

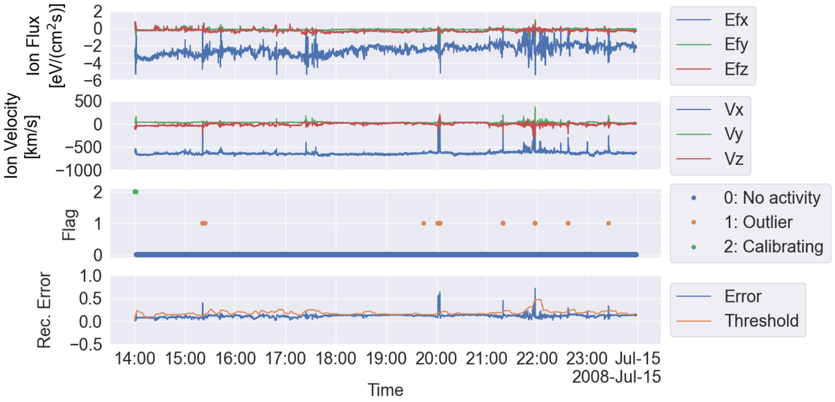

Fig. 6: The algorithm applied to an interval of dayside data from THEMIS C.

interval 0, the algorithm will over-calibrate to the ion velocities, as these have a larger value range in parts of the nightside data.

THEMIS C Dayside Data: Applying the algorithm to THEMIS C intervals, with a scaling calculated on the same interval, we find several possible interesting events; The longest of these is between 20:01:51 to 20:04:36 and 21:57:44 to 21:58:08. This second event starting at 21:57:44 has been noted in previous works as a foreshock bubble [22, 23]. The start and stop times indicated by our algorithm also correspond well to the start of expansion to end of compression, shown by Liu *et al.* [22] in Fig. 2. For this data, the calibration batch and mean window size have been increased to 30 and 180 due to the higher sampling frequency of the THEMIS C data.

For the full day, 2008-07-15, the first twelve hours of the data have values with significantly smaller magnitudes than the region in Fig. 6. When applying the algorithm to this wider data interval, with a scaling calculated on the full interval, the algorithm detects the start of the new region but does not trigger a recalibration. Lowering the threshold to 3.5 causes the algorithm to more aggressively recalibrate during the transition and find the events in Fig. 6

5 Related Work

Plasma region and Bow Shock identification: In the area of space plasma physics, bow shock crossings, which are coupled with the transition between the solar wind region and the magnetosheath are areas of active research. There have been numerous works developing methods for identifying plasma regions (e.g., solar wind and magnetosheath) and their transitions (e.g., bow shock). Both Olshevsky *et al.* [24] and Breuillard *et al.* [25] utilized neural-network-based approaches to identify different plasma regions. The bow shock transitions are either identified by where network output indicates a high probability for neighboring regions [24] or as specific labels [25]. These approaches are powerful and displayed high accuracies on the test sets. However, as the methods rely on

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18

having a training dataset, they cannot adapt to changing input data. While the algorithm presented in this paper can benefit from samples of relevant regions to create an initial model, the algorithm does not require a training set. Instead, it can be applied directly and will adapt to new data.

More recent work by Toy-Edens *et al.* [26] utilized a Gaussian Mixture Model to classify the MMS dayside region, which displayed high accuracy using a method more lightweight than a full neural network. Bakrania *et al.* [10] used a pipeline of autoencoder, PCA and Agglomerative Clustering to classify data from the ESA's Cluster spacecraft [27]. Innocenti *et al.* [12] utilizes PCA, self-organizing maps and K-means clustering to process multiple different quantities, such as B-field, particle velocity and particle density, to obtain a classification of the plasma region.

PCA-based outlier detection The two-step approach, presented by Finley *et al.* [9], analyzes a region of MMS data for interesting events by subdividing the region into smaller windows and performing a PCA on the resulting matrix of subregions. Then, a One-Classifier support vector machine (OCSVN) is used to find the windows containing outliers. This technique shows promising results; however, it is very computationally expensive, as the PCA and OCSVN have to be recomputed for each analyzed region. Zamry *et al.* [6] utilizes a PCA feature reduction coupled with a One-Class Support Vector Machine to detect anomalous sensor readings in data from a Wireless Sensor Network. Bhushan *et al.* [13] applied an Incremental PCA on streaming data with geospatial spread. They only consider finding outliers in one type of feature from one sensor type and the outliers are artificially introduced into the data. Furthermore, the PCA model is constantly updated, increasing the computational complexity. The algorithm presented in this paper limits the computational complexity by only updating the PCA model when necessary.

6 Discussion and Conclusion

In this work, we have presented a lightweight outlier detection algorithm for streaming multi-feature data. The algorithm can adapt to new data regions, enabling its deployment in applications where all the data regions are not known or fully understood. We utilize the Incremental PCA to have a lightweight algorithm that could be deployed on low-performing hardware. However, the basic principles in the algorithm can be applied using other unsupervised feature reduction techniques, where a reconstruction error can be calculated, for example, with autoencoders. Furthermore, while we have demonstrated the algorithm using space mission data, it can also be used to analyze other types of streaming data, such as measurements from a weather monitoring station. This will have to be evaluated for the specific data in question.

With optimal parameters, the algorithm can find most dayside ROIs, together with several crossings from solar wind and ion foreshock to the magnetosheath. For the nightside data, the number of components needed to be increased and the threshold slightly lowered. If we compare where our primary ROI is located in the dayside data to where it is located in the nightside data, we can understand why. In the dayside data, the ROIs are located in the Solar Wind/Ion foreshock regions, where the ion spectrum mainly consists of a tight beam of energy, as we can see in Fig. 3. This region is well-defined by only two parameters and outliers are visible in the reconstruction error. The nightside data is more varied around the marked ROI, leading to the ROI being masked by other errors. By increasing the number of components, the model can better describe the structure of the data, allowing the ROI to be found. This result indicates that the algorithm can also be used as a data mining tool to find interesting events in the dayside and nightside regions of the Earth's magnetosphere. This can help scientists find events for further studies. However, the different parts of the nightside can vary significantly and further studies will be necessary to understand how well the algorithm performs on data from the nightside region.

The algorithm can also be used on data from different spacecraft, as demonstrated by the use of THEMIS C data. In this case, the data used would be available onboard the spacecraft, indicating the algorithm's potential to be used onboard to inform the decision on which data to prioritize for downlink. In a real mission scenario, the algorithm would be applied to data which has already been subjected to a robust onboard pre-processing check for sensor failures. However, the algorithm needs to be more extensively tested using data available onboard, from multiple spacecraft, to show that the principle is robust enough for use in future space missions. Furthermore, the algorithm has to be evaluated on representative hardware before it can be used onboard during a space mission. Therefore, future work will focus on the creation and evaluation of a version of the algorithm optimized for low-performing hardware.

The FC-MinMax scaling used to scale the different feature types enables the use of multiple feature types with different magnitudes. It also opens up the possibility of adding a feature selection by weighing the different feature types differently. However, the necessity of this pre-scaling step is one of the two main weaknesses of the current algorithm, the second is how the threshold is defined. To further advance the algorithm, we will focus on this pre-scaling step and evaluate ways to integrate it into the algorithm itself. It would then be calibrated at the same time as the algorithm in the initialization stage. During a recalibration, the new data can be checked against the existing scaler to see if it needs to be updated. However, this could mean that the PCA model has to be recalculated and not just updated using the new samples.

For the second weakness, the main issue is that defining the threshold using the standard deviation can lead to large thresholds when there are large decreases in the reconstruction error, as seen in Fig. 4. This can lead to missed detections of interesting events. Furthermore, an incorrectly selected threshold can lead to excessive recalibration of the algorithm. To address these issues, the threshold could be dynamically updated either continuously or during calibration. Another option to limit excessive recalibration could be to introduce a forgetting factor to the IncrementalPCA. This would decrease the importance of earlier data and allow the new data to be more prominent in the model [11].

The algorithm presented in this paper is a promising approach to finding outliers in multi-feature data. It highlights areas of interest for further investigation and can inform decisions on which data should be prioritized for downlink. To further prioritize the data, the algorithm's output can be evaluated for the number of sequential outliers and the height of the error signal above the threshold.

References

[1] J. L. Burch *et al.*, "Magnetospheric multiscale overview and science objectives," *Space Science Reviews*, vol. 199, 2016.

[2] V. Angelopoulos, "The themis mission," in *The THEMIS Mission*, J. L. Burch *et al.*, Eds. Springer New York, 2009.

[3] R. C. Perez *et al.*, "Chapter 5 - oceanographic buoys: Providing ocean data to assess the accuracy of variables derived from satellite measurements," in *Field Measurements for Passive Environmental Remote Sensing*, N. R. Nalli, Ed., Elsevier, 2023.

[4] S.-E. Qian, "Hyperspectral satellites, evolution, and development history," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, 2021.

[5] I. Souiden *et al.*, "A survey of outlier detection in high dimensional data streams," *Computer Science Review*, vol. 44, 2022.

[6] N. M. Zamry *et al.*, "Lightweight Anomaly Detection Scheme Using Incremental Principal Component Analysis and Support Vector Machine," *Sensors*, vol. 21, 2021.

[7] V. Toy-Edens *et al.*, "Classifying 8 years of mms dayside plasma regions via unsupervised machine learning," *Journal of Geophysical Research: Space Physics*, vol. 129, 2024.

[8] C.-C. M. Yeh *et al.*, "Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016.

[9] M. G. Finley *et al.*, "Generalized Time-Series Analysis for In Situ Spacecraft Observations: Anomaly Detection and Data Prioritization Using Principal Components Analysis and Unsupervised Clustering," *Earth and Space Science*, vol. 11, 2024.

[10] M. R. Bakrania *et al.*, "Using dimensionality reduction and clustering techniques to classify space plasma regimes," *Frontiers in Astronomy and Space Sciences*, vol. 7, 2020.

[11] D. A. Ross *et al.*, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, 2008.

[12] M. E. Innocenti *et al.*, "Unsupervised classification of simulated magnetospheric regions," *Annales Geophysicae*, vol. 39, 2021.

[13] A. Bhushan *et al.*, "Incremental principal component analysis based outlier detection methods for spatiotemporal data streams," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, University of Pittsburgh, 2015.

[14] S. Raptis *et al.*, "Revealing an unexpectedly low electron injection threshold via reinforced shock acceleration," *Nature Communications*, vol. 16, 2025.

[15] L. Richard *et al.*, "Are dipolarization fronts a typical feature of magnetotail plasma jets fronts?" *Geophysical Research Letters*, vol. 49, 2022.

[16] C. Pollock *et al.*, "Fast Plasma Investigation for Magnetospheric Multiscale," *Space Science Reviews*, vol. 199, 2016.

[17] R. B. Torbert *et al.*, "The fields instrument suite on mms: Scientific objectives, measurements, and data products," *Space Science Reviews*, vol. 199, 2016.

[18] P. Kajdič *et al.*, "Transient upstream mesoscale structures: Drivers of solar-quiet space weather," *Frontiers in Astronomy and Space Sciences*, vol. 11, 2024.

[19] V. Angelopoulos *et al.*, "Bursty bulk flows in the inner central plasma sheet," *Journal of Geophysical Research: Space Physics*, vol. 97, no. A4, pp. 4027–4039, 1992.

[20] M. Sitnov *et al.*, "Explosive magnetotail activity," *Space Science Reviews*, vol. 215, 2019.

[21] *Variable descriptions*, Accessed: 2025-02-23. [Online]. Available: `https://themis.igpp.ucla.edu/var_desc.shtml`.

[22] T. Z. Liu *et al.*, "Relativistic electrons generated at earth's quasi-parallel bow shock," *Science Advances*, vol. 5, 2019.

[23] L. B. Wilson *et al.*, "Relativistic electrons produced by foreshock disturbances observed upstream of earth's bow shock," *Phys. Rev. Lett.*, vol. 117, 2016.

[24] V. Olshevsky *et al.*, "Automated classification of plasma regions using 3d particle energy distributions," *Journal of Geophysical Research: Space Physics*, vol. 126, 2021.

[25] H. Breuillard *et al.*, "Automatic Classification of Plasma Regions in Near-Earth Space With Supervised Machine Learning: Application to Magnetospheric Multi Scale 2016–2019 Observations," *Frontiers in Astronomy and Space Sciences*, vol. 7, 2020.

[26] V. Toy-Edens *et al.*, "Classifying 8 Years of MMS Dayside Plasma Regions via Unsupervised Machine Learning," *Journal of Geophysical Research: Space Physics*, vol. 129, 2024.

[27] C. Escoubet *et al.*, "Introduction the cluster mission," in *Annales Geophysicae*, Copernicus GmbH, vol. 19, 2001.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_18