Computer Physics Communications 185 (2014) 1669–1680

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Clustering of periodic multichannel timeseries data with application to plasma fluctuations

S.R. Haskey *, B.D. Blackwell, D.G. Pretty

Plasma Research Laboratory, Research School of Physics and Engineering, The Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form
3 March 2014
Accepted 10 March 2014
Available online 17 March 2014

Keywords:
Clustering
Periodic datamining
Von Mises distribution
Magnetic fluctuations
Plasma physics

ABSTRACT

A periodic datamining algorithm has been developed and used to extract distinct plasma fluctuations in multichannel oscillatory timeseries data. The technique uses the Expectation Maximisation algorithm to solve for the maximum likelihood estimates and cluster assignments of a mixture of multivariate independent von Mises distributions (EM-VMM). The performance of the algorithm shows significant benefits when compared to a periodic k-means algorithm and clustering using non-periodic techniques on several artificial datasets and real experimental data. Additionally, a new technique for identifying interesting features in multichannel oscillatory timeseries data is described (STFT-clustering). STFT-clustering identifies the coincidence of spectral features over most channels of a multi-channel array using the averaged short time Fourier transform of the signals. These features are filtered using clustering to remove noise. This method is particularly good at identifying weaker features and complements existing methods of feature extraction. Results from applying the STFT-clustering and EM-VMM algorithm to the extraction and clustering of plasma wave modes in the time series data from a helical magnetic probe array on the H-1NF heliac are presented.

1. Introduction

The identification and characterisation of plasma wave modes as a function of machine and plasma parameters is a subject of considerable interest for plasma magnetic confinement devices. As has been observed with Alfvén waves [1], high energy fusion alphas or neutral beam injection ions can interact with these modes, severely degrading their confinement and driving the modes to large amplitude [2]. This causes significant problems such as damage to the first wall [3], and may prevent fusion plasmas from reaching ignition. Diagnostics such as arrays of magnetic probes are critical for identifying and characterising the spectral and spatial nature of these modes. These diagnostics are "always on" on major experiments, generating extremely large databases of time-series data which provides a perfect opportunity for knowledge discovery using datamining techniques.

Data clustering, a recognised technique for unsupervised classification, has recently been applied to the field of plasma physics for intelligent data retrieval from large fusion device databases [4–7] and for the identification and classification of wave modes [8–10] using non-periodic clustering algorithms.

The techniques described in this paper address the problem of dealing with periodic data and can be applied to many applications where multichannel diagnostics produces periodic signals. Applications within plasma physics include interferometers, soft X-ray arrays, arrays of magnetic probes and imaging diagnostics. For simplicity we will focus on the application to magnetic probe signals where the spatial information, such as mode numbers, is encoded in the phase differences between magnetic probe signals at the frequency of the mode. These phase differences ($\Delta\psi$) are periodic, $(-\pi, \pi]$, causing problems with standard clustering techniques. Additionally, the number of probes available is often quite large giving rise to high dimensional data. These constraints require the application of specialised clustering techniques. Two options that have good memory scalability are a periodic version of the k-means algorithm and expectation maximisation (EM) using mixtures of multivariate independent von Mises distributions (EM-VMM). Minimal information is available in the literature about the application of EM to multivariate independent von Mises distributions with more than 3 variables, so this is described in detail in Section 4. Previously [8–10], clustering on timeseries data was performed using standard non-periodic clustering techniques by trigonometrically encoding the data ($\sin(\Delta\psi)$ and $\cos(\Delta\psi)$). This method has several drawbacks including artificially creating structure, encoding systematic errors in the data, and doubling the dimensionality of the problem.

* Corresponding author. Tel.: +61 428778853.
E-mail address: shaun.haskey@anu.edu.au (S.R. Haskey).

http://dx.doi.org/10.1016/j.cpc.2014.03.008
0010-4655/

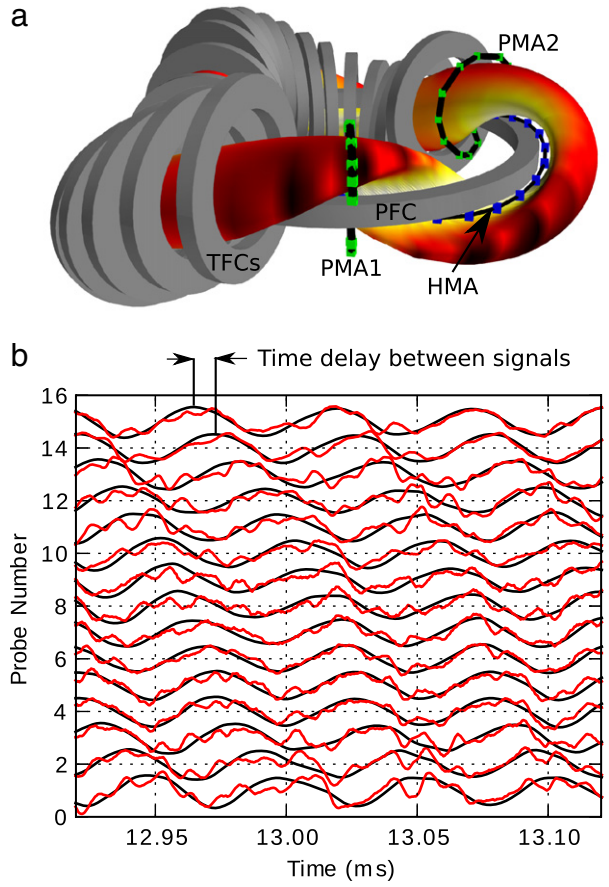

Fig. 1. (a) An overview of the H-1NF heliac including a subset of the equilibrium magnetic field coils (poloidal field coil (PFC), toroidal field coils (TFC)), the poloidal Mirnov arrays (PMA1, PMA2) and the helical Mirnov array (HMA). The surface colour represents the equilibrium magnetic field strength on the last closed flux surface. (b) Examples of the timeseries signals from the probes in the HMA, red is the raw signal and black is a bandpass filtered signal. The time delay in the signal between channels can be converted to a phase difference which represents the spatial structure of the mode. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Using several artificial datasets, we find that the EM-VMM algorithm performs better than the other available algorithms without incurring a significant computational cost. As a case study we successfully apply the EM-VMM algorithm to real data from the H-1NF heliac [11].

H-1NF is a three field-period helical axis stellarator with major radius $R = 1$ m and average minor radius $\langle r \rangle \approx 0.2$ m. The design of the machine allows access to an extensive range of magnetic configurations, making H-1NF well-suited to explore the relationship between plasma behaviour and magnetic configuration [12]. A variety of magnetic fluctuations have been observed with a recently installed helical Mirnov array (HMA) [13], which provides our experimental datasets in this paper. An overview picture of H-1NF including a subset of magnetic field coils and magnetic probe arrays as well as an example of the time trace signals from the HMA when a strong mode is present are shown in Fig. 1(a) and (b) respectively.

Additionally, a pre-processing technique for more robust identification of fluctuations in multichannel oscillatory timeseries data is described. The technique involves a combination of singular value decomposition (SVD) analysis, and an averaged short time Fourier transform followed by clustering (STFT-clustering). The STFT-clustering technique involves finding spectral features using the averaged short time Fourier transform followed by preliminary periodic clustering analysis to identify interesting features.

This paper is organised as follows: Section 2 provides an overview of the feature extraction and clustering process. Section 3 describes the STFT-clustering technique and how combining this with the SVD technique identifies features other techniques miss. Section 4 describes in detail how to apply the expectation maximisation algorithm to a mixture model of multivariate independent von Mises distributions. Section 5 compares the results of applying the periodic and standard clustering techniques to artificial data, and Section 6 shows results from applying STFT-clustering and EM-VMM to experimental data from the H-1NF heliac. Finally we provide some conclusions in Section 7.

2. Overview of the feature extraction and clustering process

For our application, we are ultimately interested in the physical nature of instabilities in plasmas, in particular, their dispersion relations. This information allows us to identify measures that can be taken to prevent these instabilities from growing to destructive amplitudes, and provides information on possible ways to use them beneficially.

Many different types of instabilities give rise to observable fluctuations in a magnetised toroidal plasma, for example ($n = 4$, $m = -3$) global Alfvén eigenmode (GAE), $(5, -4)$ GAE, etc. [1]. Their existence and aspects of their behaviour such as frequency depend on the experimental conditions and plasma parameters such as magnetic field strength and its rotational transform profile and the plasma density. For clustering purposes we assume the spatial structure of a fluctuation instance is what defines it and makes it unique from other fluctuations. Unless the plasma equilibrium is very steady, if a fluctuation exists in a shot it will have different frequencies at different times depending on the plasma parameters such as density and magnetic field strength. Therefore, frequency is not a good identifier of a particular fluctuation and is not used in the clustering process. This and other attributes of each fluctuation instance (time, plasma parameters etc.) are only used later in interpreting the nature of each cluster. Each cluster represents a collection of measurements of the same type of fluctuation that have existed during different experiment conditions which together provide a great deal of information important for interpretation of the underlying physical phenomena.

An overview of the feature extraction, clustering, and analysis process is shown in Fig. 2. The measurements available to identify these instabilities generally consist of timeseries data from arrays of experimental diagnostics such as magnetic pickup probes or multichannel interferometers. In this paper we will focus on magnetic probes but the same technique has been successfully applied to interferometer data.

The magnetic probe signal from a mode that consists primarily of one component such as a global Alfvén eigenmode [1] can be described as follows:

$$V_i \propto \cos(n\phi_{B,i} + m\theta_{B,i} - \omega t). \tag{1}$$

Here, ω is the mode frequency, m represents the poloidal mode number, n the toroidal mode number, i an index in the toroidal array of probes, and, $\phi_{B,i}$ and $\theta_{B,i}$ are the toroidal and poloidal Boozer angles [14,15] of the ith probe, respectively. Examples of the time trace signals from a magnetic probe array due to a mode are shown in Fig. 1(b).

From Eq. (1) we can see that spatial information we are interested in (n and m) is contained in the phase structure of the signal at the frequency of the perturbation (ω). Therefore, the first task is to identify the frequencies of the perturbations over discrete time intervals, and extract the phase structure of the signal at those frequencies for each of the magnetic probes in the array. To make the data independent of the choice of time origin, we calculate the phase difference between successive coils in the array. This forms a

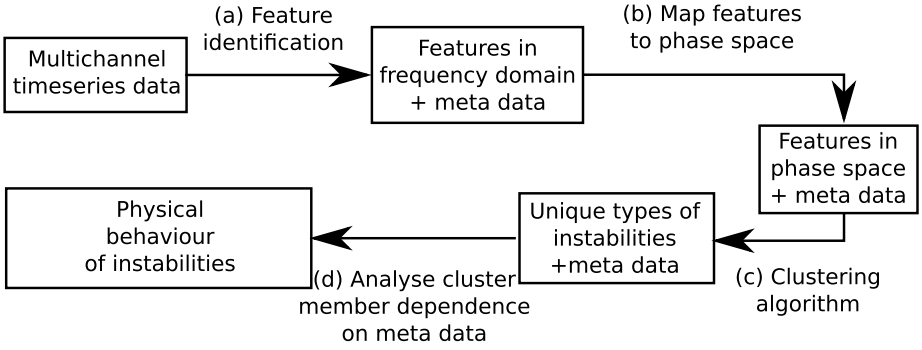

Fig. 2. Overview of the stages involved in the feature extraction and clustering process.

$N_c - 1$ set of nearest neighbour phase differences for each measurement, where N_c is the number of channels or probes in the array ((a) in Fig. 2). This maps the identified features to the $(-\pi, \pi]^{N_c-1}$ hyper-torus which we call $\boldsymbol{\Delta\psi}$-space (step (b) in Fig. 2).

Options for this extraction step include the SVD techniques [16,8,17], the STFT-clustering technique described in Section 3, or a combination of techniques. We also record any meta-data such as frequency, electron density profile, heating power and magnetic configuration that is not used in the clustering process but is useful for analysing the identified clusters dependence on physical parameters.

In the final clustering stage, we partition the identified features (measurements) into meaningful subgroups, where each subgroup consists of similar phase difference measurements, and therefore represents a different type of fluctuation ((c) in Fig. 2). Clustering methods usually follow either a hierarchical strategy, such as agglomerative hierarchical clustering, or a relocating strategy such as k-means or clustering via mixture models based on the EM algorithm [18]. In this paper, we focus on the relocating method as we have found that the hierarchical methods with reasonable run times are very memory intensive when dealing with the substantial number of instances we usually have in our datasets. The final stage of the analysis is to use the meta-data of each cluster to investigate the physical dependence of the observed features.

3. Feature extraction using the STFT-clustering pre-processing technique

The first step in the clustering process is to extract interesting measurements (or instances) from the time-series probe data which will be used in the final clustering process. This step is essentially a coarse filtering which reduces the number of instances used in the main clustering stage to a manageable level. Several methods are available to extract interesting features from the time series data [19] and map them to $\boldsymbol{\Delta\psi}$-space for clustering. One successful option that we have applied extensively is the SVD "fluctstruc" identification technique [17,8]. While this process works well, it fails to identify lower power fluctuations whose singular values are small, or are dwarfed by other fluctuations that exist at the same time.

Another option which is described in detail below is to break the probe time series data into short time intervals and apply the DFT to these intervals (equivalent to the short time Fourier transform (STFT)). By identifying peaks in the magnitudes in the STFT, we can locate data that is likely related to fluctuations. Using this approach, it is possible to select a large number of peaks for each discrete time step in a single experimental discharge (or shot) to ensure that lower power fluctuations are also included. By including the lower power peaks a significant number of measurements due to noise will also be included. To filter out the noise measurements a primary or "filtering" clustering step is performed on each shot individually. Those features that are assigned to poorly defined or very broad clusters are removed because they are likely to be noise. The primary clustering step for each shot is separate from the final clustering of the data which uses the surviving measurements from all shots. Both the primary and final clustering use the technique described in Section 4. The whole process for a single shot is shown in Fig. 3. Combining the SVD and STFT-clustering techniques allows substantially more useful measurements to be extracted from the magnetic probe array data.

3.1. Multi channel averaged STFT feature extraction using clustering as a filter

The time-series data for a single shot consisting of N_s samples from N_c diagnostic channels, sampled at $(1/\tau)$ Hz can be represented in the following $N_c \times N_s$ matrix:

$$\mathcal{S} = \begin{pmatrix} s_0(t_0) & s_0(t_0+\tau) & s_0(t_0+2\tau) & \dots & s_0(t_0+(N_s-1)\tau) \\ s_1(t_0) & s_1(t_0+\tau) & s_1(t_0+2\tau) & \dots & s_1(t_0+(N_s-1)\tau) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{N_c-1}(t_0) & s_{N_c-1}(t_0+\tau) & s_{N_c-1}(t_0+2\tau) & \dots & s_{N_c-1}(t_0+(N_s-1)\tau) \end{pmatrix}. \tag{2}$$

This assumes that each channel has the same time base or has been interpolated onto a common time base before assembling $\mathcal{S}$.

The frequency and phase content of the signals we are interested in change over time. To capture the time dependence of the spectral content of the signals, we apply the STFT to our time series data ((a) in Fig. 3). To do this, we break up the time series data in $\mathcal{S}$ into T small time chunks consisting of N_f samples each. A window function (typically Hanning) is applied to each time chunk. This data is referenced as, $x_{c,\tau}$ where $c = 1, 2, \dots, N_c$ represents the channel number, and $\tau = 0, 1, \dots, T$ identifies the time chunk. The discrete Fourier transform (DFT) of each time chunk for each channel is taken:

$$X_c[\tau, k] = \sum_{n=0}^{N_f-1} x_{c,\tau}[n] \exp(-i2\pi kn/N_f). \tag{3}$$

For our purposes 0.5 ms time chunks corresponding to $N_f = 1024$ samples at a 2 MHz sampling rate provides the best compromise between temporal and frequency resolution for data from the H-1NF heliac.

The next step is to identify the frequencies of interesting features from the multi-channel STFT of the shot. Using a single reference channel and searching for peaks in its STFT biases the reference channel, may miss features that are seen on other channels, and does not take advantage of the reduction in signal to noise that is available by using multiple channels. An attractive

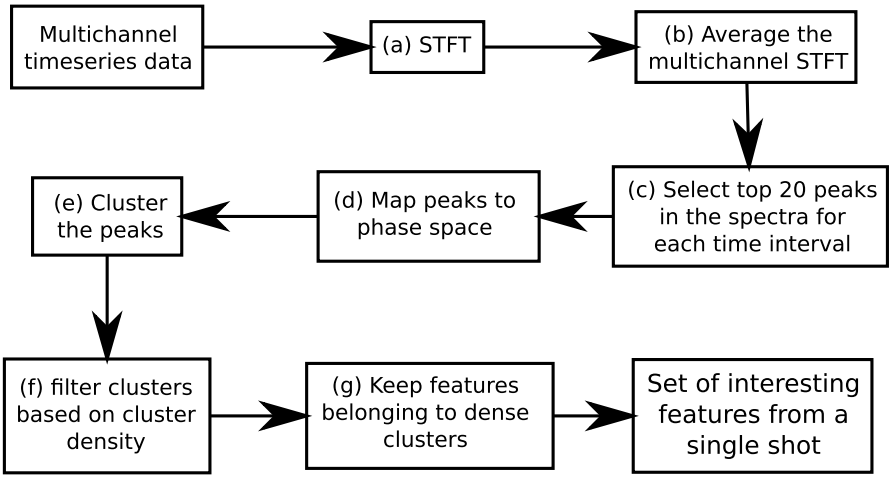

Fig. 3. Overview of the STFT-clustering process for a single shot. The retained features from each shot are combined to form the dataset for the major clustering stage.

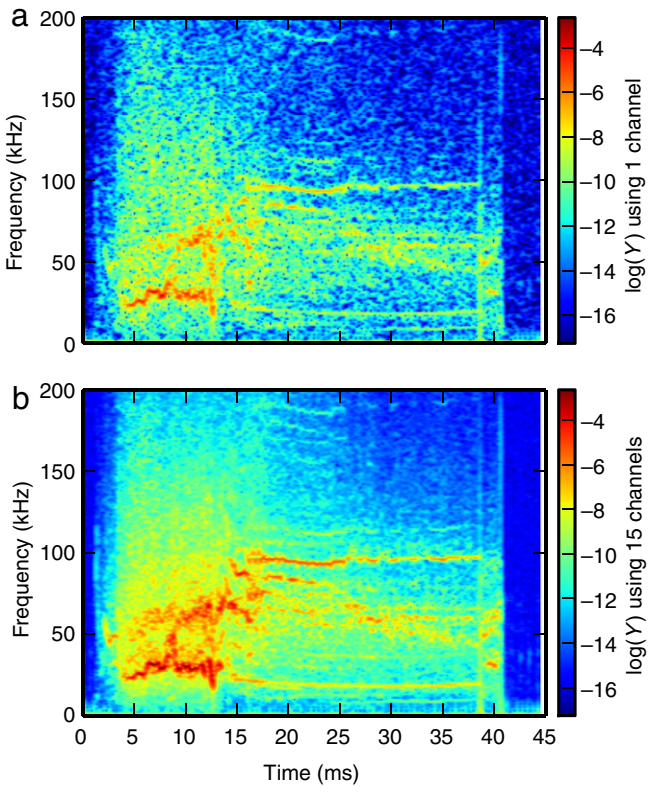

Fig. 4. Averaged STFT ($\log(Y)$) using 15 channels (b) and just a single channel (a) showing the improved signal to noise when using multiple channels, particularly for the $\approx$ 100 kHz signal at 20 ms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

alternative is to average the magnitude squared of the STFT from all channels ((b) in Fig. 3):

$$Y[\tau, k] = \frac{1}{N_c} \sum_{c=1}^{N_c} |2X_c[\tau, k]|^2. \tag{4}$$

This is similar to the method of averaged periodograms (Bartlett's method) for a single channel. Instead of splitting up a longer time window, we use multiple channels, which reduce the variance of the periodogram without the usual reduction in temporal resolution. Fig. 4(a) and (b) show the magnitude squared of a single probe channel and the average magnitude squared of the multichannel STFT respectively clearly demonstrating the improved signal to noise for the multichannel case using this method.

The next step is to identify peaks in $Y[\tau, k]$ ((c) in Fig. 3). This is achieved by comparing the amplitude of each datapoint with the amplitude of the five point moving average:

$$P[\tau, k] = Y[\tau, k] - \frac{1}{5} \sum_{j=k-2}^{k+2} Y[\tau, j]. \tag{5}$$

The 5-point moving average was optimum for this application as it was small enough to provide a local average, without covering multiple peaks at once. For applications where the frequencies are known to be well separated, a greater number of points can be included in the moving average. We then select peaks in $P[\tau, k]$ (i.e. $Y[\tau, k]$ must satisfy $Y[\tau, k-1] < Y[\tau, k] > Y[\tau, k+1]$). For each interval τ, we rank the peaks, k, by how peaked the points are by ordering $P[z, k]$ and select the largest 20 peaks to ensure that we capture all the available interesting features. This creates a list of tuples, $Z = [(\tau_1^*, k_1^*), (\tau_2^*, k_2^*), \ldots]$, which identifies peaks we are interested in. At this stage we are not concerned if we include points which may be noise because we are about to apply a clustering technique to filter them based on their structure in $\boldsymbol{\Delta\psi}$-space. We map each τ_j^*, k_j^* in Z to $\boldsymbol{\Delta\psi}$-space by extracting the $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_c)$ where $\psi_i = \arg(X_i[\tau_j^*, k_j^*])$. The elements of $\boldsymbol{\Delta\psi} = (\Delta\psi_1, \Delta\psi_2, \ldots, \Delta\psi_{N_c-1})$ are calculated as follows: $\Delta\psi_i = \psi_{i+1} - \psi_i$ ((d) in Fig. 3).

We are now ready to apply the periodic clustering based feature extraction filter to these datapoints from a single shot ((e) in Fig. 3). We apply the EM-VMM algorithm described in Section 4 with a large number of clusters (typically 16) to ensure that interesting features can be separated out into clusters, we then analyse the resulting clusters to see how well-defined they are using the average circular standard deviation $\bar{\sigma}$ which is defined in Eqs. (14) and (10) in Section 4. The phase difference measurements that belong to clusters that have $\bar{\sigma} < 20°$ are likely to be due to fluctuations so they are kept. The phase difference measurements in the other low density, broader clusters are likely to be noise so they are discarded. The cutoff value for $\bar{\sigma}$ was chosen as the point at which the rate of increase in the number of features with $\bar{\sigma}$ fell markedly (steps (f) and (g) in Fig. 3).

An example of these two techniques applied to a single shot is shown in Fig. 5. Fig. 5(a) shows the features identified using the standard SVD technique, (b) shows the features identified using the peaks in the averaged STFT (black dots), and the features which survive the cluster filtering (green circles). Fig. 5(c) shows the result of combining both the SVD and STFT-clustering techniques

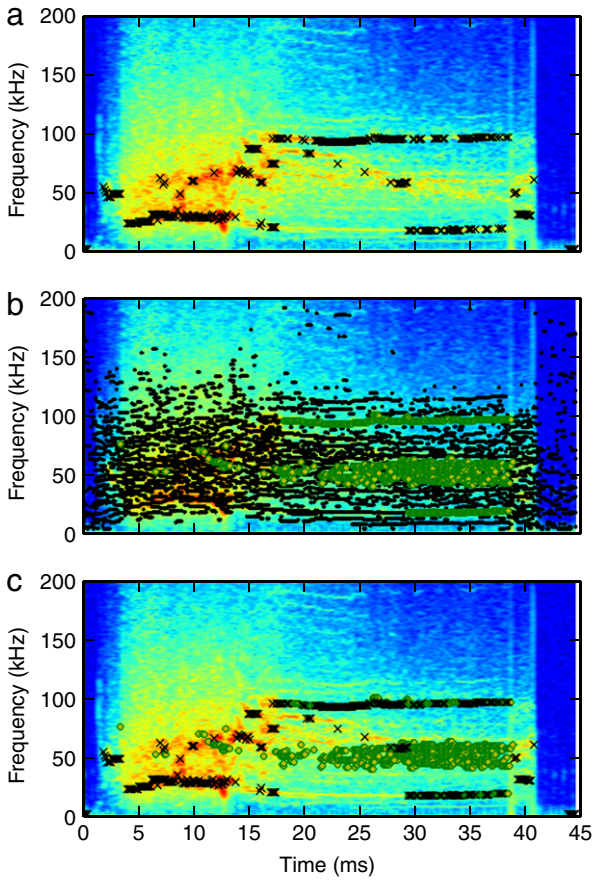

Fig. 5. (a) Black crosses represent features identified using the SVD method. (b) Green circles represent the features identified using STFT-clustering. The top 20 peaks at each timestep before filtering are marked with small black dots. (c) The features identified by the SVD method (black crosses), and the STFT clustering filtered method (green circles), demonstrating that the features identified using the two techniques complement each other. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

using an "inclusive OR" to avoid double counting features that are common to both methods. There are many common features, however, both methods also complement each other allowing us to identify a greater number of real features.

3.2. Comparison with SVD feature extraction

The SVD, STFT-clustering and combination of both methods were used to extract features in a 131 shot database which was part of a magnetic configuration scan on H-1NF. Table 1 shows the number of features that were identified using each technique as well as the number of features that are identified by both techniques (duplicates are only counted once). Using the combination of both techniques provides a significant advantage, allowing us to find approximately twice as many meaningful features as explained below.

In the second and final clustering step, to check the quality of the extracted features using these three techniques, we apply the EM-VMM clustering algorithm described in Section 4 to the entire set of features that was found throughout the scan using each method. Next, we check the number of features that is assigned to clusters below a certain $\bar{\sigma}$ value. The greater the number of features that belong to dense clusters (smaller $\bar{\sigma}$), the higher the quality of the features that have been extracted. A plot of the number of features that are in clusters with a given or smaller value of $\bar{\sigma}$ is shown in Fig. 6. The filtering criterion used in the primary clustering step,

Table 1
Comparison of the SVD and STFT-clustering feature extraction techniques using 131 shots that were part of a magnetic configuration scan. The number of features identified using each method is shown along with the number of identified features that are common to the two methods. Also shown is the number of features obtained by combining the two methods.

	Number of features
SVD	23 000
STFT-clustering	27 000
Common (% SVD, % STFT)	13 000 (57%, 48%)
Combined	37 000

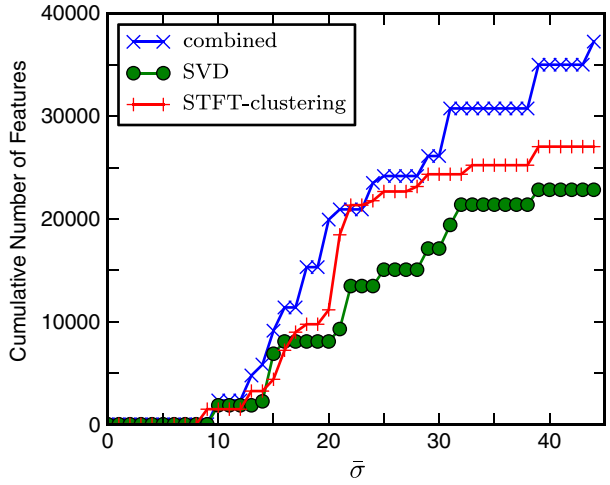

Fig. 6. Clustering the same features shown in Table 1 to check the significance of the features found by each method. Twice as many features are included in the good clusters ($\bar{\sigma} < 20^{\circ}$) using the combined method compared with only the SVD method signifying a substantial improvement in the feature identification procedure.

$\bar{\sigma} < 20^{\circ}$, is indicative of a "dense" cluster, but the ultimate criterion of cluster quality is how distinct the cluster is, as discussed in relation to Fig. 15. We find that the STFT-clustering method performs marginally better than the SVD method, and the combined method performs best, identifying 50% more features that form quality clusters compared with the SVD method.

4. Model based clustering using multivariate independent von Mises distributions

4.1. Selecting the correct model for the experimental measurements

As explained in Section 2, we characterise the spatial structure of fluctuations by the set of phase differences between adjacent Mirnov coils ($\Delta\psi_i$). A collection of phase difference measurements due to a single fluctuation has a well-defined mean value, and because it is subject to random processes such as electrical noise in the probe amplifiers, and noise generated by localised plasma turbulence will be distributed about this mean. Therefore, these measurements, which are inherently (2π) periodic, can be modelled using a circular distribution function. Of the several circular distribution functions available, we use the von Mises distribution which is one of the best known and widely used [20,21]. If the experimental measurements are generated by a variety of fluctuations, each of which produces unique patterns of phase difference between probes, we can use a mixture of von Mises distributions to model the data. The expectation maximisation algorithm [22,23] can then be used to compute the maximum-likelihood estimates (MLE) of the mixture model parameters and the cluster assignments.

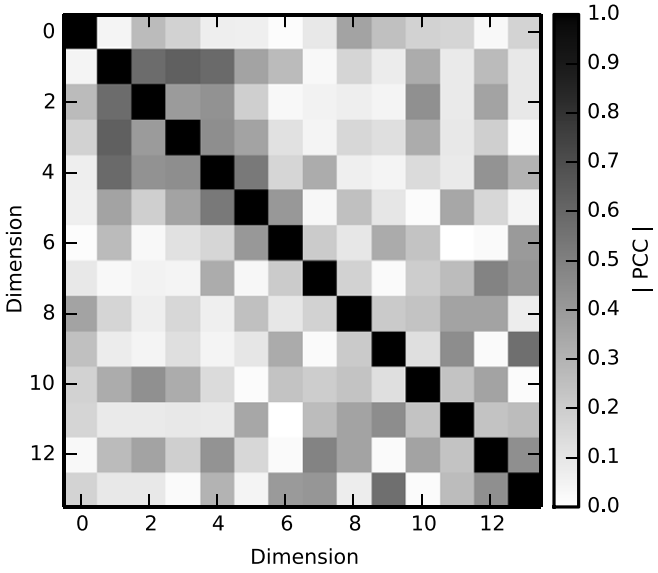

Fig. 7. Graphical representation of the absolute values of the Pearson correlation coefficient for a cluster that was obtained from the H-1NF dataset using EM with Gaussian mixtures with a full covariance matrix. The off diagonal PCCs are generally below 0.5 and have a mean magnitude of 0.21 indicating that covariance is not dominant in this cluster.

In our case, we have an array of such measurements ($\boldsymbol{\Delta\psi}$), which we model using a mixture of *multivariate* von Mises distributions (as described in [24]); however, beyond the bi-variate case the normalising constant becomes intractable when covariance is present. To overcome these difficulties, Mardia et al. [25] described a concentrated multivariate sine (CMS) model. Unfortunately, our experimental data is not sufficiently concentrated causing difficulties with this algorithm.

In the absence of practical methods that include the full covariances, we can use a mixture of multivariate independent von Mises distributions. We expect each of the variables $\boldsymbol{\Delta\psi}$ due to a single type of fluctuation to have a degree of independence due to localised noise from plasma turbulence and uncorrelated electrical noise. Covariances could however exist due to shared noise, or movement of the plasma relative to the pickup probes. To obtain an estimate of the importance of including covariance, we performed a clustering analysis on some of the H-1NF data using EM with a Gaussian mixture model. While this method does not provide the correct distribution for the periodic data, it allows us to include the full covariance matrix to test its importance, and is valid for this dataset because the clusters are well clear of the folding at $-\pi, \pi$. The Pearson correlation coefficients (PCC) between all dimensions of a cluster of interest are shown in Fig. 7. While there is some degree of correlation, the off diagonal elements are generally below 0.5 and have a mean magnitude of 0.21 in this case. This should not have a significant impact on the cluster assignments and suggests that using a multivariate independent von Mises distribution provides a good model for the individual clusters.

The rest of this section describes in detail how to solve for the parameters and cluster assignments in a mixture of multivariate independent von Mises distributions using the EM algorithm.

4.2. The von Mises distribution

The probability density function for the von Mises distribution is defined as follows [20,21]:

$$f(x;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa\cos(x-\mu)\} \tag{6}$$

where $-\pi < x \leq \pi$, $I_0(\kappa)$ is the modified Bessel function of the first kind and order 0, μ represents the mean direction and $\kappa \geq 0$ is a concentration parameter that increases with distributions that are more concentrated around the mean. We obtain a uniform distribution on $(-\pi, \pi]$ when $\kappa = 0$. This distribution can be used to model measurements of the phase differences between two probes due to a single type of fluctuation. $f(x;\mu,\kappa)$ describes the probability of obtaining the phase difference measurement x (same as $\Delta\psi$ from Section 3.1) from a particular fluctuation for which μ is the mean value of the measurements, and κ describes how concentrated the measurements are.

Given a random sample $(x_1, x_2, \ldots, x_n$: in the data following, these will be phase difference measurements) which is drawn from $f(x;\mu,\kappa)$ (i.e. they are due to a particular type of fluctuation), the maximum likelihood estimates, $\hat{\mu}$, and $\hat{\kappa}$ are given as follows [20]:

$$R = \frac{1}{n}\sum_{j=1}^{n}\exp(ix_j)$$

$$\bar{R} = |R|$$

$$\hat{\mu} = \arg(R) \tag{7}$$

$$\frac{I_1(\hat{\kappa})}{I_0(\hat{\kappa})} = \bar{R} \tag{8}$$

where $i = \sqrt{-1}$, and $I_1(\hat{\kappa})$ is the modified Bessel function of the first kind and first order. There is no analytical solution to solve Eq. (8) for $\hat{\kappa}$, so one must resort to numerical techniques, a pre-calculated lookup table, or the following approximation [20]:

$$\hat{\kappa} = \begin{cases} 1/(2-2\bar{R}), & \text{for } 0.85 < \bar{R} \\ -0.4 + 1.39\bar{R} + 0.43/(1-\bar{R}), & \text{for } 0.53 < \bar{R} < 0.85 \\ 2\bar{R} + \bar{R}^3 + (5/6)\bar{R}^5, & \text{for } \bar{R} < 0.53. \end{cases} \tag{9}$$

Each of these methods work well, providing various trade-offs between speed and accuracy. We can also calculate a circular standard deviation, which has units of radians and has more intuitive meaning than κ [26]:

$$\sigma = \sqrt{-2\ln(\bar{R})}. \tag{10}$$

4.3. The multivariate independent von Mises distribution

We can extend Eq. (6) to get the probability density function of the P-variate independent von Mises distribution as follows:

$$f_P(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\kappa}) = \left(\frac{1}{2\pi}\right)^P \frac{1}{I(\boldsymbol{\kappa})}\exp\{\boldsymbol{\kappa}\cdot c(\boldsymbol{x},\boldsymbol{\mu})\} \tag{11}$$

where:

$$-\pi < x_i \leq \pi,$$
$$\kappa_i \geq 0,$$
$$\boldsymbol{x} = (x_1, x_2, \ldots, x_P),$$
$$\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_P),$$
$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_P),$$
$$c(\boldsymbol{x},\boldsymbol{\mu})^T = (\cos(x_1-\mu_1), \cos(x_2-\mu_2), \ldots, \cos(x_P-\mu_P)),$$
$$I(\boldsymbol{\kappa}) = \prod_{p=1}^{P} I_0(\kappa_p).$$

Eq. (11) is the extension of Eq. (6) to phase difference measurements from an array with more than 2 probes. f_p gives the probability of obtaining the phase difference measurement $\boldsymbol{x}$ consisting of $N_c - 1 = P$ values from an array with N_c probes (same as $\Delta\boldsymbol{\psi}$ from Section 3.1) from a particular fluctuation for which $\boldsymbol{\mu}$ is the

mean value of the phase difference measurements between adjacent coils, and $\boldsymbol{\kappa}$ describes how concentrated the measurements are.

When dealing with a large number of dimensions, $\exp\{\boldsymbol{\kappa} \cdot c(\boldsymbol{x}, \boldsymbol{\mu})^T\}$ can become very large, while $\left(\frac{1}{2\pi}\right)^P \frac{1}{I(\boldsymbol{\kappa})}$ becomes very small causing computational problems. These problems can be overcome by rewriting Eq. (11) as follows:

$$f_P(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\kappa}) = \exp\left\{\boldsymbol{\kappa} \cdot c(\boldsymbol{x}, \boldsymbol{\mu}) - P\ln(2\pi) - \sum_{p=1}^{P} \ln(I_0(\kappa_p))\right\}. \quad (12)$$

4.4. EM and mixtures of multivariate independent von Mises distributions

The phase difference measurements from K different types of fluctuation each with their own $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ can be represented using a mixture of K, P-variate independent von Mises distributions:

$$M(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \ldots, \boldsymbol{\kappa}_K) = \sum_{k=1}^{K} p_k f_P(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\kappa}_k) \quad (13)$$

where p_k represents the mixing ratios which represent how likely the measurement is to come from the fluctuation represented by k, $p_k > 0$ and $\sum_{k=1}^{K} p_k = 1$. Given a set of n data points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ from a mixture of K, P-variate independent von Mises distributions, we can define the cluster membership in a $n \times K$ matrix, $Z = (z_{ik})$, where $z_{ik} = 1$ if $\boldsymbol{x}_i$ is a member of group k and zero otherwise.

Two useful measures of the concentration of a P-variate independent distribution are the average circular standard deviation, and the equivalent circular standard deviation:

$$\bar{\sigma} = \frac{1}{P}\sum_{i=1}^{P} \sigma_i \quad (14)$$

$$\sigma_{eq} = \left(\prod_{i=1}^{P} \sigma_i\right)^{1/P} \quad (15)$$

where σ is the circular standard deviation of each dimension (as defined in Eq. (10)). The equivalent standard deviation can be thought of as the side length of a hyper cube that has the same the volume as a hyper-cuboid whose side lengths are the circular standard deviation in each dimension. These measures are used to quantify the density of the clusters.

Our aim is to find the MLE estimates of our dataset ($\hat{p}_k$, $\hat{\boldsymbol{\mu}}_k$, $\hat{\boldsymbol{\kappa}}_k$, and $\hat{z}_{ik}$ where $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n$). The expectation maximisation algorithm [22] which seeks to maximise the log-likelihood can be used to find the MLE values. The log-likelihood of the MLE values is given by:

$$L = \sum_{k=1}^{K}\sum_{i=1}^{n} \hat{z}_{ik} \ln(f_P(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\kappa}}_k)). \quad (16)$$

The EM algorithm involves the following two steps which are iterated over until convergence criteria such as minimal changes in $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\kappa}}_k$ between iterations, are met.

Expectation (E) step: given $\hat{\boldsymbol{\mu}}_k$, $\hat{\boldsymbol{\kappa}}_k$, and $\hat{p}_k$ from the maximisation step, we calculate $\hat{z}_{ik}$:

$$\hat{z}_{ik} = \frac{\hat{p}_k f_P(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\kappa}}_k)}{\sum_{l=1}^{K} \hat{p}_l f_P(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\kappa}}_l)}. \quad (17)$$

This means, $0 \leq \hat{z}_{ik} \leq 1$, and $\sum_{k=1}^{K} \hat{z}_{ik} = 1$. This is referred to as soft cluster assignment.

The maximisation (M) step: given $\hat{z}_{ik}$ from the expectation step, we calculate $\hat{\boldsymbol{\kappa}}_k = (\hat{\kappa}_{1k}, \hat{\kappa}_{2k}, \ldots, \hat{\kappa}_{Pk})$, $\hat{\boldsymbol{\mu}}_k = (\hat{\mu}_{1k}, \hat{\mu}_{2k}, \ldots, \hat{\mu}_{Pk})$, and $\hat{p}_k$ for each cluster $k = 1, \ldots, K$ in a similar manner to the uni-variate case:

$$\hat{p}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik}}{n}$$

$$R_{jk} = \frac{1}{\sum_{i=1}^{n} \hat{z}_{ik}} \sum_{i=1}^{n} \hat{z}_{ik} \exp(ix_{ip})$$

$$\bar{R}_{jk} = |R_{jk}|$$

$$\hat{\mu}_{jk} = \arg(R_{jk}) \quad (18)$$

$$\frac{I_1(\hat{\kappa}_{jk})}{I_0(\hat{\kappa}_{jk})} = \bar{R}_{jk}. \quad (19)$$

Eq. (19), is solved for $\hat{\kappa}_{jk}$, numerically, using a pre-calculated lookup table, or using the approximation shown in Eq. (9).

We initialise the EM algorithm using the cluster assignments, $\hat{z}_{ik}$ which are obtained from a k-means run. These values are used for the maximisation step. We then alternate between the E and M steps until convergence. Finding a local maximum instead of the global one is a well known problem with the EM algorithm. To overcome this, the algorithm is run several times with different starting points and the solution which maximises the log-likelihood is chosen.

4.5. Implementation and computational requirements

Many software packages exist that include an implementation of Expectation Maximisation for Gaussian mixture models for clustering [27,28]; however, implementations using von Mises distributions appear to be rare. A related approach is implemented in the SNOB code [29,30], which solves for independent von Mises distributions using the minimum message length technique. We have implemented the algorithm described in Section 4 using the Python programming language. Using the Scipy [31] and Numpy [32] modules allows us to achieve close to compiled speeds. Each iteration takes approximately 0.5 s for 37,000 instances, 14 dimensions, and 16 clusters using a single core on a laptop with an Intel core i7 processor. After ≈30 iterations, the algorithm is usually well converged. This speed allows many clustering parameters to be tried without excessive delays. The computational cost is roughly the same as performing the clustering with the same parameters using a Gaussian mixture model with trigonometric encoding.

5. Artificial dataset

In this section, we compare the expectation maximisation algorithm using mixtures of multivariate independent von Mises distributions (EM-VMM) with the standard non-periodic datamining techniques: EM with Gaussian mixtures (EM-GMM), EM with Gaussian mixtures using trigonometric encoding (EM-GMM-trig), k-means using trigonometric encoding (k-means-trig) and a periodic k-means variant (k-means-periodic) whose distance measure is $D(\theta, \phi) = \min\{|\theta - \phi|, 2\pi - |\theta - \phi|\}$, and whose new cluster center is calculated in the same way as for the EM-VMM algorithm (Eq. (18)) with hard cluster assignment. For the EM-GMM clustering, we use the implementation in the Python scikit-learn module [27].

For this comparison we have generated three datasets—case 1, case 2 and case 3. For all of the cases, the dataset is generated using the normal distribution, and wrapping the data to $(-\pi, \pi]$

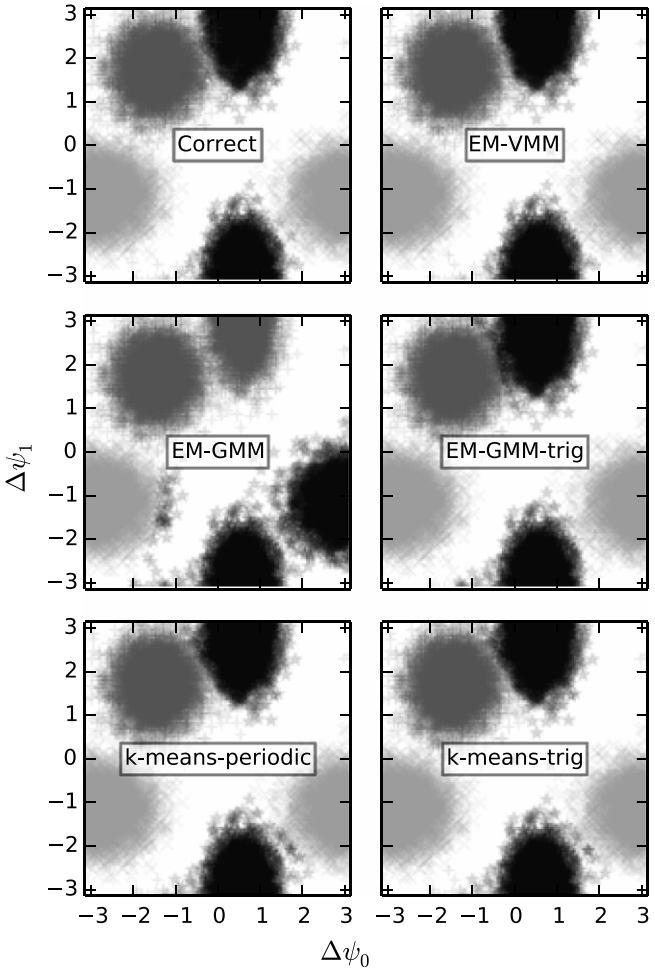

Fig. 8. The performance of the various clustering algorithms for case 1 where the algorithms must deal with the periodicity problems that occur at π and $-\pi$. Each point is a single datapoint in the artificial dataset that is used to test the clustering algorithms. The colour of each datapoint represents the cluster to which it has been assigned by various methods as labelled. The correct assignment is shown in the top left plot.

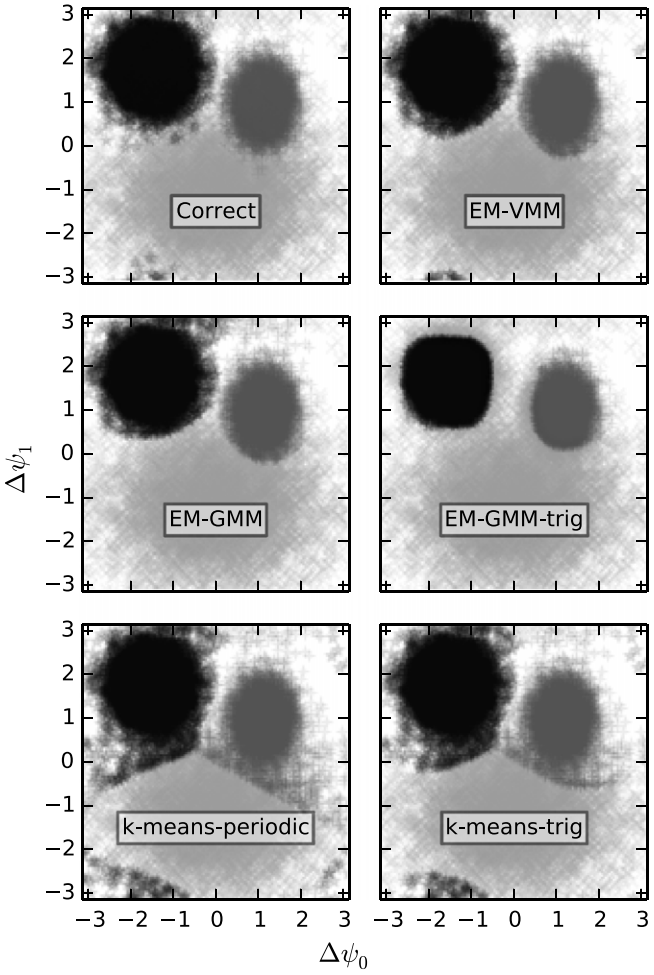

Fig. 9. The performance of the various clustering algorithms for case 2 where problems occur with the distortions caused by trigonometric encoding. The caption for Fig. 8 provides a description of what each datapoint and the colours represent.

by adding π, taking modulus 2π and subtracting π. Case 1 consists of 3 clusters in 2 dimensions. Two of the clusters have a single dimension whose mean is close to π to test the performance of the various techniques near the $-\pi, \pi$ discontinuity. Case 2 also consists of 3 clusters in 2 dimensions. For this case, the means of all dimensions are far enough away from π that periodic effects should not be important; however, one of the clusters has a large standard deviation, which is more sensitive to the distortions caused by trigonometric encoding. The third dataset consists of 4 clusters in 4 dimensions to test the performance of the algorithms on higher dimensional data. For this dataset, the means were randomly selected to be $-\pi < \mu < \pi$, and the standard deviation values are chosen to be in the range $\pi/12 < \sigma < \pi/4$ except for one cluster whose σ values are π for all dimensions to simulate noise.

The results from clustering case 1 are shown in Fig. 8, and the percentage of correct identifications is shown in Table 2. The EM-VMM, periodic *k*-means algorithm, EM-GMM-trig, *k*-means-trig perform well, with EM-VMM performing the best. EM-GMM performs poorly because it fails to deal with the periodicity problem as expected.

The results from clustering case 2 are shown in Fig. 9 and the percentage of correct identifications is shown in Table 2. Again, EM-VMM performs the best. For this case, EM-GMM also performs well because the periodicity problems are not present. Both the algorithms with trigonometric encoding (EM-GMM-trig and *k*-means-trig) perform less well due to distortions caused by the trigonometric encoding. The distortions can be seen clearly for the *k*-means-trig case in Fig. 9. The partitions separating clusters should be along straight lines for *k*-means, however, the distortions introduced by *k*-means-trig cause the partition between the clusters to follow an arc.

Table 2
Performance of the various algorithms based on percentage of correct cluster identification for cases 1–3 which represent difficult cases for periodicity, trigonometric distortion, and multi dimensionality (more in text).

	Case 1 (per)	Case 2 (dist)	Case 3 (dim)
EM-VMM	99.7%	97.1%	95.3%
EM-GMM	71.9%	96.9%	74.1%
EM-GMM-trig	99.4%	93.3%	90.6%
k-means-periodic	99.6%	92.8%	79.5%
k-means-trig	99.6%	94.4%	88.3%

For the higher dimensionality dataset in case 3, EM-VMM performs substantially better than the other algorithms. The algorithms with trigonometric encoding perform well, while EM-GMMs performs poorly due to its inability to deal with the periodic data.

The results from the artificial datasets clearly show that EM-VMM performs better than the other algorithms. This is because we are fitting an accurate model for periodic data to the datasets.

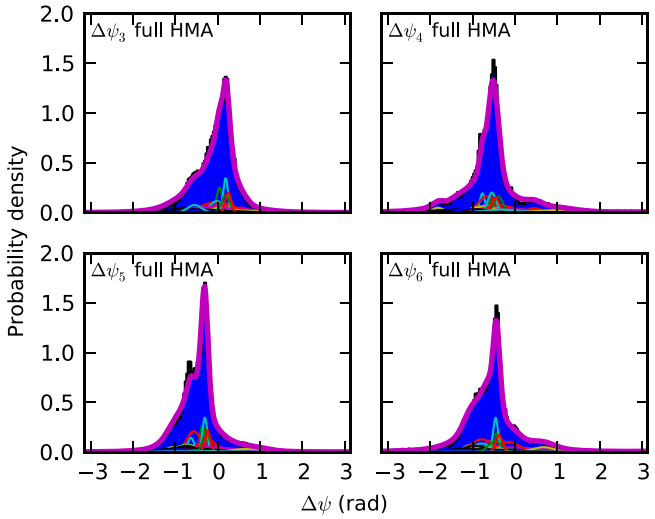

Fig. 10. Histograms for four phase differences when using the full HMA. The blue colour represents a histogram of the raw data. The thick line is the mixture of von Mises distributions that has been fitted to the data using EM. The thin lines represent the probability density functions for each of the clusters that make up the mixture model multiplied by their mixture weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

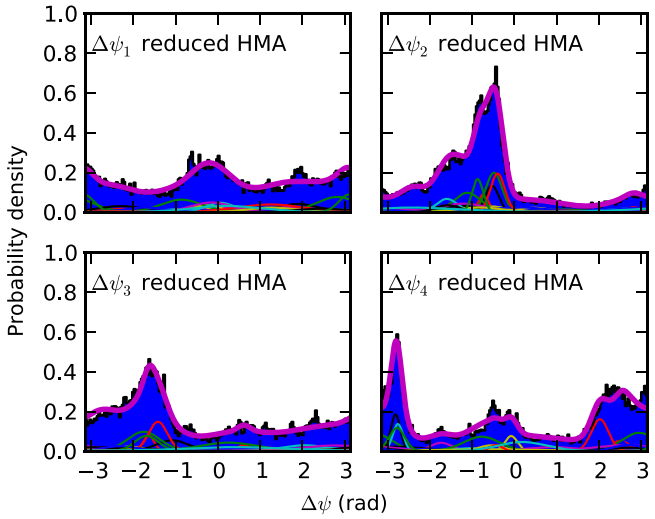

Fig. 11. Histograms for four phase differences (blue) when using every third coil in the HMA to simulate a smaller array. Compared with Fig. 10, the phase differences are generally 3 times larger (as the probes are further away from each other), and a substantial number of datapoints are close to $-\pi$ and π. The thick line is the mixture of von Mises distributions that has been fitted to the data using EM and the thin lines represent the probability density functions for each of the clusters that make up the mixture model multiplied by their mixture weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Additionally, while the *k*-means algorithm is faster than the EM algorithms, it tends to find clusters of comparable spatial extent. This represents a significant disadvantage because we expect the clusters in our experimental data to have a variety of shapes.

6. Application to the H-1NF helical Mirnov array dataset

The flexibility of the H-1NF heliac allows access to different rotational transform profiles by varying the ratio of currents in certain field coils, a parameter represented by κ_H (this has no relation to the κ concentration parameter for the von Mises distributions). The rotational transform is essentially a measure of the twist of magnetic field lines on flux surfaces [33]. It plays an important role in determining the type of magnetic fluctuations that can exist and their frequencies. A scan consisting of 131 separate shots was performed by incrementally increasing κ_H. Data from the helical Mirnov array (HMA) [13] was used to identify approximately 37,000 features in the scan using the combination of SVD and STFT-clustering extraction techniques described in Section 3.

The histograms for $\Delta\psi_3$, $\Delta\psi_4$, $\Delta\psi_5$ and $\Delta\psi_6$ for the 37,000 datapoints are shown in solid blue in Fig. 10. The histograms are normalised to show the proportion of items that fall into each bin so that the area under the histogram is 1 for comparison with the probability distribution functions. The other dimensions, which are not shown have similar histograms. The phase differences are peaked far away from the $(\pi, -\pi)$ discontinuity because the spacing of the HMA is very dense. Data from this array could be clustered using the EM-GMM algorithm without having to resort to trigonometric encoding, although this is not ideal as some important datapoints in low population clusters may still be close to π and $-\pi$.

Spacing the magnetic probes as densely as in the HMA is not always possible and can be prohibitively expensive, especially in larger fully three dimensional stellarators [33]. To simulate a less dense array, we only include every third coil in the HMA. The histograms for $\Delta\psi_1$, $\Delta\psi_2$, $\Delta\psi_3$, and $\Delta\psi_4$ of this reduced array are shown in Fig. 11. In this case, the EM-VMM algorithm has many benefits because the histograms are much less concentrated, and the clustering algorithm must be able to handle periodicity, and minimise the distortion of the phase space.

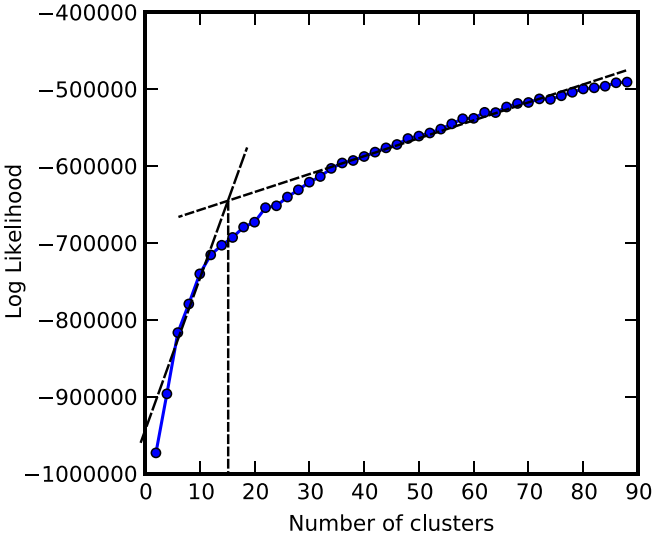

Fig. 12. Plot of log likelihood versus number of clusters. The elbow in the plot occurs at approximately 16 clusters.

One of the main difficulties in clustering is choosing the correct number of clusters. The "elbow" criterion, which identifies a flattening of some error measure versus the number of clusters, is a commonly used heuristic method for identifying the optimal number of clusters [34]. The Bayesian information criterion (BIC) [35] or Aikake information criterion (AIC) are other commonly used measures. A plot of the EM log likelihood versus number of clusters is shown in Fig. 12 suggesting that a good number of clusters is 16. The speed of the algorithm makes it easy to experiment using different numbers of clusters. Visual inspection of the clusters suggests that between 16 and 24 clusters provide the most useful separation of the data.

The results of clustering using the EM-VMM for the full HMA and the reduced HMA (every third coil) are also shown in Figs. 10

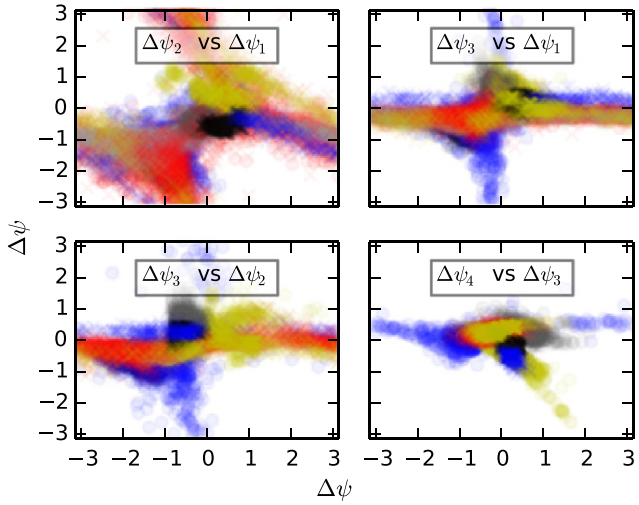

Fig. 13. Each datapoint represents one of the measured phase differences from the full HMA (each measurement consists of $N_c - 1$ numbers). The subplots are different projections of the $N_c - 1$ space needed to plot the measurements. Only the datapoints that are assigned to well-defined clusters are shown to exclude measurements that are noise. Each cluster is identified using a different colour (and marker style in the case of duplicated colours). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

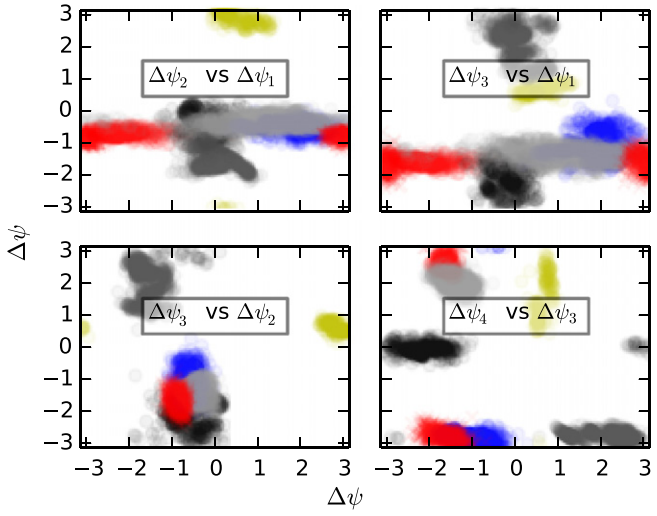

Fig. 14. Same as Fig. 13 except using the measurements from the reduced HMA ($N_c/3 - 1$). Compared with the full HMA (Fig. 13), the clusters are more separated, and closer to the $(-\pi, \pi]$ discontinuity. Each cluster is identified using a different colour (and marker style in the case of duplicated colours). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and 11. The von Mises distributions for the individual clusters are shown, along with the sum of the mixtures with the appropriate mixing proportions. The sum of the mixtures is very similar to the histograms, indicating that the von Mises mixture model fits the data well.

Plots showing the location of the clusters in the coordinate space of selected phase differences are shown in Figs. 13 and 14, for the full HMA and reduced array respectively. For the full HMA, the majority of phase differences are far from $(-\pi, \pi)$ (as seen in Fig. 10), however, there are clearly a few datapoints which are close to the discontinuity that require including periodicity to treat them properly. The plot for the reduced HMA clearly shows a great deal of spread in the location of the clusters indicating how important

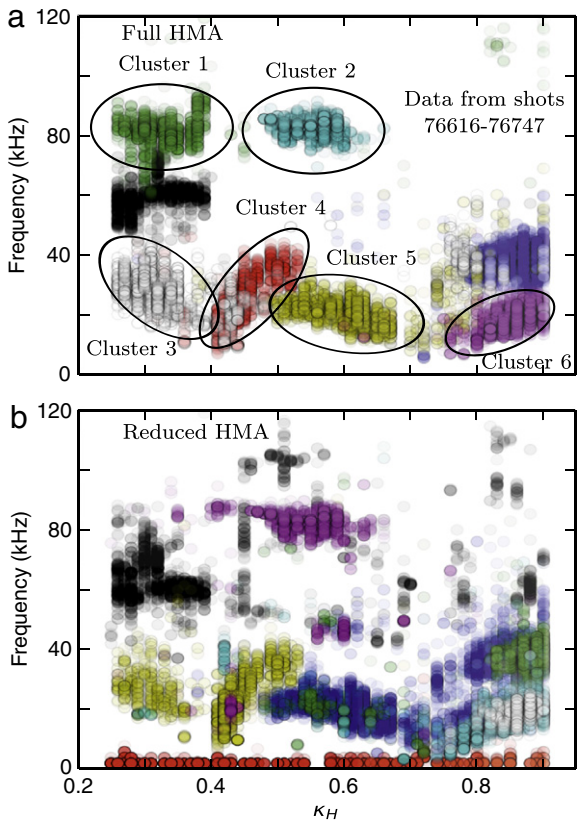

Fig. 15. Plot of dense clusters as a function of magnetic configuration (κ_H) and frequency found using EM-VMM using the full HMA (a) and reduced HMA (b). The clusters are clearly localised in κ_H, and some show a frequency dependence as κ_H changes (clusters 3–6), while others do not have such a strong dependence (clusters 1 and 2). The full HMA produces cleaner, more defined clusters when using the same clustering parameters because there is more information available to separate the clusters for the full array. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

including periodicity is for the case of a less densely spaced array. Note that the apparently reduced overlap of clusters compared to the full HMA is not indicative of better clustering, but is because the phase differences are larger by a factor of 3. The quality of clustering is discussed in relation to Fig. 15.

The final stage of the clustering analysis is to look at the cluster dependence on the meta-data that was recorded with the features. This allows us to determine the dependence of the instability on interesting physical quantities ((d) in Fig. 2). Fig. 15(a) and (b) show several of the most dense clusters plotted against frequency and κ_H for the full HMA and reduced HMA respectively. Each cluster represents a distinct type of fluctuation and by plotting the cluster member's dependence on ω and κ_H we can see what effect magnetic configuration has on the types of fluctuations that exist. The identified clusters show very well-defined behaviour in (ω, κ_H) space even though neither of these parameters were included in the clustering. The full HMA provides better defined clusters and more accurate fluctuation identification. For example, clusters 3 and 4 appear as a single cluster for the reduced HMA, consistent with the loss of information resulting from the omission of two thirds of the signals. This demonstrates the importance of having dense magnetic probe spacing for resolving different types of fluctuations.

Most clusters are limited to particular ranges of κ_H indicating that the magnetic configuration is important in determining whether or not a specific type of fluctuation can exist. For example, cluster 2 only exists for $0.45 < \kappa_H < 0.65$. Some clusters such as 3–6 show a relationship between ω and κ_H. As κ_H increases

the members of these clusters frequency increases or decreases systematically. This shows that the magnetic configuration affects the frequency of these fluctuations providing information about the fluctuations dispersion relation. This is a clear demonstration of the ability of the clustering algorithm to find interesting physical phenomena using just the phase differences between magnetic probes.

7. Conclusion

The application of the expectation maximisation algorithm to mixtures of multivariate independent von Mises distributions (EM-VMM) has been described in detail. Such a model does not distort phase space, and provides an accurate representation of multi-dimensional periodic data. This technique is ideally suited to clustering large multidimensional periodic datasets, such as those obtained from magnetic probe arrays observing instabilities in magnetic confinement plasma devices. The technique would also be useful to the multitude of applications where multichannel arrays of diagnostics record large quantities of periodic signals such as interferometers, soft X-ray arrays, magnetic probe arrays and imaging diagnostics.

The EM-VMM algorithm was shown to have superior performance on several artificial periodic datasets compared with other clustering options such as EM with Gaussian mixtures with and without trigonometric encoding, *k*-means with trigonometric encoding and a periodic *k*-means variant. The improved performance does not come with a significant increase in computational requirements. A 14 dimensional, 37,000 instance dataset takes less than a minute to converge when fitting 16 clusters, which is roughly equivalent to the time taken for EM with Gaussian mixtures and trigonometric encoding. This fast convergence is particularly useful for interactive data analysis, where several clustering parameters are varied to find the best combination.

A new method of feature extraction from multi-channel time-series data was described. The method, called STFT-clustering, identifies peaks in the multi-channel averaged short time Fourier transform of the timeseries, and clusters the data using the EM-VMM. Interesting data can then be extracted by selecting clusters with an average circular standard deviation below a threshold. This method was compared to an SVD extraction technique that is currently applied to data from the helical Mirnov array on H-1NF. The two methods complement each other, and when used together identify up to twice as many interesting features in the multichannel time-series data. The STFT-clustering method is particularly good at identifying lower power features.

Results from applying the STFT-clustering for feature extraction, and EM-VMM for clustering, to a 131 shot dataset from a magnetic configuration scan on the H-1NF heliac were presented. The combination of SVD and STFT-clustering identified 37,000 interesting features in the 15 channel helical Mirnov array (HMA) dataset. The EM-VMM clustering then successfully fitted a 14 dimensional, 16 mixture model to the data. Comparing the meta-data for each of the identified clusters showed distinct behaviour as a function of magnetic configuration and frequency, indicating the successful identification of unique, physically interesting instabilities.

Acknowledgements

The authors would like to thank Professors Markus Hegland and Donald Poskitt for valuable suggestions and comments, and the H-1NF team for continued support for experimental operations. This work was supported by the Education Investment Fund under the Super Science Initiative of the Australian Government. SRH wishes to thank AINSE Ltd. for providing financial assistance to enable this work on H-1NF to be conducted.

References

[1] W. Heidbrink, Basic physics of Alfvén instabilities driven by energetic particles in toroidally confined plasmas, Phys. Plasmas 15 (5) (2008) 055501-1–055501-15.
[2] K.L. Wong, R.J. Fonck, S.F. Paul, D.R. Roberts, E.D. Fredrickson, R. Nazikian, H.K. Park, M. Bell, N.L. Bretz, R. Budny, S. Cohen, G.W. Hammett, F.C. Jobes, D.M. Meade, S.S. Medley, D. Mueller, Y. Nagayama, D.K. Owens, E.J. Synakowski, Excitation of toroidal Alfvén eigenmodes in tftr, Phys. Rev. Lett. 66 (1991) 1874–1877. http://dx.doi.org/10.1103/PhysRevLett.66.1874.
[3] R. White, E. Fredrickson, D. Darrow, M. Zarnstorff, R. Wilson, S. Zweben, K. Hill, Y. Chen, G. Fu, Toroidal Alfvén eigenmode-induced ripple trapping, Phys. Plasmas 2 (8) (1995) 2871–2873.
[4] H. Nakanishi, T. Hochin, M. Kojima, Search and retrieval method of similar plasma waveforms, Fusion Eng. Des. 71 (1) (2004) 189–193.
[5] J. Vega, A. Pereira, A. Portas, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, et al., Data mining technique for fast retrieval of similar waveforms in fusion massive databases, Fusion Eng. Des. 83 (1) (2008) 132–139.
[6] J. Vega, Intelligent methods for data retrieval in fusion databases, Fusion Eng. Des. 83 (2) (2008) 382–386.
[7] J. Vega, A. Murari, A. Pereira, A. Portas, G.A. Rattá, R. Castro, Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases, Fusion Eng. Des. 84 (7) (2009) 1916–1919.
[8] D. Pretty, B. Blackwell, A data mining algorithm for automated characterisation of fluctuations in multichannel timeseries, Comput. Phys. Commun. 180 (10) (2009) 1768–1776.
[9] R. Jiménez-Gómez, A. Könies, E. Ascasíbar, F. Castejón, T. Estrada, L. Eliseev, A. Melnikov, J. Jiménez, D. Pretty, D. Jiménez-Rey, et al., Alfvén eigenmodes measured in the TJ-II stellarator, Nucl. Fusion 51 (3) (2011) 033001.
[10] S. Yamamoto, D. Pretty, B. Blackwell, K. Nagasaki, H. Okada, F. Sano, T. Mizuuchi, S. Kobayashi, K. Kondo, R. Jiménez-Gómez, et al., Studies of MHD stability using data mining technique in helical plasmas, Plasma Fusion Res. 5 (2010) 034-1–034-7.
[11] S. Hamberger, B. Blackwell, L. Sharp, D. Shenton, H-1 design and construction, Fusion Technol. 17 (1990) 123–130.
[12] J. Harris, M. Shats, B. Blackwell, W. Solomon, D. Pretty, S. Collis, J. Howard, H. Xia, C. Michael, H. Punzmann, Fluctuations and stability of plasmas in the H-1NF heliac, Nucl. Fusion 44 (2) (2004) 279.
[13] S.R. Haskey, B.D. Blackwell, B. Seiwald, M.J. Hole, D.G. Pretty, J. Howard, J. Wach, A multichannel magnetic probe system for analysing magnetic fluctuations in helical axis plasmas, Rev. Sci. Instrum. 84 (9) (2013) 093501. http://dx.doi.org/10.1063/1.4819250. URL http://link.aip.org/link/?RSI/84/093501/1.
[14] A.H. Boozer, Guiding center drift equations, Phys. Fluids 23 (5) (1980) 904–908. http://dx.doi.org/10.1063/1.863080. URL http://link.aip.org/link/?PFL/23/904/1.
[15] W.D. D'haeseleer, Flux coordinates and magnetic field structure: a guide to a fundamental tool of plasma structure, in: Springer Series in Computational Physics, 1991.
[16] C. Nardone, Multichannel fluctuation data analysis by the singular value decomposition method. Application to MHD modes in jet, Plasma Phys. Control. Fusion 34 (9) (1992) 1447.
[17] D. Pretty, A study of mhd activity in the h-1 heliac using data mining techniques, Ph.D. Thesis, 2007.
[18] C. Fraley, A.E. Raftery, How many clusters? which clustering method? answers via model-based cluster analysis, Comput. J. 41 (8) (1998) 578–588.
[19] T. Warren Liao, Clustering of time series data a survey, Pattern Recognit. 38 (11) (2005) 1857–1874.
[20] K.V. Mardia, P.E. Jupp, Directional Statistics, vol. 494, Wiley, 2009.
[21] R. Von Mises, über die ganzzahligkeit der atomgewichte und verwandte fragen, Phys. Z 19 (1918) 490–500.
[22] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. (1977) 1–38.
[23] G. McLachlan, D. Peel, Finite Mixture Models, Wiley-Interscience, 2004.
[24] K.V. Mardia, G. Hughes, C.C. Taylor, H. Singh, A multivariate von Mises distribution with applications to bioinformatics, Canad. J. Statist. 36 (1) (2008) 99–109.
[25] K.V. Mardia, J.T. Kent, Z. Zhang, C.C. Taylor, T. Hamelryck, Mixtures of concentrated multivariate sine distributions with applications to bioinformatics, J. Appl. Stat. 39 (11) (2012) 2475–2492.
[26] N.I. Fisher, Statistical Analysis of Circular Data, Cambridge University Press, 1995.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[29] C.S. Wallace, D.L. Dowe, MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions, Stat. Comput. 10 (2000) 73–83.
[30] C.S. Wallace, D.L. Dowe, Intrinsic classification by MML—the Snob program, in: Proc. 7th Australian Joint Conf. on Artificial Intelligence, World Scientific, 1994, pp. 37–44.

[31] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: open source scientific tools for Python (2001–). URL http://www.scipy.org/.
[32] T.E. Oliphant, Guide to NumPy, Provo, UT (Mar. 2006). URL http://www.tramy.us/.
[33] A.H. Boozer, What is a stellarator? Phys. Plasmas 5 (5) (1998) 1647–1655.
[34] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (2) (2001) 411–423.
[35] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.