Computer Physics Communications 180 (2009) 1768–1776

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

A data mining algorithm for automated characterisation of fluctuations in multichannel timeseries

D.G. Pretty *, B.D. Blackwell

Plasma Research Laboratory, Research School of Physical Sciences and Engineering, Australian National University, Canberra ACT 0200, Australia

ARTICLE INFO

Article history:
Received 21 December 2007
Received in revised form 20 February 2009
Accepted 6 May 2009
Available online 13 May 2009

PACS:
07.05.Kf
07.05.Rm
52.25.Gj
52.55.-s

Keywords:
Data mining
Plasma physics
Mirnov oscillations
Magnetic fluctuations
Mode analysis

ABSTRACT

We present a data mining technique for the analysis of multichannel oscillatory timeseries data and show an application using poloidal arrays of magnetic sensors installed in the H-1 heliac. The procedure is highly automated, and scales well to large datasets. The timeseries data is split into short time segments to provide time resolution, and each segment is represented by a singular value decomposition (SVD). By comparing power spectra of the temporal singular vectors, related singular values are grouped into subsets which define *fluctuation structures*. Thresholds for the normalised energy of the fluctuation structure and the normalised entropy of the SVD can be used to filter the dataset. We assume that distinct classes of fluctuations are localised in the space of phase differences $\Delta\psi(n, n+1)$ between each pair of nearest neighbour channels. An expectation maximisation clustering algorithm is used to locate the distinct classes of fluctuations and assign mode numbers where possible, and a *cluster tree* mapping is used to visualise the results.

1. Introduction

Plasma instabilities are ubiquitous, driven by the large energy densities, gradients and non-thermal distribution functions typical of high temperature plasma. Unstable Alfvén eigenmodes in particular pose a significant threat [1] to the successful extraction of fusion energy from magnetically confined plasma. The match between the Alfvén velocity and the speed of fusion-born alpha particles makes the interchange of energy likely, leading to (already demonstrated) growth of large amplitude Alfvén waves, and the likelihood that through non-linear effects, these waves could expel those fusion alpha particles. The associated energy is not lost altogether, but can no longer maintain the plasma temperature, thereby preventing the plasma from reaching "ignition".

This project was motivated by the need to identify and understand the nature of these modes by extracting dispersion relation data from their temporal and spatial variations. Data mining techniques are particularly useful in this situation, in which experimentalists cannot control the waves directly, and other unexpected modes may appear during the course of experiments. In view of this and the difficulty and the high cost of these experiments, large quantities of data are recorded for analysis in many international facilities. The technique described here is now being applied to most of the major international stellarator experiments.

This paper presents a data mining technique for classifying fluctuations in large databases of multichannel timeseries data. The example given is an analysis of fluctuations in magnetically confined plasmas during parameter scans in the H-1 flexible heliac [2,3]. The H-1 heliac shown in Fig. 1 is a three field-period helical axis stellarator [4] with major radius $R = 1$ m, minor radius $\langle r \rangle = 0.2$ m and a finely tunable magnetic geometry.

Experimental scans through plasma configurations via the geometric parameter κ_h, which controls the rotational transform ι (twist of the magnetic field lines shown on the outer plasma surface in Fig. 1) and shear ι' (radial derivative of rotational transform), have produced diverse spectra of magnetohydrodynamic (MHD) activity. The MHD activity is recorded via two toroidally separated poloidal arrays of Mirnov coils (induction solenoids: Mirnov arrays in Fig. 1) which sample dB/dt locally. In the example dataset presented here, 28 Mirnov coils are used for 92 distinct plasma configurations, resulting in more than 100,000 short time Fourier spectra.

The data mining process used to reduce this dataset is described in the following sections. In Section 2 we explain the preprocessing stage, which includes filtering and mapping into a high

* Corresponding author.
E-mail addresses: david.pretty@anu.edu.au (D.G. Pretty), boyd.blackwell@anu.edu.au (B.D. Blackwell).

0010-4655/$ – see front matter
doi:10.1016/j.cpc.2009.05.003

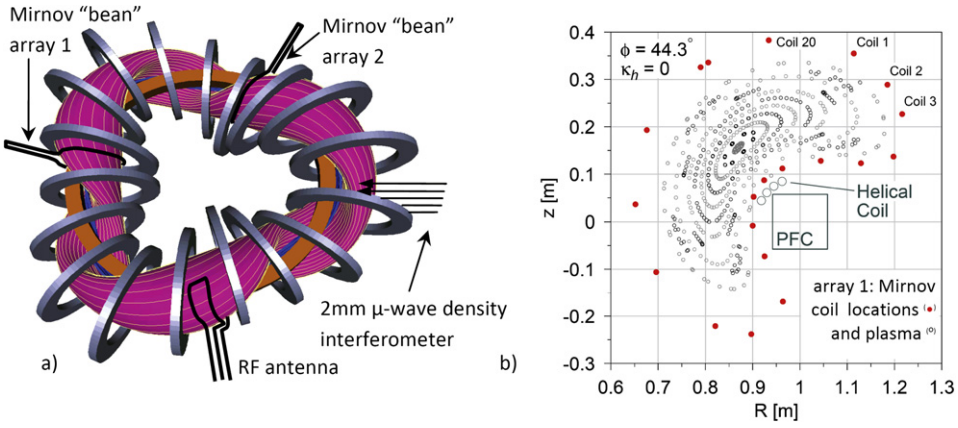

Fig. 1. (a) H-1 plasma showing location of two Mirnov "bean" arrays, RF antenna producing the plasma and interferometer; showing only 18 of 36 toroidal field coils (grey) for clarity, (b) the Mirnov coil locations within the bean-shaped array relative to the plasma and poloidal and helical magnetic coils.

dimensional phase space. In Section 3 the clustering algorithm for distinguishing classes of fluctuations is described, followed by a demonstration of a visualisation procedure. A discussion of some important aspects of the procedure follows in Section 4.

2. Preprocessing

2.1. Data preparation

We assume that each set of timeseries data can be represented as a $N_c \times N_s$ matrix:

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

where τ is the inverse of the sampling frequency, N_c is the number of channels and N_s is the number of samples. To achieve time resolution Δt, we split $\mathcal{S}$ into short time segments S' with shape $N_c \times N'_s$, where $N'_s = \Delta t/\tau$. Each row (channel) in S' has its baseline removed and is normalised to its variance. In our example dataset, the signal amplitudes depend on the plasma-coil distance which is a function of the plasma shape (magnetic configuration) controlled by κ_h. Such configuration bias is reduced through normalisation of the channels. The dataset consists of $N_d N_s/N'_s$ short time segments S', where N_d is the number of data matrices $\mathcal{S}$. Arbitrary metadata may be associated with each S', although for pre-processing and clustering only the timeseries data within S' is used.

2.2. The singular value decomposition

Each S' is represented by a singular value decomposition (SVD) [5]

$$S' = UAV^* \tag{2}$$

where the columns of U and V contain the spatial (topo) and temporal (chrono) singular vectors respectively, V^* denotes the conjugate transpose of V, and the diagonal elements of A are the $N_a = \min(N_c, N'_s)$ non-negative singular values. The convention that the singular values to be sorted in decreasing monotonic order means that A is independent of the ordering of the channels within S'. The set of topos (chronos) are an orthonormal basis of $\mathbf{R}^{N_c(N'_s)}$. The isomorphic mapping between topos and chronos $S'v_i = a_i u_i$ ensures the SVD retains one-to-one correspondence between spatial and temporal components, avoiding degenerate singular values and allowing proper reconstruction of the input signal.

Shown in Fig. 2 are singular values from a typical H-1 Mirnov dataset, acquired at a sample rate of 1 MHz using $N'_s = 1024$ ($\Delta t \sim 1$ ms) and $N_a = N_c = 28$. From the chronos power spectra we see that there are two dominant modes, each with multiple singular vectors; the 45 kHz mode is expressed by two (orthogonal) components, suggesting a travelling wave, whereas the greater number of components for the 29 kHz mode suggests more complex structure. We also see that the variance-normalisation of each channel degrades the signal to noise ratio of the system, which can also be described in terms of the normalised entropy.

We calculate the normalised entropy H of the singular values a_k in A:

$$H = \frac{-\sum_{k=1}^{N_a} p_k \log p_k}{\log N_a}, \tag{3}$$

where p_k is the dimensionless energy:

$$p_k = \frac{a_k^2}{E}, \quad E = \sum_{k=1}^{N_a} a_k^2. \tag{4}$$

The low entropy case ($H \to 0$) occurs when the system is well ordered, while the maximal entropy case ($H = 1$, all singular values equal) corresponds to pure noise in the N_c, $N'_s \to \infty$ limit. To an extent, the scalar quantity H can be used as a measure of how physically interesting the signals in S' are without any further investigation into the structure of S', though care must be taken with this interpretation. A standing wave in a system with no noise has only one non-zero singular value $a_0 = 1$ giving $H = 0$, whereas a travelling wave requires at least two singular values giving $H > 0$ for finite N_c, N'_s. A threshold value of normalised entropy may be used to filter out noisy data, however this requires some *a priori* understanding of the data in order to select an appropriate threshold. Entropy filtering was not necessary for the dataset presented here because coherent fluctuations are present in all but a few time segments.

2.3. Fluctuation structures

The orthogonality of the singular vectors implies that any mode, other than a standing wave, will be described by multiple singular values. In order to consider physical modes of the system being analysed, the singular values need to be arranged into groups, hereafter referred to as *fluctuation structures*, which describe the same mode. For the method presented here, the fluctuation structure, denoted by α, is the fundamental representation of an instance of a fluctuation. The normalised energy of a fluctuation

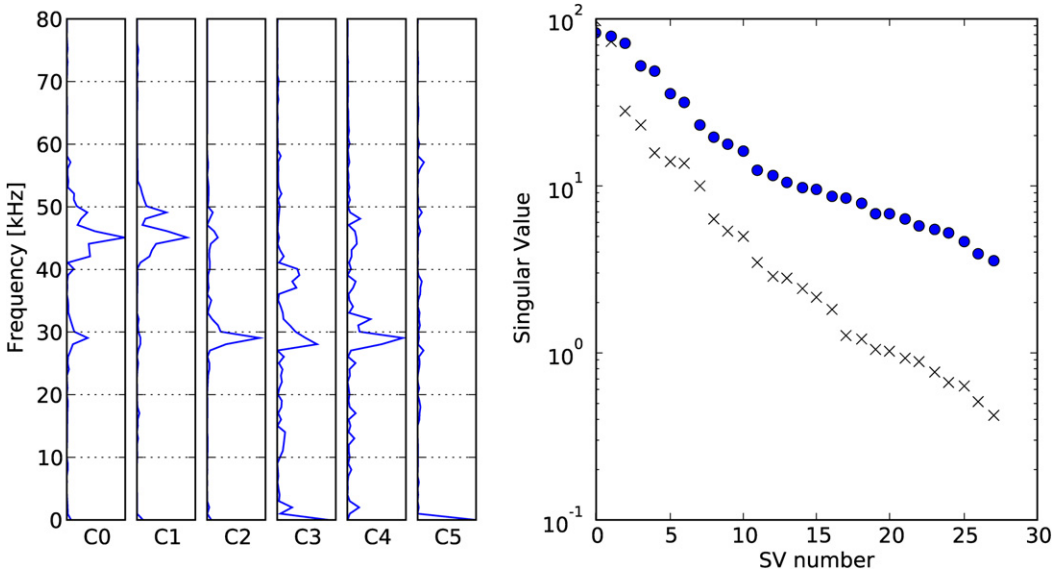

Fig. 2. Example of chronos power spectra and singular values. Singular values from both normalised (o) and unnormalised (x) S' are shown. $C0, C1, \ldots, C5$ denote the chronos from the normalised singular value $0, 1, \ldots, 5$. There are two distinct modes, one at $f \sim 45$ kHz described by SV0 and SV1; the other, weaker, signal is at $f \sim 29$ kHz and is described by SV2–4. The data is from H-1 shot #58122 at 31 ms $< t <$ 32 ms.

structure is defined as the sum if the normalised energies of its constituent singular values:

$$p_\alpha = \sum_{k\in\alpha} p_k. \tag{5}$$

A pair of chronos for an ideal travelling wave will differ only via a $\pi/2$ phase shift; this prompts a grouping of singular values into fluctuation structures according to similarity of power spectra of their chronos. The similarity between two chronos c_1 and c_2 can be quantified with the normalised average of the cross-power spectrum γ_{c_1,c_2}:

$$\gamma_{c_1,c_2} = \frac{G(c_1,c_2)^2}{G(c_1,c_1)G(c_2,c_2)}, \tag{6}$$

where $G(a,b) = \langle|\mathcal{F}(a)\mathcal{F}^*(b)|\rangle$, $\mathcal{F}$ is the Fourier transform, and $\langle\ldots\rangle$ represents the spectral average.

When allocating singular values to fluctuation structures, the observation:

$$\gamma_{a,b} > \gamma_{\min} \quad \text{and} \quad \gamma_{a,c} > \gamma_{\min} \quad \not\Rightarrow \quad \gamma_{b,c} > \gamma_{\min}, \tag{7}$$

suggests it is insufficient to require $\gamma_{a,b} > \gamma_{\min}$ for each pair of singular values a, b within a fluctuation structure. We present two methods for constructing fluctuation structures, the first of which requires manual selection of a threshold value $\gamma_{\min}$, the second of which selects an appropriate $\gamma_{\min}$ from the data in each case.

The 'prescribed threshold' method for grouping singular values, given in Algorithm 1, requires only that each constituent singular vector has $\gamma > \gamma_{\min}$ with respect to the dominant singular vector (with largest corresponding singular value) of the structure, where $\gamma_{\min}$ is a pre-defined. Iteratively, the subset of unassigned singular vectors which meet the $\gamma > \gamma_{\min}$ criterion with respect to the dominant singular vector are defined as a fluctuation structure until all singular vectors have been assigned to a fluctuation structure.

The prescribed threshold method raises the problem of selection of a suitable value of $\gamma_{\min}$. The resulting fluctuation structures for $0 \leqslant \gamma_{\min} \leqslant 1$ from the dataset of Fig. 2 are shown in Fig. 3. At $\gamma_{\min} = 0$, all singular values are grouped together as a single fluctuation structure, while at $\gamma_{\min} = 1$ each fluctuation structure contains one singular value. The key features are the two fluctuation structures $\alpha_0 = \{a_0, a_1\}$ and $\alpha_1 = \{a_2, a_3\}$ which coexist for $0.50 < \gamma_{\min} < 0.87$. After application of such analysis to a suitably sized sample of short time segments, a threshold of $\gamma_{\min} = 0.7$ was found to be appropriate for the dataset presented in the remainder of this paper.

while Number of unallocated singular values > 0 **do**
  Define a new fluctuation structure as an empty set: $\alpha_i = \{\}$
  Denote the largest unallocated singular value by a_ξ
  for Every unallocated singular value a_η **do**
    if $\gamma_{\eta,\xi} > \gamma_{\min}$ **then**
      Allocate a_η to fluctuation structure α_i
    end if
  end for
end while

Algorithm 1. The prescribed threshold method for building fluctuation structures α_i from singular values a_j. The largest unallocated singular value a_ξ will always be allocated to α_i because $\gamma_{\xi,\xi} = 1$.

For different datasets, the nature of the signals will determine the structure of Fig. 3. The effect of decreased signal to noise ratio is a reduction of range of $\gamma_{\min}$ over which a fluctuation structure remains unchanged. For example, with the addition of normally-distributed noise with variance of half the signal RMS to the data used for Fig. 3, the range over which both $\{a_0, a_1\}$ and $\{a_2, a_4\}$ are defined is reduced to $0.70 < \gamma_{\min} < 0.77$. Using modelled signals, a similar effect is seen in the case of increasing fluctuation bandwidth.

It is clear that the manual selection of a $\gamma_{\min}$ threshold is not ideal. An alternative 'automated threshold' method for grouping singular values is described in Algorithm 2 with an example shown in Fig. 4.

Here, the assumption is made that coherent chronos will be grouped together over a wide range of $\gamma_{\min}$ relative to the non-related chronos. Fig. 4 shows this grouping method for the singular values of Fig. 2. In general, there is little variation between the dominant (high energy) fluctuation structures as defined by the two grouping methods. The small singular values tend to undergo less grouping with the second process, though most of these low energy noise terms do not survive the energy thresholding. Using

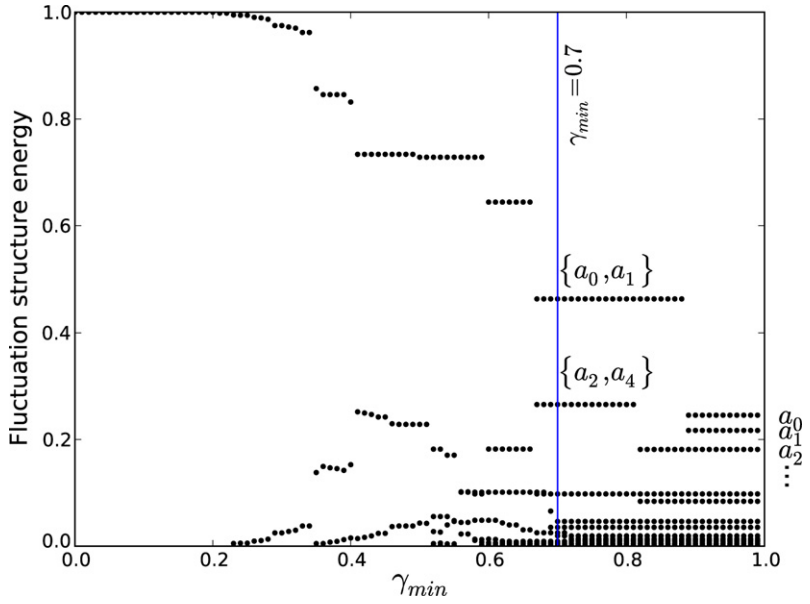

Fig. 3. The possible fluctuation structure groupings according to their energies as defined by Algorithm 1 through the range of $\gamma_{\min}$. The dataset is the same as in Fig. 2. Note that $\gamma_{\min} \lesssim 0.4$ allows unrelated singular values to be included within a fluctuation structure, whereas with $\gamma_{\min} > 0.87$ Algorithm 1 will not recognise the similarity between a_2 and a_3. A value of $\gamma_{\min} = 0.7$ has been selected for larger dataset presented in this paper after consideration of a random sample of such figures.

while Number of unallocated singular values > 0 **do**
  Define a new fluctuation structure as an empty set: $\alpha_i = \{\}$
  Denote the largest unallocated singular value by a_ξ
  Γ = sort($\gamma_{\xi,\eta}$) for unallocated singular values a_η
  $k_i = \arg\max_{k>0}(\Gamma[k+1] - \Gamma[k])$
  Set $\gamma_{\min} = \Gamma[k_i]$
  for Every unallocated singular value a_η **do**
    if $\gamma_{\eta,\xi} > \gamma_{\min}$ **then**
      Allocate a_η to fluctuation structure α_i
    end if
  end for
end while

Algorithm 2. The automated threshold method for constructing fluctuation structures. The requirement $k > 0$ removes the trivial case of $\gamma_{\min} = 0$.

a subset of the H-1 Mirnov data, it was found that less than 10% of fluctuation structures as defined by the two algorithms differed by more than 20% in normalised energy and so, for practical purposes, the results are the same.

Unless filtering is applied, both the above processes will produce many fluctuation structures built from low energy singular values which correspond to noise. A simple method of filtering is to apply an energy threshold to the singular values or fluctuation structures. Fig. 5 shows a random 10% sample of fluctuation structures produced from H-1 Mirnov coil data using Algorithm 1 with $\gamma_{\min} = 0.7$; there is a clear separation between the signal and noise populations. In this case a threshold value of $p_{\min} = 0.2$ is appropriate. The expense of having a single filter operation on the whole dataset is the calculation and storage of rejected fluctuation structures. A less expensive option is to retain a specified proportion of the signal energy; for example, in Fig. 4 the process could be terminated after the 3rd iteration if only 80% of the signal energy was to be retained.

2.4. Mapping of fluctuation structures into $\Delta\psi$-space

We regard each fluctuation structure as a point in the space $[-\pi, \pi]^{N_c}$, an N_c-dimensional torus of length 2π which we will call $\Delta\psi$-space. In this application ψ represents the electrical phase of the reconstructed fluctuation structure at the positions of the coils. Fluctuation structures which are close in $\Delta\psi$-space can be considered the same type. This interpretation arises from the expectation that the possible waves have various phase velocities and mode numbers due to the periodic boundary conditions of the physical plasma torus in which they propagate. It is also applicable to the more general case where waves are spatially localised within the system and do not have well defined mode numbers.

For each fluctuation structure α_l, filtered SVD components are recombined to obtain the filtered data matrix S'_l:

$$S'_l = U_l A_l V_l^*, \tag{8}$$

where U_l, A_l, V_l respectively contain only the topos, singular values and chronos in α_l; and the rows of the matrix S'_l contain the timeseries relating to α_l for each original input channel.

The assumption is made that the power spectra of the chronos in α_l are peaked around a single frequency ω_l, allowing scalar phase differences $\Delta\psi_{a,b}(\omega = \omega_l)$ between channels a and b to be evaluated and used to define the coordinates in $\Delta\psi$-space. To reduce the dimensionality of $\Delta\psi$-space to N_c, only phase differences between nearest neighbour channels are used, such that the coordinates of a fluctuation structure are $[\Delta\psi_{1,2}, \Delta\psi_{2,3}, \ldots, \Delta\psi_{N_{c-1},N_c}]$.

Note that in our example dataset the actual phase difference between channels depends on κ_h so we map the phase differences to a coordinate system which is independent of κ_h, namely the κ_h-averaged magnetic angles of the Mirnov coils.

2.5. An overview of the preprocessed dataset

As an overview of the preprocessed dataset, Fig. 6 shows the fluctuation structures created with Mirnov coil signals from the H-1 configuration scan described in Section 1, with $p > 0.2$ filter applied. The frequency of the datapoints is given by the peak frequency of the power spectrum of the dominant chrono in the fluctuation structure. The radial location of low-order rational magnetic surfaces, $\iota = n/m$, are also shown. These rational surfaces are important as fluctuations with toroidal and poloidal mode numbers n and m respectively can resonate with the twisted field lines. The main features of the fluctuation spectra are the resonances about $\kappa_h = 0.4$ and $\kappa_h = 0.76$, related to the $\iota = 5/4$

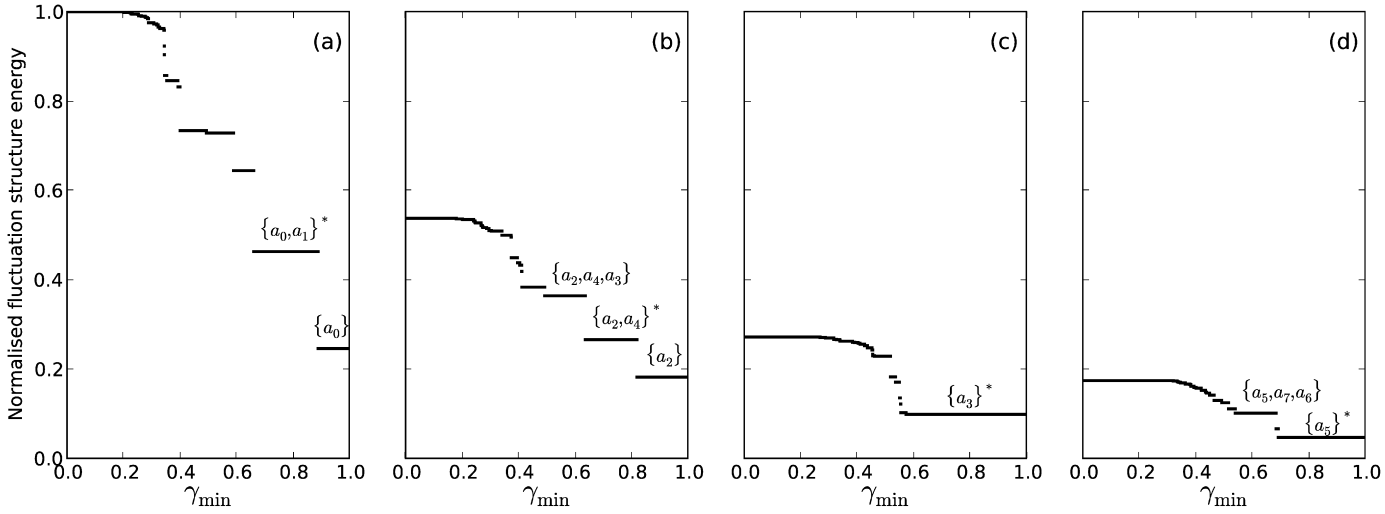

Fig. 4. The process of grouping singular values into fluctuation structures as defined by Algorithm 2. In the first iteration, figure (a), the largest singular value is a_0 and the sets of singular values $\{a_i, \ldots\}$ for which $\gamma_{a_0,a_i} > \gamma_{\min}$ are plotted against normalised energy and $\gamma_{\min}$. The best fluctuation structure for a_0 is determined to be $\{a_0, a_1\}$ because it is defined over the largest range of $\gamma_{\min}$. With a_0 and a_1 removed from the set, the process is repeated (b), and $\{a_2, a_4\}$ is found to be the most appropriate fluctuation structure for a_2.

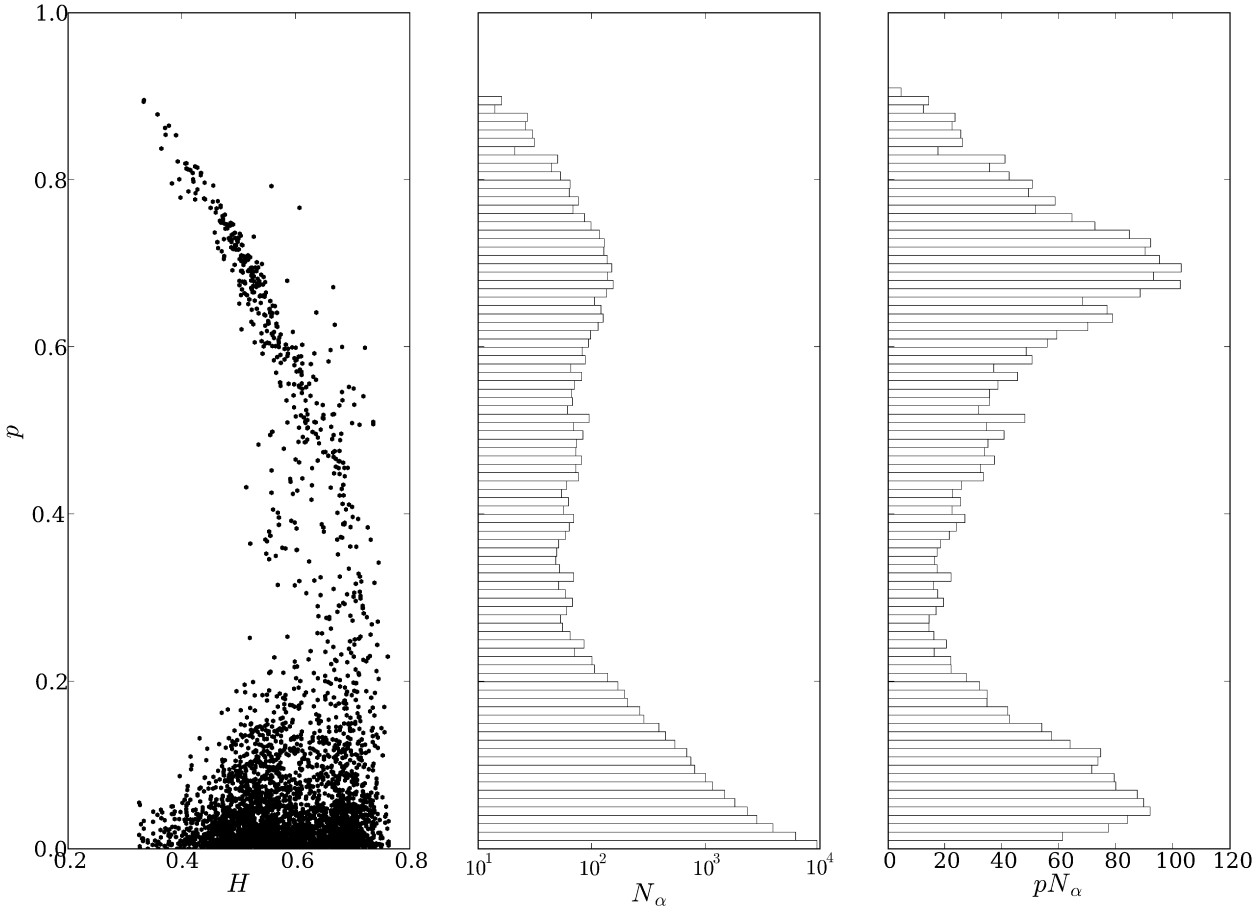

Fig. 5. A 10% random sample of the H-1 Mirnov dataset. The left panel shows p and H for the fluctuation structures, where p is given by Eq. (5) and H by Eq. (3) for the corresponding SVD. The middle panel shows the number of fluctuation structures N_α within $|\delta p| < 0.01$. The right panel shows pN_α, which is effectively the density of normalised energy; while this is not physically meaningful because the normalisation factor varies between time segment, it is a useful guide to the energy distribution among fluctuation structures.

and $\iota = 4/3$ surfaces respectively. We expect that any automated process used to locate distinct types of fluctuations would identify these features, and hopefully find some less obvious features. Indeed in Section 4 it can be seen that these two features are the first to be distinguished by the following clustering algorithm.

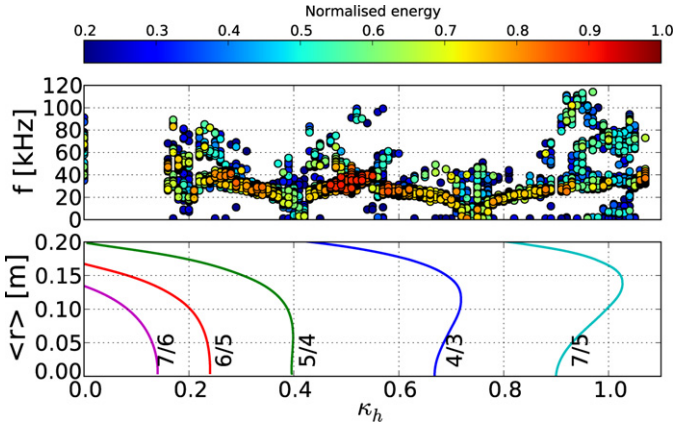

Fig. 6. Full H-1 κ_h scan dataset with 5300 fluctuation structure datapoints after application of the $p > 0.2$ filter. In the upper panel, fluctuation structures are mapped to frequency and the magnetic geometry parameter κ_h; the marker colour is proportion to the normalised energy of the fluctuation structure. In the lower panel, the average minor radial $\langle r \rangle$ locations of low order rational magnetic surfaces are shown.

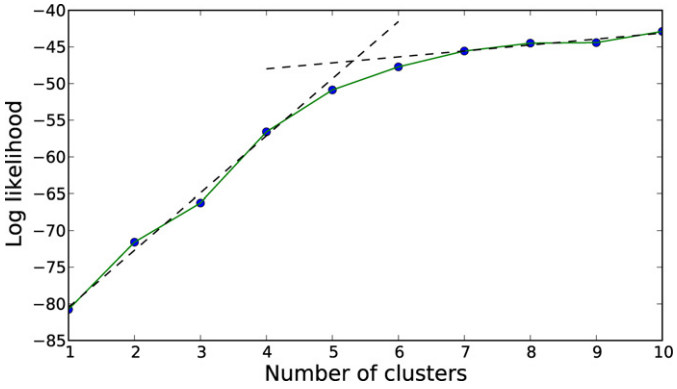

Fig. 7. Log likelihood vs. number of clusters for H-1 κ_h scan Mirnov data clustered with the EM algorithm. The elbow criterion shows the optimal number of clusters to be $N_{Cl} \simeq 5$.

3. Clustering

We aim to discover any underlying lower-dimensional model of the dataset; that is, groups of fluctuation structures which are similar throughout some range of short time segments. As discussed in Section 2.4, we assume that a class of fluctuations is localised in the N_c-dimensional $\Delta\psi$-space. For example, it is simple to understand such localisation in terms of a simple cylindrical geometry with equidistant poloidal measurements, where each mode with poloidal mode number m will be located at $\Delta\psi = 2\pi m/N_c$ in each dimension. However, we assume a generalised case in which the fluctuation may have arbitrary, including localised, structure.

Clustering algorithms generally fall into two categories, using either hierarchical or relocation methods [7]. Hierarchical methods involve the iterative merging or dividing of clusters, with each step determined by the optimization of some criterion, e.g. minimal distance between two clusters to decide which clusters to merge. Relocation methods take a prescribed number of clusters and iteratively vary the cluster parameters until a convergence criterion is met.

To simplify the clustering procedure, the data is mapped to the non-periodic space $[-1, 1]^{2N_c}$ though $\sin(\Delta\psi)$ and $\cos(\Delta\psi)$ projections allowing the use of the standard Euclidean metric. In this space, we assume that each type of fluctuation can be described by a $2N_c$-dimensional Gaussian distribution. For such a group of distributions, or mixture model, the expectation maximisation (EM) algorithm is usually used [7].

The EM algorithm is a relocation method for estimating the most likely values of latent variables in a probabilistic model [6]. The latent variables are the mean μ_i and standard deviation σ_i for each cluster i, where $i = 1, 2, 3, \ldots, N_{Cl}$ and N_{Cl} is the number of clusters. Given the initial conditions, in the form of random initial μ_i and σ_i values for a prescribed number of clusters, the EM algorithm consists of two steps which repeat until a convergence criterion is met. Firstly, the expectation step assigns to each datapoint a probability, or expectation value, of belonging to each cluster which is calculated with the Gaussian distribution function. Secondly, μ_i and σ_i are recalculated using the new expectation values as weight factors.

The 10-fold cross-validated log-likelihood ratio is used as a measure of how well the cluster assignments fit the data. The cross-validation process involves partitioning the dataset into random subsamples and comparing results from each subset to avoid oversensitivity to outliers in the data. The likelihood is the conditional probability of obtaining the cluster means and standard deviations given the observed data. Because the EM algorithm can only guarantee a local maximum in likelihood we use a Monte Carlo approach, with multiple repetitions with different randomised initial conditions for each N_{Cl}.

The identification of the optimal number of clusters, or of those which are important, is a difficult and often an ill-posed problem. A commonly used heuristic is the 'elbow' criterion, which identifies a distinct flattening of the plot of W versus N_{Cl} with the optimal number of clusters, where W is some convenient error measure [10]. Shown in Fig. 7 are the EM clustering results for the H-1 dataset with the likelihood function as the error measure. The elbow criterion suggests $N_{Cl} \simeq 5$ to be optimum.

Other methods for finding the optimal number of clusters include clustering gain [9] for hierarchical clustering and the Bayesian information criterion (BIC) [11] for EM clustering. However for the case of low N_{Cl}, and when automation is not essential, i.e. the optimal N_{Cl} in not required as input for successive iterations of a parent algorithm, a manual inspection of the range of clustering results is often more informative than a nominally optimal set of clusters.

3.1. Visualisation

In the absence of a reliable method to identify the optimum number of clusters, we have found inspection of a dendrogram, or *cluster tree*, mapping to be a practical method for identifying the important clusters. The cluster tree displays clusters for each N_{Cl} below some maximum value $N_{Cl,max}$, with all clusters for a given N_{Cl} forming a single tree level. Each child cluster is mapped to the cluster on the parent level with which it has the largest fraction of common datapoints. Cluster branches which do not fork over a significant range of N_{Cl} are deemed to be well defined, and the point where well defined clusters start to break up suggests that N_{Cl} is too high. While this procedure is clearly a subjective one, it is effective and does not depend on the type of clustering algorithm used.

The cluster tree for our example dataset applied to the EM algorithm is shown in Fig. 8. Each cluster has been defined only by its phase structure ψ and mapped back to κ_h and frequency $f = (2\pi)^{-1}\omega_l$. The base of the tree ($N_{Cl} = 1$) shows all the data within a single cluster (EM:A); as we climb up the tree different classes of fluctuation are isolated. For example, the branch starting at cluster EM:B contains fluctuations with toroidal mode number $n = 5$ and poloidal mode number $m = 4$ which occur at configurations near the $\iota = 5/4$ resonance ($\kappa_h \simeq 0.4$). Similarly, the branch containing cluster EM:C is due to the $\iota = 4/3$ resonance near $\kappa_h = 0.75$. Other clusters include fluctuations which occur at

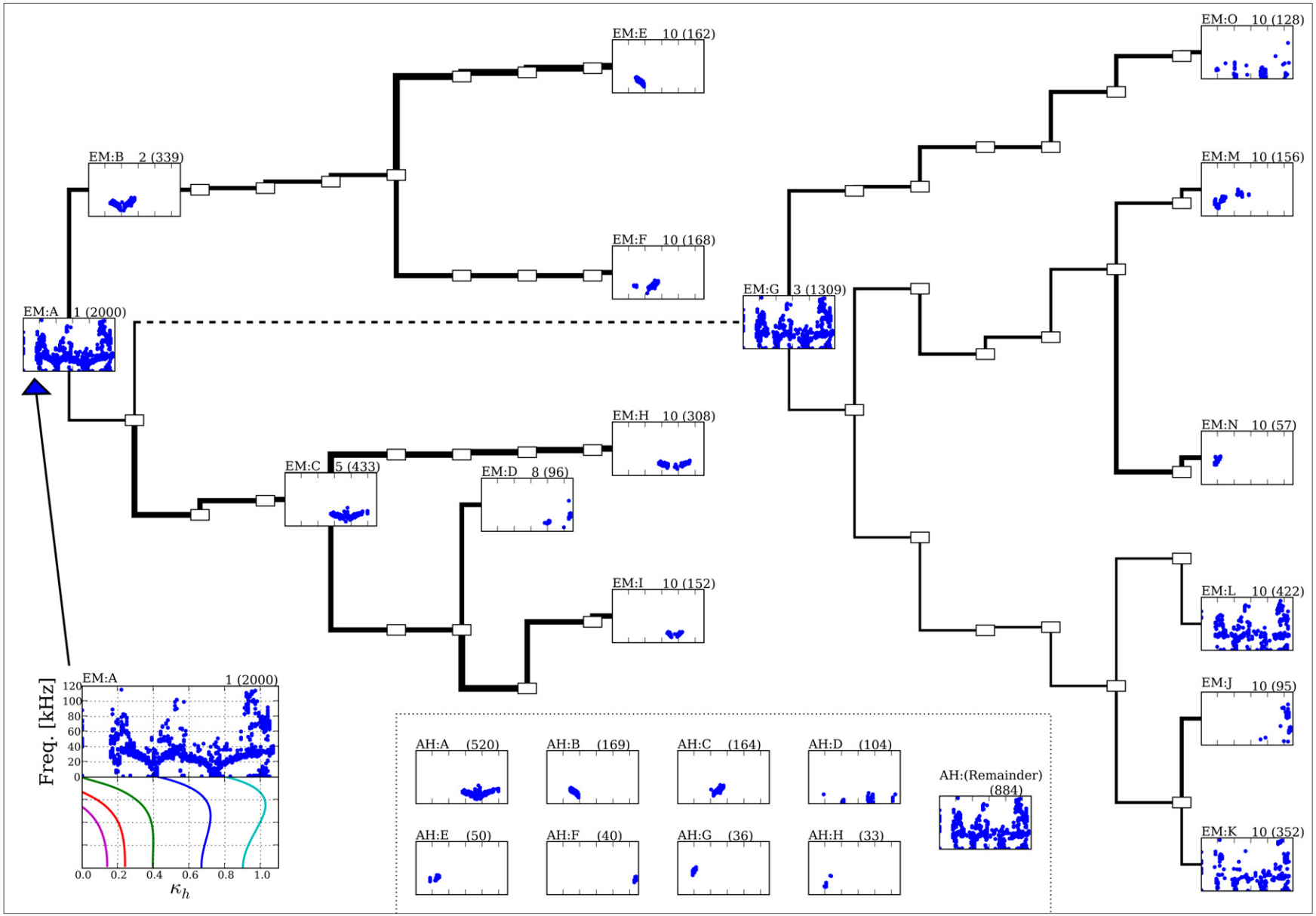

Fig. 8. Cluster tree of the example dataset. The figure in the bottom left corner is equivalent to Fig. 6; its upper panel shows the fluctuation structures mapped to f and κ_h, the numbers 1 (2000) at the top right are the tree level, N_{Cl}, and cluster population respectively, EM:A is a cluster label used for reference. For clarity, only a subset of clusters within the tree have their contents displayed and EM:G has been displaced to prevent overlap. Vertical parent-child distance is proportional to the distance between cluster means, while line thickness is inversely proportional to the Gaussian width of the cluster. Several clusters produced by the agglomerative hierarchical (AH: labels) method are also shown, these are essentially equivalent to the $N_{Cl} = 10$ level EM clusters, see Table 1 for comparison.

higher order resonances, as well as low frequency $n, m = 0$ modes (cluster EM:O branch) and weakly defined residual clusters (EM:K and EM:L) which would be resolved at a higher level of the tree than is shown here.

In order to illustrate some of the clusters in a physically meaningful way, poloidal phase-angle plots for a single poloidal Mirnov array are shown in Fig. 9. The centre line corresponds to the cumulative mean phase of the coil pairs. The lines above and below are the cumulative cluster standard deviations of the coil pairs, where $\sigma^2_{1,n} = \sum_{j=1}^{n-1} \sigma^2_{j,j+1}$ for $\psi_{1,n} = \sum_{j=1}^{n-1} \Delta\psi_{j,j+1}$. Here, the magnetic angles have been evaluated for a flux surface at $r = 0.1$ m to better represent the broad radial structure expected for these modes. In our particular case of the modes in H-1, this is the desired end result of the classification of data, as it shows the mode numbers of each cluster clearly. Independently of the physical interpretation, it is also a convenient way to visualise the spread of the cluster in each dimension. Finally, the relatively smooth increase in phase with angle indicates a travelling wave – a standing wave would should phase jumps of π. This is consistent with the fluctuation structure having two main components, one "sine-like" and the other "cosine-like" as discussed earlier in relation to Fig. 2.

4. Discussion

The data mining algorithm presented here is potentially useful in numerous other domains where spatio-temporal data is used. However, there is a limitation to the nature of the fluctuations amenable to this analysis due to the SVD. The SVD is not effective in distinguishing different modes coexisting with the same frequency or spatial structure because the modes would share a chrono or topo, whereas the SVD requires orthogonal components to distinguish modes. The assumption that such coexisting modes are not present is also important in assigning a single frequency ω_l to a fluctuation structure, i.e.: two modes with the same spatial structure will also share chronos, but only one frequency would be recorded. The SVD was selected as an initial approach to feature extraction because it is a standard linear method in which energy is conserved and the decorrelation of temporal and spatial components produces intuitive basis vectors on which coherent structures are easily discernible. Future work may involve other linear and non-linear approaches based on the SVD such as principle- and independent-component analysis [18].

The EM clustering method described here assumes clusters to be described by Gaussian distributions, which is a poor assumption given the mapping to $\sin(\Delta\psi)$ and $\cos(\Delta\psi)$ projections. To check if the imposed Gaussian distributions significantly influence the cluster outcomes we have also applied to the dataset the agglomerative hierarchical (AH) clustering algorithm [8] which is an agglomerative rather than relocation algorithm. The initial condition for AH clustering is that each fluctuation structure defines a cluster. Utilising the same Euclidean metric used for the EM algorithm, the two closest clusters are combined. This joining process is iterated until $N_{Cl} = 1$, giving a (prohibitively large) cluster tree. Compression of the AH cluster tree can be achieved by filtering out clusters with small populations, allowing for a visualisation similar to the EM cluster tree in Fig. 8; a set of AH clusters which are essentially equivalent to the $N_{Cl} = 10$ level of the EM cluster tree are also shown in Fig. 8. The clusters resulting from the EM and AH methods have been found to be essentially the same apart from the weakly defined clusters (EM:K,L) and the 'remainder' (AH:Rem). A quantitative comparison between populations of EM and AH clusters in Fig. 8 is shown in Table 1. Alternatives to these clustering procedures, such as support vector machines [18], will are being considered for future work.

In an ideal dataset the mapping of the system to the detectors would be fixed. In practice, however, it is possible that the detector or system geometry is not constant. Variation of detector or system configuration will result in distortion of the data in the clustering space, leading to inaccurate or difficult to interpret results. These effects can be mitigated through additional pre-processing, or modification of the clustering metric. As mentioned in Section 2.4, for the dataset presented here an extra preprocessing step was applied in order to compensate the variation of magnetic geometry with κ_h. For fusion devices with more dynamically variable plasma geometries, for example the Shafranov shift in high-pressure plasmas, this may pose a more difficult problem.

Table 1
A comparison of populations of clusters produced by the EM ($N_{Cl} = 10$) and AH algorithms. Clusters are shown in Fig. 8.

Cluster	AH:A	AH:B	AH:C	AH:D	AH:E	AH:F	AH:G	AH:H	AH:(Rem.)	total
EM:H	307								1	308
EM:I	152									152
EM:E		161							1	162
EM:F		3	155						10	168
EM:O				88					40	128
EM:M					50			28	78	156
EM:J	3					40			52	95
EM:N							36		21	57
EM:K	33	3	6	16				5	289	352
EM:L	25	2	3						392	422
total	520	169	164	104	50	40	36	33	884	2000

It is important to consider the scalability and computational requirements of the algorithm. Given fixed values of N_c and Δt, the size of S' remains constant and the preprocessing stage has complexity $\mathcal{O}(N_{\mathcal{S}})$, where $N_{\mathcal{S}}$ is the number of timeseries datasets $\mathcal{S}$. The scalability of the clustering stage depends on the algorithm used, for the EM case we have $\mathcal{O}(N_{Cl}N_\alpha)$, which gives $\mathcal{O}(N_{\mathcal{S}})$ for constant N_{Cl}. The AH clustering algorithm is less desirable in terms of scalability as it has complexity $\mathcal{O}(N_\alpha^2)$ due to distance calculation between each pair of fluctuation structures.

We have implemented the preprocessing and visualisation stages using the python language with the Scipy and Matplotlib libraries [12]. The preprocessing of our dataset, 4600 S' arrays (28 by 1024), takes around 2 hours using a 1.9 GHz Intel Pentium M processor. The results are stored in MySQL tables; a table of fluctuation structure properties excluding $\Delta\psi$-space mapping is around 5 Mb in size, with the 3.6×10^6 rows of the $\Delta\psi$ mapping table taking around 30 Mb, using optimal data types. For clustering, we have used the EM algorithm from the WEKA suite of data mining tools [13,14] which runs at about $0.05 \times N_{Cl} \times N_\alpha$ CPU seconds using 2.2 GHz AMD Opteron processors. For each N_{Cl}, 100 randomised initial conditions were used; the results with maximal log-likelihood are selected as the best clusters.

The physical nature of the fluctuations in our example dataset is not yet completely understood. The dependence of spectra on plasma density n measured by the interferometer in Fig. 1 and ι suggests a dispersion relation similar to that of the global Alfvén eigenmode (GAE) [15,16]. However, the observed frequencies are smaller than the expected GAE frequencies by a factor of around 1/3 [17]; an experimental campaign is presently being undertaken in order to resolve this difference.

Future directions for this work, apart from those mentioned above, include the application of the technique to databases of other fusion devices. This work has started for the Heliotron J [19] and TJ-II [20] devices, with planning underway for LHD [21] and the historical database of W7-AS [22]. While the work presented in this paper used the prescribed threshold method of Algorithm 1, initial results from these other devices have successfully applied the more scalable automated threshold method of Algorithm 2. An important development will be the design and implementation of a framework in which to relate clustering results from Mirnov sig-

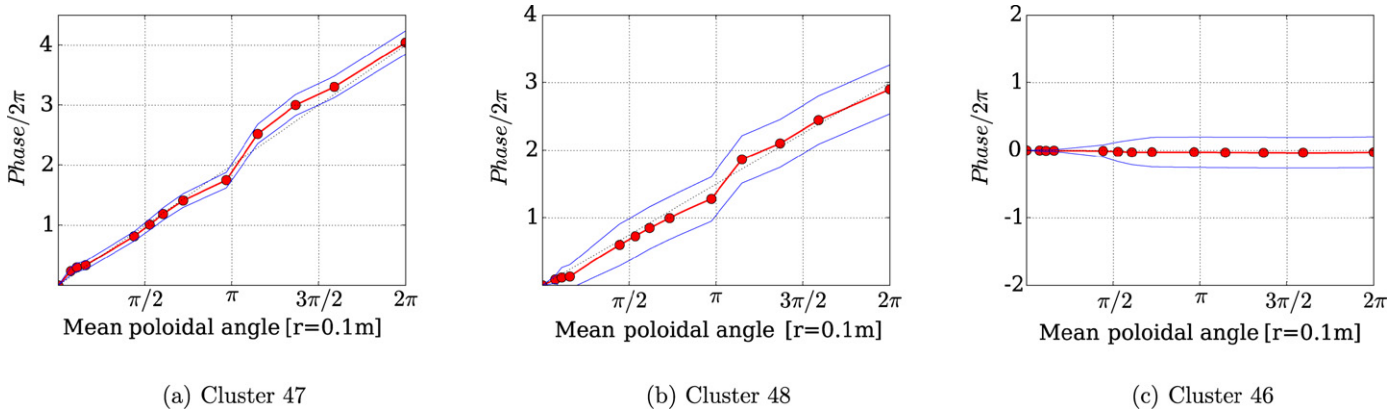

Fig. 9. Variation of phase around one of the poloidal Mirnov arrays, plotted against mean magnetic poloidal coordinate. The centre line is the cumulative mean phase of the coil pairs, with standard deviation shown above and below. The mode numbers shown here are supported by Fourier analysis of the data.

nals to other machine data through techniques such as Bayesian inference. To facilitate these plans, an open-source project, PyFusion [23], has been developed to allow researchers to share data analysis, and data mining, code across fusion devices.

5. Conclusion

We have presented a highly automated data mining process for the characterisation of fluctuations in multichannel timeseries data. In the implementation shown here, the manual interaction is restricted to two tasks: the selection of a cross-power threshold $\gamma_{\min}$ and the choice of appropriate filter parameters. For the former we have shown an alternative method which does not require manual intervention, while the latter is an operation applied to the dataset as a whole.

Given an appropriate choice of clustering algorithm, the data mining process scales well, with complexity $\mathcal{O}(N_S)$. We have used the procedure here with magnetic fluctuation data from configuration scans in the H-1 heliac, identifying different modes in parameter space. The process should be easily adaptable to other types of multichannel oscillatory timeseries data.

Acknowledgements

The authors would like to thank the H-1 team for continued support of experimental operations as well as J. Harris, F. Detering and M. Hegland for useful discussions. This work was performed on the H-1NF National Plasma Fusion Research Facility established by the Australian Government, and operated by the Australian National University, with support from the Australian Research Council Grants DP0344361 and DP0451960.

References

[1] A. Fasol, C. Gormenzano, H.L. Berk, B. Breizman, et al., Progress in the ITER physics basis, Nucl. Fusion 47 (2007) S264–S284, doi:10.1088/0029–5515/47/6/S05, Chapter 5: Physics of energetic ions.
[2] S.M. Hamberger, B.D. Blackwell, L.E. Sharp, D.B. Shenton, H-1 design and construction, Fusion Technol. 17 (1990) 123–130.
[3] J.H. Harris, et al., Fluctuations and stability of plasmas in the H-1NF heliac, Nucl. Fusion 44 (2004) 279–286.
[4] B.D. Blackwell, Results from helical axis stellarators, Phys. Plasmas 8 (2001) 2238–2244.
[5] T. Dudok de Wit, A.-L. Pecquet, J.-C. Vallet, R. Lima, The biorthogonal decomposition as a tool for investigating fluctuations in plasmas, Phys. Plasmas 1 (1994) 3288–3300.
[6] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Royal Stat. Soc. 39 (1977) 1–38.
[7] C. Fraley, A.E. Raftery, How many clusters? Which clusters? Answers via model-based cluster analysis, Comput J. 41 (1998) 578–588.
[8] W.H.E. Day, H. Edelsbrunner, Efficient algorithms for agglomerative hierarchical clustering methods, J. Classification 1 (1984) 7–24.
[9] Y. Jung, H. Park, D.-Z. Du, B.L. Drake, A decision criterion for the optimal number of clusters in hierarchical clustering, J. Global Optim. 25 (2003) 91–111.
[10] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Statist. Soc. B 63 (2) (2001) 411–423.
[11] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[12] http://python.org,, http://scipy.org,, http://matplotlib.sourceforge.net.
[13] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, 2005.
[14] http://www.cs.waikato.ac.nz/ml/weka.
[15] K.L. Wong, A review of Alfvén eigenmode observations in toroidal plasmas, Plasma Phys. Control. Fusion 41 (1999) R1–R56.
[16] D.A. Spong, R. Sanchez, A. Weller, Shear Alfvén continua in stellarators, Phys. Plasmas 10 (2003) 3217–3224.
[17] D.G. Pretty, PhD Thesis, Australian National University, 2007.
[18] V. Cherkassky, F. Mulier, Learning From Data: Concepts, Theory, and Methods, second ed., John Wiley & Sons, 2007.
[19] T. Obiki, et al., Goals and status of Heliotron J, Plasma Phys. Control. Fusion 42 (2000) 1151–1164.
[20] C. Alejaldre, et al., TJ-II project; a flexible heliac stellarator Fusion Technol. 17 (1990) 131–139.
[21] A. Iiyoshi, et al., Overview of the Large Helical Device project, Nucl. Fusion 39 (1999) 1245–1256.
[22] H. Renner, et al., Initial operation of the Wendelstein 7AS advanced stellarator, Plasma Phys. Control. Fusion 31 (1989) 1579–1596.
[23] http://pyfusion.googlecode.com.