Automatic Analysis of Multichannel Time Series Data Applied to MHD Fluctuations

D. G. Pretty[1)], B. D. Blackwell[1)], F. Detering[1)], M. Hegland[3)], J. Howard[1)], D. Oliver[1)], M. J. Hole[2)] and J. H. Harris[4)]

[1)] *Plasma Research Laboratory, and*
[2)] *Department of Theoretical Physics, Research School of Physical Sciences and Engineering, and*
[3)] *Mathematical Sciences Institute, all of the Australian National University, ACT, 0200, AUSTRALIA*
[4)] *Oak Ridge National Laboratory, Tn, USA*

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. In a preprocessing step, 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, 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 are used to filter the dataset. We assume that distinct classes of fluctuations are localised in the space of phase differences (n, n+1) between each pair of nearest neighbour channels. An expectation maximisation (EM) clustering algorithm is used to locate the distinct classes of fluctuations, and a cluster tree mapping is used to visualise the results. Different classes of fluctuations in H-1 distinguished by this procedure are shown to be associated with MHD activity around separate resonant surfaces, with corresponding toroidal and poloidal mode numbers. Equally interesting are some clusters that don't exhibit this behaviour.

Keywords: Data mining, plasma physics, Mirnov oscillations, magnetic fluctuations

1 Motivation

Rotational transform parameter scans have been undertaken in the H-1NF heliac [1, 2, 3]. Using 28 Mirnov coils over 92 discharges, with each shot incrementally changing the magnetic geometry, a large set of timeseries data was produced.

The motivation for the present work has been to find an algorithm which, with minimal human interaction, can group together all similar fluctuations over a large number of shots. Here we present our method, showing that it can be used to discover interesting spectral features and to map classes of fluctuations to any known parameter of the dataset, e.g.: magnetic geometry.

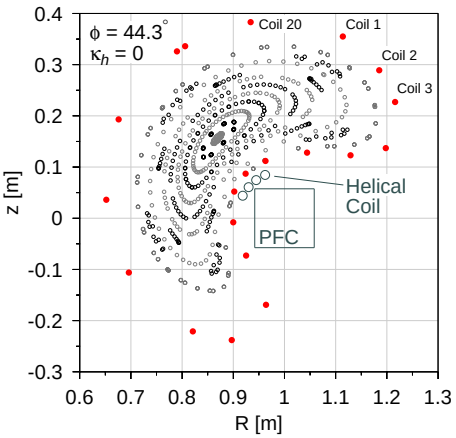

Fig. 1 One of two poloidal Mirnov coil arrays installed in the H-1 heliac

2 Data

Two toroidally-separated poloidal Mirnov coil arrays were used in the experimental campaign described here, one array is shown in figure 1. An additional linear array of 5 coils is also installed. In all, 28 Mirnov coils are used in this dataset.

The experimental parameter scanned is κ_h, the ratio of current in the helical winding coil to that in the toroidal and poloidal field coils. To a good approximation, κ_h scales linearly with rotational transform on axis (ι_0). Due to the precisely controllable coil power supplies, very reproducible plasmas can be formed with an accuracy in vacuum field rotational transform of 1 part in 1000. The range of configurations used here is shown in figure 2, from a monotonic profile with $\iota_0 = 1.12, \iota_a = 1.28$ ($\kappa_h = 0$) to a reverse-shear profile with $\iota_0 = 1.45, \iota_a = 1.46, \iota_{\min} = 1.41$ ($\kappa_h = 1.07$).

The time evolution of Mirnov spectra and line-averaged electron density for shot typical of this dataset are shown in figure 3. The RF power is essentially constant at 60 kW, producing peak density of $\bar{n}_e = 10^{18}\mathrm{m}^{-3}$. The

author's e-mail: david.pretty@anu.edu.au

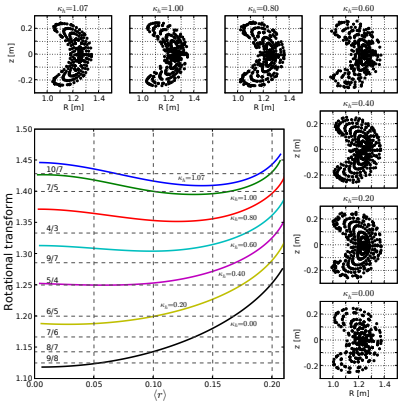

Fig. 2 Rotational transform profiles used in this campaign. Poincaré plots shown for the various magnetic configurations.

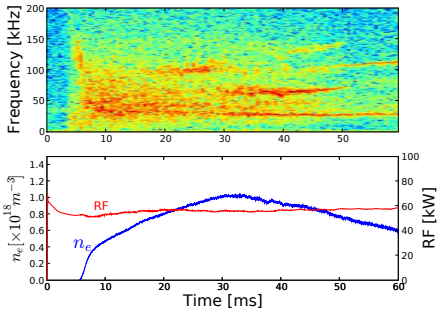

Fig. 3 Typical shot for this campaign (at $\kappa_h = 1.0$), showing Mirnov spectra in the top panel and line-averaged electron density in the lower panel. Heating power (60 kW RF) is essentially constant throughout the discharge.

Mirnov spectra shows multiple co-existing modes, with the higher frequency features ($f > 50$ kHz) showing Alfvénic density scaling ($f \propto n_e^{-1/2}$).

3 Data Pre-processing

Data mining procedures generally have 3 stages: firstly the data needs to be pre-processed into a format suitable for the main algorithm; secondly the main algorithm (neural network, clustering, association rules, etc) is applied to the data, finally the results are visualised and interpreted. The pre-processing used here involves separating out different mode components from the timeseries data and mapping them to a phase space in which the main clustering algorithm is used to distinguish different classes of fluctuation.

For each shot, we split the timeseries data into short time segments, in this case we use $\Delta t = 1$ ms. For each short time segment, we take the singular value decomposition (SVD) [4] of all Mirnov channels. Noise is filtered out of the system by placing a threshold on the normalised energy of a singular value ($p_k = a_k^2/E$, $E = \sum_{k=1}^{N_{\rm Ch}} a_k$ for s.v. a_k and $N_{\rm Ch} < N_s$ where $N_{\rm Ch}$ is the number of channels and N_s is the number of samples in Δt). Noisy short time segments can be filtered with a threshold value of normalised entropy ($H = -\sum_k p_k \log p_k / \log N_{\rm Ch}$).

A mode can be be described by several singular values, for example a rotating mode will have two orthogonal bases (i.e. sine and cosine topos (spatial basis vector) with chronos (temporal basis vector) also with $\pi/2$ phase difference). We assume that singular vectors whose chronos have similar power spectra belong to the same mode, grouping together sets of singular values with normalised cross-power above a threshold value γ:

$$\gamma_{a,b} = \frac{G(a,b)^2}{G(a,a)G(b,b)}, \quad G(a,b) = \langle |\mathcal{F}(a)\mathcal{F}^*(b)| \rangle, \quad (1)$$

where $\mathcal{F}$ is the Fourier transform, and $\langle \ldots \rangle$ represents the spectral average. An examination of a randomly selected subset of SVDs showed that a threshold of $\gamma = 0.7$ is suitable for the present dataset. Each group of singular values defines a *fluctuation structure*.

For each fluctuation structure, we take the inverse SVD using only the allocated subset of singular values (others are set to zero) to return timeseries data for each coil representing the given fluctuation. From these timeseries we take the phase differences between nearest neighbour coils to produce coordinates in $N_{\rm Ch}$-dimensional phase space ("$\Delta\phi$-space") in which the clustering algorithm operates.

4 Clustering

We aim to discover any underlying lower-dimensional model of the dataset, i.e.: to group data points into classes or modes which can then be mapped back to other properties, such as magnetic geometry. We assume that distinct modes of fluctuation will be localised in the $\Delta\phi$-space defined by the nearest neighbour phase-differences. This $\Delta\phi$-space localisation can be easily understood in the simplified case of poloidally equispaced Mirnov arrays in cylindrical geometry – here a mode with poloidal mode number m will be localised in the region of $\Delta\phi(i, i+1) = 2\pi m/N_M$, where N_M is the number of Mirnov coils in the array. Clearly the heliac geometry is not so simple, however the clustering algorithm can find arbitrary phase structures and there is no need to interpret the phase structure before clustering.

We use the expectation maximisation (EM) clustering algorithm which finds the most likely values of latent variables in a probabilistic model [6]. Here we model the data by $N_{\rm Cl}$ $N_{\rm Ch}$-dimensional Gaussian distributions in $\Delta\phi$-space, with mean and standard deviation for each cluster as the set of latent variables. To ensure the results are not biased by the assumption of Gaussian cluster shape, we compare the EM clustering results with results from an ag-

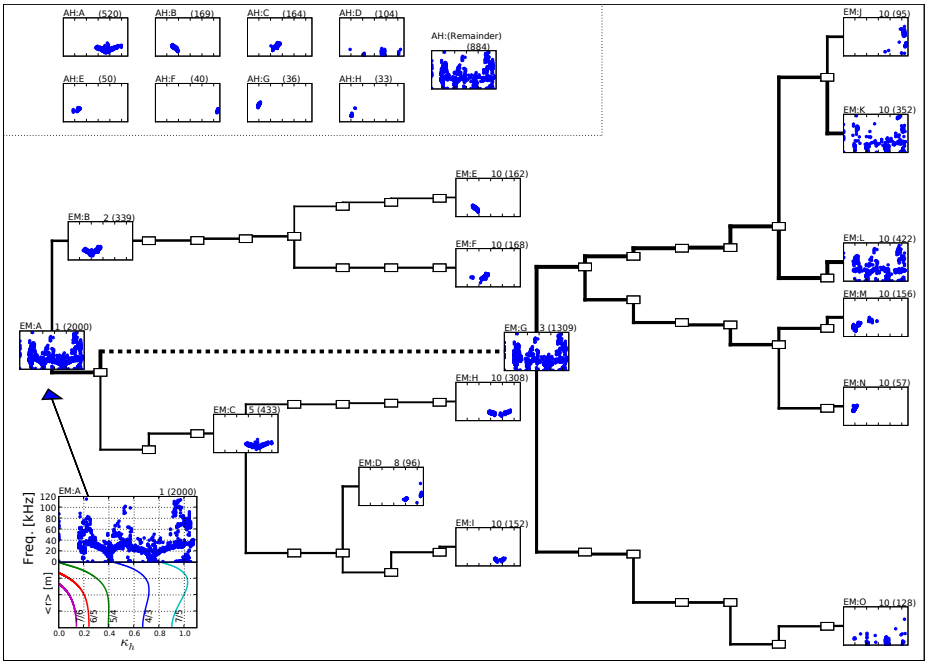

Fig. 4 A Cluster tree representation of the κ_h scan data. The figure in the bottom left corner contains the whole dataset; its upper panel shows the fluctuation structures mapped to f and κ_h, the numbers 1 (2000) at the top right are the tree level, N_{Cl} , and cluster population respectively. The lower panel shows the contours of low-order rational surface within the plasma minor radius (y-axis) for the configurations. 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 AH clusters are also shown for comparison.

glomerative hierarchical (AH) clustering algorithm. Good agreement is found between results from both methods.

5 Results

Shown in figure 4 is a cluster tree graph with plots showing frequency vs. κ_h of selected clusters, each point plotted in a graph corresponds to a fluctuation structure. The root of the tree (left side) has the trivial case of $N_{\mathrm{Cl}} = 1$, such that the whole dataset is contained in a single cluster. The highest level branch shown here (right side) has $N_{\mathrm{Cl}} = 10$. The localisation of the clusters in the $f - \kappa_h$ projection shown here is due to the relation of the phase structure to the magnetic configuration – neither frequency nor κ_h are included in the clustering metric.

Vertical displacement at a node in the tree is proportional to the distance in phase space between the parent and child clusters. Results from AH clustering are also shown for comparison. The AH clusters show good agreement with EM clusters at the $N_{\mathrm{Cl}} = 10$ end of the tree.

Near the root of the tree we find well defined clusters which have resonant structure in the $f - \kappa_h$ projection. For the EM:B cluster, the frequency minimum occurs at $\kappa_h = 0.4$, corresponding to the $\iota = 5/4$ rational surface in the low-shear region of the plasma. The EM:C cluster has frequency minimum at $\kappa_h \simeq 0.72$, where the $\iota = 4/3$ surface is located in the region of zero-shear. Both modes show $f \sim |\iota - n/m|$ resonant behavior, as do other clusters for higher order rational surface configurations. Analysis of the phase structure of these modes has shown that the dominant Fourier components are those expected for the relevant rational surface, i.e.: $(n, m) = (4, 3)$ for the $\iota = 4/3$ mode [7].

Cluster EM:O has very well defined $(n, m) = (0, 0)$ structure and also exists in configurations where low order rational surfaces exist in low-shear regions. Cluster EM:J appears to be a helical Alfvén eigenmode (HAE) coupled between the $(7, 5)$ and $(10, 7)$ resonance, this $\delta_n = 3$, $\delta_m = 2$ is consistent with the relatively large $(n, m) = (3, 2)$ Fourier component of the heliac magnetic geometry.

6 Discussion

The clustering occurs only in $\Delta\phi$-space, and is unbiased by the κ_h and frequency coordinates plotted in cluster tree figure 4. The clusters can also be mapped to any other known plasma properties.

Complications arise in the case of H-1 configuration scans due to the changing shape of magnetic field with κ_h. The coil coordinates have been mapped to κ_h-averaged magnetic angles to account for this.

The process has been kept general enough for it to be applicable to any set of geometrically ordered timeseries data. While the simple phase difference clustering metric works fine in this case, an alternative metric may be required if timeseries from other diagnostics are included.

The scalability depends on the clustering algorithm used. The EM method scales well with the number of datapoints N_{dp} ($N_{dp} \times N_{\mathrm{Cl}}$), while the AH method scales poorly (N_{dp}^2). Variations to the EM clustering algorithm and alternative ways to quantify the results are areas of ongoing investigation.

Data mining methods apart from the clustering analysis described here are also being investigated. "Association rule mining" is one such approach. It is designed to discover temporal patterns in the entire spectrograms such as shown in figure 3. In this context we search for rules (frequent patterns) in the form of "if a strong mode at $f \simeq 10\,\mathrm{kHz}$ occurs from 10-20ms then a mode at $f \simeq 20\,\mathrm{kHz}$ from 40-50ms will occur (with e.g.: 50% probability)".

7 Acknowledgements

The authors thank the H-1 team for continued support of experimental operations. 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 Grant DP0344361 and DP0451960.

References

[1] S. M. Hamberger, B. D. Blackwell, L. E. Sharp and D. B. Shenton, H-1 design and construction. Fusion Technol. 17 (1990) 123–130

[2] J. H. Harris et al, Fluctuations and stability of plasmas in the H-1NF heliac. Nucl. Fusion 44 (2004) 279–286

[3] B. D. Blackwell, Results from helical axis stellarators, Phys. Plasmas 8 (2001) 2238–2244

[4] T. Dudok de Wit, A.-L. Pecquet, J.-C. Vallet and R. Lima, The biorthogonal decomposition as a tool for investigating fluctuations in plasmas. Phys. Plasmas. 1 (1994) 3288–3300

[5] A. Dempster, N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. 39 (1977) 1–38

[6] W. H. E. Day and H. Edelsbrunner, Efficient Algorithms for Agglomerative Hierarchical Clustering Methods, J. Classification 1 (1984) 7–24

[7] D. G. Pretty, PhD Thesis, Australian National University (2007)