A novel unsupervised machine learning algorithm for automatic Alfvénic activity detection in the TJ-II stellarator

Enrique de Dios Zapata Cornejo, David Zarzoso, S.D. Pinches, Andres Bustos, Alvaro Cappa, Enrique Ascasibar

A novel unsupervised machine learning algorithm for automatic Alfvénic activity detection in the TJ-II stellarator

E. Zapata Cornejo[1,2],D. Zarzoso[1], S. D. Pinches[2], A. Bustos[3], A. Cappa[4], E. Ascasibar[4]

[1] Aix-Marseille University, CNRS, Centrale Marseille, M2P2 UMR 7340, Marseille
[2] ITER Organization, Route de Vinon-sur-Verdon, CS 90 046, 13067 St. Paul Lez Durance Cedex, France.
[3] Department of Technology-CIEMAT, Av. Complutense 40, 28040 Madrid, Spain
[4] National Fusion Laboratory—CIEMAT, Av. Complutense 40, 28040 Madrid, Spain

Abstract. A novel sparse encoding algorithm is developed to detect and study plasma instabilities automatically. This algorithm, called Elastic Random Mode Decomposition, is applied to the Mirnov coil signals of a dataset of 1291 discharges of the TJ-II stellarator, enabling the identification of the Alfvénic activity. In the presented approach, each signal is encoded as a collection of basic waveforms called atoms, drawn from a signal's dictionary. Then the modes are identified using clustering and correlations with other plasma signals. The performance of the proposed algorithm is dramatically increased by using elastic net regularization and taking advantage of GPU architectures, hence the signal size and the number of dictionary elements are no longer limiting factors for encoding complex signals. Once the modes are retrieved from the shots, they can be easily analyzed with standard clustering techniques, thereby describing the physical mode characteristics of this subset of TJ-II shots. The clustering features consider the relationship with the plasma current I_p, the diamagnetic energy W, and inverse squared root electronic density $1/\sqrt{n}$, profiling different subtypes of Alfvénic activity. The proposed algorithm can potentially create large databases of labeled modes with unprecedented detail.

Keywords: MHD instabilities, Alfvénic activity, unsupervised learning, TJ-II, clustering, sparse coding, elastic net, random dictionaries, Gabor atoms

1. Introduction

Achieving self-sustained and controlled nuclear fusion reactions for future energy production is one of the most important and challenging problems to be solved in the years to come. Both physics and engineering constraints set the operational limits of future fusion reactors. From the physical point of view, fulfilling the ignition condition while maintaining the plasma equilibrium is not the only requisite. The extreme conditions inside fusion plasmas create very strong gradients of temperature, density, and particle distribution functions. When resonance conditions are satisfied, these gradients act as sources of free energy that can drive instabilities leading to a subsequent reduction of the plasma confinement. In this way, instabilities can limit the performance of current and future fusion devices. For this reason, understanding the physics of plasma instabilities is a matter of key importance in fusion, which requires improving the diagnostic analyses, detection and plasma control. Therefore, machine learning (ML) and artificial intelligence (AI) algorithms are potential candidates to open up new possibilities for improved instability detection and control, as they are revolutionizing how data is used across many fields.

The plasma instability's effect on the magnetic field can be understood as a perturbation $\delta\mathbf{B}$ added to the equilibrium configuration, whose amplitude increases over time. This perturbation can be fully characterized if we define it as a superposition of Fourier modes $\delta\mathbf{B}(\rho,\theta,\varphi,t) = \sum_{\mathrm{m,n}} \delta\mathbf{B}_{\mathrm{mn}}(\rho)e^{\gamma t}\cos(\mathrm{m}\vartheta + \mathrm{n}\varphi - 2\pi f t)$, where m and n are the poloidal and toroidal mode numbers, associated to plasma coordinate angles ϑ and φ respectively. The perturbation depends on the radial position ρ. Furthermore, the mode can evolve in time with a growth rate γ and a frequency f. Therefore, among all possibilities, instabilities admit wavelike behavior. Hannes Alfvén proposed the existence of field-fluid coupled magnetohydrodynamics waves in 1942 [1]. There are the underlying waves of many instabilities in plasmas and it is well known, from their dispersion relation $\omega = k_{||}v_A$, that their frequency is proportional to the magnetic field, and inversely proportional to the squared root of the plasma density $f \propto B/\sqrt{n}$.

The Alfvén wave-instabilities are extensively studied because energetic particles (EPs) can transfer energy to plasma waves. Gradients in EP distribution come either from external heating, like Ion Cyclotron Resonance Heating (ICRH) or Neutral Beam Injection (NBI), or from the nuclear fusion reactions, like alpha particles. With the energy coming from EP, the instabilities can increase their amplitude (positive γ), causing thereby particle and energy losses or eventually wall damages or disruptions. It is well known that multiple physical mechanisms can cause Alfvén instabilities, so the family of observed Alfvén waves in tokamaks and stellarators is quite numerous. Some of them, that can be observed in TJ-II are TAEs (toroidicity-induced Alfvén eigenmodes), HAE (helicity-induced Alfvén eigenmodes), and BAE (beta-induced Alfvén eigenmodes).

Any plasma oscillation translates into a magnetic signal that can measured using magnetic diagnostics, like the Mirnov coils, and they can be further studied using spectral analysis algorithms. However, identifying and labeling the Alfvénic modes in the spectrograms remains in practice a manual task that must be supervised by experts in the field. Therefore, fast and robust instability detection algorithms can be quite useful for both plasma control and data mining of massive datasets. In the case that labels are available in the dataset then an algorithm can be trained to predict these class labels automatically. Once the algorithm has good accuracy in classifying data observations according to these labels, it can be employed in the detection of one of the classes. In consequence, this type of training is known as supervised learning, because it requires solved examples.

To this date and to the best of our knowledge, several works have approached the detection of EP modes using supervised machine learning with data from different tokamaks: COMPASS [2], NSTX [3], and DIIID [4–7]. All of these works mainly use models that are trained on labeled spectrogram images, with various neural network architectures. The previous work from Bustos et al. [8] on TJ-II stellarator, used supervised deep learning to extract modes automatically from Mirnov coils signals spectrograms. They carefully annotated modes manually on ∼500 spectrograms to create a training dataset. The trained model was capable of segmenting the spectrogram images, recovering 989 modes on the full dataset. These works motivated us to research an unsupervised labeling technique of mode signals. Because deep learning networks can potentially boost their performance by employing more training data [9], the capability to create a bigger labeled database of modes with intershot analysis is crucial to generate enough data to train real-time accurate models.

This paper is structured as follows. In section 2, the possibility of using unsupervised learning as a new signal representation is discussed. Afterward, the TJ-II's diagnostic data used in this work is described in section 3. Section 4 is devoted to describing the new algorithm proposed to label the Alfvén modes and to explaining why it is a step forward in terms of speed and performance. In section 5, we describe the results obtained by applying the algorithm to a dataset of 1291 TJ-II shots, using statistics of mode distributions and clustering of the results to demonstrate the algorithm's performance. Finally, future work and conclusions are presented in section 6.

2. Unsupervised learning for Alfvén modes detection

In classification problems, when labels are not available, we can still try to find patterns in data and predict those patterns. For example, clustering algorithms applied to a dataset of experimental observations could be used to label the data automatically. Although expert knowledge is always needed to interpret the cluster grouping.

In the case of most plasma signals, the data is collected as time series. The approach of this work for clustering time series is based on the observation that finding modes in a signal is equivalent to creating a new signal representation in which only the modes are presented. It is important to mention that, previous work [10–14] used unsupervised learning to mine plasma signals but don't specifically address the problem on labeling Alfvénic activity or use sparse coding.

Signal representations can be obtained by a projection over a basis. Consequently, algorithms chose different base expansions and different projection recipes. Therefore, they have various advantages and disadvantages depending on the signal's nature. Paraphrasing Mallat on [15], page 1; "*there is not an ideal transform adapted to all signals*". For instance, the Fourier transform is a linear projection of a signal $S(t)$ over a basis of sines and cosines, but it loses all the time information. The short-time fast Fourier transform (STFFT) allows us to calculate time-frequency representations (spectrograms); it offers computational efficiency, but it has difficulties in analyzing quick transients. Moreover, none of these time-frequency representations automatically identify the mode components of the signal.

We aim to find a representation that keeps the mode features but discards undesired features like noise, facilitating the performance of the clustering algorithm. We assume the following *ansatz*: in which the signal recovered on a pickup coil can be written as a linear combination of terms of the form $s_i = W(t-\tau_i)\cos(2\pi f_i t + \phi_i)$, where $W(t-\tau_i) = \exp\left(-\frac{(t-\tau_i)^2}{2\sigma^2}\right)$ is a Gaussian time window function of fixed standard deviation σ, which modulates the amplitude. The $\cos(2\pi f_i t + \phi_i)$ reconstructs the oscillatory structure. These terms are usually referred to as *atoms*. The linear combination of all these atoms can be constructed by setting a vector of coefficients c_i that we have to fit to characterize the reconstruction of the full signal

$$\boldsymbol{s} \approx \sum_i c_i \cdot s_i = \sum_i c_i \cdot \exp\left(-\frac{(\boldsymbol{t}-\tau_i)^2}{2\sigma^2}\right)\cos(2\pi f_i t + \phi_i). \tag{1}$$

When the selected window function $W(t-\tau_i)$ is Gaussian, the waveform in Eq.(1) is also known as the real part of a Gabor's atom. Then, selecting multiple parameters (τ_i, f_i, ϕ_i) will characterize different time oscillations which can represent parts of the signal. It's worth mentioning the possibility of changing the atoms' Eq.1. For instance, a first Gaussian derivative function could be used to detect sawteeth or pellet injections. In addition, the ansatz could be extended to model mode numbers $W(t-\tau_i)\cos(m\theta_j + n\phi_j - 2\pi f_i t)$ using multiple Mirnov signals, although it would multiply the amount of data to be processed as the phase difference between various Mirnov coil signals is needed.

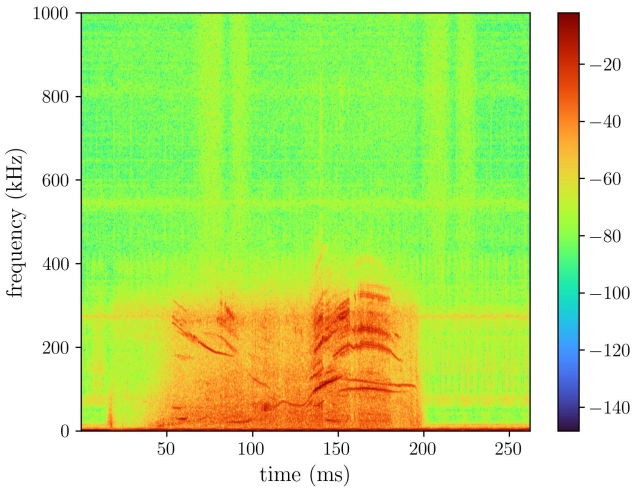

Figure 1. Spectrogram of shot 38399, using MIR5C coil signal sampled at 2MHz, having complex mode activity between 0 and 400 kHz.

This collection of atoms composes a signal, as words in a dictionary represent sentences. The terms *dictionary* and *atoms* were proposed by Mallat and Zhang [16] in 1993. The *dictionary* $\mathbf{D}$ is usually stored in matrix form. Each column vector of $\mathbf{D}$ represents an atom's time series. Therefore the matrix product $\mathbf{D}\cdot\boldsymbol{c}$ will reconstruct the signal $\mathbf{D}\cdot\boldsymbol{c} = \boldsymbol{s}$. For this reason, the vector of weights $\boldsymbol{c}$ is also known as *code*. If the sparsity of the basis vector $\boldsymbol{c}$ is promoted during the optimization, it is possible to cluster neighbor atoms according to the parameters (τ_i, f_i). Then, the time-frequency groups determined by the cluster will be the most significant components of the signal, i.e. the plasma modes.

3. Data

The signals used in this study: magnetic field variation, electronic density, plasma current, and magnetic energy, were obtained from magnetic sensors and a microwave interferometer. Specifically, the variation in the magnetic field was registered in one of the Mirnov coils of the straight array, known as MIR5C. This coil captures the voltage induced by the magnetic field fluctuations. Notably, the spectrograms generated using MIR5C are exceptionally clear, because of its position inside the vessel, very close to the plasma. Fig.1 illustrates the spectrogram of a typical shot observed in TJ-II.

The plasma current is measured in kA using a Rogowski coil. This current is the contribution of the bootstrap current, the currents induced by one or both NBIs and the current driven by oblique ECRH injection (ECCD). Moreover, due to the proximity of the TJ-II's main field conductors to the plasma, an unwanted Ohmic contribution

due to the ripple in the currents flowing through the conductors induces current fluctuations in the plasma [17]. A diamagnetic loop measures the energy stored in the plasma. It is worth noting that the diamagnetic loop is also affected by the plasma current, part of the current fluctuations will be present in the energy's time traces.
The microwave interferometer is important in measuring the plasma density along a specific line of sight. The integrated line approximately crosses the magnetic axis with an inclination of 18.7° with the vertical. This measurement, which also is involved in the machine's operation, will be used in this work to find correlations with frequency.
The sampling frequencies for the different measurements vary. The Mirnov coil is sampled most frequently, typically at 1 MHz, whereas the magnetic energy is collected at the least frequency (≈ 10 kHz). Therefore, linear interpolation is needed to calculate relationships between variables.

The stellarator has two heating systems: two NBI (Neutral Bean Injection) units, and two ECRH (Electron Cyclotron Resonant Heating) beams. Fig.1 shows the spectrogram of magnetic fluctuations registered by the MIR5c pickup coil in a NBI heated plasma (no ECRH in this case). When both NBI's are launched consecutively, a complex mode activity appears between 0 and 400 kHz. Not all modes that can be observed are necessarily Alfvénic. The Alfvenic activity in TJ-II has been widely observed in the frequency range of 50kHz to 400kHz [18–26] along with other types of instability.

4. The Elastic Random Mode Decomposition algorithm (ERMD)

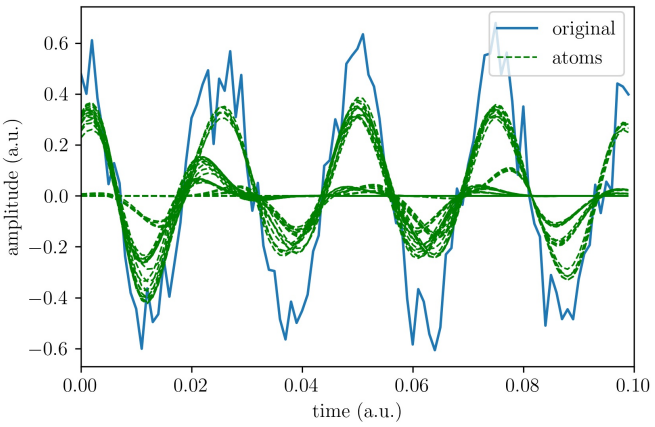

Figure 2. Examples of atoms that minimize reconstruction error, corresponding code values c_i will be significantly greater than zero.

To construct the dictionary matrix $\mathbf{D}$, we can use Gabor's atom as in Eq.(1), assuming a uniform distribution of parameters (τ_i, f_i, ϕ_i) time, frequency, and phase. By taking random samples of these distributions we can assign an atom to each column of the matrix $\mathbf{D}_{:,i}$. Then, each atom ($\mathbf{D}_{:,i}$ column) constitutes a time series's feature. It should be noted that the number of atoms in the basis is not predetermined, and the variance of the parameters (τ_i, f_i, ϕ_i) has to be enough to represent the signal accurately. This method for constructing the dictionary is derived from an algorithm called Sparse Random Mode Decomposition (SRMD), proposed by Richardson et al. [27]. The complete algorithm is available as open-source code. In Figs. 2-3-4 an application of the algorithm on a synthetic signal is shown. The atoms in Fig.2 have a similar frequency and phase to the original signal, the code values will adjust to set a proper amplitude for each atom to represent the sign. On the other side, atoms in Fig.3 cannot represent the local features of the time signal, SRMD will adjust their amplitudes to be close to zero. In Fig.4 we can see that very few atoms can reconstruct the right frequency, and increasing the number of atoms improves the capability of representation of the signal's features.

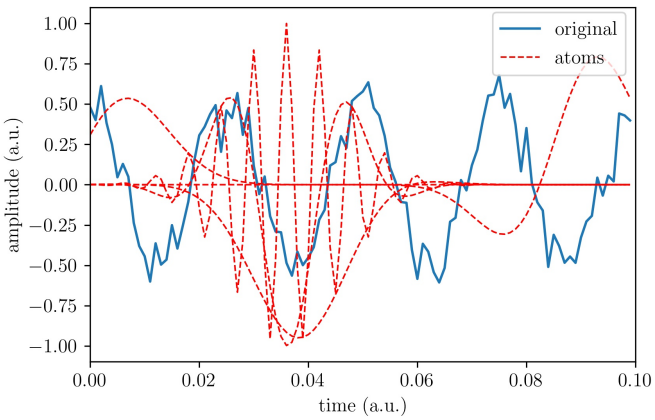

Figure 3. Examples of atoms that don't minimize reconstruction error, respective code values c_i need to be close to zero.

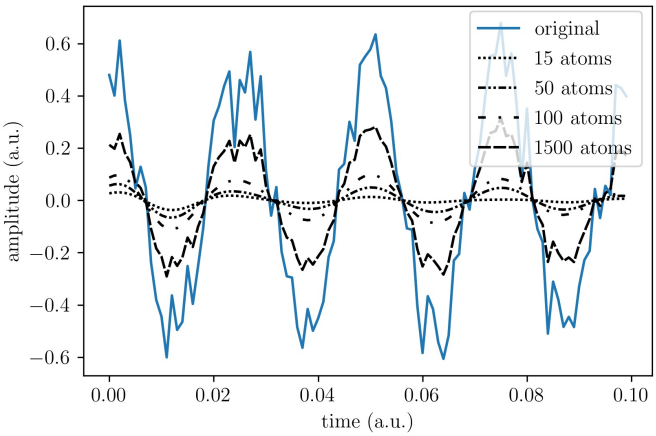

Figure 4. Examples of reconstruction improvement by adding atoms to the code.

The motivation for finding a sparse code for the dictionary is not only algebraic. Indeed, a closer look at the spectrograms with mode activity reveals that MHD mode

signals are very sparse in the frequency domain. Imposing sparsity in the dictionary representation will help to encode and cluster the MHD activity. Finding a sparse code $\boldsymbol{c}$ such that $\mathbf{D}\cdot\boldsymbol{c}\approx\boldsymbol{s}$ constitutes an optimization problem. In Appendix A, we explain in detail the different optimization options and their properties. It's worth mentioning that sparsity is an important regularization in machine learning, and it is gaining attention in plasma physics [28] with applications as different as calculating mode numbers [29] or optimizing stellarators [30].

In [27], a Basis Pursuit De-Noised optimizer (BPDN) was used the obtain the code vector; the optimization problem is formulated as follows: $\boldsymbol{c}=\{\arg\min_{\boldsymbol{c}}\{||\boldsymbol{c}||_1\}$ subject to $||\mathbf{D}\cdot\boldsymbol{c}-\mathbf{s}||_2^2\leq\sigma\}$. The norm $||\boldsymbol{c}||_1$ is responsible for promoting sparsity in the code $\boldsymbol{c}$ [31]. However, the optimizers that solve BPDN require the allocation of the full dictionary matrix $\mathbf{D}$. This difficult the analysis of Mirnov signals for complete shots, which usually comprise several million samples. To solve this problem, we propose the use of a linear neural network to encode the atoms dictionary using the *elastic net* regularization [31, 32] in the loss function.

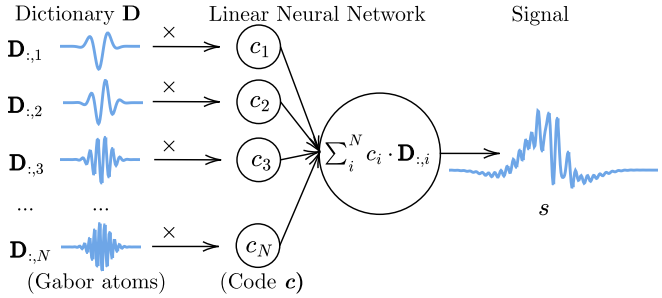

Figure 5. Diagram of the proposed neural network architecture: dictionary columns $\mathbf{D}_{:,i}$ are inputs of the neural network with weights c_i.

We can use a single-layer neural network with linear activation function $\boldsymbol{y}=\mathbf{X}\cdot\boldsymbol{w}+\boldsymbol{b}$, and $\boldsymbol{b}=\mathbf{0}$. We train the network in such a way that the training is equivalent to solving our signal representation $\boldsymbol{s}=\mathbf{D}\cdot\boldsymbol{c}$. In other words, the neural network architecture is equivalent to the decoder equation

$$\begin{pmatrix}\text{Decoder}\\ \text{equation}\end{pmatrix}\quad \boldsymbol{s}=\mathbf{D}\cdot\boldsymbol{c}, \tag{2}$$

and training the neural network is equivalent to solving the optimization problem of the encoder equation

$$\begin{pmatrix}\text{Encoder}\\ \text{equation}\end{pmatrix}\quad \boldsymbol{c}=\arg\min_{\boldsymbol{c}}\left\{||\mathbf{D}\cdot\boldsymbol{c}-\mathbf{s}||_2^2+\lambda_1||\boldsymbol{c}||_1+\lambda_2||\boldsymbol{c}||_2\right\}. \tag{3}$$

It is worth noting the large asymmetry between coding and decoding. Because the decoder equation is linear, whereas the encoder equation is nonlinear. Fig.5 illustrates schematically the architecture of the employed neural network. Even though the activation function is linear, nonlinearity is present in the Gabor's atom, which acts as a nonlinear kernel. The deepness is replaced by the large length of the input layer. In this work, PyTorch [33] with Lightning modules have been used to implement this model. By using this neural network, we can benefit from parallel computing on GPUs.

Note that the encoder equation is different than the BPDN formulation, the loss function inside the encoder equation is known as *elastic net* [31]. There are two main differences though. First, the problem formulation is unconstrained (it does not need a noise parameter σ). Second, there are two regularization terms λ_1 and λ_2 associated with the modules of the code $\boldsymbol{c}$ in ℓ_1 norm $||\boldsymbol{c}||_1$ and ℓ_2 norm $||\boldsymbol{c}||_2$ respectively, while in BPDN only the norm ℓ_1 is used. The norm $||\boldsymbol{c}||_2$ is responsible for promoting the presence of correlated groups in the code $\boldsymbol{c}$ [31].

In this work, the use of elastic net is proposed based on a fundamental observation: the random atoms used in the construction of the random dictionary can be correlated. This idea is illustrated in Fig.6. It can be seen that atoms with the same frequency and phase but different amplitude are highly correlated, while atoms out of phase are not correlated at all. Most importantly, atoms with connecting tails in phase have some correlation. The improvement of adding ℓ_1 regularization can be appreciated by comparing Fig.7 and Fig.8. It is important to emphasize that the atom's correlations can be exploited to promote the time continuity of the encoded modes, which is an essential physical feature of the signal. More details on optimization formulations are explained in Appendix A.

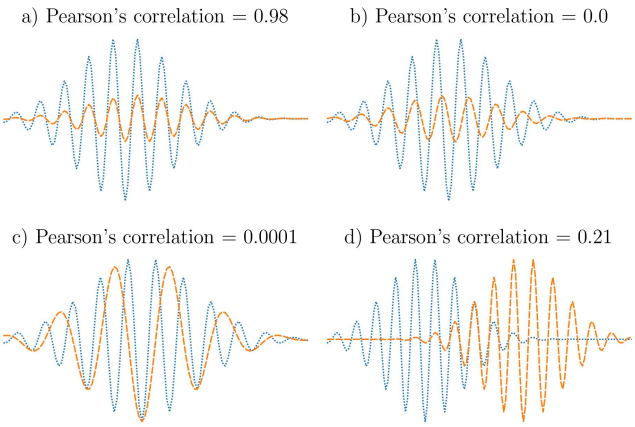

Figure 6. Varying correlations between different pairs of atoms: Overlapping needed to reconstruct the signal can correlate.

Two major improvements in memory usage have been implemented. First, it is possible to use the atom formula given by Eq.1 and the random number generator seed to allocate dictionary atoms only when a signal's batch is allocated in the GPU. This is a breakthrough in memory optimization; we can refer to this method as *in-GPU* dictionary construction. Second, we can adapt the gradient descent to optimize the weights of the full signal while working by batches. This technique is known as *gradient accumulation*. The gradient update is worked out for each batch until a given number of batches is reached, then all updates are combined indicating the direction of the gradient descent. This method works assuming that the

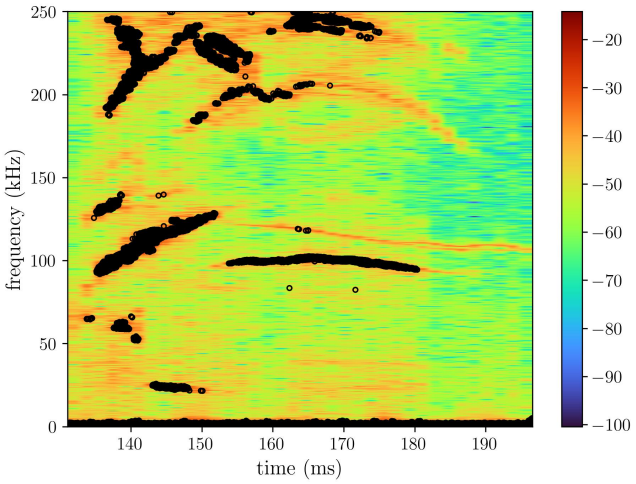

Figure 7. Result of ℓ_1 encoding: setting $\lambda_1 \neq 0$ and $\lambda_2 = 0$ in Elastic Net regularization loss.

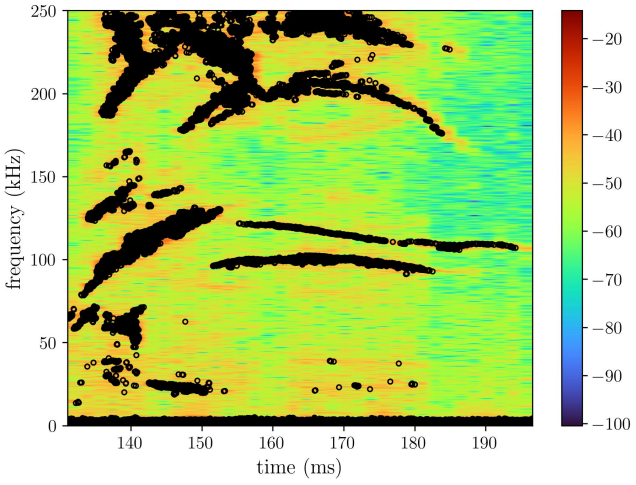

Figure 8. Result of mixed ℓ_1 and ℓ_2 encoding: setting $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$ in Elastic Net regularization loss.

averaged gradient direction of all batches considered is in the right direction. Gradient accumulation allows the encoding of high-resolution signals, and to the best of our knowledge, it is used to encode signals for the first time in this work. It is worth mentioning that we used PyTorch's Adam optimizer with its default parameters.

Finally, a strong scaling test has been carried out to quantify the speed-up of the proposed encoding algorithm. The result can be seen in Figs. 9.a-9.b. The test was carried out using 8 NVIDIA Tesla P100 GPUs, each one with 16384 MB of memory. Execution times are averaged over 5 epochs, the error bars cannot be appreciated in the figure. The result shows that the time execution scales linearly with the number of atoms in the dictionary, as is evidenced in Fig.9. Also, the speed-up saturates over 8 GPUs due to data communication, as reported in Fig.9. Therefore, an epoch can be completed in less than 5 minutes. On the contrary, the encoding can be done on the CPU with the SPGL1 algorithm [34, 35], dividing the signal into 10-15 chunks, running times take from 40 to 60 minutes, and chunks codes need to be re-weighted at the end. If we compare the performance with SPGL1's running times, the improvement achieved using the neural network is remarkable. More details about the optimization algorithms can be found in Appendix A.

It is to be noted that the elastic net regularization has been used before for images in dictionary learning algorithms (when code and dictionary need to be learned simultaneously [36, 37]). And recent works accelerated image encoding using a linear neural network [38]. In addition, there are some use precedents with deterministic Gabor's dictionaries [39][40] but not with dictionaries of random features. The use of elastic net regularization in these publications consistently improves performance, when identifying correlated atoms (modes) is needed.

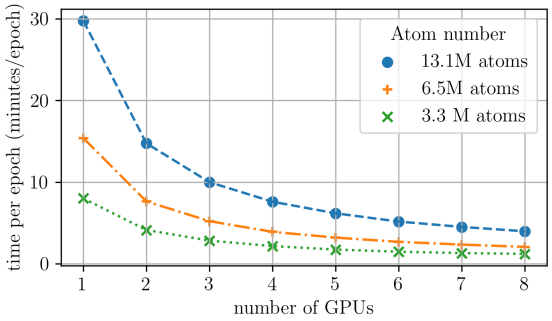

Figure 9. a Strong Scaling Benchmark. Each experiment has been repeated 5 times, error bars can not be appreciated.

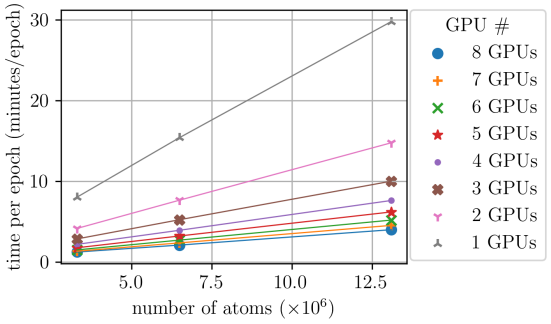

Figure 9. b Linear scaling with number of atoms. Each experiment has been repeated 5 times, error bars can not be appreciated.

The different atoms are grouped using an unsupervised cluster algorithm. The resulting clusters are identified as modes. In addition, improvements have been made in the clustering algorithm. The clustering algorithms based on density are adequate for our application as encoded atoms tend to concentrate on the time-frequency mode structures, and they are equipped to handle noise points and non-linearly separable clusters of arbitrary shapes and sizes. The clustering algorithm initially proposed in [27] for

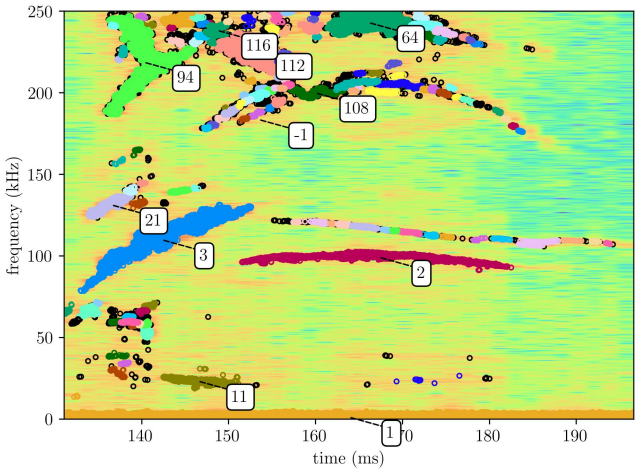

Figure 10. Cluster result using HDBSCAN (Annotated the centers of 10 largest clusters and noise (-1).)

grouping the atoms in modes is DBSCAN [41]. However, DBSCAN has two limitations. First, the most sensitive hyperparameter ϵ, which controls the maximum neighbor distance, varies between shots. Second, DBSCAN assumes constant density clusters, which is not true for the results of encoding. In contrast, hierarchical DBSCAN (HDBSCAN) [42] does not assume constant density and is capable of determining the varying density structure. Moreover, HDBSCAN does not have a sensitive hyperparameter to modify on each shot. For these reasons, HDBSCAN is more adequate than DBSCAN. An example of an HDBSCAN's result is shown in Fig.10.

We propose the adoption of the name *Elastic Random Mode Decomposition* (ERMD) to refer to our new algorithm. ERMD is summarized as follows:

Algorithm 1 *Elastic Random Mode Decomposition* (ERMD) for Alfvénic mode classification

INPUTS: signal $\boldsymbol{s}$, time $\boldsymbol{t}$.
OUTPUT: code $\boldsymbol{c}$, best model code $\boldsymbol{c}_{\text{best}}$, atom's time position $\boldsymbol{\tau}$, frequencies $\boldsymbol{f}$, phases $\boldsymbol{\phi}$, cluster labels $\boldsymbol{l}$ and Alfvénic character $\boldsymbol{a}$. All with the same vector length.
PARAMETERS: $\{\lambda_1,\ \lambda_2,\ \sigma,\ N_{max},\ N_f,\ B_s,\ \eta,\ max_{epoch},\ Q,\ Q_c\ \text{min}_{size},\ \rho_{thr}.\ \}$ See table (1).
STEPS:

(i) Standardize signal $\frac{\boldsymbol{s}-\mu(\boldsymbol{s})}{\sigma(\boldsymbol{s})}$ and time $\frac{\boldsymbol{t}-\mu(\boldsymbol{t})}{\sigma(\boldsymbol{t})}$ arrays.

(ii) Apply high-pass filter (Optional: it can help to detect high-frequency modes)

(iii) Set a global seed for all GPUs, it is necessary because sampling features space $\{\tau_i,\ f_i,\ \phi_i\}$ has to be deterministic.

(iv) Set the number of atoms in the random dictionary N as a multiple of signal length $\boldsymbol{s}_l$. $N = N_f \times \text{lenght}(\boldsymbol{s})$. It is useful to set a maximum number of atoms N_{max} to avoid out-of-memory problems.

(v) Repeat in parallel until reach max_{epoch}:

(a) For each batch of time signals, load the data (segment of size B_s) to an assigned GPU and compute a segment of matrix $\mathbf{D}_{\text{batch}}$ in the assigned GPU. Use seed and atom equations to calculate the value of corresponding atoms in $\mathbf{D}_{\text{batch}}$ (in-GPU dictionary construction)

$$\mathbf{D}_{\text{batch}} = [\boldsymbol{s}_i(\boldsymbol{t}_{\text{batch}})] = \left[exp\left(-\frac{(\boldsymbol{t}_{\text{batch}}-\tau_i)^2}{2\sigma}\right)\cdot\cos\left(2\pi f_i \boldsymbol{t}_{\text{batch}}+\phi_i\right)\right]. \tag{4}$$

(b) Forward pass for each batch: Calculate elastic net regression loss

$$\mathcal{L}_{\text{batch}} = ||\mathbf{D}_{\text{batch}}\cdot\boldsymbol{c}-\boldsymbol{s}_{\text{batch}}||_2^2+\lambda_1||\boldsymbol{c}||_1+\lambda_2||\boldsymbol{c}||_2, \tag{5}$$

then calculate each batch gradient $g_{batch}(\boldsymbol{c})$ and keep it.

(c) When the number of batches processed completes the epoch: accumulate gradient (the sum of all batches gradients determine the descend direction), and actualize code weights.

$$\boldsymbol{c}_{\text{new}} = \boldsymbol{c}_{\text{old}} - \eta \sum_{\text{batches}} g_{\text{batch}}(\boldsymbol{c}) \tag{6}$$

(d) Keep the best code $\boldsymbol{c}_{\text{best}}$: model with best loss $\sum\mathcal{L}_{\text{batch}}$ so far.

(vi) Return $\boldsymbol{c}$, $\boldsymbol{c}_{\text{best}}$, $\boldsymbol{\tau}$, $\boldsymbol{f}$, $\boldsymbol{\phi}$.

(vii) Code thresholding: group atoms code values $\boldsymbol{c}_{\text{best}}$ by quantiles, using the specified (number of quantiles Q) and keep those above the (quantile cut Q_c).

(viii) Cluster modes using HDBSCAN with desired Min. cluster size min_{size}. Return $\boldsymbol{l}$.

(ix) For each identified mode, except noise clusters (-1), there are two options:

(a) Calculate Pearson's correlation between mode frequencies and $1/\sqrt{n}$. If the correlation is greater than the correlation cut ρ_{thr}, label the mode as 1 (Alfvénic) otherwise 0 (No Alfvénic). Return $\boldsymbol{a}$.

(b) Use mutual information (MI) instead, this can help to retrieve more modes. The details can be found in Appendix B. Return $\boldsymbol{a}$.

(x) Store result $\boldsymbol{c}$, $\boldsymbol{c}_{\text{best}}$, $\boldsymbol{\tau}$, $\boldsymbol{f}$, $\boldsymbol{\phi}$, $\boldsymbol{l}$, $\boldsymbol{a}$.

The relevant hyperparameters and used values are summarized in the table 1. Among them, the most important parameters are the regularization terms λ_1 and λ_2 which adjust the promotion of sparsity and regularization. The values of λ_1 and λ_2 depend on the signal-to-noise ratio and they have an important effect on the mode retrieval. Unfortunately, there is no rule to adjust these parameters as they heavily depend on the data and

the hyperparameter search space has multiple local minima. Using hyperparameter optimization is difficult, as there can be multiple objectives (minimize reconstruction error, minimize the number of modes, or minimize the number of epochs). Instead, it is advisable to first determine λ_1 with $\lambda_2 = 0$ and then add λ_2 as needed. All in all, it can be found values of λ_1 and λ_2 that work reasonably well with all shots. Nevertheless, to obtain the best result, avoiding the loss of modes, or fully eliminating broadband noise, it is necessary to fine-adjust λ_1 and λ_2.

Parameter	Value(s)
λ_1	0.00005
λ_2	0.001
Atom standard deviation σ	256 samples
Maximum number of atoms N_{max}	14×10^6
Feature factor N_f	20
Batch size B_s	32-64 samples
Learning rate η	0.00008
Maximum epochs $\max_{epoch}$	5
Number of quantiles Q	1000
Quantile cut Q_c	999
Minimum cluster size $\min_{size}$	10
Correlation threshold ρ_{thr}	0.9

Table 1. ERMD Hyperparameters

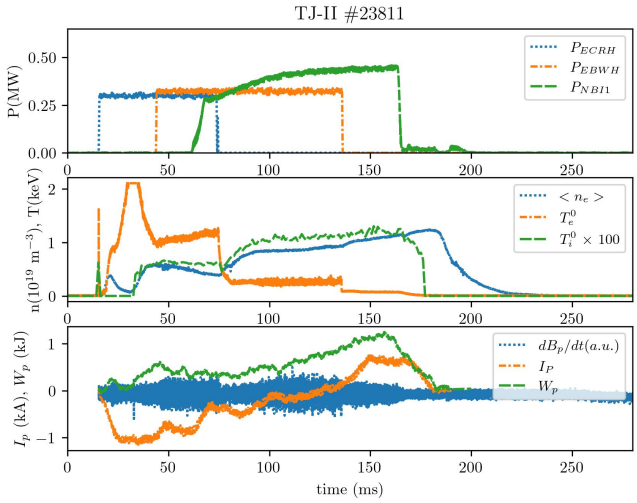

Figure 11. Shot 23811 signals, from top to bottom: heating power, electronic density $< n >$ and temperatures T_e^0, $T_i^0 \times 100$, and the MIR5C signal, I_p, and W_p.

5. Results of unsupervised feature extraction of Alfvén activity

5.1. Performance

To evaluate the performance of the algorithm, we run ERMD on 1000 shots with a fixed set of parameters. The hyperparameters were first selected by trial and error to obtain good performance on a selection of shots. About the execution times, it took 206.8 hours to run the 1000 shots, so this gives an average of 4.8 shots per hour approximately. The shots have different numbers of samples so this number can fluctuate from 3 to 7 shots per hour. In total, we accumulated 1291 shots in the database.

As an example of a typical result, the encoding and clustering of shot 23811 are shown in Fig.12. Three high-frequency modes are correctly identified, with the assigned numbers 1,4, and 6. Some noise atoms are correctly identified as noise by HDBSCAN. Finally, in Fig.13 the two higher-frequency modes are identified as Alfvénic, as they are strongly correlated with the $1/\sqrt{n}$ signal. However, part of the broadband noise structure is misidentified with fragmented low-frequency modes, this behavior is found to

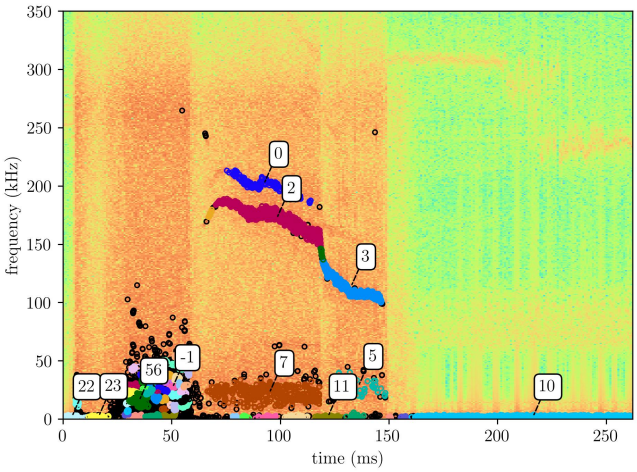

Figure 12. Cluster results of shot 23811. Marked the center of top 10 numerous clusters. Black circles (-1) are noise.

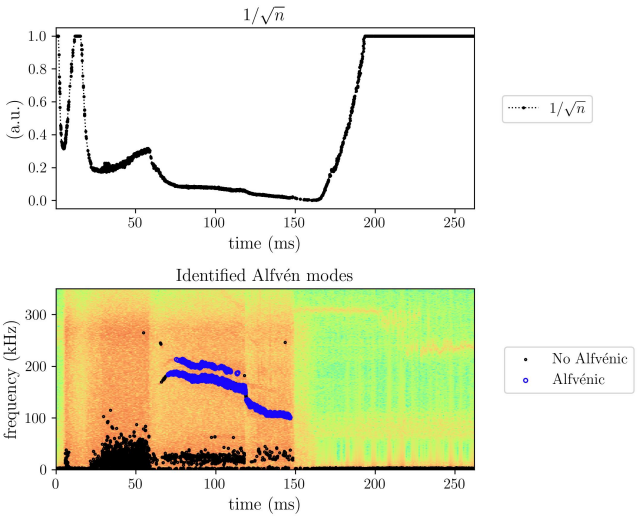

Figure 13. Result of labeling shot 23811: modes identified as Alfvénic (big blue circles) follow $1/\sqrt{n}$.

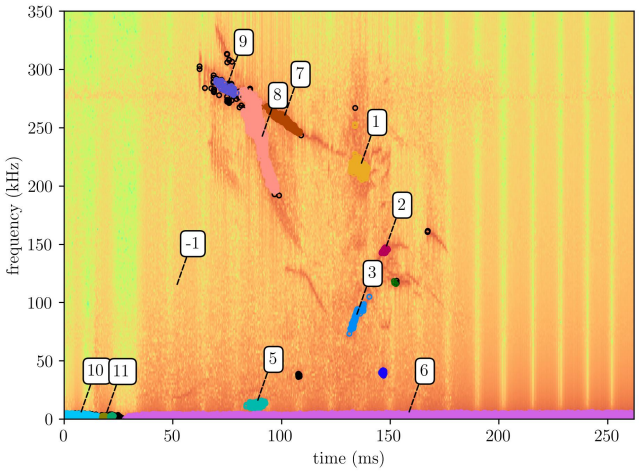

Figure 14. Cluster results of shot 38393: ERMD handles crossing modes (7,8,9). Noise (-1) is distributed between low and high frequencies.

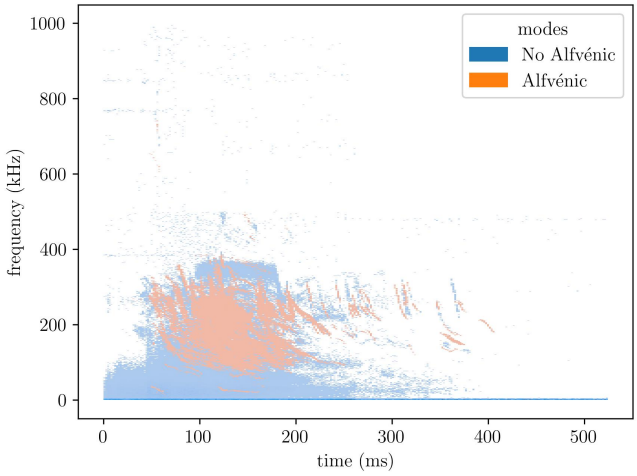

Figure 15. Time-frequency distributions of Alfvénic mode atoms of this set of 1291 shots.

be common but does not affect the labeling of Alfvénic modes if the correlation threshold is set high. It is to be noted, how the amplitude of broadband noise increases when heating systems are activated, this can be seen comparing power inputs in Fig.11 with the spectrogram in Fig.12 and Fig.13.

Because each mode has a different density of atoms, the algorithm is capable of labeling crossing modes, as is shown in Fig.14. This task is very difficult for algorithms based on the segmentation of the spectrogram. Therefore, this is a remarkable advantage of this algorithm which is reinforced with the addition of the density signal $1/\sqrt{n}$ to the analysis. However, there is a trade-off: the density-based clustering also tends to divide one mode into different sub-modes. In consequence, ERMD results might not be reliable to measure mode duration Δt.

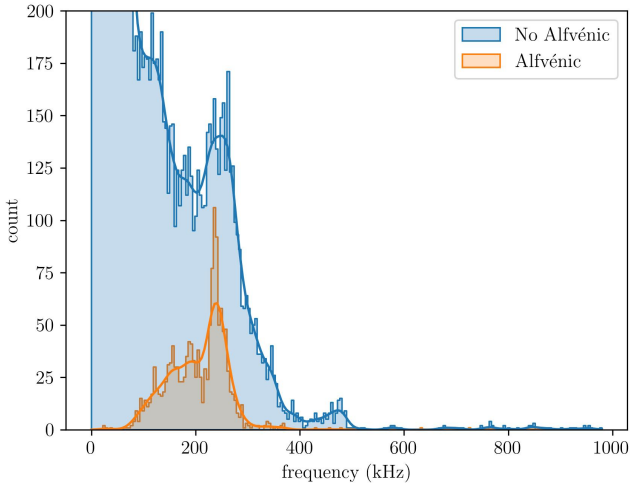

Figure 16. Frequency distribution of Alfvénic mode activity in the analyzed shots.

The frequency distribution of atoms in Alfvén modes is checked for agreement with the frequencies expected from the literature. As can be seen in Fig.15 and Fig.16. Alfvén frequencies range from 100 kHz to 400 kHz, overlapping with the mentioned interval 50kHz to 400kHz [18–26]. Most Alfvénic modes appear between 50 ms and 200 ms, and the majority of modes (80%) in this set of shots are chirping down.

The use of the algorithm is not exempt from difficulties, as some modes can be missed in this encoding (Fig.14). The risk of missing modes is in general low except for two cases: when there are several modes stacked in parallel frequencies or when the signal-to-noise ratio is very low. Because the algorithm is based on minimizing the signal reconstruction error with some regularization, these two errors will persist without fine-tuning the parameters for each shot. The reconstruction error is small in the case of modes with less amplitude, this makes the encoding difficult as other frequencies with greater amplitude are encoded first. When the signal-to-noise ratio is very low, the regularization needs to have very shot-specific values to separate modes from broadband noise, like in some modes of Fig.14.

More difficulties have been found in the presence of sawteeth. These fast transient signals generate broad frequency spectral leakage, as it happens in STFFT. When minimizing the signal reconstruction error, many different atoms can fit these transients. Therefore, the sparsity assumption does not hold anymore and the regularization parameters should be changed to address this difficulty, reducing both λ_1 and λ_2. This phenomenon could be tackled in principle by adding sawteeth atoms to the dictionary.

5.2. Profiling TJ-II modes

Once the atoms representing the Mirnov coil signal have been conveniently extracted and clustered into modes, a

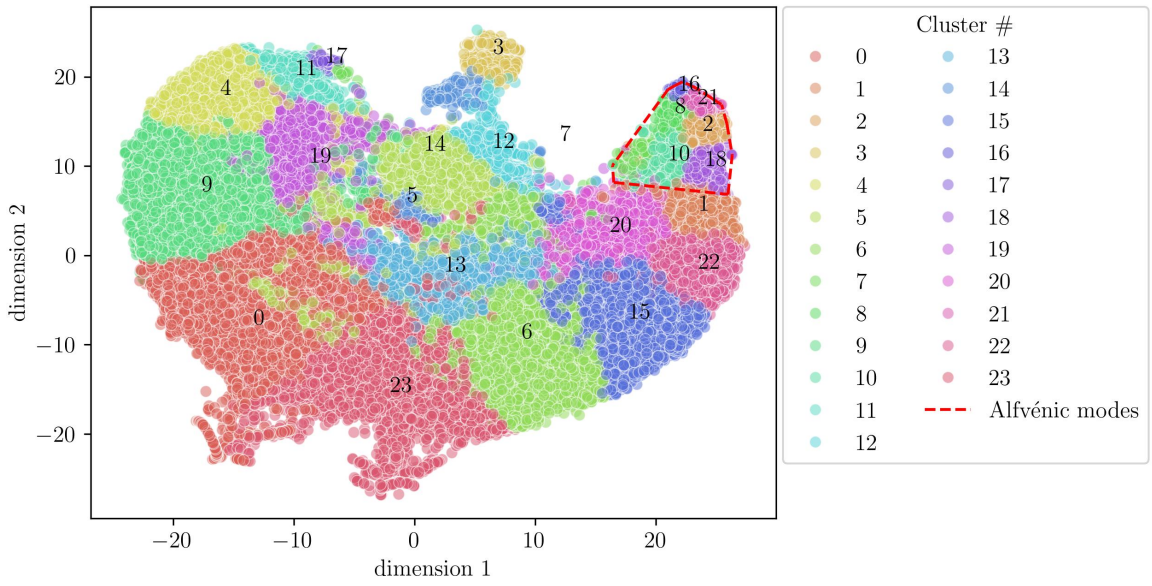

Figure 17. T-SNE embedding of mode features. Clusters 2, 21, 16, 8, 10, and 18 are mostly Alfvénic modes the rest can be noise or a mix of different modes.

second clustering can be applied to extract higher-level patterns. The plasma signals used in addition to the frequency of the modes are the current I_p, the magnetic energy W, and the density $1/\sqrt{n}$. An example of these signals is plotted in Fig.11. Profiling each cluster will help to understand the result of the ERMD run, and the nature of the modes collected. The clustering method selected is K-means, which differs from HDBSCAN. K-means is based on minimizing within-cluster variance, it does not use density like HDBSCAN, and it has great scalability with the number of samples [43].

If we project the modes obtained by ERMD into a lower dimensional embedding, using the *T-distributed Stochastic Neighbor Embedding* (T-SNE) algorithm [44], we can mark the cluster numbers and visualize which groups are similar by their separation distance and neighborhood.

The features used for the projections are the same as the features used for clustering: the mutual information (MI) of the mode's frequencies with the density MI(f,$1/\sqrt{n}$), current MI(f,I_p), and energy of the plasma MI(f,W). In stellarators, and particularly in TJ-II, small changes in plasma current can have a measurable impact on the spectrum of shear Alfvén waves and therefore, evaluating the correlations between changes in mode frequency and plasma current is necessary. In addition, the modes's frequency chirp $\Delta f = f_{\max} - f_{\min}$, the variation of density $\Delta(1/\sqrt{n}) = (1/\sqrt{n})_{\max} - (1/\sqrt{n})_{\min}$, the variation of current $\Delta I_p = I_{p\,\max} - I_{p\,\min}$, and plasma energy $\Delta W = W_{\max} - W_{\min}$ are included to give directionality and magnitude of parameters change.

The result of T-SNE is shown in Fig.17, all discovered modes are represented in a two-dimensional space. The numbers of the labels are positioned in the plot by taking the mean values of dimension 1 and dimension 2 in each group. Finally, the class label that results from the K-means clustering is used to assign a color to the projection points and interpret the projection result.

The interpretation of the cluster analysis is summarized in the table 2. For the sake of clarity we provide some examples of clusters in Fig.18. We can observe that the clustering algorithm is capable of separating the Alfvénic class from other types of modes with the given variables. The region of Alfvénic modes is highlighted with a dashed red line polygon, in the upper right corner of Fig.17. Groups 8, 16, 21, and 2 are mostly composed of Alfvénic modes, and clusters 10, 18, and 7 have a mixed composition, as their position is close to the border of the polygonal region.

The groups neighboring cluster 0 in the opposite region to the Alfvénic class are noise. We consider noise broadband frequency structures mislabeled in some cases by ERMD. In addition, we consider as noise the very low-frequency structures misidentified by modes that can be seen in Fig.12, which also occurs when broadband noise is present. The MI values of these groups are close to 0, as they do not have any relationship with fluctuating plasma signals, which helps discard errors of the algorithm.

If we look at the time-frequency distributions of modes in Fig.18 we can appreciate variate behavior. Note that the density gets very high in some values due to the presence of repeated shots. Moreover, even if the time variable has not been used, the modes appear to be grouped around specific times at the beginning of the pulse when NBI is engaged. We can observe as well that the less Alfvénic activity, the more probable the presence of noise. Though some noise might be present, the identification of Alfvénic activity is satisfactory. In addition, some clusters appear to have different natures, like clusters 7 and 17, in which modes are correlated with the current, in this case, modes appear to have lower frequencies.

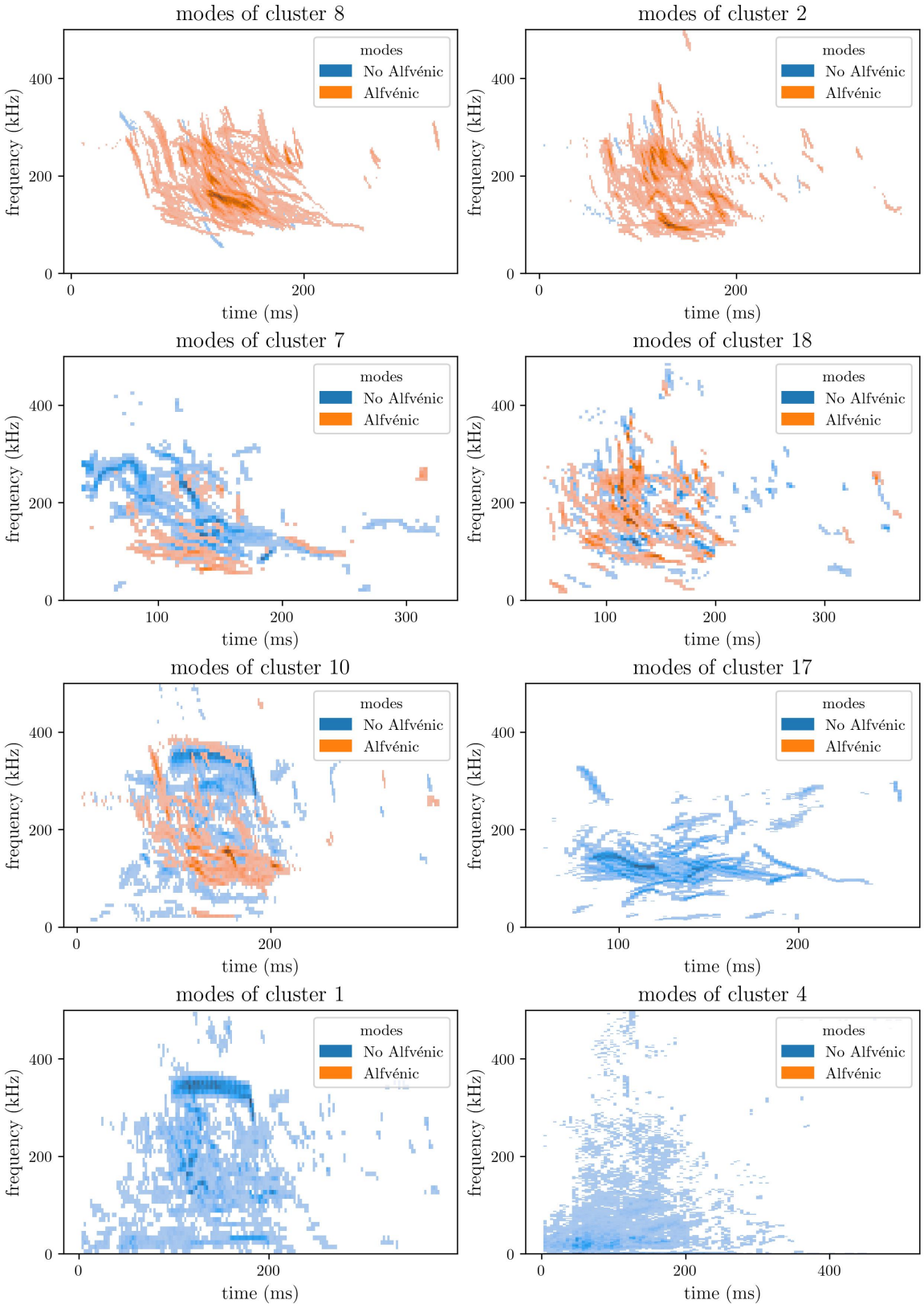

Figure 18. Time-frequency histograms of mode distributions from different clusters.

Cluster Number	Interpretation
8, 16, 21	Strongly driven Alfvénic modes occurring when heating is engaged
2	Alfvénic modes correlated with magnetic energy
7	Mix of Alfvénic and Non Alfvénic modes around 100 kHz correlated with current
18	Mix of Alfvénic and Non Alfvénic modes (probably False negatives)
10	Mix of Alfvénic and Non Alfvénic modes
17	Anti-Alfvénic modes strongly correlated with current
1	Non Alfvénic modes and mid-frequency noise
4-5, 9, 11-13, 19-20, 22-23	Noise + low-frequency modes (No-Alfvénic)
0, 3, 6, 12, 14-15	Broadband and low frequency noise produced by heating systems

Table 2. Clusters interpretation

It is of interest that this clustering technique can be used for looking for patterns in shot databases and eventually finding unexplored plasma activity. However, a detailed physical interpretation of each cluster requires further analysis based on complementary simulations and first physical principles. For instance, now that the presence of modes is detected it is possible to run equilibrium reconstruction and stability analysis to determine the physical nature of the modes. This could add labels like HAE, TAE, GAE or MIAE to the clusters groups. The mode identification problem in TJ-II from the physics point of view has been addressed in [23–25].

It is worth mentioning that the MI information can be replaced by Pearson's correlation or even combined, and the results of the cluster are comparable. Still, we prioritize the use of MI as it is capable of capturing nonlinear relationships, and it is more robust to outliers than Pearson's. The interested reader can find an alternative to clustering in Appendix C. It is possible to create categorical features using MI and constructing contingency tables to analyze the mode's relationship with other plasma signals.

Other mode variables can be created and added to the clustering. However, we follow the parsimony principle, i.e. adding variables until the result is interpretable, and then stop including features. The same principle is followed when selecting the number of clusters, i.e. the number of clusters is increased until the granularity is fine enough to describe Alfvénic subgroups.

6. Conclusions, further applications and future work

In this paper, we successfully illustrated the practical use of sparse coding in plasma instabilities detection. Overall, the results shown using our new proposed algorithm called ERMD are promising. However, it has some practical limitations, mostly the computational cost, and the hyperparameter tuning. Nevertheless, ERMD allows us to work on the time domain with raw signals, it can extract the Alfvénic pattern in a TJ-II sample of shots, and it is capable of outperforming other algorithms in less computing time.

This novel approach required algorithmic innovations as well. The major contributions to the sparse coding field are (1) the use of an in-GPU dictionary construction of Gabor's atoms, (2) the use of accumulated gradient descend for parallelizing the optimization algorithm in multiple GPUs with small signal batches, (3) the improvement of mode identification by using a variable density based clustering HDBSCAN, and (5) the acknowledge of the multicollinearity in random atoms dictionaries by using elastic net regularization to improve mode decomposition.

The use of MI information to capture linear and nonlinear relationships between variables is another innovation in mode detection, that provides more robustness against noise and outliers. Moreover, the clustering technique used for profiling the modes retrieved by ERMD has proven to be useful in summarizing mode activity in the 1291 shots. It complements the ERMD algorithm as it can be used to remove noisy clusters from the ERMD result.

Mirnov signals have a very high sampling rate, up to 2 MHz, which is the main factor slowing down the ERMD runs. Nevertheless, the algorithm can label more than 700 shots in a week, with consistency that cannot be provided by a human team. Though the technique is useful for inter-shot analysis, a ML surrogate model of ERMD could in principle detect Alfvén activity using raw Mirnov coil signals in real-time. Subsequent work might investigate the training of machine learning models using the created database for the detection of the Alfvénic class or the clustering sub-classes. In addition, researching the use of dictionary learning (learning dictionary elements while encoding) opens new possibilities to process plasma signals.

ERMD is adapted to analyze high-resolution, non-stationary time signals in any field. Future applications can be as variate as mode decomposition of heart electrocardiograms, seismic analysis, animal calls, or musical analysis. Owing to the broad application of ML and signal processing, our algorithm could impact other scientific disciplines as well.

Acknowledgments

We thank Pedro Pons-Villalonga (CIEMAT) for his help understanding TJ-II Mirnov coils system. This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200-EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This work has also received financial support from the AIM4EP project (ANR-21-CE30-0018), funded by the

French National Research Agency (ANR). The project that gave rise to this publication benefited from funding from the Aix-Marseille University Initiative of Excellence - A*Midex, Investissements d'Avenir programme AMX-19-IET-013. In addition, the project Horizonte Europa EU-LAC ResInfra Plus, No. 101131703 has contributed to this work. The views and opinions expressed herein do not necessarily reflect those of the ITER Organization.

References

1. Alfvén, H. Existence of Electromagnetic-Hydrodynamic Waves. en. *Nature* **150.** Number: 3805 Publisher: Nature Publishing Group, 405–406. ISSN: 1476-4687. doi:10.1038/150405d0 (Oct. 1942).
2. Škvára, V. *et al.* Detection of Alfvén Eigenmodes on COMPASS with Generative Neural Networks. *Fusion Science and Technology* **76.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15361055.2020.1820805, 962–971. ISSN: 1536-1055. doi:10.1080/15361055.2020.1820805 (Nov. 2020).
3. Woods, B. J. Q. *et al.* Machine Learning Characterization of Alfvénic and Sub-Alfvénic Chirping and Correlation With Fast-Ion Loss at NSTX. *IEEE Transactions on Plasma Science* **48.** Conference Name: IEEE Transactions on Plasma Science, 71–81. ISSN: 1939-9375. doi:10.1109/TPS.2019.2960206 (Jan. 2020).
4. Jalalvand, A. *et al.* Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks. en. *Nuclear Fusion* **62.** Publisher: IOP Publishing, 026007. ISSN: 0029-5515. doi:10.1088/1741-4326/ac3be7 (Dec. 2021).
5. Jalalvand, A. *et al.* Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks. *Nuclear Fusion* **62,** 026007. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/ac3be7 (Feb. 2022).
6. Kaptanoglu, A. A. *et al.* Exploring data-driven models for spatiotemporally local classification of Alfvén eigenmodes. *Nuclear Fusion* **62,** 106014. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/ac8a03 (Oct. 2022).
7. Garcia, A. V. *et al. Alfvén eigenmode detection using Long-Short Term Memory Networks and CO $_2$ Interferometer data on the DIII-D National Fusion Facility* in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Gold Coast, Australia, June 2023), 1–8. ISBN: 978-1-66548-867-9. doi:10.1109/IJCNN54540.2023.10191760.
8. Bustos, A., Ascasíbar, E., Cappa, A. & Mayo-García, R. Automatic identification of MHD modes in magnetic fluctuation spectrograms using deep learning techniques. en. *Plasma Physics and Controlled Fusion* **63.** Publisher: IOP Publishing, 095001. ISSN: 0741-3335. doi:10.1088/1361-6587/ac08f7 (July 2021).
9. Hestness, J. *et al.* Deep Learning Scaling is Predictable, Empirically. Publisher: [object Object] Version Number: 1. doi:10.48550/ARXIV.1712.00409 (2017).
10. Vega, J. *et al.* Overview of intelligent data retrieval methods for waveforms and images in massive fusion databases. en. *Fusion Engineering and Design* **84,** 1916–1919. ISSN: 09203796. doi:10.1016/j.fusengdes.2008.11.097 (June 2009).
11. Pretty, D. & Blackwell, B. A data mining algorithm for automated characterisation of fluctuations in multichannel timeseries. en. *Computer Physics Communications* **180,** 1768–1776. ISSN: 00104655. doi:10.1016/j.cpc.2009.05.003 (Oct. 2009).
12. Yamamoto, S. *et al.* Studies of MHD Stability Using Data Mining Technique in Helical Plasmas. en. *Plasma and Fusion Research* **5,** 034–034. ISSN: 1880-6821. doi:10.1585/pfr.5.034 (2010).
13. Blackwell, B. D. *et al.* Observations of Alfvénic MHD Activity in the H-1 Heliac. en.
14. Haskey, S., Blackwell, B. & Pretty, D. Clustering of periodic multichannel timeseries data with application to plasma fluctuations. en. *Computer Physics Communications* **185,** 1669–1680. ISSN: 00104655. doi:10.1016/j.cpc.2014.03.008 (June 2014).
15. Mallat, S. *A Wavelet Tour of Signal Processing* en. ISBN: 978-0-12-374370-1. doi:10.1016/B978-0-12-374370-1.X0001-8 (Elsevier, 2009).
16. Mallat, S. & Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41.** Conference Name: IEEE Transactions on Signal Processing, 3397–3415. ISSN: 1941-0476. doi:10.1109/78.258082 (Dec. 1993).
17. Mulas, S. *et al.* Validating neutral-beam current drive simulations in the TJ-II stellarator. *Nuclear Fusion* **63,** 066026. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/acca92 (June 2023).
18. Jiménez-Gómez, R. *et al.* Alfvén eigenmodes measured in the TJ-II stellarator. en. *Nuclear Fusion* **51,** 033001. ISSN: 0029-5515. doi:10.1088/0029-5515/51/3/033001 (Feb. 2011).

19. Melnikov, A. V. *et al.* Alfvén eigenmode properties and dynamics in the TJ-II stellarator. en. *Nuclear Fusion* **52.** Publisher: IOP Publishing and International Atomic Energy Agency, 123004. ISSN: 0029-5515. doi:10.1088/0029-5515/52/12/123004 (Nov. 2012).
20. Melnikov, A. *et al.* Effect of magnetic configuration on frequency of NBI-driven Alfvén modes in TJ-II. *Nuclear Fusion* **54,** 123002. ISSN: 0029-5515, 1741-4326. doi:10.1088/0029-5515/54/12/123002 (Dec. 2014).
21. Melnikov, A. *et al.* Detection and investigation of chirping Alfvén eigenmodes with heavy ion beam probe in the TJ-II stellarator. *Nuclear Fusion* **58,** 082019. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/aabcf8 (Aug. 2018).
22. Eliseev, L. G. *et al.* Experimental observation of the geodesic acoustic frequency limit for the NBI-driven Alfvén eigenmodes in TJ-II. en. *Physics of Plasmas* **28,** 072510. ISSN: 1070-664X, 1089-7674. doi:10.1063/5.0049225 (July 2021).
23. Varela, J., Spong, D. & Garcia, L. Analysis of Alfven eigenmodes destabilization by energetic particles in TJ-II using a Landau-closure model. *Nuclear Fusion* **57,** 126019. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/aa83c4 (Dec. 2017).
24. Sun, B., Ochando, M. & López-Bruna, D. Alfvén eigenmodes including magnetic island effects in the TJ-II stellarator. *Nuclear Fusion* **55,** 093023. ISSN: 0029-5515, 1741-4326. doi:10.1088/0029-5515/55/9/093023 (Sept. 2015).
25. Cappa, Á. *et al.* Stability analysis of TJ-II stellarator NBI driven Alfvén eigenmodes in ECRH and ECCD experiments. *Nuclear Fusion* **61,** 066019. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/abf74b (June 2021).
26. Ghiozzi, A. *et al.* Modeling of frequency-sweeping Alfvén modes in the TJ-II stellarator. *Nuclear Fusion* **64,** 036005. ISSN: 0029-5515, 1741-4326. doi:10.1088/1741-4326/ad1c93 (Mar. 2024).
27. Richardson, N., Schaeffer, H. & Tran, G. SRMD: Sparse Random Mode Decomposition. Publisher: arXiv Version Number: 2. doi:10.48550/ARXIV.2204.06108 (2022).
28. Kaptanoglu, A. A., Hansen, C., Lore, J. D., Landreman, M. & Brunton, S. L. Sparse regression for plasma physics. en. *Physics of Plasmas* **30,** 033906. ISSN: 1070-664X, 1089-7674. doi:10.1063/5.0139039 (Mar. 2023).
29. Klein, A. *et al.* A sparsity-based method for the analysis of magnetic fluctuations in unevenly-spaced Mirnov coils. en. *Plasma Physics and Controlled Fusion* **50,** 125005. ISSN: 0741-3335, 1361-6587. doi:10.1088/0741-3335/50/12/125005 (Dec. 2008).
30. Kaptanoglu, A. A., Qian, T., Wechsung, F. & Landreman, M. Permanent-Magnet Optimization for Stellarators as Sparse Regression. en. *Physical Review Applied* **18,** 044006. ISSN: 2331-7019. doi:10.1103/PhysRevApplied.18.044006 (Oct. 2022).
31. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. en. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67,** 301–320. ISSN: 1369-7412, 1467-9868. doi:10.1111/j.1467-9868.2005.00503.x (Apr. 2005).
32. Tibshirani, R. Regression Shrinkage and Selection via The Lasso: A Retrospective. en. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73,** 273–282. ISSN: 1369-7412, 1467-9868. doi:10.1111/j.1467-9868.2011.00771.x (June 2011).
33. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019).
34. Van Den Berg, E. & Friedlander, M. P. Probing the Pareto Frontier for Basis Pursuit Solutions. en. *SIAM Journal on Scientific Computing* **31,** 890–912. ISSN: 1064-8275, 1095-7197. doi:10.1137/080714488 (Jan. 2009).
35. Van Den Berg, E. & Friedlander, M. P. Sparse Optimization with Least-Squares Constraints. en. *SIAM Journal on Optimization* **21,** 1201–1229. ISSN: 1052-6234, 1095-7189. doi:10.1137/100785028 (Oct. 2011).
36. Joneidi, M. *Functional Brain Networks Discovery Using Dictionary Learning with Correlated Sparsity* en. arXiv:1907.03929 [eess, q-bio, stat]. July 2019.
37. Shen, B., Liu, B.-D. & Wang, Q. Elastic net regularized dictionary learning for image classification. en. *Multimedia Tools and Applications* **75,** 8861–8874. ISSN: 1380-7501, 1573-7721. doi:10.1007/s11042-014-2257-y (Aug. 2016).
38. Sathish, R., Sheet, D. & Khare, S. *Linear Neural Network as a Fast Solver for Dictionary Learning* en. preprint (Feb. 2023). doi:10.36227/techrxiv.21988181.v1.

39. Kelly, J. W., Degenhart, A. D., Siewiorek, D. P., Smailagic, A. & Wei Wang. *Sparse linear regression with elastic net regularization for brain-computer interfaces* en. in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, San Diego, CA, Aug. 2012), 4275–4278. ISBN: 978-1-4577-1787-1 978-1-4244-4119-8. doi:10.1109/EMBC.2012.6346911.
40. Fakhry, M. & Gallardo-Antolín, A. Elastic net regularization and gabor dictionary for classification of heart sound signals using deep learning. en. *Engineering Applications of Artificial Intelligence* **127,** 107406. ISSN: 09521976. doi:10.1016/j.engappai.2023.107406 (Jan. 2024).
41. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, Oregon, Aug. 1996), 226–231.
42. Campello, R. J. G. B., Moulavi, D., Zimek, A. & Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data* **10,** 5:1–5:51. ISSN: 1556-4681. doi:10.1145/2733381 (July 2015).
43. Arthur, D. & Vassilvitskii, S. *k-means++: the advantages of careful seeding* in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (Society for Industrial and Applied Mathematics, USA, Jan. 2007), 1027–1035. ISBN: 978-0-89871-624-5.
44. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605. ISSN: 1533-7928 (2008).
45. Elad, M. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing* en. ISBN: 978-1-4419-7010-7 978-1-4419-7011-4. doi:10.1007/978-1-4419-7011-4 (Springer New York, New York, NY, 2010).
46. Tillmann, A. M. On the Computational Intractability of Exact and Approximate Dictionary Learning. *IEEE Signal Processing Letters* **22.** Conference Name: IEEE Signal Processing Letters, 45–49. ISSN: 1558-2361. doi:10.1109/LSP.2014.2345761 (Jan. 2015).
47. Chen, S. S., Donoho, D. L. & Saunders, M. A. Atomic Decomposition by Basis Pursuit. en. *SIAM Review* **43,** 129–159. ISSN: 0036-1445, 1095-7200. doi:10.1137/S003614450037906X (Jan. 2001).
48. Murphy, K. P. *Probabilistic Machine Learning: An introduction* (MIT Press, 2022).
49. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Physical Review E* **69.** Publisher: American Physical Society, 066138. doi:10.1103/PhysRevE.69.066138 (June 2004).
50. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. ISSN: 1533-7928 (2011).
51. Ross, B. C. Mutual Information between Discrete and Continuous Data Sets. en. *PLOS ONE* **9.** Publisher: Public Library of Science, e87357. ISSN: 1932-6203. doi:10.1371/journal.pone.0087357 (Feb. 2014).

Appendix A: Optimization schemes for signal coding

In this subsection, we formulate the signal representation by starting with the most general constrained minimization problem and subsequently introducing step-by-step different alternative approaches to achieve sparsity in the codification, a detailed discussion can be found in Elad's book [45].

It is assumed that the dictionary matrix has more columns than rows, so the system of equations $\mathbf{D}\cdot\boldsymbol{c}=\boldsymbol{s}$ will be undetermined as there will be more unknowns than equations. If the system is consistent, meaning $\boldsymbol{s}$ can be linearly generated by columns of $\mathbf{D}$, there will be infinite solutions. Therefore, to cope with infinite solutions we aim to find the solution $\mathbf{c}$ with minimum norm. The problem can be formulated as a constrained optimization problem $(P_{\mathcal{L}})$,

$$(P_{\mathcal{L}}) \equiv \min_{\boldsymbol{c}} \quad \mathcal{L}(\boldsymbol{c}) \quad \text{subject to} \quad \mathbf{s}=\mathbf{D}\cdot\boldsymbol{c}, \tag{7}$$

where the loss function $\mathcal{L}(\boldsymbol{c})$ controls the solutions we can obtain. Choosing $\mathcal{L}(\boldsymbol{c})$ to be the Euclidean squared norm $||c||_2^2$ is the common choice,

$$(P_2^2) \equiv \min_{\boldsymbol{c}} \quad ||\boldsymbol{c}||_2^2 \quad \text{subject to} \quad \mathbf{s}=\mathbf{D}\cdot\boldsymbol{c} \tag{8}$$

as it can be shown [45] (page 4), the problem P_2^2 has a unique solution because the objective functional is convex and the solution can be obtained [45] using the Moore–Penrose pseudoinverse matrix $\mathbf{X}^+$.

Now, we can formulate an *unconstrained* version of the problem (P_2') which has equivalent results, and it is stated as the minimization of the signal's mean squared reconstruction error (MSE), so choosing the cost function as $\mathcal{L}(x)\equiv||\mathbf{D}\cdot\boldsymbol{c}-\mathbf{s}||_2^2$ results in

$$(P_2^{2'}) \equiv \quad \min_{\boldsymbol{c}} \quad ||\mathbf{D}\cdot\boldsymbol{c}-\mathbf{s}||_2^2. \tag{9}$$

The unconstrained formulation can be more familiar to the reader because it is by definition the ordinary least squared (OLS) estimator for linear regression. Remarkably, the unconstrained version does not enforce obtaining a solution of $\mathbf{s}=\mathbf{D}\cdot\boldsymbol{c}$, nor does it assume any property of the solution vector $\boldsymbol{c}$ components. Above all, we can

get a result that is as good as the optimization algorithm's performance.

However, the solution of the OLS estimator is not adequate for creating a basis; because generally every $\boldsymbol{c}$'s components $[c_1, c_2, ..., c_N]$ can be different than zero, and the explanatory power of the code is lost. We need to vanish as many code coefficients $[c_1, c_2, ..., c_N]$ as possible; that can be achieved using l_0 norm instead of l_0 in the constrained problem $P_2^{2'}$. The norm l_0 is defined as the number of vector components different than zero, so minimizing l_0 and the MSE simultaneously can lead to a suitable basis. Hence, we can define the problem P_0 as

$$(P_0) \equiv \min_{\boldsymbol{c}} ||\boldsymbol{c}||_0 \quad \text{subject to} \quad \boldsymbol{s} = \mathbf{D} \cdot \boldsymbol{c}. \tag{10}$$

This problem is significantly difficult to solve, since (P_0) is an $\mathcal{NP}$ hard problem [46], so trying to solve it directly is not common practice. Nevertheless, we can relax the constraints if we use the ℓ_1 norm which is defined as the sum of the vector's components. Consequently, minimization of the ℓ_1 norm will shrink some vector components close to zero. This *optimization strategy* [47] is called *basis pursuit* (BP) , it will be noted as P_1

$$(P_1) \equiv \min_{\boldsymbol{c}} ||\boldsymbol{c}||_1 \quad \text{subject to} \quad \boldsymbol{s} = \mathbf{D} \cdot \boldsymbol{c}. \tag{11}$$

Additionally, we need to take into account that most signals will not be pure and they will have noise. That can be modeled by adding an σ error parameter to the functional in the following way

$$(BPDN) \equiv \min_{\boldsymbol{c}} ||\boldsymbol{c}||_1 \quad \text{subject to} \quad ||\mathbf{D} \cdot \boldsymbol{c} - \mathbf{s}||_2^2 \leq \sigma. \tag{12}$$

The later version is called basis pursuit denoised (BPDN). Though it was first proposed [47] in its unbound version (P_1'), it can be formulated as follows:

$$(P_1') \equiv \min_{\boldsymbol{c}} \left\{ ||\mathbf{D} \cdot \boldsymbol{c} - \mathbf{s}||_2^2 + \lambda_1 ||\boldsymbol{c}||_1 \right\}. \tag{13}$$

This problem is also found in machine learning known as *Lasso* regression.

Another alternative for regularization is using ℓ_2 norm (no squared). This functional takes the form

$$(P_2') \equiv \min_{\boldsymbol{c}} \left\{ ||\mathbf{D} \cdot \boldsymbol{c} - \mathbf{s}||_2^2 + \lambda_2 ||\boldsymbol{c}||_2 \right\} \tag{14}$$

in its unconstrained version. As in this case, the equivalent problem in machine learning is named *Ridge regression.* which is useful for avoiding shrinkage of the solution vector when data have multicollinearities.

The effect of ℓ_1 regularization on a group of correlated variables is the selection of one variable, discarding the rest, even if these discarded variables have explanation power or are part of the solution. If there are correlated features and a sparse solution that conserves groups of correlated variables is desired, adding ℓ_2 to the functional, like in

$$\begin{pmatrix} ELASTIC \\ NET \end{pmatrix} \equiv \min_{\boldsymbol{c}} \left\{ ||\mathbf{D} \cdot \boldsymbol{c} - \mathbf{s}||_2^2 + \lambda_1 ||\boldsymbol{c}||_1 + \lambda_2 ||\boldsymbol{c}||_2 \right\}, \tag{15}$$

is a good alternative. This problem is known in the machine learning community as elastic net regression.

About the decision of using constrained or unconstrained versions of the optimization problems, it usually resolves in favor of the unconstrained version, because more and faster algorithms are available with guaranteed performance and studied convergence. For instance: solvers [34, 35] can be used to solve the basis pursuit problem with a random dictionary as with SRMD [27]. However, a major difficulty shared by all alternatives to ERMD is the allocation of the matrix in memory. The features in matrix $\mathbf{D}$ must have the same number of samples as the signal. Moreover, because we have signals sampled with 1MHz with a pulse duration of 1 second and thousands of features are necessary to represent the modes accurately, the matrix size can easily exceed all available memory in the computing node (of the order of hundred terabytes).

A first valid idea to address this issue is to divide the signal into chunks and process each chunk sequentially on one CPU, or in a parallel process on several computing nodes. But this approach has one inconvenience, namely each signal chunk has a different signal-to-noise ratio, and the weights of vector $\boldsymbol{c}$ may be unevenly set along chunks. In addition, there are discontinuities in the reconstruction of the joins between chunks. The gradient accumulation proposed in this paper tackles this issue.

Appendix B: Finding relationships between frequency and density

As we mentioned in the introduction, the Alfvén waves' frequencies f follow a linear dependency with $1/\sqrt{n}$. Once the encoded signal is clustered, we have a database of modes that can be mined for finding Álfvén instabilities.

The linear relationship can be captured using Pearson's correlation $r = cov(x, y)/(\sigma_x \sigma_y)$, while assumptions for the model are true. Pearson's model assumes a linear relationship between variables X and Y, in addition, the variables have to be normally distributed, with equal variance (homoscedastic), and without outliers. If these assumptions are True and Pearson's value $r(f, 1/\sqrt{n})$ is close to 1, we can say that we are identifying an Alfvénic mode. These conditions are not true for all clustered modes.

For n observations of random variables $\mathcal{X}$,$\mathcal{Y}$ and continuous distributions, the Pearson's correlation can be calculated as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{16}$$

It is commonly known that correlation does not imply causation; however, correlation does not even imply correlation in many circumstances. Pearson's formula is very sensitive to outliers and noise, and clustered modes can often have outliers. Then the normality assumption and equal noise variance may not hold true. In addition, the plasma conditions can be modified very fast, in this case, Alfvén's instabilities frequency follows the evolution of $1/\sqrt{n}$, with some non-linear distortion. Another effect that might distort the relationship is the fact that the measured

density is integrated along one line of sight, and the changes in density might be localized outside this line of sight.

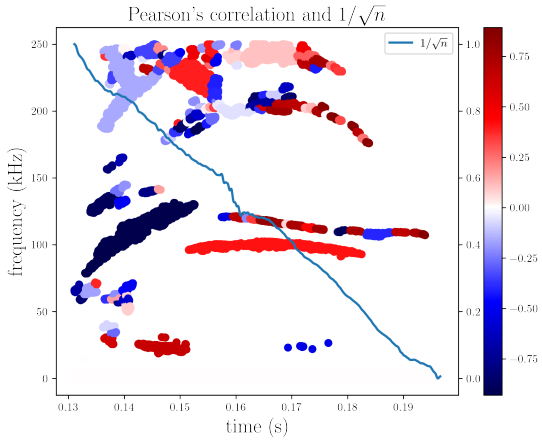

Figure 19. Pearson correlation of modes frequencies and density

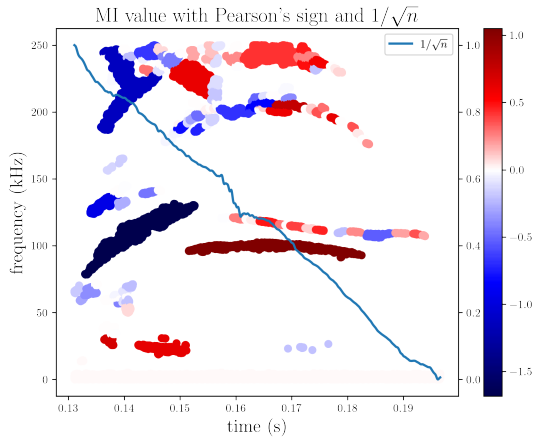

Figure 20. MI of modes frequencies and density.

To avoid false correlations the threshold value can be set high ≈0.9, but many Alfvén modes will be lost. To avoid this limitation, we select mutual information (MI) to estimate if the frequencies are related to the density variation. Mutual information ranges from 0 to ∞, 0 meaning full statistical independence, then the higher the value the stronger the dependence. MI is known as "generalized correlation" [48](pages 213ff) as it is capable of capturing linear and nonlinear dependencies on continuous and categorical data, with better resistance to outliers, and fewer assumptions.

For continuous distributions of random variables X,Y the MI can be understood as the information gain $I(X;Y) = D_{\mathrm{KL}}(\mu(x,y)||\mu_x(x)\mu_y(y))$ obtained if the two variables are not independent [48](pages 213ff). The Kullback-Leibler (KL) divergence, which measures the "distance" between probability distributions, will be 0 if the distribution $\mu(x,y)$ is independent $\mu(x,y) = \mu_x(x)\mu_y(y)$ (separable), in other words: KL divergence will be 0 if both variables don't have a relationship. MI can be estimated using the marginals densities $\mu_x(x)$, $\mu_y(y)$ and joint distribution $\mu(x,y)$ [49] in the formula

$$I(X;Y) = \int\int \mu(x,y)\log\left(\frac{\mu(x,y)}{\mu(x)\mu(y)}\right)\mathrm{d}x\,\mathrm{d}y. \qquad (17)$$

The integration of continuous variables is performed by binning the data, we have used scikit-learn implementation [50][49][51].

MI does have one limitation though. It does not distinguish between positive and negative correlation. As we found that the values obtained by Pearson's correlation and MI are similar, we take the sign from Pearson's correlation, to determine the Alfvénic character.

As we can see in figures 19 and 20. The results of using MI and Pearson's correlation are comparable. In both cases, at time 0.15 s we identify a possible Alfvén mode crossing other modes at ≈225 kHz approximately and another one at ≈35 kHz. However, at time 0.17 s, subtle differences can be appreciated. For instance, a time 0.17 s and ≈100 kHz there is a mode that shows mild correlation but high mutual information. This is caused by the bias of MI when there are a high number of observations, it tends to overestimate the relationship between variables. There is some level of relationship expected only *by chance*. When the number of observations is small, the opposite might occur. MI could underestimate relationships. This bias might be useful though, because it can avoid establishing a relationship with few and noisy observations. This can be appreciated on the long mode at ≈120 kHz from 0.15 to 0.19 s, which is broken into smaller density groups by the encoding. It can be seen in this case that Pearson's values are higher than MI, but there is not a clear relationship with $1/\sqrt{n}$. This example from shot 38399 was selected to show how complex can be encoding and finding Álfvén eigenmodes, in other shots this relationship is clearer. Selecting a different value cut to consider modes Alfvénic significantly varies the result. Therefore, to compare both methods, we show the result for the same threshold value of 0.9 in Fig.21 and Fig.22.

In total, when using Pearson's correlation method, 1315 modes are retrieved. While using the MI with sign method recovers 1433 Alfvénic modes, 8.9% more modes. This can be appreciated in Fig.21 and Fig.22. This result gives us an average of Alfvénic mode 1.1 per shot. On the other hand, if we examine the negative class we have 35472 (Pearson's) and 35354 (MI) non-Alfvénic modes respectively. As we can see, the dataset is heavily imbalanced towards the negative class, only 3-4 % of the mode structures are Alfvénic. Given this imbalance of classes, a 8.9% more Alfvénic modes is a relevant result. Because each mode counts for training models in imbalanced datasets. Therefore, because of its superior performance in handling noise false positives, and

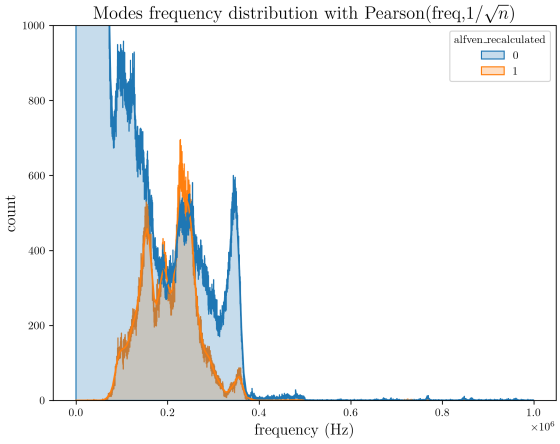

Figure 21. Distributions of Alfvénic atoms frequencies using Pearson's for labeling

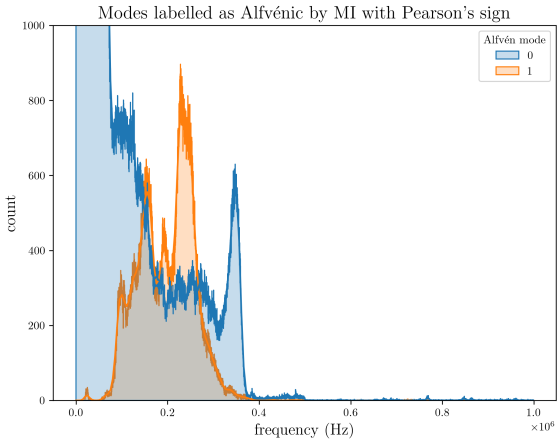

Figure 22. Distributions of Alfvénic mode frequencies using MI for labeling

keeping a higher atom count in the frequency interval, the MI method was selected to continue the analysis.

Appendix C: Modes contingency tables using categorical variables

In this section, the relationships between variables are explored with categorical features as an additional descriptive effort. For this purpose, contingency tables (or crosstabulations, or cross-tables) are calculated using new categorical variables; using the following indicator functions $\mathbf{1}_{MI(f,1/\sqrt{n})>0.9}(\text{modes})$, $\mathbf{1}_{MI(f,I_p)>0.9}(\text{modes})$ and $\mathbf{1}_{MI(f,W)>0.9}(\text{modes})$, in a shorter notation; $\mathbf{1}_{1/\sqrt{n}}$, $\mathbf{1}_{I_p}$, $\mathbf{1}_W$.

The contingency table shows the multivariate frequency distribution of the Alfvénic mode class, quantifying the dependence on these new categorical variables, and interpreting relative frequencies as a probability if the table is normalized. First, all modes labeled as noise by HDBSCAN (-1) are discarded. Then, MI is used to establish the relationship between the modes frequency chirp f and other diagnostic signals like I_p, W, $1/\sqrt{n}$. Once MI is calculated, if the value is greater than 0.9 the indicator variables $\mathbf{1}_{1/\sqrt{n}}$, $\mathbf{1}_{I_p}$ and $\mathbf{1}_W$, are set to 1, otherwise to 0: for all atoms in the mode. Finally, the table needs to be grouped to count modes instead of atoms if necessary. The contingency table can be calculated indicating the relative frequency of having an Alfvénic mode *related* with the given plasma signals.

If we study the result in Table 3: we can count 1433 Alfvénic modes in total, and 35354 non-Alfvénic. This analysis includes all frequencies, and because many low-frequency atoms are required to describe the signal: 35111 low-frequency modes are not related to any of the variables. We observe that, if there is no relationship with $1/\sqrt{n}$, there is no Alfvénic mode count, as should be expected. Meanwhile, 515 Alfvénic modes are related to variations of the density exclusively. Remarkably, 533 Alfvénic modes are related to changes in plasma energy as well, and 308 Alfvénic modes are related to changes in all variables; density, plasma energy, and current simultaneously.

It is worth mentioning that the χ^2 statistical test has been used to find evidence of the relationships of variables in the cross-table, rejecting the null hypothesis in all combinations. However, this dependence is obvious if we observe the large difference between the values of the first and second columns. We can conclude that there is a strong relationship between the Alfvén class variable and all the plasma signals considered.

$\mathbf{1}_{1/\sqrt{n}}$	$\mathbf{1}_{I_p}$	$\mathbf{1}_W$	Alfvén=0	Alfvén=1
0	0	0	35111	0
0	0	1	44	0
0	1	0	78	0
0	1	1	22	0
1	0	0	46	515
1	0	1	13	533
1	1	0	26	77
1	1	1	14	308

Table 3. Alfvénic modes' contingency table

If the low frequencies are included, $P(\text{Alfvén} = 0|\ \mathbf{1}_{1/\sqrt{n}}, \mathbf{1}_{I_p}, \mathbf{1}_W) = 0.95$, but the remaining 5% is distributed over 15 possibilities, which difficults the interpretation of the result. What's more, even if all modes labeled as noise by the cluster algorithm has been discarded, there is a high chance that low-frequency modes are just noise. Because most Alfvénic modes appear over 100 kHz, a normalized contingency table is calculated for atoms belonging to modes over 100 kHz.

Now, if we examine the Table 4 of relative frequencies (in %): Almost 6% of modes over 100 kHz are Alfvénic modes related to the plasma energy, and 6.4% of modes are Alfvénic modes related to the plasma current as well.

By using MI to construct the categorical variables, all

$\mathbf{1}_{1/\sqrt{n}}$	$\mathbf{1}_{I_p}$	$\mathbf{1}_W$	f(Alfvén=0) (%)	f(Alfvén=1) (%)
0	0	0	80.168150	0.000000
0	0	1	0.519288	0.000000
0	1	0	0.927300	0.000000
0	1	1	0.272008	0.000000
1	0	0	0.408012	5.984174
1	0	1	0.123640	6.441642
1	1	0	0.272008	0.952028
1	1	1	0.160732	3.771019

Table 4. Normalized contingency table for modes over 100kHz in (%)

nonlinear relationships (and all linear) between variables can be detected. However, when using MI to establish relationships, the positive or negative relation is not captured. Capturing nonlinear relationships is more general and therefore more important than knowing the sign of the relationship when studying modes in all shots. Moreover, the sign of the relationship can be calculated later.

There is a strong nonlinear dependence between mode frequency and the plasma signals, this is because most Alfvénic modes in this dataset are strongly driven when the heating systems are connected. Examining the histograms of Fig.23, we can find a resemblance with the results of K-means clustering. In fact; some modes of cluster 2 appear to be correlated with the density only; modes correlated with energy and density appear in clusters 21 and 8. And non-Alfvénic modes of cluster 17 are strongly correlated with the current. However, to establish a complete equivalency between the cluster analysis and the contingency tables of categorical features it is not possible, as we arbitrarily selected a criterion of the category creation. The treatment of the full variance in the cluster analysis appears more general than the use of contingency tables.

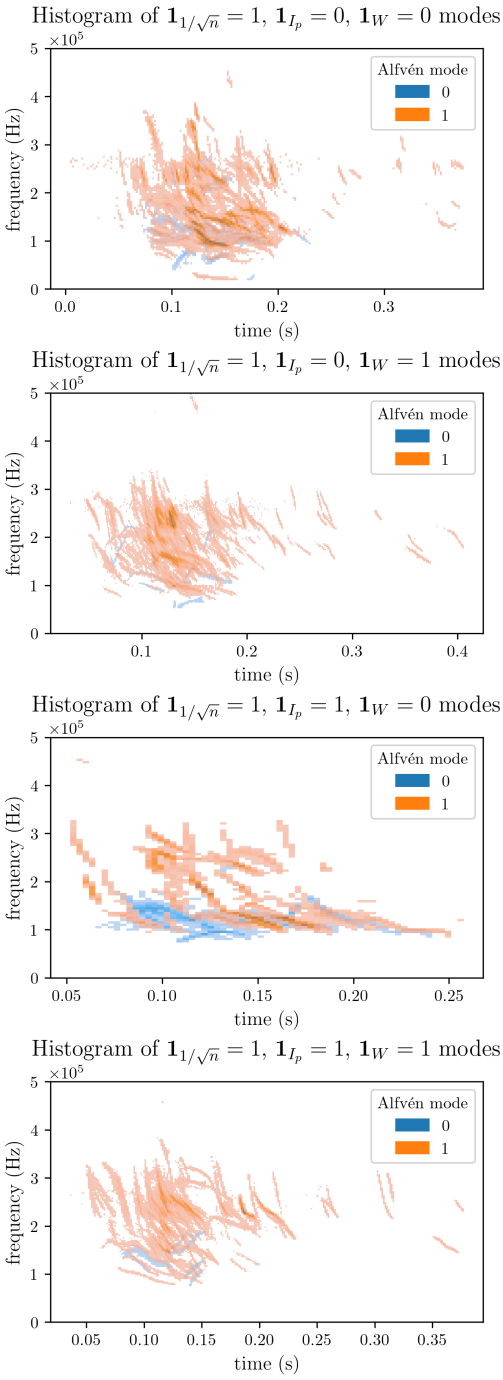

Figure 23. Contingency table modes