AIP | Review of Scientific Instruments

A probabilistic model for the identification of confinement regimes and edge localized mode behavior, with implications to scaling laws

Geert Verdoolaege and Guido Van Oost

Citation: Rev. Sci. Instrum. **83**, 10D715 (2012); doi: 10.1063/1.4733307
View online: http://dx.doi.org/10.1063/1.4733307
View Table of Contents: http://rsi.aip.org/resource/1/RSINAK/v83/i10
Published by the American Institute of Physics.

A probabilistic model for the identification of confinement regimes and edge localized mode behavior, with implications to scaling laws[a]

Geert Verdoolaege[b] and Guido Van Oost
Department of Applied Physics, Ghent University, 9000 Gent, Belgium

(Presented 9 May 2012; received 8 May 2012; accepted 12 June 2012; published online 18 July 2012)

Pattern recognition is becoming an important tool in fusion data analysis. However, fusion diagnostic measurements are often affected by considerable statistical uncertainties, rendering the extraction of useful patterns a significant challenge. Therefore, we assume a probabilistic model for the data and perform pattern recognition in the space of probability distributions. We show the considerable advantage of our method for identifying confinement regimes and edge localized mode behavior, and we discuss the potential for scaling laws. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4733307]

I. INTRODUCTION

Pattern recognition methods are very useful for extracting structure from fusion data sets with typically large statistical uncertainties and many variables. Indeed, the patterns may reflect the underlying physics of the plasma, while real-time data interpretation is indispensable for plasma control.

In this work, we advocate the essential role played by the uncertainty in the data and we show its importance for pattern recognition in fusion data. According to this viewpoint, the fundamental object resulting from a measurement act is a probability distribution. Because pattern recognition is essentially based on geometrical concepts such as distance, we use *information geometry* as a mathematical framework to calculate distances between distributions, specifically the *geodesic distance*.

There are several advantages of our approach. First, the measurement uncertainty is intrinsically part of the data description and is carried along with any further processing of the data, such as pattern recognition. Second, the full probability distribution of a measurement carries much more information than the measurement value weighted by its error bar. Hence, conclusions drawn from the data become more reliable and pattern recognition becomes more effective. Third, dependency structure, such as correlations between measurements or physical variables, contains even more information and can be modeled through multivariate distributions.

In this paper, we consider the visualization and classification of confinement modes and edge localized mode (ELM) classes in an international database and we discuss the potential of our framework for scaling laws in fusion devices.

II. A GEOMETRIC-PROBABILISTIC PATTERN RECOGNITION FRAMEWORK

In the field of information geometry, a probability density family is interpreted as a Riemannian differentiable manifold,[1] or *information manifold*. A point on the manifold corresponds to a specific probability density function within the family and the family parameters provide a coordinate system on the manifold. The Fisher information acts as a metric tensor, allowing the calculation of the geodesic distance (GD) as a natural and theoretically motivated similarity measure between probability distributions.[2,3]

In this work we use a simple probability model, namely, the univariate Gaussian distribution, parameterized by its mean μ and standard deviation σ. A closed-form expression exists for the corresponding GD, which permits a fast evaluation. More details of our formalism will be published in Ref. 4.

III. THE INTERNATIONAL TOKAMAK PHYSICS ACTIVITY (ITPA) DATABASE

We employ measurements from the ITPA Global H-mode Confinement Database (DB3, version 13f), henceforth referred to as "the ITPA database".[5,6] The data have been used extensively for determining scaling laws for the energy confinement time, mainly as a function of a set of eight plasma and engineering parameters: plasma current (I_{p}), vacuum toroidal magnetic field, total power loss from the plasma (P_{loss}), central line-averaged electron density, plasma major radius, plasma minor radius, elongation and effective atomic mass. We use the same eight variables to visualize and discriminate between confinement regimes and ELM classes.

The ITPA database lists typical error estimates of measurements for the various plasma and engineering variables, although it is acknowledged that the reliability of these estimates may vary across the database. We assume that the error bars represent a single standard deviation. If this is the only information that is available on the underlying probability distribution, then according to the principle of maximum entropy the distribution is Gaussian with mean the measurement itself and standard deviation the error bar. We also suppose that, for stationary plasma conditions, all variables are statistically independent and so the joint distribution factorizes. It is important to note that our

[a]Contributed paper, published as part of the Proceedings of the 19th Topical Conference on High-Temperature Plasma Diagnostics, Monterey, California, May 2012.

[b]geert.verdoolaege@ugent.be.

formalism has no difficulties with the heterogeneous sources of the measurements, coming from different tokamaks and possibly with *different error bars for essentially the same physical quantities*. The reason is that the error estimates are automatically embedded in the probabilistic data description.

For each entry, or sample, in the database, some characteristics of the confinement regime were determined by experts. Specifically, the confinement was labeled as OHM (Ohmic), L or RI, which we summarize through an "L-mode" class for brevity. In addition, we define an "H-mode" class as consisting of plasmas in pure H-mode (H), with frequent L-H transitions (LHLHL), H-mode with small ELMs (HSELM), high-frequency small ELMs (HSELMH), large ELMs (HGELM), or high-frequency large ELMs (HGELMH).

IV. VISUALIZATION AND IDENTIFICATION OF CONFINEMENT REGIMES AND ELM CLASSES

We now discuss the application of our framework to visualization and classification of confinement data from the ITPA database. The real-time automated identification of confinement regimes (usually L- and H-mode) and ELM types has important applications in plasma control and will be an important tool for ITER. In addition, extracted patterns from confinement data, such as clusters of similar plasma regimes, can contribute to the understanding of the physics. Furthermore, the concept of regime identification is closely related to the establishment of scaling laws for the L to H transition power threshold and the energy confinement time.

A. Visualization of confinement data

An important tool for the identification of patterns in the ITPA database is the visualization of the data through a scatter plot in the natural two-dimensional Euclidean space. Since the original data dimensionality is eight, the data visualization involves a dimensionality reduction procedure. To this end we used metric multidimensional scaling (MDS), searching for a configuration of points in the Euclidean plane yielding minimal distortion of all pairwise distances.[4]

Figure 1 shows approximately isometric projections of the ITPA data into the Euclidean plane, obtained via MDS. For each sample in the database the basic confinement regime, i.e., L-mode or H-mode, was obtained from the database and is indicated in the figure. For Figures 1(a) and 1(b) the measurement uncertainty was not considered and MDS was carried out on the basis of simple Euclidean distances in the original data space. On the contrary, the MDS in Figure 1(c) is based on GDs between Gaussian product distributions. It can be clearly noticed that the projections obtained with the GD, which take into account the measurement error, exhibit considerably more structure compared to the Euclidean case. In particular, it is much easier to visually discriminate between the L- and H-mode clusters.

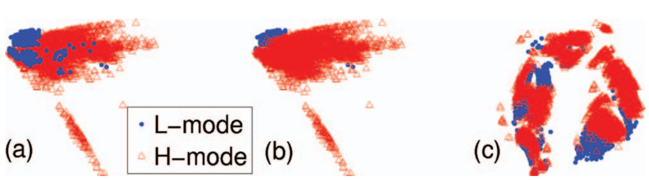

FIG. 1. Two-dimensional projections of the ITPA data using MDS, with indicated L- and H-mode clusters. (a) Using the Euclidean distance without measurement error and with the L-mode points on top. (b) The same, but with the H-mode points on top for better visibility. (c) Using the GD with measurement error.

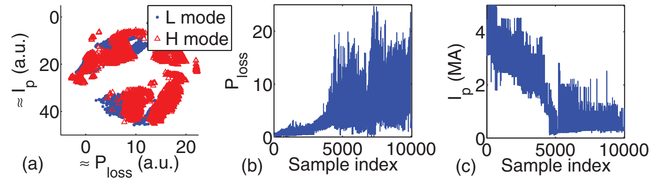

FIG. 2. (a) MDS projection using the GD with a rotated coordinate system. When moving along the horizontal (vertical) axis towards the right (upwards), the points that one meets correspond to actual measurements in the database of P_{loss} (I_{p}) as indicated in panel (b) ((c)).

For the MDS projection using the GD, it turns out that there is a coordinate system where the horizontal axis roughly corresponds to P_{loss} and the vertical axis to I_{p}. This is indicated in Figure 2(a), with the coordinate system rotated about 45° clockwise with respect to the horizontal and vertical directions in Figure 1(c). The trend of the measurements of P_{loss} and I_{p} along their respective axes is shown in Figures 2(b) and 2(c), respectively. This shows that MDS recognizes P_{loss} and I_{p} as two of the most fundamental variables (similar to principal components) governing the confinement.

Finally, in Figure 3 the MDS projection using the GD is again plotted with the database samples divided according to the confinement regimes mentioned in Sec. III. There is a considerable overlap between the different ELM classes, but this may be partly caused by the restrictive two-dimensional projection.

B. Classification of confinement regimes and ELMs

We now turn to classification of confinement data. At this stage we do not intend to present a dedicated platform for L–H or ELM classification for application in the field, although the proposed methods can certainly be used for that purpose. Rather, the objective of the experiments discussed here is to show the added value of the intrinsic probabilistic

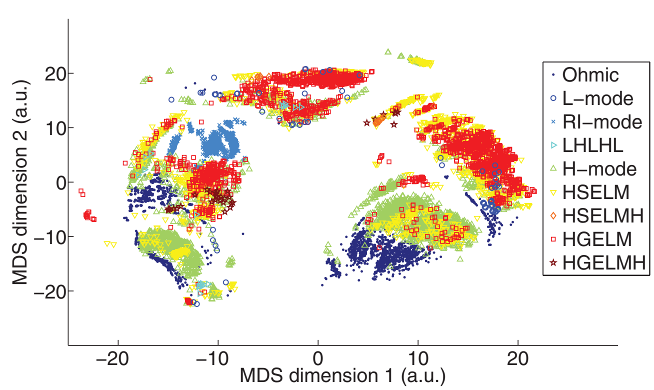

FIG. 3. MDS projection with indication of the confinement regimes and ELM classes.

TABLE I. Mean and standard deviation of the correct classification rates (%) for the classification of confinement regimes using a kNN classifier for different sets of variables, and Euclidean and geodesic distance measures.

Variables	Mode	Euclidean	GD
All eight	L	97.4 ± 0.7	97.2 ± 0.9
	H	96.6 ± 0.6	98.1 ± 0.5
I_p, P_{loss}	L	89.6 ± 1.5	94.5 ± 1.3
	H	92.0 ± 0.9	94.7 ± 0.9

description of the data and to provide a benchmark for visualization and classification algorithms.

We performed a series of classification experiments with two classes (L-mode and H-mode) using 5% of the data for training. We carried out k-nearest neighbor (kNN) classification with $k = 1$, effectively assigning a point to be classified to the class that its nearest neighbor belongs to. Both the Euclidean distance and the GD were used as a similarity measure. All experiments were repeated 100 times, each time using different training and test sets, and the mean and standard deviation of the correct classification rates for L-mode and H-mode were calculated. The results are shown in Table I. If all eight plasma parameters are used, the performance of the Euclidean distance and the GD is excellent but relatively similar. In order to more clearly show the advantage of the probabilistic approach, we repeated the experiments using only I_p and P_{loss}, which were identified above as two fundamental degrees of freedom. This results in generally lower recognition rates, but the superior performance of the GD can be clearly noticed now.

We also used kNN to see if the ELM classes can be identified in the original data space. Although about 90% of the large ELMs could be recognized, the small ELMs had a recognition rate of only 56% with the Euclidean distance and 61% using the GD. Clearly, more information is needed to reliably classify ELM behavior.

V. POTENTIAL FOR SCALING LAWS

Given the superior performance of our geometric-probabilistic approach for classification of confinement modes, an advantage may also be obtained in establishing scaling laws from the ITPA data using a regression analysis. Genuine regression on Riemannian information manifolds has not been considered before, but we performed a crude test to evaluate its potential. We projected the data using MDS based on the GD into an eight-dimensional Euclidean space. After an additional logarithmic transformation of each variable, we then performed linear regression in the projected space, and compared with linear regression in the original data space, considered as a Euclidean space. In the original data space a normalized coefficient of determination $R^2 \approx 0.3$ was obtained, while in the projected space this resulted in $R^2 \approx 0.4$. Therefore, a considerable potential for verifying and possibly improving scaling laws can be expected.

VI. CONCLUSION

We have indicated the importance for pattern recognition in fusion data of reliable estimates of measurement uncertainty and we have highlighted the fundamental character of probability distributions for describing the measurement act. We have shown the appropriateness of information geometry and the geodesic distance for visualization and classification of probabilistic confinement data in the ITPA database. It is remarkable that even the approximate and limited information in the ITPA database on the underlying probability distribution is beneficial to pattern recognition. Finally, we have noted the possibility of our framework for recognition of ELM behavior and the potential for scaling laws. Both topics will be the subject of future work.

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Transactions of Mathematical Monographs Vol. 191 (American Mathematical Society, New York, 2000).
[2] G. Verdoolaege and P. Scheunders, J. Math. Imaging Vis. **43**, 180 (2011).
[3] G. Verdoolaege and P. Scheunders, Int. J. Comput. Vis. **95**, 265 (2011).
[4] G. Verdoolaege, G. Karagounis, A. Murari, J. Vega, G. Van Oost, and JET-EFDA Contributors, "Modeling fusion data in probabilistic metric spaces: Applications to the identification of confinement regimes and plasma disruptions," Fusion Sci. Technol. (in press).
[5] D. McDonald *et al.*, Nucl. Fusion **47**, 147 (2007).
[6] See http://efdasql.ipp.mpg.de/hmodepublic for further information and access to the database.