Identification of confinement regimes in tokamak plasmas by conformal prediction on a probabilistic manifold

Geert Verdoolaege[1], Jesús Vega[2], Andrea Murari[3], and Guido Van Oost[1]

[1] Department of Applied Physics, Ghent University, Ghent, Belgium,
geert.verdoolaege@ugent.be
[2] Laboratorio Nacional de Fusion, Asociacion EURATOM-CIEMAT, Madrid, Spain
[3] Associazione EURATOM-ENEA sulla Fusione, Consorzio RFX, Padova, Italy

Abstract. Pattern recognition is becoming an increasingly important tool for making inferences from the massive amounts of data produced in magnetic confinement fusion experiments. However, the measurements obtained from the various plasma diagnostics are typically affected by a considerable statistical uncertainty. In this work, we consider the inherent stochastic nature of the data by modeling the measurements by probability distributions in a metric space. Information geometry permits the calculation of the geodesic distances on such manifolds, which we apply to the important problem of the classification of plasma confinement regimes. We use a distance-based conformal predictor, which we first apply to a synthetic data set. Next, the method yields an excellent classification performance with measurements from an international database. The conformal predictor also returns confidence and credibility measures, which are particularly important for interpretation of pattern recognition results in stochastic fusion data.

Keywords: Magnetic confinement fusion, probabilistic manifold, conformal predictor

1 Introduction

Controlled nuclear fusion research aims at the development of a clean, safe and virtually inexhaustible source of energy. One promising route towards the realization of this objective involves the confinement of a hot hydrogen isotope plasma using an appropriate magnetic field configuration in a so-called tokamak device. Different regimes of plasma confinement have been established, primarily the low-confinement mode, or 'L-mode' and the high-confinement mode or 'H-mode'. The H-mode is characterized by an enhanced plasma temperature, density and average confinement time of the plasma energy and particles, and it has been shown to result from a so-called 'transport barrier' near the plasma boundary. The H-mode has become the basis of the reference plasma scenario for next-step devices such as ITER, which is presently under construction in Southern France.

The actual physical mechanism responsible for the transition of a plasma into H-mode is still a subject of active research. However, from a practical point of view, it is possible to identify the confinement mode from a set of conventional plasma diagnostic signals. On the one hand, this represents an important tool for real-time plasma control, e.g. in ITER, in order to stabilize the complex plasma and magnetic configurations and maximize the performance [8, 5, 7]. On the other hand, a data-driven study of the primary physical variables that determine the confinement regime can contribute substantially to the physical understanding of plasma confinement.

Recently, machine learning and pattern recognition methods have shown a substantial value for such data-driven investigations by extracting patterns of interest from fusion data; see e.g. [9] for a recent overview at the JET tokamak. However, pattern recognition for fusion data is a veritable challenge, for several reasons that are discussed in this paper. In this work, we present a novel integrated framework that tackles the various pattern recognition challenges related to fusion data and we apply this to the conformal prediction of plasma confinement regimes. The application of conformal predictors in fusion data analysis was considered before, e.g. in [10]. Our framework involves a representation of complex stochastic data sets that allows an efficient recognition of interesting data patterns and to adapt pattern recognition methods to maximally profit from this representation. Adopting a *probabilistic* approach, the framework is especially suitable for analyzing measurements that are subject to a great deal of uncertainty, such as it can be the case in fusion devices. Hence, we respect the inherent probabilistic nature of the data by developing pattern recognition methods that are able to deal with probability distributions.

Pattern recognition essentially relies on geometric concepts such as distance, dispersion and projection. Therefore, our proposed method is based on *information geometry*, which provides a natural geometric structure to probability spaces. In information geometry a metric tensor on probabilistic manifolds is defined, allowing the calculation of a geodesic distance between probability distributions.

In this paper, we present results from classification experiments using a conformal predictor based on the geodesic distance geometry of the data and we compare to the Euclidean distance. We first demonstrate the main principles on a synthetic data set and then apply this to confinement mode classification in an international database. A key observation in this work is that, in addition to being the proper formalism for treating uncertainty, the probabilistic description of data actually *improves* the performance of classification and visualization techniques. 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. This, in turn, is essential for extrapolating present design characteristics and aspects of plasma performance to next-step devices such as ITER and fusion reactors.

This paper is organized as follows. In Section 2 we present the motivation for our probabilistic-geometric framework and the details of the approach. Section 3 discusses the application of conformal predictors in relation to our modeling framework, to confinement mode discrimination. Section 4 concludes the paper.

2 A geometric-probabilistic pattern recognition framework

2.1 Pattern recognition for fusion data

Pattern recognition for fusion data is hampered by several data characteristics. First, the databases are vast and learning patterns from large data amounts often requires considerable computational resources. Second, fusion plasmas are described by tens to hundreds of variables. Therefore, it is essential to reduce the data dimensionality, for instance by projecting the data into a lower-dimensional space, preferably with a minimum of distortion. Third, there is a considerable redundancy between measured quantities due to complex, *nonlinear* interactions. Fourth, measurements in fusion devices are often affected by substantial uncertainties, both stochastic and systematic. These may be introduced at various levels, from the measurement hardware to the calibration, incomplete or inaccurate physical modeling assumptions and inherently stochastic plasma behavior. Classic pattern recognition or statistical methods deal with this uncertainty in a way that somewhat decouples the measurement value from its fundamental uncertainty property. That is, the primary object of interest is usually considered to be the measurement value itself, while its associated error bar or, more comprehensively, its probability distribution, is treated as a side-effect of the measurement process, which ultimately influences the reliability of inferences. In contrast, in the approach presented in the present work, the fundamental object is the probability distribution itself, taking full advantage of the extra information that it contains over the measurement value alone. Information on the probability distribution that underlies a measurement can be obtained by fitting a distribution to a set of repeated measurements. Alternatively, error analysis can be applied to derive a measure for the dispersion of the underlying probability distribution. In some cases Gaussian error propagation suffices, but if the forward model is complicated, possibly involving multiple diagnostics, a more advanced error analysis may be appropriate. A suitable framework to do this is integrated data analysis (IDA) using Bayesian probability theory [3, 11], which has become a well-trusted tools in fusion data analysis. Dependencies between physical variables can be captured in a multivariate distribution, although many nonlinear relations are too complex to be efficiently described by tractable probability models. Such relations are better characterized by a dedicated regression analysis.

Finally, in view of the stochastic character of fusion data and data patterns, it is important to associate reliable estimates of confidence or reliability to the results of pattern recognition tasks, such as classification and regression. This requirement can be adequately fulfilled by conformal predictors, which provide a natural measure of confidence and credibility and are based on the straightforward and informative concept of a nonconformity measure.

2.2 The geometry of probability distributions

For the purpose of the classification and visualization methods presented in this work, a notion of similarity between data points is needed. In a probabilistic description this translates to a similarity measure between probability distributions. This can be obtained within the field of *information geometry*, where probability density families are interpreted as (Riemannian) differentiable manifolds [1]. A point on the manifold corresponds to a specific probability density function (PDF) within the family and the family parameters provide a coordinate system on the manifold. The Fisher information can be regarded as a metric tensor (Fisher-Rao metric) on such a manifold. Once the metric is known, one can establish and solve the geodesic equations, allowing the calculation of the *geodesic distances* on the manifold [13, 12]. Thus, the geodesic distance (GD) is a natural and theoretically motivated similarity measure between probability distributions. Given a probability model $p(\boldsymbol{x}|\boldsymbol{\theta})$ for a vector-valued variable $\boldsymbol{x}$, labeled by an N-dimensional parameter vector $\boldsymbol{\theta}$, the components of the Fisher information matrix $g_{\mu\nu}$ are defined through the relations

$$g_{\mu\nu}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^\mu \partial\theta^\nu} \ln p(\boldsymbol{x}|\boldsymbol{\theta})\right], \quad \mu, \nu = 1 \ldots N.$$

2.3 The geometry of the univariate Gaussian distribution

In this paper we model the data using a simple univariate Gaussian model. The Fisher-Rao metric for the Gaussian distribution, parameterized by its mean μ and standard deviation σ, can be given via the quadratic line element [2]:

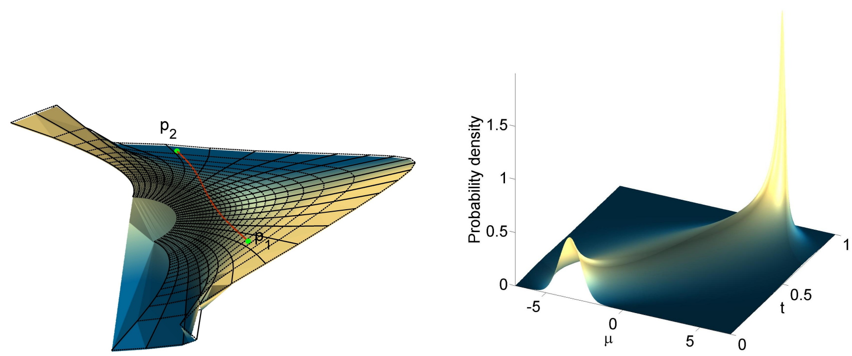
$$\mathrm{d}s^2 = \frac{1}{\sigma^2}\mathrm{d}\mu^2 + \frac{1}{\sigma^2}\mathrm{d}\sigma^2.$$

A closed-form expression exists for the GD, permitting a fast evaluation. Indeed, for two univariate Gaussian distributions $p_1(x|\mu_1, \sigma_1)$ and $p_2(x|\mu_2, \sigma_2)$, parameterized by their mean μ_i and standard deviation σ_i $(i = 1, 2)$, the GD is given by [2]

$$\mathrm{GD}(p_1||p_2) = \sqrt{2}\ln\frac{1+\delta}{1-\delta}, \quad \delta \equiv \left[\frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2 + 2(\sigma_1+\sigma_2)^2}\right]^{1/2}.$$

For illustration purposes, an (approximately) isometric embedding of the Gaussian manifold in three-dimensional Euclidean space is shown in Figure 1a. Every point on this surface represents a Gaussian distribution, characterized by a mean and a standard deviation. The Euclidean distance in the three-dimensional Euclidean space between any two points on the surface, equals the true GD between the corresponding distributions. An example of a geodesic between two arbitrary Gaussians is drawn on the surface and the evolution of the distribution along the geodesic is visualized in Figure 1b.

Finally, in the case of multiple independent Gaussian variables it is easy to prove that the square GD between two sets of products of distributions is given by the sum of the squared GDs between corresponding individual distributions [2].



(a) (b)

Fig. 1: (a) Embedding of the univariate Gaussian manifold and geodesic between two arbitrary Gaussians p_1 ($\mu_1 = -4$, $\sigma_1 = 0.7$) and p_2 ($\mu_2 = 3$, $\sigma_2 = 0.2$). The full lines are curves of constant mean, the dashed lines are curves of constant standard deviation. (b) Visualization of the distributions along the geodesic, parameterized by t. Each slice along the t-axis shows the distribution at the corresponding point on the geodesic.

3 Conformal prediction for confinement regime identification

We now consider the application of a conformal predictor in conjunction with our geometric-probabilistic modeling framework to the classification of confinement regimes. We are interested in distinguishing between L-mode and H-mode plasmas, so in this paper we consider a two-class classification problem. A subset of the data of size $n-1$ was used for 'training' the classifier, i.e. the class label of this bag of samples was assumed to be known. In all experiments we used a conformal predictor with a nonconformity measure based on the distance of the sample to be classified to its nearest neighbors of both classes in the bag. Specifically, the nonconformity score α_i of a given sample i was calculated as

$$\alpha_i = \frac{\text{Distance to sample } i\text{'s nearest neighbor in the bag with the same label}}{\text{Distance to sample } i\text{'s nearest neighbor in the bag with a different label}}.$$

The nonconformity measure for sample i was computed with respect to both classes, assuming membership of sample i of each of the classes $j = 1, 2$ in turn. By doing this for each sample in the bag and for the sample to be classified, a ranking could be determined of the nonconformity scores. Then, for each class j a p-value was calculated based on this ranking, namely:

$$p_j = \frac{\#\{i = 1, \ldots, n | \alpha_i \geq \alpha_n\}}{n}.$$

Here, it is assumed that the index n corresponds to the sample to be classified, while $i < n$ refers to the samples in the bag used for 'training'. The sample to be classified was assigned to the class with the largest corresponding p-value. The largest p-value itself is referred to as the *credibility*, while the complement of the other p-value is the *confidence* of the classification task. This mode of classification is the conformal predictor equivalent of a k-nearest-neighbor classifier with $k = 1$. As a similarity measure we first considered the Euclidean distance between the sample feature vectors and we compared its performance to that of the geodesic distance between the feature vectors, this time treated as probabilistic features.

3.1 ITPA database

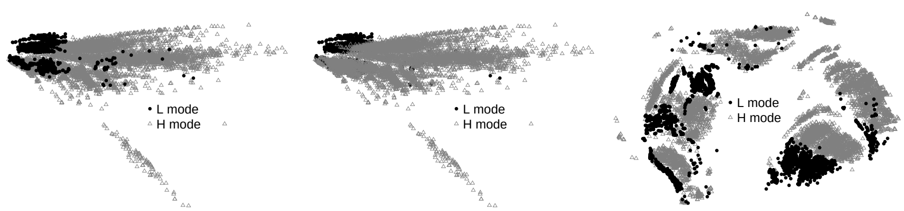
In this work for confinement mode identification we employed measurements from the International Tokamak Physics Activity (ITPA) Global H-mode Confinement Database (DB3, version 13f), henceforth referred to as the 'ITPA database' [6, 4]. The ITPA database contains more than $10,000$ validated measurements of various global plasma and engineering variables at one or several time instants during discharges in 19 tokamaks. 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, vacuum toroidal magnetic field, total power loss from the plasma, central line-averaged electron density, plasma major radius, plasma minor radius, elongation and effective atomic mass. The objective is then to extrapolate to ITER conditions. We used the same eight variables to discriminate between, roughly, L- and H-mode plasmas. Specifically, all database entries with a confinement mode labeled as H, as well as related high-confinement regimes labeled HGELM, HSELM, HGELMH, HSELMH and LHLHL, were considered to belong to the H-mode class. In contrast, discharges labeled with L, OHM and RI were assigned to the non-H-mode class, or L-mode for brevity.

The ITPA database lists typical error estimates of measurements for the various plasma and engineering variables. It should be noted that this represents very limited information on the probability distribution underlying each quantity. Furthermore, the interpretation of the error estimates is not always unambiguous and in some cases it is not clear to what extent the estimates are sufficiently reliable for subsequent analysis. Nevertheless, let us assume for the moment that the error bars pertain to a statistical uncertainty in the data, specifically that they represent a single standard deviation. According to the principle of maximum entropy the underlying probability distribution is Gaussian with mean the measurement itself and standard deviation the error bar. Let us also suppose that, for stationary plasma conditions, all variables are statistically independent and so the joint distribution factorizes. Put differently, the joint distribution for the eight variables mentioned above is assumed to be just the product of the individual univariate Gaussian distributions. Clearly, this is a strong assumption and it is imposed here mainly for keeping the calculations tractable. It is also important to note that our 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.

3.2 Data visualization

A first step towards the identification of patterns in the ITPA database consists of 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. This is done using metric multidimensional scaling (MDS), searching for a configuration of points in the Euclidean plane yielding minimal distortion of all pairwise distances.

Figure 2 shows two approximately isometric projections of the ITPA data into the Euclidean plane, obtained via MDS. For Figure 2a (and 2b) 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 2c 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.



(a) (b) (c)

Fig. 2: 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.

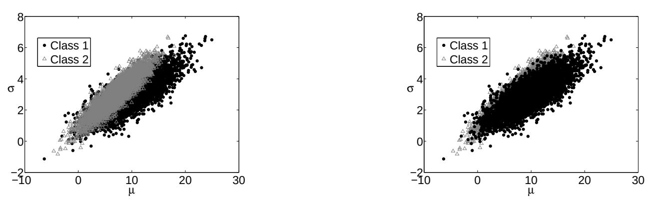
3.3 Confinement mode classification

We next show the results of a series of classification experiments with two classes (L-mode and H-mode). The first experiment applies the previously discussed concepts and methods to a synthetic data set. In a second experiment, actual measurements are used.

Synthetic data. The synthetic data set consisted of two clusters of distributions on a univariate Gaussian probabilistic manifold. Specifically, a first cluster of 5000 Gaussian distributions was created, each distribution with its mean $\mu_{1,i}$ and standard deviation $\sigma_{1,i}$ (the subscript 1 refers to the label of the cluster). The values $\mu_{1,i}$ and $\sigma_{1,i}$, in turn, were sampled from a bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$, modeling the distribution of the cluster points. A relatively strong correlation was assumed between the mean and standard deviation of the univariate Gaussians, mimicking the situation in the ITPA database. Similarly, the 5000 univariate Gaussians in the second cluster were drawn from a bivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. We used the following parameters:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 10 \\ 3 \end{bmatrix}, \qquad \Sigma_1 = \begin{bmatrix} 16 & 3.2 \\ 3.2 & 1 \end{bmatrix},$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} 7 \\ 3 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} 9 & 2.7 \\ 2.7 & 1 \end{bmatrix}.$$

The samples belonging to both clusters are plotted in Figure 3 (necessarily in a Euclidean space). The classification task is relatively difficult, due to the strong overlap of the classes in the feature space.



(a) (b)

Fig. 3: (a) Synthetic data points for a classification problem involving two classes. (b) The same, but with the points belonging to the second class on top, for better visibility.

A bag of samples was created, consisting of a random part of 5% of the data, for which the label was assumed to be known (training samples). The remaining part of the data was used for testing the classifier. The success rates, credibility and confidence were calculated and are shown in Table 1. The success rates are given for each confinement mode and their average is also mentioned. For comparison, we also performed the experiments using regular kNN classification ($k = 1$) and the results are given in a similar fashion in Table 1. It can be observed that higher rates are obtained with the GD, compared to the Euclidean distance. In addition, the conformal predictor obtains slightly higher success rates compared to the kNN classifier. The confidence levels are generally high, but the credibility is relatively low, owing to the considerable overlap of the classes in the configuration space. Finally, we observed that higher values of k slightly deteriorated the results.

Measured data. The classification experiment for the measured data was implemented in a similar way as the experiment using synthetic data. Again, 5% of the data was used for the bag of training samples and the results are also shown in Table 1. Similar trends can be observed as in the experiment with synthetic data.

4 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. A reliable analysis of error propagation is clearly not only essential for a correct interpretation of measurement results, but is also very useful for pattern recognition. Conformal predictors provide an excellent means for, in turn, assessing the reliability of pattern recognition results for fusion data. It is remarkable that even the approximate and extremely limited information in the ITPA database on the underlying probability distribution is beneficial to the classification task. In our experiments this leads to noticeable but relatively small differences in classification success rates. However, it is important to note that the experiments indicate that more precise information regarding the nature of the error bars, as well as the possibly non-Gaussian probability distribution, could be very useful for optimizing the classification. Hence, future work will concentrate on obtaining such information in this and other applications in fusion data analysis. Our results also suggest a strong potential of our framework for regression, which has an important application for establishing scaling laws in fusion.

References

1. Amari, S., Nagaoka, H.: Methods of information geometry, Transactions of mathematical monographs, vol. 191. American Mathematical Society, New York (2000)
2. Burbea, J., Rao, C.: Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivariate Anal. 12(4), 575–596 (1982)
3. Fischer, R., Dinklage, A., Pasch, E.: Bayesian modelling of fusion diagnostics. Plasma Phys. Control. Fusion 45(7), 1095–1111 (2003)
4. http://efdasql.ipp.mpg.de/hmodepublic
5. Lukianitsa, A., Zhdanov, F., Zaitsev, F.: Analyses of ITER operation mode using the support vector machine technique for plasma discharge classification. Plasma Phys. Control. Fusion 50(6, article no. 065013) (2008)

Table 1: Success rates (SR) for the classification of synthetic and real data using kNN and a conformal predictor (CP) based on Euclidean and geodesic distance measures. The mean credibility (CR) and confidence (CO) are also mentioned, with the respective standard deviations in parentheses.

			SR (%)					
			kNN		CP			
Experiment	Distance	Mode	By mode	Average	By mode	Average	CR	CO
Synthetic	Euclidean	L	67.7	69.2	68.2	70.0	0.50	0.85
		H	70.6		71.8		(0.18)	(0.09)
	GD	L	68.9	70.2	69.4	71.2	0.50	0.85
		H	71.4		72.9		(0.18)	(0.09)
Real	Euclidean	L	93.2	92.0	94.9	93.2	0.38	0.99
		H	90.7		91.5		(0.25)	(0.02)
	GD	L	94.3	93.4	95.8	94.6	0.37	0.99
		H	92.4		93.4		(0.25)	(0.02)

6. McDonald, D., et al.: Recent progress on the development and analysis of the ITPA global H-mode confinement database. Nucl. Fusion 47(3), 147–174 (2007)
7. Meakins, A., McDonald, D., JET-EFDA Contributors: The application of classification methods in a data driven investigation of the JET L-H transition. Plasma Phys. Control. Fusion 52(7, article no. 075005) (2010)
8. Murari, A., Vagliasindi, G., Zedda, M., Felton, R., Sammon, C., Fortuna, L., Arena, P., JET-EFDA Contributors: Fuzzy logic and support vector machine approaches to regime identification in JET. IEEE Trans. Plasma Sci. 34(3), 1013–1020 (2006)
9. Murari, A., Vega, J., Mazon, D., Rattà, G., Svensson, J., Palazzo, S., Vagliasindi, G., Arena, P., Boulbé, C., Faugeras, B., Fortuna, L., Moreau, D., JET-EFDA Contributors: Innovative signal processing and data analysis methods on JET for control in the perspective of next-step devices. Nucl. Fusion 50(5, art. no. 055005) (2010)
10. Vega, J., Murari, A., Pereira, A., González, S., Pastor, I.: Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson scattering. Rev. Sci. Instrum. 81(10, article no. 10E118) (2010)
11. Verdoolaege, G., Fischer, R., Van Oost, G., JET-EFDA Contributors: Potential of a Bayesian integrated determination of the ion effective charge via bremsstrahlung and charge exchange spectroscopy in tokamak plasmas. IEEE Trans. Plasma Sci. 38(11), 3168–3196 (2010)
12. Verdoolaege, G., Scheunders, P.: Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination. Int. J. Comput. Vis. 95(3), 265–286 (2011)
13. Verdoolaege, G., Scheunders, P.: On the geometry of multivariate generalized Gaussian models. J. Math. Imaging Vis. 43(3), 180–193 (2011)