Classification and dimensionality reduction of international tokamak confinement data on a probabilistic manifold

Geert Verdoolaege, Giorgos Karagounis and Guido Van Oost
Department of Applied Physics, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

Abstract

Pattern recognition for fusion data greatly contributes to a better understanding of the measurements and the physics of fusion plasmas. Through a geometric description of probability it is shown that consideration of the inherent uncertain nature of the data significantly improves the visualization of global confinement data and the identification of confinement regimes. The framework can be extended to the development of scaling laws for ITER.

1. Introduction

Pattern recognition techniques are most useful in nuclear fusion research for learning structures of interest directly from the data, either off-line or in real time. This paper addresses the visualization of tokamak plasma confinement data through dimensionality reduction and the automated identification of confinement regimes. Our description of the data is intrinsically probabilistic and we use a geometric framework to study the probability distributions. The geometric description of probability has many applications, for instance in image texture analysis [1] and for the prediction of disruptions in tokamak plasmas [2]. The key observation in this work is that through the inclusion of probabilistic information, the performance of data visualization and classification algorithms is drastically improved. The techniques underlying our classification system can be adopted for the real-time recognition of confinement modes. However, the final objective of the present work is different, namely the development of scaling laws for predicting the characteristics of ITER plasma confinement.

2. Geometric-probabilistic description of global confinement data

2.1. Probabilistic nature of global confinement data

Measurement uncertainty is a fundamental property, rather than a side-effect, of the measurement process, and it should be taken advantage of. As such, a measurement can be regarded as a sample from a latent probability distribution. Here we employ measurements from the International Tokamak Physics Activity (ITPA) Global H-mode Confinement Database (DB3) [4, 5]. The 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.

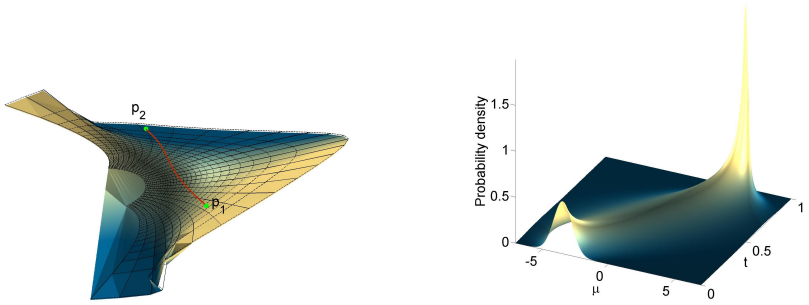
Let us assume for now that the error bars pertain to a statistical uncertainty in the data, specifically that they represent a single standard deviation. According to the principle of maximum entropy the underlying probability distribution is Gaussian with mean the measurement itself and standard deviation the error bar. Let us also suppose that, for stationary plasma conditions, all variables are statistically independent and so the joint distribution factorizes.

2.2. Geometry of the univariate Gaussian distribution

For the purpose of dimensionality reduction and classification of global confinement data, we need a notion of similarity between data points. In a probabilistic description this translates to a similarity measure between probability distributions. In the framework of *information geometry* a family of probability distributions forms a Riemannian manifold with the Fisher information playing the role of a unique metric tensor [6]. The coordinates on the manifold are the parameters of the distribution family. Given a metric, one can calculate geodesics and geodesic distances (GDs) on the manifold. 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 [7]

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

An (approximately) isometric embedding of the Gaussian manifold in three-dimensional Euclidean space is shown in Figure 1a, with an example geodesic drawn between two arbitrary Gaussians. The evolution of the distribution along the geodesic is visualized in Figure 1b.



(a) (b)

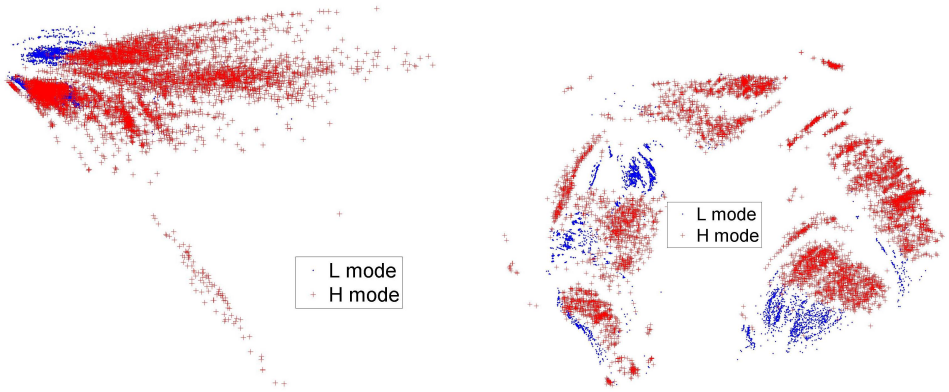
Figure 1: (a) Embedding of the univariate Gaussian manifold and geodesic between two arbitrary Gaussians p_1 ($\mu_1 = -4$, $\sigma_1 = 0.7$) and p_2 ($\mu_2 = 3$, $\sigma_2 = 0.2$). The full lines are curves of constant mean, the dashed lines are curves of constant standard deviation. (b) Visualization 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.

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 [7].

3. Visualization and classification of confinement data

3.1. The DB3 database

The DB3 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 (P_{loss}), central line-averaged electron density ($\bar{n}_{\mathrm{e}}$), plasma major radius, plasma minor radius, elongation and effective atomic mass. The objective is 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, HGELM, HSELM, HGELMH, HSELMH and LHLHL were considered to belong to the H-mode class, while discharges labeled with L, OHM and RI were assigned to the non-H-mode class, or L-mode for brevity.



(a) (b)

Figure 2: Two-dimensional isometric embeddings of the DB3 data with indicated L- and H-mode clusters. (a) Using the Euclidean distance without measurement error. (b) Using the GD with measurement error.

3.2. Data visualization

A first step towards the identification of patterns in the DB3 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 [8].

Figure 2 shows two approximately isometric projections of the DB3 data into the Euclidean plane, obtained via MDS. For Figure 2a 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 2b is based on geodesic distances between Gaussian product distributions. It can be clearly noticed that the projections obtained with the geodesic distance, which take into account the measurement error, exhibit considerably more structure compared to the Euclidean case. In particular, it is much more easy to visually discriminate between the L- and H-mode clusters. This suggests an important potential of our framework for regression, which is another form of structure or pattern recognition.

Mode	Correct classification rate			
	Euclidean no errors	Euclidean with errors	GD with errors	GD with randomized errors
L	85.1	87.7	91.0	82.4
H	88.6	89.4	93.0	86.0

Table 1: Correct classification rates (%) of confinement regimes using a kNN classifier.

3.3. Confinement mode classification

We next performed a series of classification experiments in order to discriminate between L- and H-mode plasmas. A random sample of 5% of the database was taken as training data for which the class label (L or H) was assumed to be known. We used a simple k-nearest-neighbor (kNN) classifier with $k = 1$, thus assigning a sample to the class that has the nearest element to the sample. This is where the notion of a distance measure between samples comes in, which can be the Euclidean or the geodesic distance. By using all eight variables we obtained little differentiation between classification results with or without inclusion of the uncertainty information. Therefore, in order to more clearly show the benefit of our method in the case of confinement mode classification, we present results obtained using only measurements of $\bar{n}_e$ and P_{loss}; see Table 1. The correct classification rate for both classes is clearly better if the measurement error is considered, even using the Euclidean distance. The best results are obtained with the GD, since it properly takes into account the geometry of the probabilistic manifold. The results of a final experiment are also shown, wherein randomized values of the error bars were used, although still within the same range as before. In particular, if a certain error bar that is given in the database was $x.y \times 10^z$ ($x < 10$), then the corresponding randomized error bar was taken as $u \times 10^z$, with u a uniformly sampled number from the interval $[1, 10]$. This proves that it is indeed the specific uncertainty information mentioned in the database that contains useful information.

4. Conclusion

In this paper we have shown that error estimates for global confinement data contain valuable information for pattern recognition tasks. We have demonstrated this through dimensionality reduction for data visualization and confinement mode classification. It is remarkable that even

the approximate and limited information in the DB3 database on the underlying probability distribution is already beneficial for data visualization and classification. We therefore advocate the necessity of obtaining at all times reliable estimates of measurement uncertainty through a dedicated error analysis. The present work suggests an important potential of the geometric-probabilistic framework for regression on probabilistic manifolds, with the aim to formulate scaling laws for ITER that respect the inherent probabilistic nature of the data.

References

[1] G. Verdoolaege, P. Scheunders, Int. J. Comput. Vis. **95**, 265 (2011)
[2] G. Verdoolaege, G. Karagounis, A. Murari, J. Vega, G. Van Oost, JET-EFDA Contributors, Fusion Sci. Technol. **62**, 356 (2012)
[3] G. Verdoolaege, R. Fischer, G. Van Oost, JET-EFDA Contributors, IEEE Trans. Plasma Sci. **38**, 3168 (2010)
[4] D.C. McDonald *et al.*, Nucl. Fusion **47**, 147 (2007)
[5] http://efdasql.ipp.mpg.de/HmodePublic
[6] S. Amari, H. Nagaoka, 'Methods of information geometry,' American Mathematical Society (2000)
[7] J. Burbea, C.R. Rao, J. Multivariate Anal. **12**, 575 (1982)
[8] A.J. Izenman, 'Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning,' Springer (2008)