Relating Probability Distributions Using Geodesic Least Squares Regression: Application to Edge-Localized Modes in Fusion Plasmas

Geert Verdoolaege[1,2,a)], Aqsa Shabbir[1,3] and JET Contributors*

EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK
[1]*Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium*
[2]*Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium*
[3]*Max Planck Institute for Plasma Physics, D-85748 Garching, Germany*

[a)]Corresponding author: geert.verdoolaege@ugent.be

Abstract. Geodesic least squares regression (GLS) is a new robust, but simple regression technique based on minimization of the Rao geodesic distance on a probabilistic manifold. It is particularly useful in the presence of large or unknown sources of uncertainty, as it relates probability distributions rather than individual measurements. The GLS method is employed here to estimate the dependence between the probability distributions of two important characteristics of a repetitive instability occurring in the boundary region of fusion plasmas, namely the edge-localized mode (ELM). Specifically, we study the relation between the plasma energy loss following an ELM and the time since the previous ELM. GLS is shown to produce consistent results, whether using measurements on individual ELMs or averaged quantities, even in the presence of questionable modeling assumptions. The method is illustrated using the pseudosphere as an intuitive model of the Gaussian manifold.

INTRODUCTION

Regression analysis traditionally mostly focuses on estimating trends in (multidimensional) Euclidean data spaces. Lately, regression in various types of non-Euclidean spaces has been explored as well. The recently developed method of geodesic least squares regression (GLS) aims to seek parametric dependencies between probability distributions on a probabilistic manifold [1]. This might appear as an unnecessary complication at first, since regression analysis is itself a probabilistic technique and there seems to be no reason to add the extra modeling level involving a probabilistic manifold. However, as shown in [1], and briefly reviewed and expanded upon in this paper, regression on probabilistic manifolds has the potential to increase the robustness of standard regression analysis. Here, the term 'robustness' is used to indicate that the GLS method produces consistent results, with a relatively weak dependence on questionable model assumptions. It does this by taking the variability (uncertainty) on the regression variables into account in a different way, compared to traditional methods, thereby increasing the flexibility. In this sense, GLS is a relatively straightforward extension of ordinary least squares regression (OLS) from a Euclidean data space to a Riemannian probabilistic manifold.

In this paper, a regression problem is addressed with relevance to magnetic confinement fusion. Specifically, the relation is studied between two plasma properties associated to a type of instability that is very common in the boundary plasma of fusion machines of the tokamak design, namely the edge-localized mode, or ELM [2]. This repetitive instability causes expulsion of particles and energy from the plasma, hence posing a potential threat to the integrity of various material components facing the plasma. Here, we estimate the relation between the loss of plasma energy following an ELM and the time since the previous ELM. The latter will be referred to as the *waiting time*. Previous studies already pointed out the relation between these quantities [3], but recent work suggests that the

*See the author list of "Overview of the JET results in support to ITER" by X. Litaudon et al., to be published in Nuclear Fusion special issue: Overview and summary reports from the 26th Fusion Energy Conference (Kyoto, Japan, 17–22 October 2016)

correlation is not as strong as suspected before [2]. This appears to be primarily due to the high variability of energies and waiting times from one ELM to another, even under stationary plasma conditions.

The goal of this paper is, first, to show that common regression techniques can yield inconsistent estimates of the scaling parameters in the present application, depending on the way the data are represented. This can be caused by incorrect model assumptions, e.g. due to an unexpected uncertainty source of a stochastic or systematic nature, or an inadequate regression model. In contrast, GLS is shown to be relatively insensitive to specific assumptions about data and model uncertainties. As a result, the method is able to uncover trends between highly uncertain quantities, with minimal modeling assumptions. This is because, on the one hand, GLS relates the probability distributions of the quantities of interest, rather than their averages, as is the usual practice with standard regression analysis. On the other hand, GLS leaves room for unexpected uncertainty sources. A second objective of the paper is to stress the importance of a probabilistic description of stochastic plasma phenomena (e.g. ELMs), as opposed to analyzing only averages over multiple occurrences. We show that average trends may draw an overly simplistic picture, hiding the practical consequences of the variability of the phenomenon.

In order to clarify the benefits of GLS regression, in this paper we use a simple but intuitively effective model for the manifold of univariate Gaussian distributions, namely the *pseudosphere*. By means of the pseudosphere model, we illustrate several aspects of collections of distributions, as well as the outcome of a regression analysis. Thus, we first provide a short overview of GLS regression and we introduce the pseudosphere model together with some illustrations for facilitating interpretation of the model. Then we introduce the application related to regression between ELM-related properties in fusion plasmas. We demonstrate the enhanced robustness of GLS compared to OLS, again using the pseudosphere for illustration purposes.

GLS AND THE PSEUDOSPHERE

Geodesic least squares regression

The simplest form of GLS regression was detailed in [1], involving a linear model. In general, for a single response variable y, GLS regression is a generalization of OLS, where observations of y and the predictions by the regression model are replaced by a corresponding observed and modeled probability distribution p_{obs}, resp. p_{mod}. p_{mod} is the distribution of y, conditional on n measurements x_{ij} of p predictor variables x_j, assuming that the data strictly follow the regression model (deterministic and stochastic components) ($i = 1, \ldots, n$, $j = 1, \ldots, p$). On the other hand, p_{obs} aims to characterize the measurements y_i of y by making as few assumptions as possible, in an attempt to 'let the data speak for themselves'.

In this paper, we choose a linear model and all variables are assumed to be affected by independent Gaussian noise. In the application to ELMs in fusion plasmas, the independent Gaussian assumption will turn out to be a rather crude approximation, which can be improved at a later stage. We also assume that some estimate of the standard deviations is available from experiments. This could be modeled as prior knowledge in a Bayesian framework, but for the time being we regard the estimates as exact numbers.

Provided the assumptions of the model are true, the distribution of the response variable y, corresponding to measurement i, conditional on its standard deviation σ_y, as well as the mutually independent measurements x_{ij}, their standard deviations $\sigma_{x,j}$ and the coefficients β_k ($k = 0, \ldots, p$) of the linear model (β_0 is the cut-off), is given by

$$p_{\text{mod}}\left(y|\{x_{ij}\},\{\beta_k\},\sigma_{y_i},\{\sigma_{x_{ij}}\}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\text{mod},i}} \exp\left[-\frac{(y-\mu_{\text{mod},i})^2}{2\sigma^2_{\text{mod},i}}\right], \qquad \begin{array}{l} i = 1, \ldots, n, \\ j = 1, \ldots, p, \\ k = 0, \ldots, p. \end{array} \tag{1}$$

Here, we have defined

$$\mu_{\text{mod},i} \equiv \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{2}$$

$$\sigma^2_{\text{mod},i} \equiv \sigma^2_{y_i} + \beta_1^2\sigma^2_{x_{i1}} + \ldots + \beta_p^2\sigma^2_{x_{ip}}. \tag{3}$$

Notice that, to provide for some generality, we have allowed the standard deviations to be different from one measurement to another. For instance, in many applications the error bar on a measurement corresponds to a fixed relative error (a percentage error). In that case the error bar itself, e.g. a standard deviation, is proportional to the measurement and is therefore different for every point. For a normally distributed measurement we then have, in the case of the response

variable (and similar for the predictor variables): $\sigma_{y,i} = r_y|y_i|$, with r_y the constant relative error. Furthermore, (3) implies that, according to the model, the uncertainty on the predictor variables propagates through the linear model expression (2), hence contributing to $\sigma_{\text{mod},i}$. This is strictly only justified for independent predictor variables, although a generalization for dependent variables could be envisaged.

We wish to work on the Gaussian probabilistic manifold, therefore we choose a normal distribution for the observed distribution p_{obs} as well. In fact, just as (2) and (3) predict a distinct Gaussian model for each data point, we assign an individual $p_{\text{obs},i}$ to each point, with mean given by the measurement y_i and unknown standard deviation $\sigma_{\text{obs},i}$, to be estimated from the data. Again, in principle, $\sigma_{\text{obs},i}$ can be different for each point, although in practice it is clear that we will need to introduce some regularization to render the model identifiable. For instance, in case of a fixed relative error r_{obs} we should write $\sigma_{\text{obs,i}} = r_{\text{obs}}|y_i|$ (more complicated relations between σ_{obs} and the response variable would be possible too). In this paper we regularize by assuming a common r_{obs} for all data points, to be estimated from the data.

Hence, each measurement of the response variable y is regarded as a normal probability distribution, with mean given by that measurement. This is an essential difference with standard regression methods, as will be clarified later. Moreover, the extra parameters $\sigma_{\text{obs},i}$ provide the method with additional flexibility, since each $\sigma_{\text{obs},i}$ is not *a priori* required to equal the corresponding $\sigma_{\text{mod},i}$. As a result, unexpected sources of uncertainty can be accommodated more easily, compared to standard regression methods [1].

Next, similar to the well-known class of minimum distance estimation methods, we minimize a similarity measure between the observed and modeled distribution, for each data point. The final key ingredient of the GLS method lies in the specific choice of similarity measure, which is the *Rao geodesic distance* (GD) based on the Fisher information metric on the probabilistic manifold associated to the univariate normal distribution [1]. Owing to the independence assumption of the measurements, we can write this in terms of products of the corresponding marginal distributions:

$$\begin{aligned}\left\{\hat{\beta}_k, \hat{r}_{\text{obs}}\right\} &= \underset{\beta_k, r_{\text{obs}} \in \mathbb{R}}{\text{argmin}}\, \text{GD}\left[\prod_{i=1}^{n} p_{\text{obs}}\left(y|y_i, r_{\text{obs}}\right), \prod_{i=1}^{n} p_{\text{mod}}\left(y|\{x_{ij}\}, \{\beta_k\}, \sigma_{y_i}, \{\sigma_{x_{ij}}\}\right)\right] \\ &= \underset{\beta_k, r_{\text{obs}} \in \mathbb{R}}{\text{argmin}} \sum_{n=1}^{n} \text{GD}^2\Big[p_{\text{obs}}\left(y|y_i, r_{\text{obs}}\right), p_{\text{mod}}\left(y|\{x_{ij}\}, \{\beta_k\}, \sigma_{y_i}, \{\sigma_{x_{ij}}\}\right)\Big]. \end{aligned} \tag{4}$$

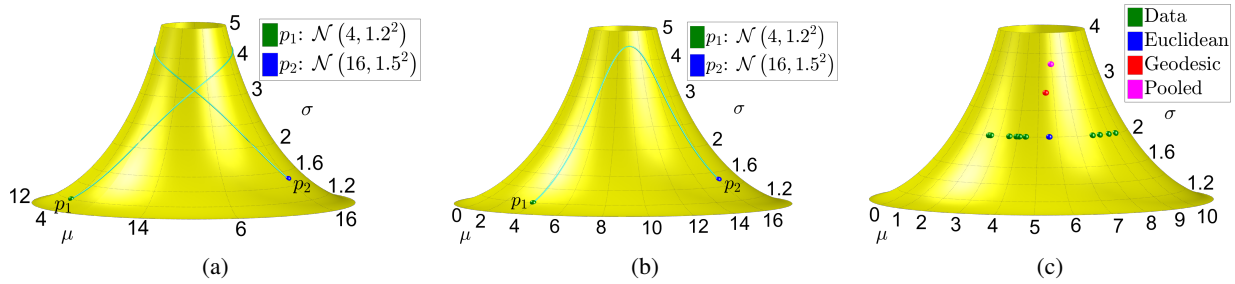
The last equality entails a considerable simplification, thanks to the property that the squared GD between products of distributions can be written as the sum of squared GDs between the corresponding factors [4]. Conveniently, an analytic expression exists for the (Rao) GD between univariate normal distributions [1]. Note that the regression parameters β_k occur both in the mean and the variance of the modeled distribution, as per (2) and (3).

The reformulation of classic regression analysis in terms of regression on a probabilistic manifold implies a certain recursiveness. Indeed, at the top level one can consider the distribution of the data as a whole on the manifold. However, when zooming into a single data point, one finds that it too represents a probability distribution, corresponding to a fixed distribution of the predictor variables.

Regression on the pseudosphere

The geometrical view on regression analysis can be illustrated by visualizing the probabilistic manifold. We here work exclusively with the univariate normal distribution, corresponding to hyperbolic geometry equipped with the Poincaré metric [4]. Various useful immersions (injective) or embeddings (not necessarily injective) of the hyperbolic geometry in Euclidean space have been developed in the past, including the Poincaré half-plane, the Poincaré disk, the Klein disk and a type of pseudosphere called the *tractroid*. We employ the pseudosphere model because its metric induced by the three-dimensional Euclidean metric is precisely the Poincaré metric, although the model is only valid for standard deviations σ greater than 1. In other words, distances on the pseudosphere correspond to distances on the univariate Gaussian manifold, enabling an intuitive understanding of the GLS method.

To get a feeling of the relation between the normal probability distribution and the geometry of the pseudosphere, Figure 1(a) shows (one blade of) the tractroid, with a periodicity of 2π in the mean μ. Two normal distributions are displayed on the surface, $p_1 = \mathcal{N}\left(4, 1.2^2\right)$ and $p_2 = \mathcal{N}\left(16, 1.5^2\right)$, and the geodesic between them is drawn. The geodesic winds around the surface due to the periodicity and for visualization purposes we often compress the surface along the μ coordinate, as in Figure 1(b) (compression factor 1/5). As always in this geometry, the geodesic passes through a region of increased standard deviation relative to that of the points it connects. That this yields the shortest



(a) (b) (c)

FIGURE 1: The pseudosphere (tractroid) ((a) periodicity 2π, (b) periodicity 10π) with two points p_1 and p_2, and the geodesic connecting them. (c) The pseudosphere with 10 normal distributions ('Data'), their Euclidean and geodesic centroids, as well as the point estimated from the pooled data.

route is intuitively clear from the shape of the surface. By way of another example, Figure 1(c) shows 10 normal distributions on the pseudosphere, all with the same standard deviation $\sigma = 2.0$, but with varying means. The average of all means is 5.6 and the corresponding point (same standard deviation) is also plotted ('Euclidean' mean). This is to be contrasted with the Fréchet mean of the distributions, which is the geodesic centroid corresponding to a normal distribution with $\mu = 5.5$ and $\sigma = 2.7$. For comparison, the point corresponding to the average and standard deviation of data obtained from 20 samples from each of the 10 distributions ('pooled data') is also shown ($\mu = 5.9$ and $\sigma = 3.3$).

REGRESSION OF ELM-RELATED PROPERTIES

In this section we apply GLS regression to study the scaling relation between two key properties of ELMs in the JET tokamak. In [3] a linear relation was observed between relative ELM energy loss and normalized ELM frequency. More recently, in [2] a linear dependence between unnormalized plasma energy loss and inter-ELM time was also noticed. However, it is important to mention that these results were obtained for quantities that were averaged over a large number of ELM occurrences, under stationary plasma conditions. Usually the motivation is that this averaging produces more robust results, with smaller error bars. But in the case of ELMs there also lies great importance in the characteristics of individual ELMs. This is because of the significant variability from one ELM to another. As a consequence, each ELM burst may pose a different threat to the plasma-facing components or may influence the plasma properties in a different way. Hence, in addition to studies of average ELM properties, it is also essential to investigate other characteristics of the corresponding probability distributions, such as their variance and skewness. This is a purpose for which GLS regression is particularly well suited, as we intend to demonstrate in this work.

Distributions of ELM properties

We perform the analysis on a subset of the database established in [2]. It concerns a set of 32 plasma pulses (discharges) in JET after it was equipped with wall components fabricated from the same materials that will be used in the future ITER device—the so-called ITER-like wall (ILW). These are plasmas in a high-confinement regime (H-mode) with type I ELM activity. No nitrogen-seeded discharges were considered. In each pulse a time window was selected with stationary plasma conditions and the ELMs in that window were considered. Next, for each ELM the time Δt since the previous ELM was registered, i.e. the waiting time, as well as the measured drop of the plasma stored energy ΔE following the ELM, obtained from magnetic measurements [2]. The latter is taken here as a measure of the energy carried out of the plasma by the ELM, although there may be considerable uncertainty due to eddy currents and plasma motion influencing the measurements during and shortly following the ELM occurrence. This adds to other uncertainty sources, such as intrinsic fluctuation of the ELM characteristics and measurement error. Furthermore, it is known that, at JET, estimates from magnetics of the energy drop for relatively fast ELMs (typically $\Delta t < 10$ ms) become too unreliable to be of practical use. Hence, such fast type I ELMs were excluded from the database.

It should be noted that the database contains measurements under various plasma conditions (different field and current, density, gas fueling, triangularity, etc.), which may contribute to even more uncertainty. Together, this results in a relatively noisy measurement set, in which discovering reliable patterns is a challenge.[1]

[1] There is even a hint of multiple qualitatively different subsets of ELMs, possibly related to faster and slower ELMs, as discussed in [2].

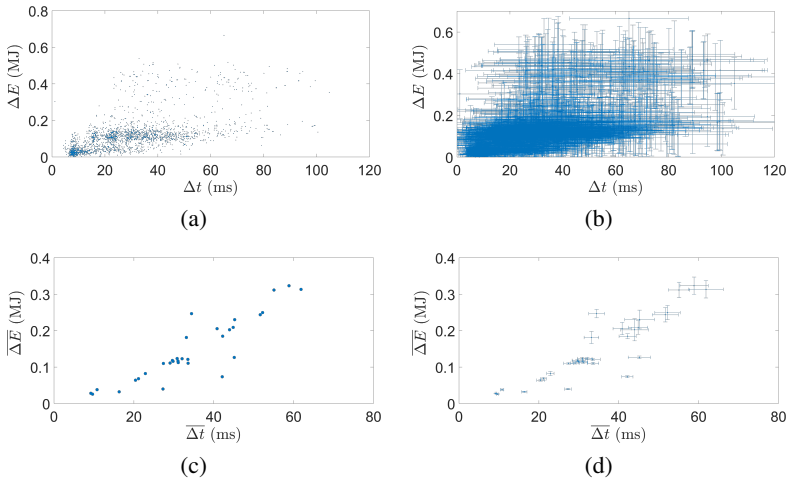


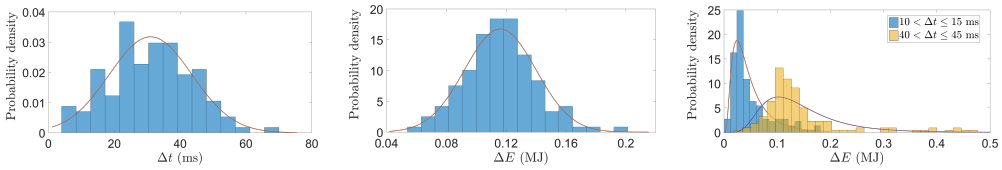
FIGURE 2: (a) Scatter plot of ELM energy ΔE vs. waiting time Δt for the individual ELM bursts in the database. (b) The same plot with error bars added. (c) Scatter plot of average ELM energy $\overline{\Delta E}$ vs. average waiting time $\overline{\Delta t}$ for the discharges in the database. (d) Same with error bars.

Figure 2(a) shows a scatter plot of ELM energy loss ΔE vs. waiting time Δt for all ELMs in the 32 plasmas in the database. Overall, there appears to be a, possibly nonlinear, relation between the two quantities, but there seems to be a saturation for long waiting times. In addition, the scatter of energies is large and its characteristics, i.e. the associated probability distribution, depend on Δt (heteroscedasticity). This becomes even more apparent when adding the error bars (explained in detail below), as in Figure 2(b).

A common approach for dealing with nonlinearly related variables and heteroscedasticity is to transform the data, often to the logarithmic domain. Perhaps a more preferable strategy is to construct a model that can capture the structure of the data, e.g. by means of generalized linear modeling. However, designing an adequate probabilistic model may require considerable skill, particularly for higher-dimensional regression problems. In fact, our purpose here is to show that, even if some of the model assumptions are poorly fulfilled by the data, GLS can still provide consistent estimates of the regression parameters, in contrast with some common established techniques.

Next, we turn to the relation between *average* ELM waiting time and energy, where the averages are taken over all ELMs in each single plasma pulse. This averaging over multiple ELMs is common practice in analyzing experimental results in fusion science. The advantage is that fluctuations related to microscopic physics are averaged out, facilitating the analysis when average trends are of interest. However, averaging is not always desirable when conclusions pertaining to individual ELMs or individual plasmas are sought. For instance, while in [3] and [2] a strong correlation was found between average ELM waiting time and plasma energy loss, under a range of plasma conditions, it was also seen in [2] that the correlation between waiting times and energies corresponding to individual ELM bursts varies across plasmas, and in general is much lower (Pearson correlation $|\rho| \lessapprox 0.4$) than the correlation observed between average quantities ($\rho \approx 0.9$). A similar discrepancy between an analysis based on either individual or average ELM quantities also emerges in the present study. Indeed, Figure 2(c) shows a scatter plot of the average loss of the plasma stored energy $\overline{\Delta E}$ following an ELM, against the average time $\overline{\Delta t}$ since the previous ELM, for each of the 32 plasmas in the data set. The smaller range of average waiting times and energies, compared with the values observed for individual ELMs, is remarkable. Moreover, in contrast with the pattern suggested by Figure 2(a) ($\rho \approx 0.6$), a linear relation between $\overline{\Delta t}$ and $\overline{\Delta E}$ appears to be quite plausible ($\rho \approx 0.9$). In Figure 2(d) the error bars on the average quantities were added.

At this point we need to go into more detail regarding the error bars displayed in Figure 2. In Figure 2(b) the error bars on Δt and ΔE for a single ELM denote the sample standard deviation obtained from all ELMs in the corresponding discharge. It should be noted that the error bars for both Δt and ΔE are appreciable. As a result, a proper regression analysis should take into account the uncertainty on both the predictor variable Δt and the response variable ΔE. This already rules out standard OLS as a good candidate for estimating the regression parameters, since OLS considers the predictor variables as infinitely precise quantities [1]. Furthermore, by drawing the error bars symmetrically around the measurement, we have implicitly assumed a symmetric underlying distribution (zero skewness). To verify this



(a) (b) (c)

FIGURE 3: Histograms of (a) waiting time Δt and (b) plasma energy loss ΔE for the ELMs in JET pulse #83642 (9–13.5 s). (c) Histograms of ΔE for all ELMs in the database with $\Delta t \in [10, 15]$ ms and $\Delta t \in [40, 45]$.

assumption, in Figure 3(a), resp. (b), histograms are drawn of Δt and ΔE for JET pulse #83642 (9–13.5 s). A normal distribution was fitted to the waiting times ($\mu = 31$ ms, $\sigma = 13$ ms) and the energies ($\mu = 0.116$ MJ, $\sigma = 0.024$ MJ), and it is clear that this provides a good fit. This is the case for the other discharges in the database as well, justifying our earlier assumptions about the error bars. Accordingly, the error bar on an average quantity in a specific plasma pulse α is defined here as the standard deviation of that quantity over the $n_{\mathrm{ELM},\alpha}$ ELMs in that pulse, divided by $\sqrt{n_{\mathrm{ELM},\alpha}}$. This explains the error bars in Figure 2(d).

Let us now consider the subset of ELMs in the entire database that have their waiting time (time since the previous ELM) in a certain small interval, say $\Delta t \in]10, 15]$ ms. This subset contains ELMs from various discharges and the distribution of the corresponding energies is given in Figure 3(c). The histogram related to another interval with larger waiting times, $\Delta t \in]40, 45]$ ms, is also given. The underlying distribution is clearly skewed and the shape is quite different from the one shown in Figure 3(b). A log-normal model is fitted in Figure 3(c) (i.e. the logarithm of the energy is taken to be approximately normally distributed), although the fit is not so good for larger waiting times. Nevertheless, this indicates that a generalized linear model with a logarithmic link function may also work well to perform the regression analysis.

Thus, although the normal distribution for the ELM energies (and waiting times) in the individual discharges in the present database is more directly related to the ELM physics, a skewed model, such as the log-normal distribution, may be more suitable for the role of the likelihood in the regression analysis. The advantage is likely to be more important for regression on the individual ELM quantities, and less for the average quantities. This is because the variance of the distribution of the average quantities in any pulse α is smaller by a factor $n_{\mathrm{ELM},\alpha}$ than the variance of the quantities for individual ELMs in that pulse, where the median of $n_{\mathrm{ELM},\alpha}$ in the database is 56 and the minimum 18. Nevertheless, in this paper we will continue working with a normal likelihood model, hence demonstrating the robustness of GLS in the presence of an inferior regression model.

Scaling of plasma energy loss with ELM waiting time

We now carry out regression analysis to estimate the dependence of the plasma energy loss ΔE following an ELM on the time Δt since the previous ELM. We do this for the average and the individual ELM quantities. Guided by Figure 2(c), we assume a linear relationship and, as mentioned before, we will continue working with the ubiquitous Gaussian probability model. In view of the possible saturation effect deduced from Figure 2(a), as well as the non-Gaussianity seen in Figure 3(c), this may be an overly simplistic model. Nevertheless, we will show that GLS gives consistent results when comparing the regression on the average quantities with the analysis that uses the individual ELM data. Furthermore, the only structure assumed on σ_{obs} is that it is proportional to the measurement of the response variable: $\sigma_{\mathrm{obs},i} = r_{\mathrm{obs}}|y_i|$, assuming a fixed relative error r_{obs} to be estimated from the data. This can be motivated by the sample standard deviation $s_{\Delta t}$ ($s_{\overline{\Delta t}}$) of the (average) waiting time in a discharge being approximately proportional to Δt ($\overline{\Delta t}$) itself, and similar for the energies (see e.g. Figure 2(b) ((d))).

We compare GLS regression with OLS and with a standard Bayesian method. For the latter we choose the likelihood given in (1)–(3), but we use an unknown standard deviation σ_{u} instead of σ_{mod}. Again, we assume that $\sigma_{\mathrm{u},i} = r_{\mathrm{u}}|y_i|$, for a fixed relative error r_{u} to be estimated. In addition, we use uninformative prior distributions for the regression parameters and a Jeffreys prior for r_{u}. This factor is then marginalized out of the posterior, which comes down to fitting a t-distribution to the data (shifted to sample mean zero). The t-distribution has heavier tails than a Gaussian, hence accommodating outliers. There are other possibilities for modeling the variance of the likelihood, which take into account the proposed σ_{mod}, but we will not consider them here. Finally, we compare with an analysis that is in all aspects the same as GLS, except that the Kullback-Leibler divergence (KLD) is used as a similarity

TABLE 1: Estimates of the cut-off β_0 and slope β_1 of the regression line obtained using OLS, Bayesian posterior mean (Bayes), KLD and GLS, on (a) the average and (b) the individual ELM quantities.

Method	β_0 **(kJ)**	β_1 **(kJ/ms)**
OLS	−50	5.7
Bayes	−10	3.4
KLD	−13	3.5
GLS	−21	4.6

(a)

Method	β_0 **(kJ)**	β_1 **(kJ/ms)**
OLS	24	3.2
Bayes	−12	2.1
KLD	−20	4.1
GLS	−22	4.2

(b)

measure, instead of the GD. Specifically, the cost function is based on $\mathrm{KLD}[p_{\mathrm{obs}}\|p_{\mathrm{mod}}]$, i.e. the entropy of p_{obs} relative to p_{mod}.

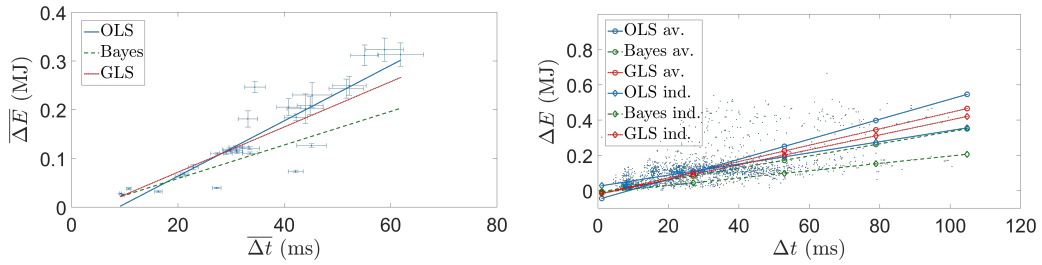
The results of the regression analysis are summarized in Table 1(a) for the average quantities and in Table 1(b) for the individual quantities. For the Bayesian analysis the posterior mean is given. At this point GLS does not directly return an uncertainty measure on the parameter estimates, therefore neither did we provide one with the other methods. The regression lines are drawn together with the average data in Figure 4(a), and a comparison with regression on the quantities for individual ELMs is shown in panel (b) (for clarity the results of KLD are not included).

In evaluating the results, we are particularly interested in comparing the estimates on the average quantities with those on the individual measurements. We consider the degree of correspondence of the respective estimates as a measure of robustness of the regression techniques. Furthermore, since estimating the cut-off parameter of a regression line is usually more difficult than estimation of the slope, and also of less physical and practical interest, we concentrate on the slope.

The results in Table 1 and Figure 4 show that GLS yields the most robust estimates of the regression parameters, while OLS performs poorly. The Bayesian and KLD methods result in very similar estimates in case of the average quantities. This is because, in the limit of a large number of (individual) measurements and provided p_{obs} is close to normal, the KLD estimates approach the maximum likelihood results, which in this case are close to the Bayesian estimates. The correspondence between the Bayesian and KLD methods is lost for regression on individual quantities, possibly because each single measurement represents only a poor approximation to the true underlying distribution. In fact, the Bayesian and KLD methods each yield relatively different results for the average and the individual quantities, rendering them less robust than GLS in the present context. This observation may indicate that the GD is better suited as a similarity measure between probability distributions than the KLD. For instance, it is known that $\mathrm{KLD}[p_1\|p_2] \approx 1/2\,\mathrm{GD}^2[p_1\|p_2]$ only for infinitesimally close points p_1 and p_2 on the manifold.

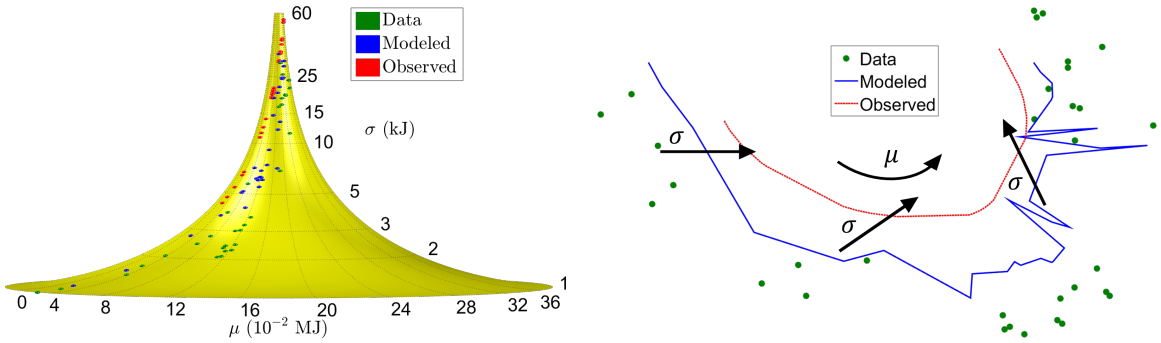
As touched upon before, it is possible that a more accurate Bayesian model will perform better on these data. A likelihood model allowing for skewness could be one improvement, a more informative prior for the standard deviation another. However, the point here is that GLS succeeds in producing consistent results even with the simplest model and without requiring any fine-tuning of model parameters. In contrast, it may be difficult to find a good Bayesian model for complex data sets, especially in multiple dimensions. Furthermore, similar model improvements are equally possible for GLS and still the method would be more flexible than the corresponding Bayesian approach, since GLS minimizes the distance between distributions.

Finally, the average plasma energy losses and their uncertainty are plotted as points (normal distributions) on the pseudosphere in Figure 5(a), together with the predictions p_{mod} by the regression model and the corresponding observed distributions p_{obs}. This plot emphasizes the fact that, apart from the mean of the plasma energy loss, GLS also predicts its standard deviation. The typical relative error on $\overline{\Delta E}$ determined from the measurements is about 5%. The modeled relative error is slightly larger, typically 8%, as it also includes the variability on the waiting times (see (3)). The observed relative error is still greater, 17%, because it takes into account the overall variation of $\overline{\Delta E}$ around the regression line. This can be seen even more clearly in Figure 5(b), which shows the projection of the points in panel (a) on the Euclidean plane using metric multidimensional scaling (MDS). Hence, σ_{obs} ($\times \sqrt{n_{\mathrm{ELM},\alpha}}$ for individual bursts) should be used to predict the variability of the plasma energy loss following an ELM at an envisaged ELM frequency under arbitrary plasma conditions (similar to those in the database). This is considerably larger than the variability deduced from a series of ELMs in an individual discharge. It may be worthwhile to try to model this extra variability by including the dependence of $\overline{\Delta E}$ on additional (e.g. global) plasma parameters.



(a) (b)

FIGURE 4: (a) Linear regression results on the average quantities using OLS, Bayesian posterior mean and GLS. (b) Comparison with the regression results on the individual quantities.



(a) (b)

FIGURE 5: (a) The pseudosphere with the data for the average energy loss $\overline{\Delta E}$, as well as the corresponding estimates by GLS using the modeled and observed standard deviation. (b) The same data projected on the Euclidean plane using MDS, with the approximate directions of increasing mean and standard deviation indicated.

Conclusion

Due to the significant variability of ELM properties in tokamak plasmas across different ELM bursts, it is important to study the characteristics of their probability distributions in addition to average trends. Geodesic least squares regression is designed to study arbitrary dependencies between probability distributions, while allowing for unexpected sources of uncertainty, including systematic errors. In a comparison with other regression techniques, GLS was shown to yield more consistent estimates of the dependence between ELM-induced energy loss and waiting time, when considering either the pulse-averaged or the individual ELM quantities in a set of JET-ILW plasmas.

ACKNOWLEDGMENTS

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1] G. Verdoolaege, Entropy **17**, 4602–4626 (2015).
[2] A. Shabbir *et al.*, Nucl. Fusion **57**, p. 036026 (2017).
[3] A. Herrmann, Plasma Phys. Control. Fusion **44**, 883–903 (2002).
[4] J. Burbea and C. Rao, J. Multivariate Anal. **12**, 575–596 (1982).