Robust Regression on Noisy Data for Fusion Scaling Laws[a)]

Geert Verdoolaege[1,2, b)]
[1)] *Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium*
[2)] *LPP–ERM/KMS, B-1000 Brussels, Belgium*

(Dated: 27 June 2014)

We introduce the method of geodesic least squares (GLS) regression for estimating fusion scaling laws. Based on straightforward principles, the method is easily implemented, yet it clearly outperforms established regression techniques, particularly in cases of significant uncertainty on both the response and predictor variables. We apply GLS for estimating the scaling of the L-H power threshold, resulting in estimates for ITER that are somewhat higher than predicted earlier.

I. INTRODUCTION

One of the main activities in analyzing data from fusion experiments consists of fitting deterministic relations reflecting physical dependencies between plasma variables. This is an essential instrument for evaluating theoretical predictions and for extrapolating key quantities to future devices via scaling laws. Ordinary least squares (OLS) regression is commonly used for this purpose, primarily owing to its simplicity[1]. However, frequently the data are contaminated by significant stochastic uncertainty caused by (non-Gaussian) measurement noise and plasma fluctuations. This can be further complicated by nonlinear relations and data outliers. Moreover, OLS does not properly handle situations with significant measurement uncertainty on the predictor variables, which is often the case in fusion science. In this paper we introduce a new regression method, *geodesic least squares* (GLS), that is able to cope with significant departure from the classic assumptions underlying OLS. We demonstrate its performance using synthetic data and we revisit the scaling law for the power threshold for the L-H transition, using data from a multimachine database.

II. GEODESIC LEAST SQUARES

The method that we propose is a clear generalization of the standard regression method using OLS. To see this, we focus on the very simple example of a linear relation $\eta = b\xi$, with b a constant. In reality we observe stochastic (noisy) variables x and y, where we assume Gaussian noise in this case:

$$y = \eta + \epsilon_y = b\xi + \epsilon_y, \qquad \epsilon_y \sim \mathcal{N}\left(0, \sigma_y^2\right), \qquad (1)$$

$$x = \xi + \epsilon_x, \qquad \epsilon_x \sim \mathcal{N}\left(0, \sigma_x^2\right). \qquad (2)$$

Hence, we explicitly allow for the challenging case of uncertainty on the predictor variable ξ. We record N observations x_i $(i = 1, \ldots, N)$, mutually independent and identically distributed (iid), and y_i, also iid. We assume known σ_x and σ_y, which are the same for all measurements (homoscedasticity). According to the regression model, conditionally on x_i each variable y_i has a normal distribution:

$$p(y_i|x_i) = \mathcal{N}\Big(bx_i, \tilde{\sigma}_y^2\Big) \equiv \mathcal{N}\Big(bx_i, \sigma_y^2 + b^2\sigma_x^2\Big), \qquad (3)$$

where we have defined $\tilde{\sigma}_y$. In our simple example this follows from standard Gaussian error propagation rules. However, for nonlinear regression laws the conditional distribution for y_i has to be obtained by integrating out (marginalizing) the unknown true values ξ_i. Nevertheless, the Gaussian error propagation laws may be used in the nonlinear case as well, to *approximate* the conditional distribution $p(y_i|x_i)$ by a normal distribution.

The well-known *maximum likelihood method* (ML)[1] can be used to estimate the slope b that best fits the data, by maximizing the probability (3) for all observations. Hence, we seek an estimate $\hat{b}$ that maximizes the *likelihood function* $\mathcal{L}$:

$$\mathcal{L}\Big(b\,\Big|\{x_i\},\{y_i\}\Big) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\tilde{\sigma}_y} \exp\left[-\frac{(y_i - bx_i)^2}{2\tilde{\sigma}_y^2}\right]. \qquad (4)$$

Put differently, we estimate b such that it maximizes the probability of observing the measured data. From (4) it is clear that, in the case of a Gaussian error distribution and neglecting the error bar on x_i, ML is equivalent to minimizing the sum of squared distances between y_i and bx_i, which is nothing but OLS.

Interestingly, the right-hand side of (4) can also be seen as the probability density for bx_i, given y_i (in accordance with the Bayesian view on probability). Now, the first key difference between OLS and our method, is that we do not assume that both y_i and bx_i are distributed with the same standard deviation $\tilde{\sigma}_y$. As a result, we not only minimize the distance between each y_i and its corresponding bx_i, but rather we aim at minimizing the difference between their entire *probability distributions*. For a Gaussian distribution, this involves comparing means *and* standard deviations. As a result, the regression analysis is based not only on measurement points, but also

[a)] Contributed paper published as part of the Proceedings of the 20th Topical Conference on High-Temperature Plasma Diagnostics, Atlanta, Georgia, June, 2014.

[b)] geert.verdoolaege@ugent.be

includes information on the distribution of the data. This leads to more reliable regression results, as will be demonstrated in the experiments.

The second key difference with OLS is that we do not use the Euclidean distance to measure the discrepancy between distributions, since it turns out to be unsuitable for that purpose[2]. Rather, we employ the *Rao geodesic distance* (GD) as a similarity measure in probability spaces. The GD is defined in the context of the theory of *information geometry*, which is a geometric approach to probability theory[2,3]. In information geometry, a probability density family is interpreted as a (Riemannian) differentiable manifold (multidimensional surface). 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, a well-known concept in statistics, plays the role of a unique metric tensor (Fisher-Rao metric) on such a manifold, which can be used to derive geodesics and the geodesic distance between two points on the manifold. For a probability model $p(\boldsymbol{x}|\boldsymbol{\theta})$ describing a vector $\boldsymbol{x}$, parameterized by an m-dimensional vector $\boldsymbol{\theta}$ with components θ_i $(i = 1, \ldots, m)$, the entries g_{ij} of the Fisher information matrix are the following:

$$g_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln p(\boldsymbol{x}|\boldsymbol{\theta})\right], \quad i, j = 1 \ldots m.$$

Here, $\mathbb{E}$ signifies the expectation. In this paper we use a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$, but it is important to note that GLS can easily be implemented with any distribution model. In the Gaussian case an analytic expression for the Fisher-Rao metric is available, which in turn allows a closed-form expression for the GD between two normal distributions[2]. Indeed, for two univariate normal 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[4]

$$\mathrm{GD}(p_1||p_2) = \sqrt{2} \ln \frac{1+\delta}{1-\delta} = 2\sqrt{2} \tanh^{-1} \delta,$$
$$\delta \equiv \left[\frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2 + 2(\sigma_1+\sigma_2)^2}\right]^{1/2}. \quad (5)$$

In the case of multiple independent normal 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 the corresponding individual distributions[4].

As an illustrative example of why the GD is more suitable as a similarity measure between PDFs, as compared to the Euclidean distance, we consider the case of two Gaussians with PDFs $p_1(x|2, 0.3)$ and $p_2(x|6, 0.5)$, drawn in Figure 1(a). In Figure 1(b) two distributions $p_3(x|2, 1.5)$ and $p_4(x|6, 2)$ are displayed with the same respective means, but larger standard deviations compared to the first case. Now, whereas p_1 and p_2 are easy to distinguish, the distributions p_3 and p_4 overlap to a much larger extent. This difference in the level of 'distinguishability' should, of course, be reflected in the distance between the distributions. That is, the distance between p_1 and p_2 should be larger than that between p_3 and p_4. From the expression in (5) it can be seen that the GD fulfills this requirement; indeed: $\mathrm{GD}(p_1||p_2) = 5.7$ and $\mathrm{GD}(p_1||p_2) = 2.1$. However, on the contrary, the Euclidean distance between p_1 and p_2, calculated as

$$\sqrt{(\mu_1-\mu_2)^2 + (\sigma_1-\sigma_2)^2} = 4.0050,$$

is *smaller* than the Euclidean distance between p_3 and p_4, which is 4.0311. Also, as suggested by this example, the GD is more sensitive to differences in the standard deviations, compared to the Euclidean distance. Hence, the Euclidean distance does not properly take into account the intrinsically non-Euclidean character of probability distributions.

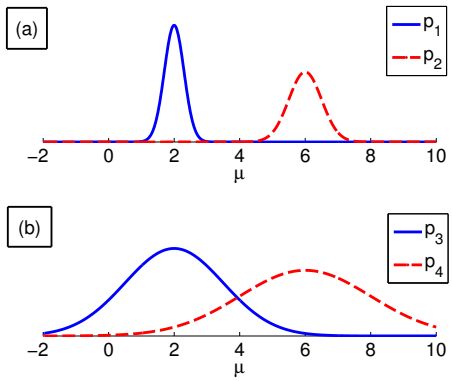

FIG. 1. A pair of normal distributions with relatively small standard deviations in (a), compared to a pair of distributions with the same respective means, but larger standard deviations in (b). The distance between p_1 and p_2 is larger than the distance between p_3 and p_4.

Continuing the case of a Gaussian conditional distribution of the dependent regression variable, we now introduce an extra parameter σ_{obs}, with the purpose of modeling the uncertainty in the observations of the response variable y. We then estimate b and σ_{obs} by minimizing the sum of squared GDs between the *modeled* distributions, with mean bx_i and standard deviation $\tilde{\sigma}_y = \sqrt{\sigma_y^2 + b^2\sigma_x^2}$, and the *observed* distributions, with mean y_i and standard deviation σ_{obs}. Owing to the added flexibility offered by the extra parameters σ_{obs}, GLS is less sensitive to incorrect model assumptions, as demonstrated in the next section.

III. NUMERICAL SIMULATIONS

A. Effect of outliers

We first tested the effect of outliers on the performance of GLS, concentrating on estimation of the slope of a

TABLE I. Monte Carlo estimates of the mean and standard deviation for the slope parameter in linear regression with errors on both variables and one outlier.

Original	GLS	OLS	MLE	TLS	ROB
$b = 3.00$	3.031	3.528	3.696	4.61	2.992
	$\pm$ 0.035	$\pm$ 0.038	$\pm$ 0.049	$\pm$ 0.11	$\pm$ 0.041

regression line with a single independent variable. To this end, a data set was generated consisting of ten points labeled by coordinates ξ_i and η_i $(i = 1, \ldots, 10)$, with the ξ_i chosen unevenly between 0 and 50 and $\eta_i = 3\xi_i$. Then, Gaussian noise was added to all coordinates according to (1) and (2), with $\sigma_y = 2.0$ and $\sigma_x = 0.5$. Finally, an outlier was created by doubling the value of y_8.

We next estimated b by means of GLS and compared the estimates with those obtained by OLS, maximum likelihood estimation (MLE) using the model in (3), total least squares (TLS)[5], which is a typical errors-in-variables technique, and a robust method (ROB) based on iteratively reweighted least squares (bisquare weighting)[6]. In all cases we assumed knowledge of the values of σ_x and σ_y. In order to get an idea of the variability of the estimates, Monte Carlo sampling of the data-generating distributions was performed and the estimation was carried out 100 times.

The results are given in Table I, mentioning the sample average and standard deviation of b over the 100 runs for each of the methods. GLS is seen to perform similar to the robust method. The average σ_{obs} was 5.43 with a standard deviation of 0.24. On the other hand, the modeled value of the standard deviation in the conditional distribution for y_i was $\sqrt{\sigma_y^2 + 9\sigma_x^2} = 2.5$. Hence GLS succeeds in ignoring the outlier by increasing the estimated variability of the data.

B. Effect of logarithmic transformation

We next tested the effect of a logarithmic transformation, which is often used to transform a power-law regression model into a linear form. However, the logarithm alters the data distribution, which may lead to misguided inferences from OLS[7]. Therefore the flexibility offered by GLS is expected to be beneficial in this case, as it allows the observed distribution to deviate from the modeled distribution. To this end, we performed a regression experiment with a power law deterministic model and additive Gaussian noise on all variables. In accordance with the typical situation of fitting fusion scaling laws to multimachine data, the noise standard deviation was taken proportional to the simulated measurements, corresponding to a given set of relative error bars. As a result, in the logarithmic space the distributions were only approximately Gaussian, with the standard deviation given by the constant relative error on the original measurement (homoscedasticity). Ten points were chosen with independent coordinates ξ_i unevenly spread between 0 and 60. A power law was proposed to relate the unobserved ξ_i and η_i:

$$\eta_i = b_0 \xi_i^{b_1}, \quad i = 1, \ldots, 10.$$

Then, Gaussian noise was added to both coordinates, corresponding to a substantial relative error of 40%. We finally took the natural logarithm of all observed values x_i and y_i, enabling application of the same linear regression methods that were used in the previous experiments. In this particular experiment we chose $b_0 = 0.8$ and $b_1 = 1.4$, but we found that other values yield similar conclusions. Again, 100 data replications were generated, allowing calculation of Monte Carlo averages.

TABLE II. Monte Carlo estimates of the mean and standard deviation for the parameters in a log-linear regression experiment with proportional additive noise on both variables.

Parameter	Original	GLS	OLS	MLE	TLS	ROB
b_0	0.80	0.94	2.2	1.75	0.99	2.72
		$\pm$ 0.47	$\pm$ 2.3	$\pm$ 0.58	$\pm$ 0.70	$\pm$ 0.77
b_1	1.40	1.39	1.19	1.21	1.41	1.17
		$\pm$ 0.11	$\pm$ 0.16	$\pm$ 0.10	$\pm$ 0.14	$\pm$ 0.11

The averages and standard deviations over all 100 runs are given in Table II. Again, the results show that GLS is robust against the flawed model assumptions, performing similar to TLS.

IV. POWER THRESHOLD SCALING

We finally applied our method to the estimation of the scaling law for the power threshold for the L-to-H transition in tokamaks[8]. We assumed dependence of the power threshold on the line-averaged electron density $\bar{n}_{\text{e}}$ $(10^{20}\ \text{m}^{-3})$, the toroidal magnetic field B_{t} (T) and the plasma surface area (m^2):

$$P_{\text{thr}} = b_0 \bar{n}_{\text{e}}^{b_1} B_{\text{t}}^{b_2} S^{b_3}. \tag{6}$$

We employed data from seven devices in a multi-machine database, overall containing 645 measurements of power, density, magnetic field and surface area (subset IAEA02 [9]).

A. Linear scaling

We first followed the standard practice of transforming to the logarithmic scale to estimate the coefficients b_i $(i = 0, \ldots, 3)$ via linear regression. In the GLS method we introduced additional parameters, approximately describing the relative errors for the power threshold (one for each device), similar to the parameter σ_{obs} in the example above. The estimation results are shown in Table III, with the estimated relative errors on P_{thr} varying

TABLE III. Estimates of the regression coefficients b_i and predictions for ITER in log-linear scaling for the H-mode threshold power.

Method	b_0	b_1	b_2	b_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	0.059	0.73	0.71	0.92	48	80
GLS	0.065	0.93	0.64	1.02	62	117

TABLE IV. Estimates of the regression coefficients b_i and ITER predictions in nonlinear scaling of the H-mode threshold.

Method	b_0	b_1	b_2	b_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	0.051	0.85	0.70	1.00	62	111
GLS	0.048	0.96	0.59	1.05	64	124

between 21% and 48%. The predictions for ITER are also shown, for two typical densities (0.5 and 1.0×10^{20} m^{-3}).

B. Nonlinear scaling

Finally, we show the results of a nonlinear regression in the original data space, i.e. without logarithmic transformation. Whereas this prevents an analytic solution using OLS, it should be noted that OLS for nonlinear regression is not particularly more complicated, while for GLS there is no conceptual difference compared to the linear case. Indeed, the distribution of the right-hand side in (6) can be approximated by a Gaussian with mean $\mu_{\mathrm{mod}} = b_0 \bar{n}_{\mathrm{e}}^{b_1} B_{\mathrm{t}}^{b_2} S^{b_3}$ and standard deviation σ_{mod}, given by

$$\sigma_{\mathrm{mod}}^2 = \sigma_{P_{\mathrm{thr}}}^2 + \mu_{\mathrm{mod}}^2 \left[b_1^2 \left(\frac{\sigma_{\bar{n}_{\mathrm{e}}}}{\bar{n}_{\mathrm{e}}} \right)^2 + b_2^2 \left(\frac{\sigma_{B_{\mathrm{t}}}}{B_{\mathrm{t}}} \right)^2 + b_3^2 \left(\frac{\sigma_S}{S} \right)^2 \right].$$

Hence, the error bars depend on the measurements (heteroscedasticity). Nevertheless, we introduced an approximation assuming constant error bars for all measurements from a single machine. This assumption may be relaxed in the future. The results of the scaling and predictions are given in Table IV. It is interesting to note that the results for GLS are similar to those derived on the logarithmic scale (Table III), indicating that, indeed, GLS is less susceptible to flawed model assumptions. Furthermore, the results for OLS and GLS are now in the same range, with slightly lower predictions by OLS. Nevertheless, both methods suggest higher power thresholds than those obtained in earlier studies in the same database ($\hat{P}_{\mathrm{thr},0.5} = 44$ MW [8]).

V. CONCLUSION

Several important scaling laws have been established in the past providing essential design constraints for next-step fusion devices. With the present study, on the one hand we have aimed to indicate that continuing efforts in this area are still useful. We have shown that geodesic least squares regression provides a simple but robust alternative to standard methods. In specific relation to the scaling of the L-H power threshold, we have noted predictions that are consistently higher than reported earlier. On the other hand, regression analysis is routinely performed in fusion science for the purpose of model building and prediction in the context of new physics studies. With the GLS method, we aim to provide a reliable tool to the fusion community for scaling studies in demanding circumstances (e.g. large uncertainties). For this purpose, future work will involve improving and generalizing GLS (including error bars on predictions) and implementation in a publicly accessible software package.

[1] G. Casella and R. Berger, *Statistical Inference*, 2nd ed. (Cengage Learning, Belmont, CA, 2002).
[2] G. Verdoolaege *et al.*, Plasma Phys. Control. Fusion **54** (2012).
[3] S. Amari and H. Nagaoka, *Methods of Information Geometry* (American Mathematical Society, New York, 2000).
[4] J. Burbea and C. Rao, J. Multivariate Anal. **12**, 575 (1982).
[5] I. Markovsky and S. Van Huffel, Signal Process. **87**, 2283 (2007).
[6] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods* (Wiley, New York, 2006).
[7] D. McDonald *et al.*, Plasma Phys. Control. Fusion **48**, A439 (2006).
[8] J. Snipes *et al.*, in *Proceedings of the 19th IAEA Fusion Energy Conference*, CT/P-04 (Lyon, France, 2002).
[9] http://efdasql.ipp.mpg.de/threshold (2002).