A New Methodology for Scaling Laws with Arbitrary Error Distributions: Case Study for the H-Mode Power Threshold

UNIVERSITEIT GENT — FACULTY OF ENGINEERING AND ARCHITECTURE — LPP ERM - KMS — IPP

G. Verdoolaege[1,2] and J.-M. Noterdaeme[1,3]

[1]Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium

[2]Laboratoire de Physique des Plasmas de l'ERM – Laboratorium voor Plasmafysica van de KMS (LPP–ERM/KMS), Ecole Royale Militaire – Koninklijke Militaire School, B-1000 Brussels, Belgium

[3]Max Planck Institute for Plasma Physics, Boltzmannstr. 2, 85748 Garching, Germany

Abstract

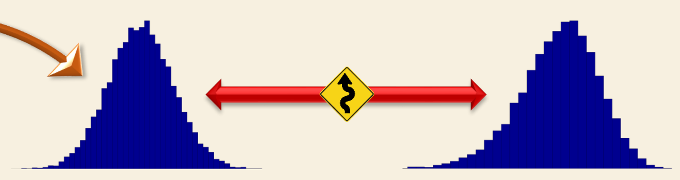
In regression analyses for deriving scaling laws in the context of fusion studies, usually standard regression methods have been applied, of which ordinary least squares (OLS) is the most popular. However, concerns have been raised with respect to several assumptions underlying OLS in its application to fusion data. More sophisticated statistical techniques are available, but they are hardly known or used in the fusion community and, moreover, the predictions by scaling laws may vary significantly depending on the particular regression method used. Given the ubiquity and importance of scaling laws in fusion research, it is natural to approach their estimation with dedicated statistical tools. We have developed a new regression method for this purpose, which we call **geodesic least squares regression (GLS)**, that is robust in the presence of significant uncertainty on both the data and the regression model [1,2]. The method is based on probabilistic modeling of all variables involved in the scaling expression, using adequate probability distributions and a natural similarity measure between them (geodesic distance). In this work we revisit the scaling law for the power threshold for the L-to-H transition in tokamaks, using data from the multi-machine ITPA database. The prediction of the power threshold for ITER is higher than that obtained with OLS on the same database, suggesting caution in interpreting results predictions by established scaling laws.

Motivation

- In fusion science, regression analysis is used:
  - As an aid to build and validate theoretical models from data to find **parametric dependencies**
  - As a statistical tool to formulate **scaling laws** for the purpose of **extrapolation**
- **Ordinary least squares regression (OLS)** is the workhorse
- Often, multiple assumptions underlying OLS are not fulfilled [3,4,5]
- There may be various reasons:
  - Considerable measurement uncertainty: statistical and systematic
  - Uncertainty on response (dependent, y) and predictor (independent, x_j) variables
  - Model uncertainty: linear, power law, semi-empirical, …

    Power law: $y = b_0 x_1^{b_1} x_2^{b_2} \dots x_m^{b_m}$
  - Heterogeneous data and error bars, correlations, non-Gaussian probability distributions
  - Atypical observations (outliers)
  - Near-collinearity of predictor variables
  - Data transformations, e.g.

    $\ln y = \ln b_0 + b_1 \ln x_1 + \dots + b_m \ln x_m$
- **Inferior regression analysis counteracts other efforts!**
- A **flexible**, **robust** and **user-friendly** regression tool is needed



Geodesic least squares regression (GLS)

OLS → Formulate model with parameters $\vec{\theta} = [\theta_1, \dots, \theta_p]^{\mathrm{T}}$: $y = f(x_1, \dots, x_m|\vec{\theta})$ ← **GLS**

Take n measurements y_i and $\vec{x}_i = [x_{i1}, \dots, x_{im}]^{\mathrm{T}}$

Minimize error

Zero error on the x_{ij}!

Least squares estimation
$$\hat{\vec{\theta}} = \arg\min_{\vec{\theta}} \sum_{i=1}^{n} \left[y_i - f(\vec{x}_i|\vec{\theta})\right]^2$$

or

Maximum likelihood estimation (MLE) (e.g. Gaussian likelihood)
$$\hat{\vec{\theta}} = \arg\max_{\vec{\theta}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - f(\vec{x}_i|\vec{\theta})\right]^2}{\sigma^2}\right\}$$

Includes errors on the x_{ij}!

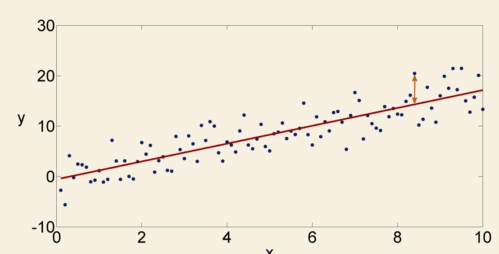
Modeled distribution (e.g. Gaussian)
$$p_{\mathrm{mod}}(y|\vec{x}_1, \dots, \vec{x}_n, \vec{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{mod}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left[y - f(\vec{x}_i|\vec{\theta})\right]^2}{\sigma_{\mathrm{mod}}^2}\right\}$$

σ_{mod} includes uncertainty on y and $f(\vec{x}_i|\vec{\theta})$

Minimize distance

Observed distribution (e.g. Gaussian)
$$p(y|y_1, \dots, y_n) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{obs}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{[y - y_i]^2}{\sigma_{\mathrm{obs}}^2}\right\}$$

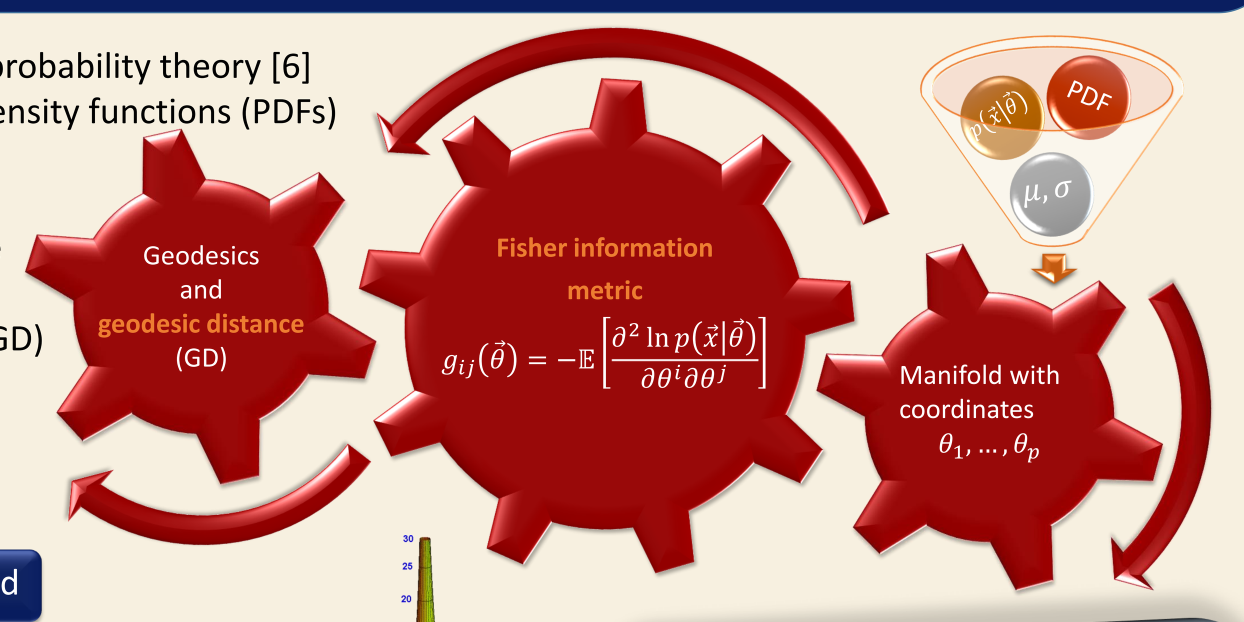
σ_{obs} to be estimated from the data



Which distance measure between probability distributions?

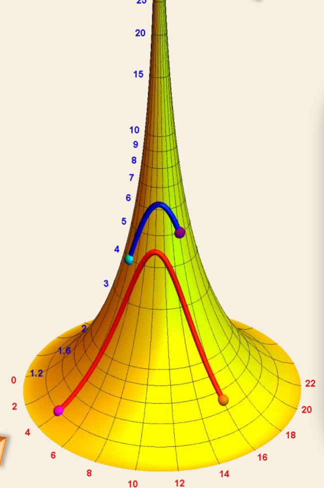
Information geometry

- Geometric approach to probability theory [6]
- A family of probability density functions (PDFs) forms a metric space, or **manifold**
- **Fisher information** is the metric tensor
- **Rao geodesic distance** (GD) is the shortest distance between points (PDFs)



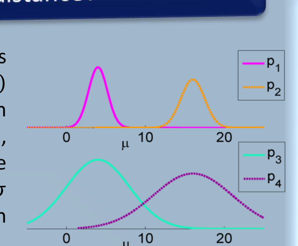
Example: Gaussian manifold

- $p_1(x|\mu_1, \sigma_1) \leftrightarrow p_2(x|\mu_2, \sigma_2)$
- $\mathrm{GD}(p_1, p_2) = 2\sqrt{2}\tanh^{-1}\delta$,
$$\delta = \left[\frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2 + 2(\sigma_1+\sigma_2)^2}\right]$$
- The **pseudosphere** (tractroid) is a model for the manifold of univariate Gaussian distributions, respecting the true geometry (μ in red, σ in blue)



Why a geodesic distance?

Intuitively, the Gaussians $p_3(x|4, 4.0^2)$ and $p_4(x|16, 5.0^2)$ are closer (more overlap) than $p_1(x|4, 1.2^2)$ and $p_2(x|16, 1.5^2)$, although the respective means are the same. The reason is that σ does not behave like a Euclidean coordinate.



Indeed, $\mathrm{GD}(p_3, p_4) = 2.4 < \mathrm{GD}(p_1, p_2) = 5.3$, whereas with the Euclidean distance $\mathrm{ED}(p_3, p_4) = 12.04 > \mathrm{ED}(p_1, p_2) = 12.00$. The distributions are mapped on the pseudosphere on the left.

Numerical simulations

1. Atypical observations (outliers)

$$p_{\mathrm{mod}}(y|x_1, \dots, x_{10}, b) = \frac{1}{\sqrt{2\pi(\sigma_y^2 + b^2\sigma_x^2)}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{10}\frac{[y - b\,x_i]^2}{\sigma_y^2 + b^2\sigma_x^2}\right\}$$

Minimize Rao GD — Estimate b, σ_{obs}

$$p(y|y_1, \dots, y_n) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{obs}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{10}\frac{[y - y_i]^2}{\sigma_{\mathrm{obs}}^2}\right\}$$

- Linear model with a single predictor and Gaussian noise:

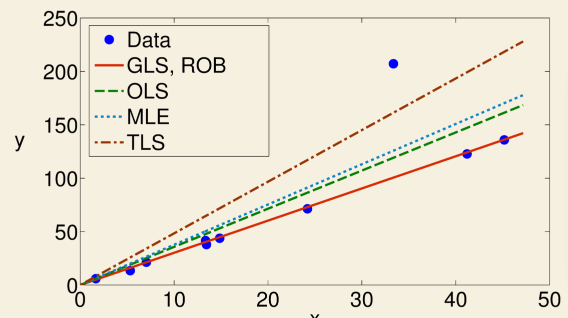
$$0 \le \xi_i \le 50, \quad i = 1, \dots, 10 \qquad \begin{cases} x_i = \xi_i + \epsilon_{x,i}, & \epsilon_{x,i} \sim \mathcal{N}(0, \sigma_x^2) \\ y_i = \eta_i + \epsilon_{y,i}, & \epsilon_{y,i} \sim \mathcal{N}(0, \sigma_y^2) \end{cases}$$
$$\eta_i = b\xi_i, \quad b = 3.00$$
$$\sigma_x = 0.5, \quad \sigma_y = 2.0$$

- Introduce an outlier: $y_i \to 2 \times y_i$, $i \in [8, 10]$ uniformly
- 100 Monte Carlo runs
- GLS captures outlier by estimating an average

$$\hat{\sigma}_{\mathrm{obs}} = 4.36\,(\pm 0.32) > \sigma_{\mathrm{mod}} = \sqrt{\sigma_y^2 + b^2\sigma_x^2} = 2.5$$

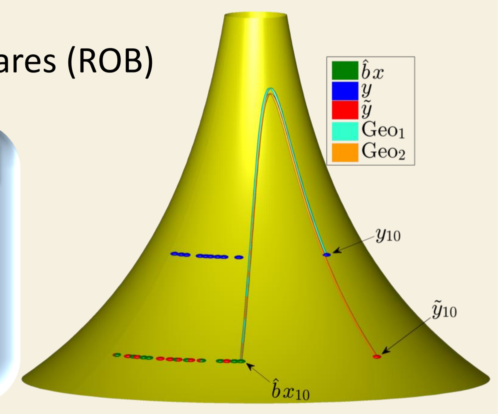
- Comparison with
  - OLS
  - Total least squares (TLS)
  - Maximum likelihood estimation (MLE)
  - Robust (iteratively re-weighted) least squares (ROB)

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



Regression on the pseudosphere

On the right is the pseudophere with superimposed the estimates for the model with outlier (for one specific data set in the simulation). The points $\hat{b}x$ signify the modeled distributions ($\sigma_{\mathrm{mod}} = 2.5$, outlier at $\hat{b}x_{10}$). y denotes the observed distributions ($\hat{\sigma}_{\mathrm{obs}} = 4.36$) and $\tilde{y}$ are the same but shifted to $\sigma_{\mathrm{mod}} = 2.5$. The geodesic Geo_1 ($\mathrm{GD} = 5.13$) is indeed shorter than Geo_2 ($\mathrm{GD} = 5.85$).



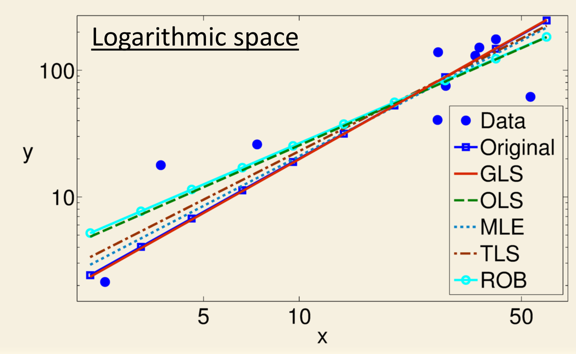
2. Logarithmic transformation

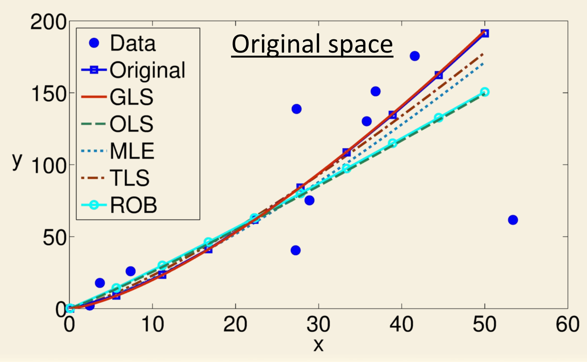
- Power law model with Gaussian noise (40 %):

$$0 \le \xi_i \le 60, \qquad i = 1, \dots, 10$$
$$\eta_i = b_0 \xi_i^{b_1}, \quad b_0 = 0.80, \quad b_1 = 1.40$$

- Transform to logarithmic space

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





Power threshold scaling

- Statistical analysis of established power threshold scaling has revealed several **flawed assumptions** [3]:
  - Negligible uncertainty on predictor variables compared to response variable
  - Equal relative error on variables in all devices and experiments
  - Normal distribution of logarithmic quantities
- Regression **model uncertainty**:
  - Additional predictor variables [4]
  - Non-power law forms [5]

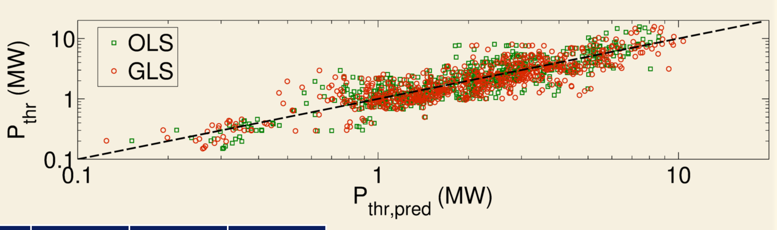
Parameter	GLS	OLS	MLE	TLS	ROB
b_0	**0.053**	0.059	0.053	0.027	0.055
b_1	**0.93**	0.73	1.22	0.99	0.74
b_2	**0.64**	0.71	0.40	0.86	0.72
b_3	**1.02**	0.92	1.15	1.15	0.94
$P_{\mathrm{thr},0.5}$	**62**	**48**	79	101	50
CI 1σ	–	$+3.7 / -3.5$	–	–	–
CI 95%	–	$+7.6 / -6.6$	–	–	–
$P_{\mathrm{thr},1.0}$	**118**	**80**	183	200	83
CI 1σ	–	$+7.4 / -6.8$	–	–	–
CI 95%	–	$+15 / -12$	–	–	–

1. Linear regression on logarithmic scale

- Classic power law:

$$P_{\mathrm{thr}} = b_0 \bar{n}_e^{b_1} B_t^{b_2} S^{b_3}$$
$$\Rightarrow \quad \ln P_{\mathrm{thr}} = \ln b_0 + b_1 \ln \bar{n}_e + b_2 \ln B_t + b_3 \ln S$$

- **ITPA H-mode threshold database** [7], subset IAEA02 [8]: 645 measurements from 7 devices
- Logarithmic variables assumed Gaussian: single standard deviation = relative error from database
- One σ_{obs} for each device: 21% → 48%
- **Power threshold estimates are higher with GLS**



2. Nonlinear regression on logarithmic scale

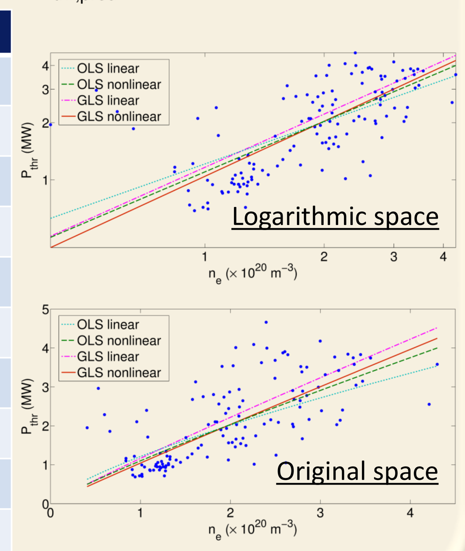
- Gaussian approximation of modeled distribution:

$$\mu_{\mathrm{mod}} = b_0 \bar{n}_e^{b_1} B_t^{b_2} S^{b_3}$$
$$\sigma_{\mathrm{mod}}^2 = \sigma_{\mathrm{thr}}^2 + \mu_{\mathrm{mod}}^2\left[b_1^2\left(\frac{\sigma_{\bar{n}e}}{\bar{n}_e}\right)^2 + b_2^2\left(\frac{\sigma_{Bt}}{B_t}\right)^2 + b_3^2\left(\frac{\sigma_S}{S}\right)^2\right]$$

Calculate average for each device

- **GLS maintains predictions**, OLS changes
- GLS better captures the pattern: e.g. C-Mod @ $B_t \approx 5.2$ T, $S \approx 7.0\ \mathrm{m}^2$

Parameter	GLS	OLS	MLE
b_0	**0.048**	0.051	0.033
b_1	**0.96**	0.85	1.29
b_2	**0.59**	0.70	0.37
b_3	**1.05**	1.00	1.24
$P_{\mathrm{thr},0.5}$	**64**	**62**	81
CI 1σ	–	±5	–
CI 95%	–	±10	–
$P_{\mathrm{thr},1.0}$	**124**	**111**	198
CI 1σ	–	±12	–
CI 95%	–	±23	–



Conclusion

- Regression for fusion scaling laws requires **dedicated tools**
- Regression methodology needs to be **flexible** and **robust**
- **Geodesic least squares** regression fulfils these requirements
- GLS is **user-friendly** and offers a **unified solution** to a variety of regression problems
- Power threshold estimates are **higher with GLS than OLS**
- Future development: **error bars** on GLS estimates and predictions
- GLS will be implemented in a **public software package**

References

[1] G. Verdoolaege *et al.*, Plasma Phys. Control. Fusion **54**, 124006, 2012
[2] G. Verdoolaege *et al.*, Rev. Sci. Instrum. **85**, 11E810, 2014
[3] D.C. McDonald *et al.*, Plasma Phys. Control. Fusion **48**, A439, 2006
[4] A. Murari *et al.*, Nucl. Fusion **52**, 063016, 2012
[5] A. Murari *et al.*, Nucl. Fusion **53**, 043001, 2013
[6] S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS, New York, 2000
[7] Y.R. Martin *et al.*, J. Phys.: Conf. Ser. **123**, 012033, 2008
[8] J.A. Snipes *et al.*, Fusion Energy 2002 (Proc. 19th Int. Conf. Lyon), IAEA, Vienna, CT/P-04