EUROfusion

Robust analysis of trends in noisy data

G. Verdoolaege[1,2], A. Shabbir[1,3] and G. Hornung[1]

[1]Department of Applied Physics, Ghent University, Ghent, Belgium

[2]Laboratory for Plasma Physics, Royal Military Academy (LPP–ERM/KMS), Brussels, Belgium

[3]Max-Planck-Institut für Plasmaphysik, Garching D-85748, Germany

Belgian Physical Society, 18-05-2016

UNIVERSITEIT GENT

LPP ERM - KMS

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.

Abstract

Detection and quantification of trends of key quantities in terms of a set of 'predictor' variables is a common task for model building and experimental planning in many areas of science, such as astronomy, geology, ecology and also in nuclear fusion science. The standard way to handle the corresponding regression analysis problem is by means of a linear or power-law regression function and ordinary least squares (OLS) to perform the fit. However, OLS is a very simple technique that is not suitable in the presence of complex uncertainties on the measured data. Its assumptions can be overly simplifying, e.g. when the measurements originate from multiple diagnostics or experiments, when the predictor variables are affected by considerable uncertainty, or when the data contain outliers. This often leads to erroneous estimates of the regression parameters, which, moreover, greatly depend on the adequateness of the proposed regression function. Furthermore, the measurements used in the regression analysis are often averages over a time window or over multiple occurrences of the phenomenon under study. Effectively, this means that potentially valuable information in the data is discarded. Whenever a measured quantity is subject to considerable fluctuation or measurement noise it can be very beneficial to consider the probability distribution of the quantity instead of its average. We have developed the method of ***geodesic least squares regression*** (GLS) that does not depend on the overly simplifying assumptions of OLS, by exploiting the full probability distribution of the regression variables. In the present contribution, the method is applied to regression analysis of plasma energy confinement, resulting in strongly improved robustness with respect to uncertainty in both the data set and in the regression model.

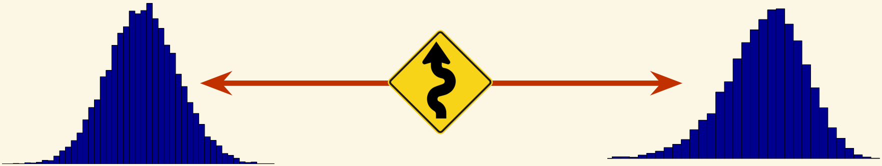
Scaling laws

- Scaling laws in fusion science:
  - Evaluate theoretical predictions
  - Estimate parametric dependencies
  - Extrapolate to future devices
- Terminology:
  - ***Scaling*** law: scale to larger sizes, magnetic fields, etc.
  - Often power law: $y = b_0 x_1^{b_1} x_2^{b_2} \dots x_p^{b_p}$
  - ***Regression analysis***: probabilistic/statistical framework for estimation with confidence intervals
  - Scaling law estimation $\subset$ regression analysis $\subset$ parameter estimation
- Applications in astronomy, geology, ecology, ...

Challenges for fusion scaling laws

- Large (non-Gaussian?) stochastic uncertainties (noise)
- Systematic measurement uncertainties
- Uncertainty on response (y) ***and*** predictor (x_j) variables
- Uncertainty on regression model (nonlinear?)
- Near-collinearity of predictor variables
- Atypical observations (outliers)
- Heterogeneous data and error bars
- Logarithmic transformation in power laws:

$$\ln(y) = \ln(b_0) + b_1 \ln(x_1) + b_2 \ln(x_2) + \ldots + b_p \ln(x_p)$$



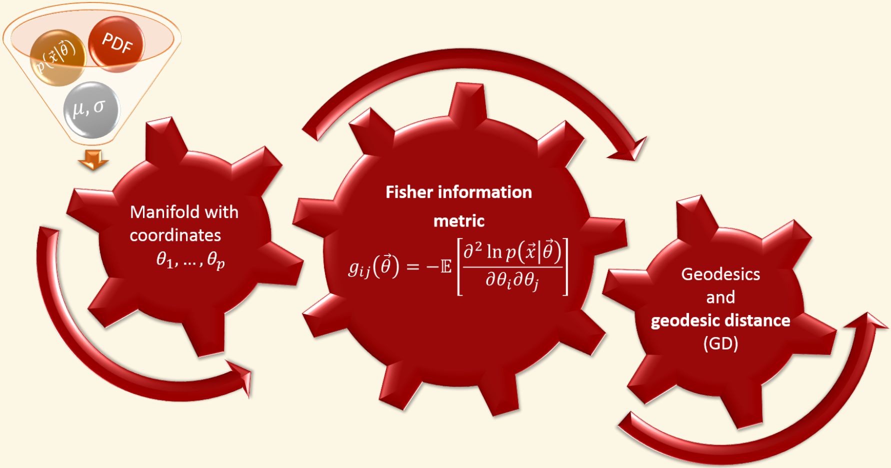
The minimum distance approach

- Need ***robust regression*** considering all uncertainties
- Parameter estimation $\rightarrow$ distance minimization: expected $\leftrightarrow$ measured:
  - Ordinary least squares (***OLS***)
  - Maximum likelihood (ML) / maximum ***a posteriori*** (MAP):

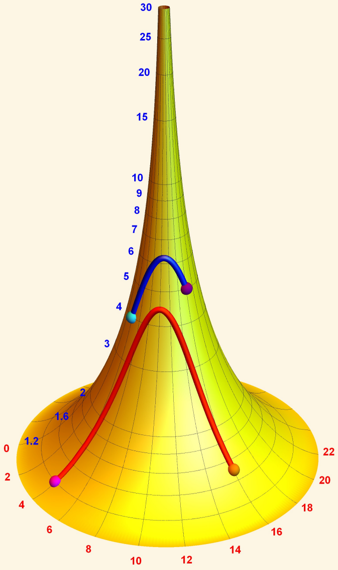
$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{[y - f(x,\theta)]^2}{\sigma^2} \right\}$$

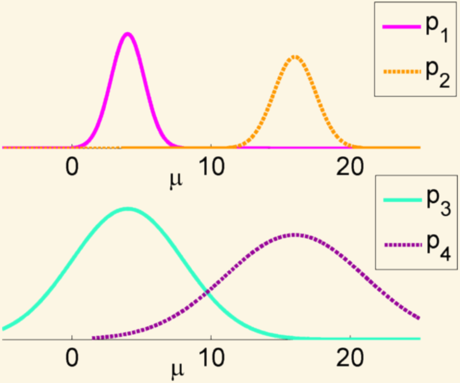
- Measurement $\rightarrow$ ***probability distribution***
- ***Minimum distance estimation***: Hellinger divergence, Kullback-Leibler divergence, ...
- Firm mathematical basis: ***information geometry***
  $\Longrightarrow$ regression on ***probabilistic manifolds***

Information geometry



The Gaussian probability space





Estimation through distance minimization

$$\frac{1}{\sqrt{2\pi\left(\sigma_y^2+\sum_{j=1}^m \beta_j^{\,2}\sigma_{x,j}^2\right)}}\exp\left\{-\frac{1}{2}\frac{\left[y-\left(\beta_0+\sum_{j=1}^m \beta_j x_{ij}\right)\right]^2}{\sigma_y^2+\sum_{j=1}^m \beta_j^{\,2}\sigma_{x,j}^2}\right\}$$

$$\updownarrow$$

Rao geodesic distance (GD)

$$\frac{1}{\sqrt{2\pi}\,\sigma_{\text{obs}}}\exp\left[-\frac{1}{2}\frac{(y-y_i)^2}{\sigma_{\text{obs}}^{\;2}}\right]$$

- Minimize GD between ***modeled*** (p_{mod}) and ***observed*** (p_{obs}) ***distributions***
- To be estimated: $\sigma_{\text{obs}}, \beta_0, \beta_1, \ldots, \beta_m$
- iid data: minimize sum of squared GDs

  $\Longrightarrow$ ***geodesic least squares (GLS)*** regression

Numerical experiment: L-H power threshold

- Log-linear model:

$$P_{\mathrm{thr}} = \beta_0 \bar{n}_{\mathrm{e}}^{\beta_1} B_{\mathrm{t}}^{\beta_2} S^{\beta_3}$$
$$\Longrightarrow \ln P_{\mathrm{thr}} \approx \ln \beta_0 + \beta_1 \ln \bar{n}_{\mathrm{e}} + \beta_2 \ln B_{\mathrm{t}} + \beta_3 \ln S$$

  - P_{thr}: L-H power threshold (MW)
  - $\bar{n}_{\mathrm{e}}$: central line-averaged electron density ($10^{20}\ \mathrm{m}^{-3}$)
  - B_{t}: toroidal magnetic field (T)
  - S: plasma surface area (m^2)

- ITPA Power Threshold Database: ***2002 version***

  (J. Snipes *et al.*, IAEA FEC 2002, CT/P-04)

- Data + error bars from 7 tokamaks: $>$ 600 entries
- $p_{\mathrm{mod}} \approx \mathcal{N}(\mu_{\mathrm{mod}}, \sigma^2_{\mathrm{mod}})$:

$$\mu_{\mathrm{mod}} = \ln \beta_0 + \beta_1 \ln \bar{n}_{\mathrm{e}} + \beta_2 \ln B_{\mathrm{t}} + \beta_3 \ln S$$
$$\sigma^2_{\mathrm{mod}} = \beta_1^2 \sigma^2_{\ln \bar{n}_{\mathrm{e}}} + \beta_2^2 \sigma^2_{\ln B_{\mathrm{t}}} + \beta_3^2 \sigma^2_{\ln S}$$

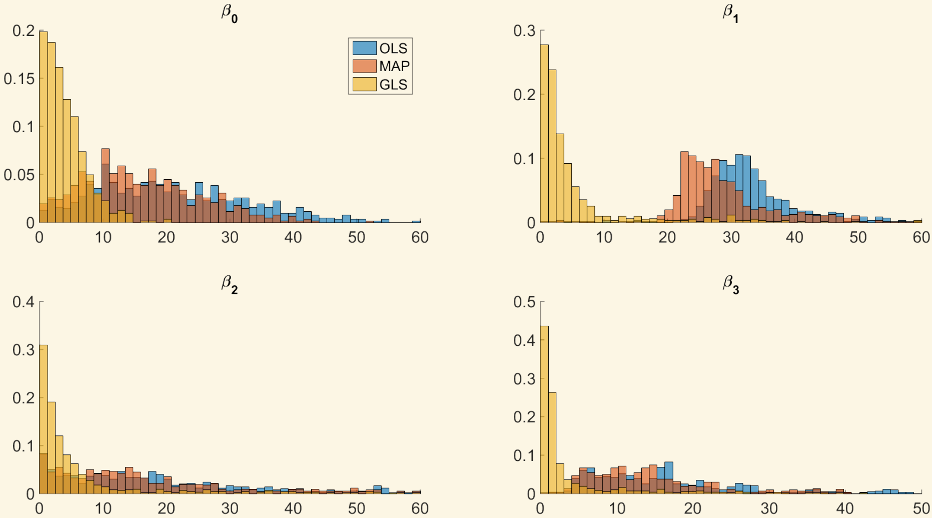
Synthetic regression models

$$\ln P_{\mathrm{thr}} = \ln \beta_0 + \beta_1 \ln \bar{n}_{\mathrm{e}} + \beta_2 \ln B_{\mathrm{t}} + \beta_3 \ln S$$

- β_0: $1, 1.1, \ldots, 20$
- β_1, β_2, β_3: $0.1, 0.2, \ldots, 2$
- Percentage errors:
    - P_{thr}: ***15%***
    - $\bar{n}_{\mathrm{e}}$: ***20%***
    - B_{t}: ***5%***
    - S: ***15%***
- 10 trials per parameter set

Experimental results

Percentage error on parameter estimates



Energy confinement scaling in tokamaks

$$\tau_{\mathrm{E}} = \beta_0 \; I_{\mathrm{p}}^{\beta_1} \; B_{\mathrm{t}}^{\beta_2} \; \bar{n}_{\mathrm{e}}^{\beta_3} \; P_{\mathrm{loss}}^{\beta_4} \; R^{\beta_5} \; \kappa^{\beta_6} \epsilon^{\beta_7} \; M_{\mathrm{eff}}^{\beta_8}$$

- τ_{thr}: thermal energy confinement time (s)
- I_{p}: plasma current (*MA*)
- B_{t}: toroidal magnetic field (*T*)
- $\bar{n}_{\mathrm{e}}$: central line-averaged electron density ($10^{20}\ \mathrm{m}^{-3}$)
- P_{loss}: thermal power loss (MW)
- R: plasma major radius (m)
- κ: plasma elongation
- ϵ: inverse aspect ratio
- M_{eff}: effective atomic mass

- ITPA Global H-mode Confinement Database

  (D.C. McDonald *et al.*, Nucl. Fus. **47**, pp. 147–174, 2007)

- ‘Standard set’: $>$ 1200 entries from 6 tokamaks

Comparison of trends

$$\tau_{\mathrm{E}} = \beta_0 \; I_{\mathrm{p}}^{\beta_1} \; B_{\mathrm{t}}^{\beta_2} \; \bar{n}_{\mathrm{e}}^{\beta_3} \; P_{\mathrm{loss}}^{\beta_4} \; R^{\beta_5} \; \kappa^{\beta_6} \epsilon^{\beta_7} \; M_{\mathrm{eff}}^{\beta_8}$$

Log-linear

Meth.	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
OLS	0.030	***0.80***	***0.57***	***0.39***	-0.70	2.3	***0.52***	***0.33***	***0.34***
GLS	0.035	***0.58***	***0.77***	***0.44***	-0.78	2.5	***0.90***	***0.84***	***0.42***

Nonlinear

Meth.	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
OLS	0.034	***0.56***	***0.53***	***0.56***	-0.69	2.7	***0.74***	***0.85***	***0.15***
GLS	0.042	***0.50***	***0.77***	***0.37***	-0.74	2.5	***0.99***	***1.0***	***0.45***

Results

$$\tau_{\mathrm{E}} = \beta_0 \; I_{\mathrm{p}}^{\beta_1} \; B_{\mathrm{t}}^{\beta_2} \; \bar{n}_{\mathrm{e}}^{\beta_3} \; P_{\mathrm{loss}}^{\beta_4} \; R^{\beta_5} \; \kappa^{\beta_6} \epsilon^{\beta_7} \; M_{\mathrm{eff}}^{\beta_8}$$

- Weaker dependence on I_{p}
- Stronger dependence on B_{t}
- Stronger dependence on κ
- Stronger dependence on ϵ (minor radius)
- ITER predictions:

Log-linear:

- OLS: ***5.6 s***
- GLS: ***4.2 s***

Nonlinear:

- OLS: ***5.9 s***
- GLS: ***3.7 s***

Conclusions and future work

- Geodesic least squares regression: ***flexible*** and ***robust***
- ***Consistent results***
- ***Easy*** to use, ***fast*** optimization
- Application to scaling laws in fusion, astronomy, ecology, etc.
- To be implemented in publicly accessible software package