UNIVERSITEIT GENT

FACULTY OF ENGINEERING AND ARCHITECTURE

LPP ERM - KMS

Geodesic least squares regression with applications to astrophysical data

G. Verdoolaege

Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium

Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium

Abstract

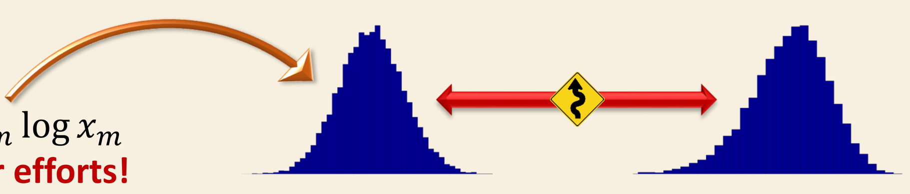
Regression analysis on astrophysical data is a relatively challenging activity and it has long been recognized that many of the assumptions of ordinary least squares regression (OLS) are not valid in several applications in the field. Accordingly, various techniques from the domains of frequentist statistics and Bayesian probability theory have been proposed to address the shortcomings of OLS. We here present a new regression method called '**geodesic least squares regression**' (GLS), which has recently been developed and applied in the field of magnetic confinement fusion (MCF) [1,2]. The main difference with standard techniques is that, for the case of a single response variable, the distribution of the response conditional on a value suggested by the regression model ('modeled distribution'), need not be the same as its distribution conditional on an actual measurement ('observed distribution'). Then, instead of minimizing the difference between modeled and observed values of the response variable, which is the goal of standard OLS, GLS aims at minimizing the distance between the modeled and observed distributions. To this end, we use the Rao geodesic distance (GD) on the probabilistic manifold corresponding to the distribution in the regression model, equipped with the Fisher information metric. The method can handle errors in all variables, is robust against data outliers and uncertainty in the regression model, and can be used with arbitrary distribution models and regression functions. After introducing GLS and demonstrating its advantages on a synthetic data set, we show results of fitting MCF scaling laws as well as the baryonic Tully-Fisher relation in astronomy.

Motivation

- In many areas of science, regression analysis is used:
  - As an aid to build and validate **theoretical models** from data and 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
- 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: $y = b_0 x_1^{b_1} x_2^{b_2} \ldots x_m^{b_m}$
  - Heterogeneous data and error bars, correlations, non-Gaussian probability distributions
  - Atypical observations (outliers)
  - Near-collinearity of predictor variables
  - Data transformations, e.g.

    $\log y = \log b_0 + b_1 \log x_1 + \ldots + b_m \log x_m$



- **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 $\vec{\theta} = [\theta_1, \ldots, \theta_p]^{\mathrm{T}}$: $y = f(x_1, \ldots, x_m|\vec{\theta})$ ← GLS

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

Minimize error

Zero error on the x_{ij}!

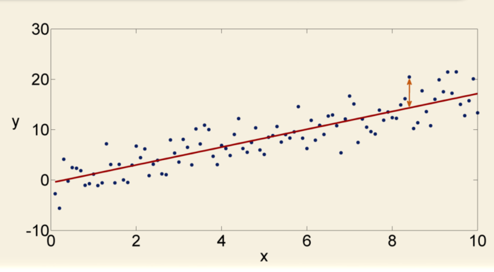
Least squares estimation

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

or

Maximum likelihood estimation (MLE) (e.g. Gaussian likelihood)

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



Includes errors on the x_{ij}!

Modeled distribution (e.g. Gaussian)

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

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

Minimize distance

Observed distribution (e.g. Gaussian)

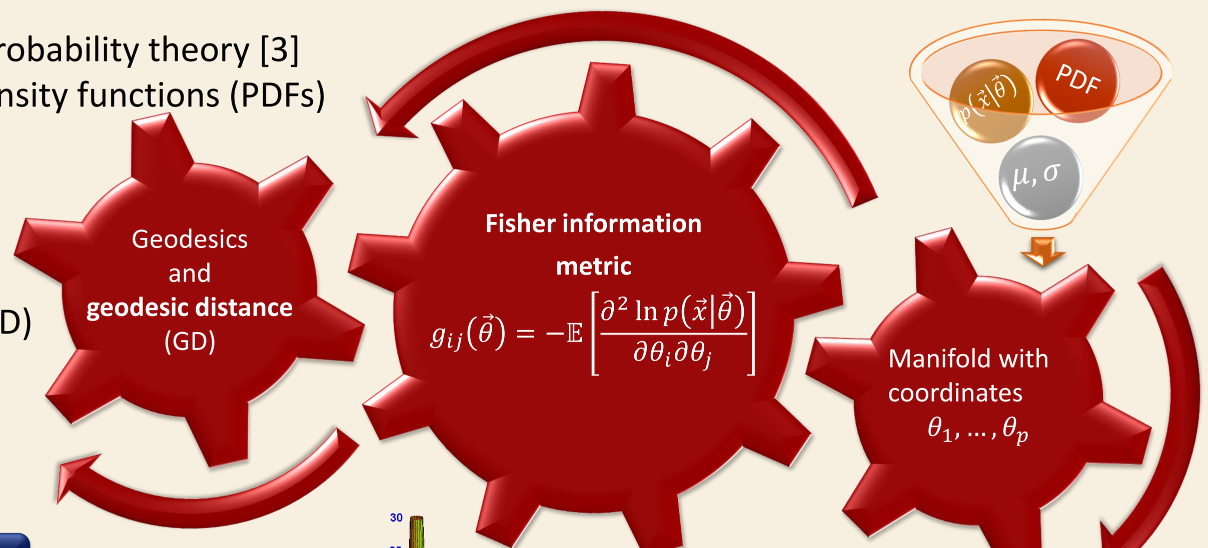
$$p(y|y_1, \ldots, y_n) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{obs}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{[y - y_i]^2}{\sigma_{\mathrm{obs}}^2}\right\}$$

σ_{obs} to be estimated from the data

Which distance measure between probability distributions?

Information geometry

- Geometric approach to probability theory [3]
- 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)

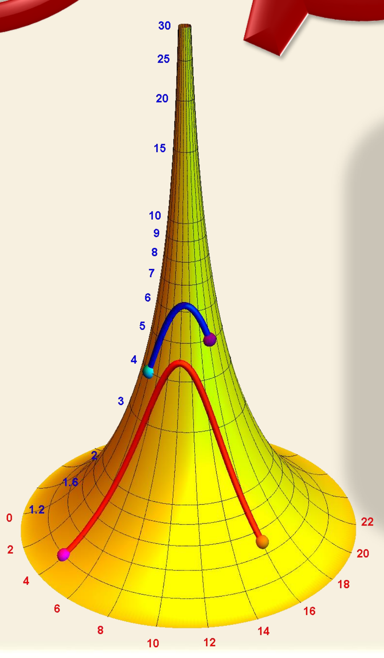


Example: Gaussian manifold

- $p_1(x|\mu_1, \sigma_1) \leftrightarrow p_2(x|\mu_2, \sigma_2)$
- $\mathrm{GD}(p_1, p_2) = 2\sqrt{2}\tanh^{-1}\delta$,

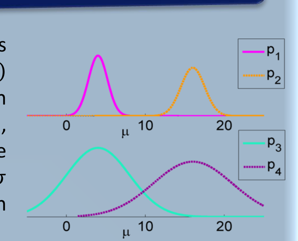
$$\delta = \left[\frac{(\mu_1-\mu_2)^2 + 2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2 + 2(\sigma_1+\sigma_2)^2}\right]$$

- 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.



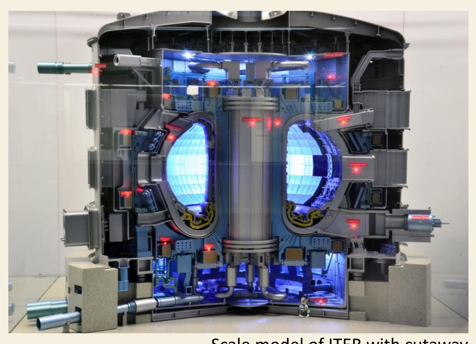
Indeed, $\mathrm{GD}(p_3, p_4) = 2.4 < \mathrm{GD}(p_1, p_2) = 5.3$, whereas with the Euclidean distance $\mathrm{ED}(p_3, p_4) = 12.04 > \mathrm{ED}(p_1, p_2) = 12.00$. The distributions are mapped on the pseudosphere on the left.

Scaling in magnetic confinement fusion

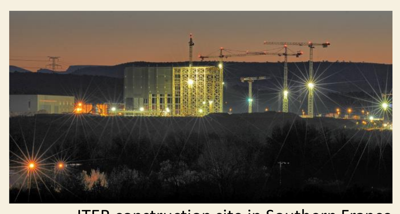
H-mode power threshold

- Controlled nuclear fusion: clean, safe, limitless energy
- Magnetic confinement fusion: tokamaks (ITER), stellarators, …
- High confinement mode (H-mode): threshold P_{thr} on input power
- Scaling with classic power law:

$$P_{\mathrm{thr}} = \beta_0 \bar{n}_{\mathrm{e}}^{\beta_1} B_{\mathrm{t}}^{\beta_2} S^{\beta_3}, \quad \begin{cases} \bar{n}_{\mathrm{e}} = \text{average plasma density } (10^{19}\ \mathrm{m}^{-3}) \\ B_{\mathrm{t}} = \text{toroidal magnetic field (T)} \\ S = \text{plasma surface area } (\mathrm{m}^2) \end{cases}$$



Scale model of ITER with cutaway

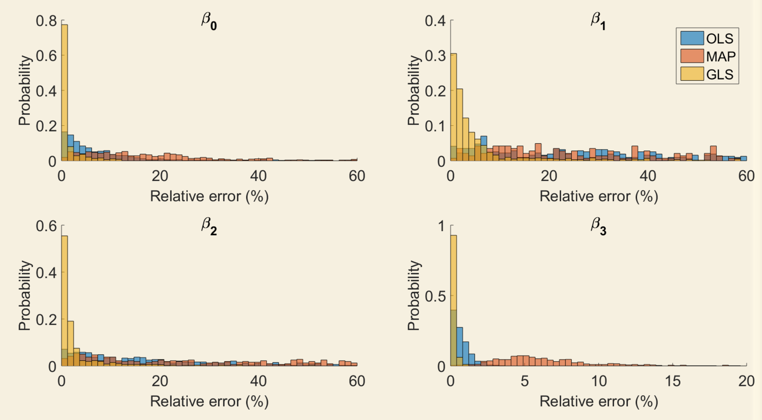


ITER construction site in Southern France

- **ITPA H-mode threshold database** [4]: 645 measurements from 7 tokamaks
- Logarithmic variables assumed Gaussian: single standard deviation = relative error from database

1. Synthetic data: outliers

- Linear regression: $\eta = \beta_0 + \beta_1 \bar{n}_{\mathrm{e}} + \beta_2 B_{\mathrm{t}} + \beta_3 S$
- Artificial data sets: $\begin{cases} \beta_0 = 1, 1.1, \ldots, 20 \\ \beta_1, \beta_2, \beta_3 = 0.1, 0.2, \ldots, 2 \end{cases}$
- Gaussian noise: 4% on $\bar{n}_{\mathrm{e}}$, 1% on B_{t}, 3% on S, 15% on P_{thr}

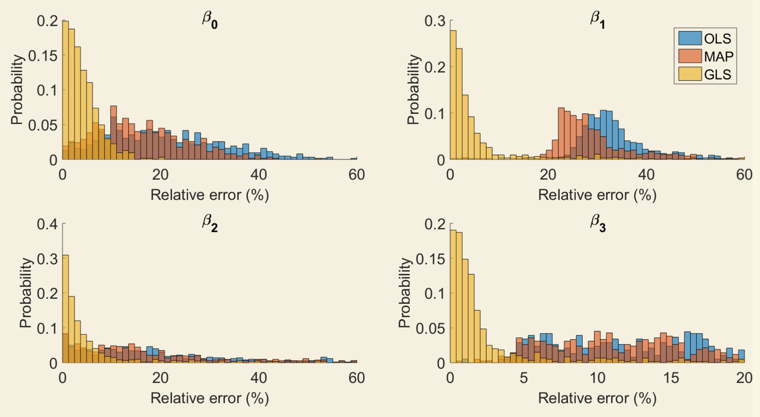


2. Synthetic data: logarithmic space

- Logarithmic transformation:

$$\eta = \log\beta_0 + \beta_1 \log\bar{n}_{\mathrm{e}} + \beta_2 \log B_{\mathrm{t}} + \beta_3 \log S$$

- Gaussian noise: 20% on $\bar{n}_{\mathrm{e}}$, 5% on B_{t}, 15% on S, 15% on P_{thr}



3. Real data: comparison of loglinear with nonlinear

Loglinear (10 bootstraps, 95% conf. interv.)

Method	β_0	β_1	β_2	β_3	$P_{\mathrm{thr}\,0.5}$ (MW)
GLS	0.043 ±0.004	0.66 ±0.07	0.80 ±0.06	0.95 ±0.03	48 ± 5
OLS	0.051 ±0.006	0.49 ±0.07	0.87 ±0.06	0.84 ±0.04	38 ± 4
MAP	0.045 ±0.005	0.57 ±0.08	0.87 ±0.07	0.90 ±0.04	46 ± 5

Nonlinear (10 bootstraps, 95% conf. interv.)

Method	β_0	β_1	β_2	β_3	$P_{\mathrm{thr}\,0.5}$ (MW)
GLS	0.040 ±0.004	0.72 ±0.07	0.75 ±0.08	0.98 ±0.03	52 ± 4
OLS	0.027 ±0.008	0.77 ±0.09	1.0 ±0.1	1.04 ±0.07	70 ± 20
MAP	0.046 ±0.004	0.64 ±0.07	0.79 ±0.08	0.93 ±0.03	44 ±4

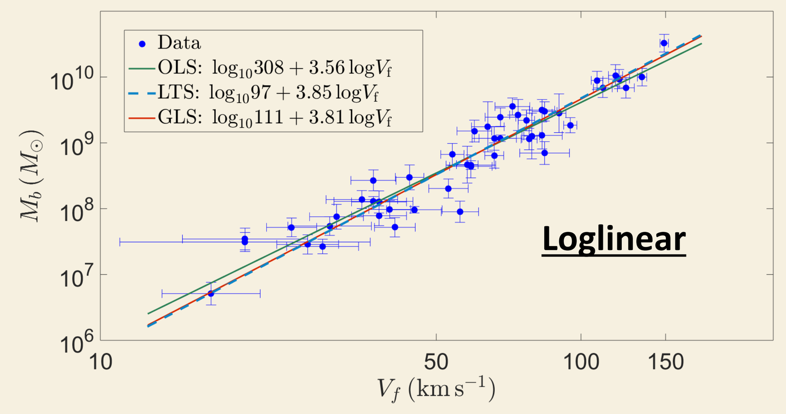
Baryonic Tully-Fisher relation in astronomy

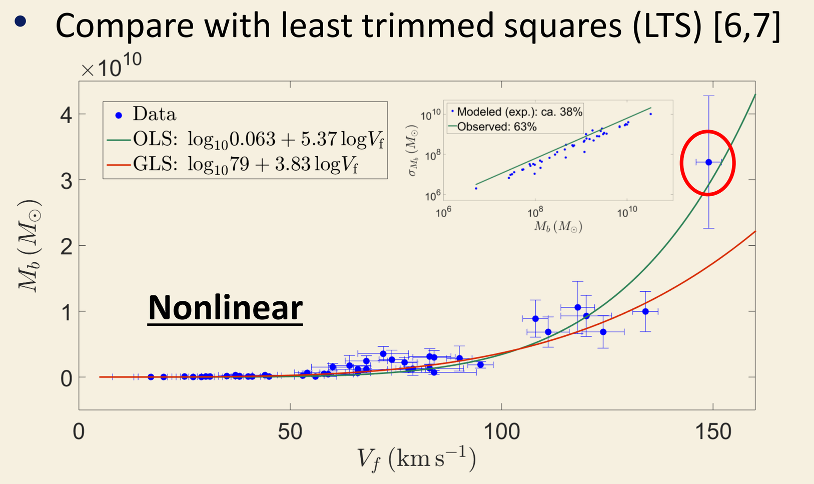
- Relation between rotational velocity and baryonic mass of galaxies
- Various purposes:
  - Distance indicator
  - Constraints on galaxy formation models
  - Test for alternatives to ΛCDM (e.g. MOND) via slope and scatter

NGC 3198

$$M_{\mathrm{b}} = \beta_0\, V_{\mathrm{f}}^{\beta_1} \quad \begin{cases} V_{\mathrm{f}} = \text{rotational velocity in flat part of rotation curve } (\mathrm{km\ s^{-1}}) \\ M_{\mathrm{b}} = M_* + M_{\mathrm{gas}} = \text{total baryonic mass } (M_\odot) \\ \beta_0, \beta_1 = \text{constants} \end{cases}$$

- Data from McGaugh [5] (gas-rich galaxies)
- Compare with least trimmed squares (LTS) [6,7]

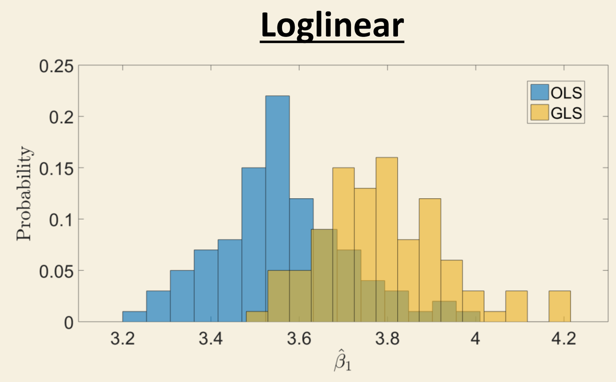




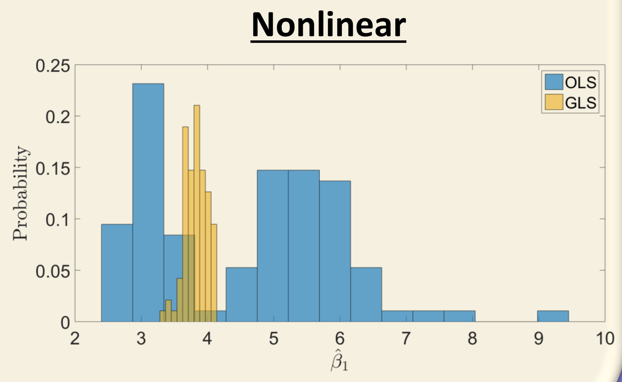
100 bootstraps, 95% conf. interv.

Method	β_0	β_1
OLS loglin.	370 ±400	3.56 ±0.29
OLS nonlin.	2×10^3 $\pm8\times10^3$	4.6 ±2.7
GLS loglin.	140 ±150	3.80 ±0.28
GLS nonlin.	110 ±220	3.82 ±0.34

Loglinear



Nonlinear



Conclusion

- **Geodesic least squares** regression is **flexible** and **robust**
- GLS is **simple** but **powerful** due to strong mathematical foundations
- GLS offers **unified solution** to various regression problems
- Probability distributions **more informative** for regression
- Loglinear regression can be biased w.r.t. nonlinear analysis
- Future development: more accurate **error bars** on GLS estimates and predictions
- GLS will be implemented in a **public software package**

References

[1] G. Verdoolaege, Entropy **17**, 4602, 2015.

[2] G. Verdoolaege *et al.*, Nucl. Fusion **55**, 113019 2015.

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS, New York, 2000.

[4] J.A. Snipes *et al.*, Fusion Energy 2002 (Proc. 19th Int. Conf. Lyon), IAEA, Vienna, CT/P-04.

[5] S. McGaugh, Astron. J. **143**, 40, 2012.

[6] M. Cappellari *et al.*, Mon. Not. R. Astron. Soc. **432**, 1709, 2013.

[7] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.