EUROfusion

Pattern recognition in nuclear fusion data by means of geometric methods in probabilistic spaces

Geert Verdoolaege

Department of Applied Physics, Ghent University, Ghent, Belgium

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

ECEA 2017, November 21 – December 1, 2017

GHENT UNIVERSITY

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.

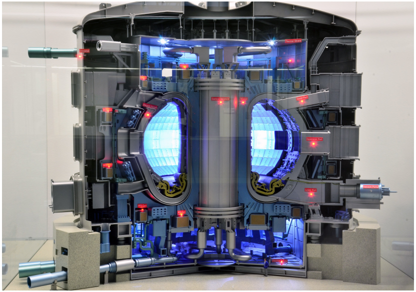
Overview

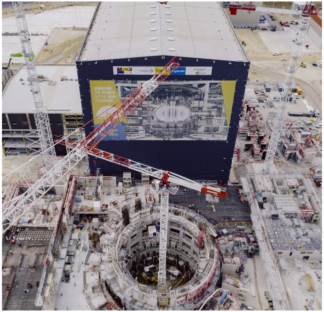
1. Stochastic uncertainty in fusion plasmas
2. Pattern recognition in probabilistic spaces
3. Geodesic least squares regression
4. Application in fusion science: edge-localized plasma instabilities
5. Application in astronomy: Tully-Fisher scaling
6. Conclusion

Overview

Fusion energy

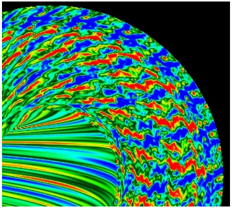
- ‘Star on earth’
- Clean, safe, inexhaustible energy source
- Magnetic confinement fusion: tokamak, stellarator, . . .
- Confine hot hydrogen isotope plasma with magnetic fields
- *ITER*: next-generation international tokamak
- Complex physical system, turbulent transport
- Difficult to probe $\rightarrow$ uncertainty in measurements and models





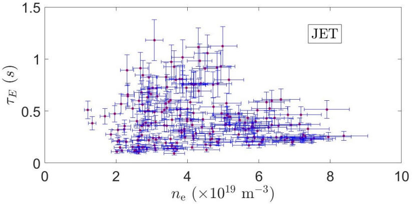
Uncertainty in fusion plasmas

- Sources of statistical uncertainty:
    - Fluctuation of system properties
    - Measurement noise



Plasma turbulence (PPPL)

Edge-localized modes (MAST)

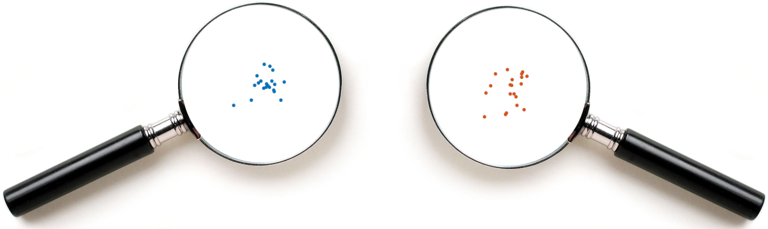


Overview

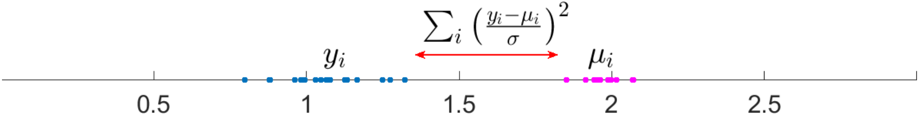
Difference/distance between points

Patterns $\leftrightarrow$ distances

Zooming in...



Mahalanobis distance



Information geometry

- Family of probability distributions $\rightarrow$ differentiable manifold
- Parameters $=$ coordinates
- Metric tensor: ***Fisher information*** matrix

Parametric probability model: $p\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \Longrightarrow$

$$g_{\mu\nu}\left(\boldsymbol{\theta}\right) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^\mu \partial\theta^\nu} \ln p\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\right], \quad \mu,\nu = 1,\ldots,m$$

$\boldsymbol{\theta}$ = m-dimensional parameter vector

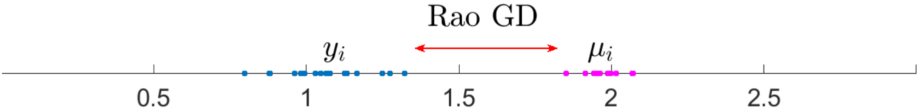
- Line element:

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}\theta^\mu \mathrm{d}\theta^\nu$$

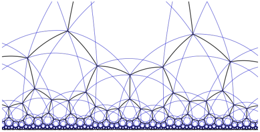
- Minimum-length curve: ***geodesic***
- ***Rao geodesic distance*** (GD)

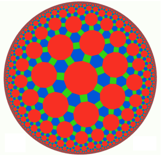
Pattern recognition in probabilistic spaces

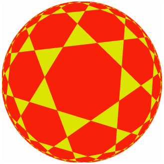
- Pattern recognition:
  - Classification, clustering
  - Regression analysis
  - Dimensionality reduction, visualization
- Observation/prediction (structureless number)
  $\rightarrow$ distribution (structured object)
- More information, more flexibility



The univariate Gaussian manifold







- PDF:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- Line element:

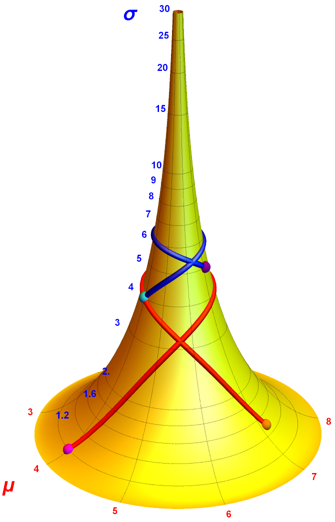
$$ds^2 = \frac{d\mu^2}{\sigma^2} + 2\frac{d\sigma^2}{\sigma^2}$$

- Hyperbolic geometry: Poincaré half-plane, Poincaré disk, Klein disk, ...
- Analytic geodesic distance

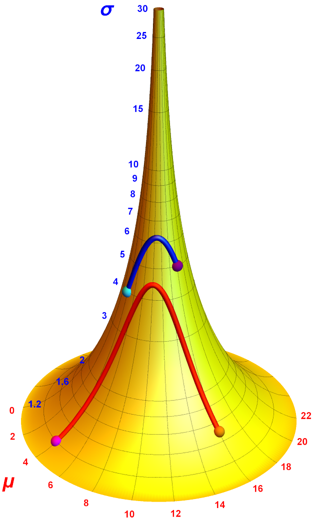
https://www.youtube.com/watch?v=i9IUzNxeH4o

The pseudosphere (tractroid)

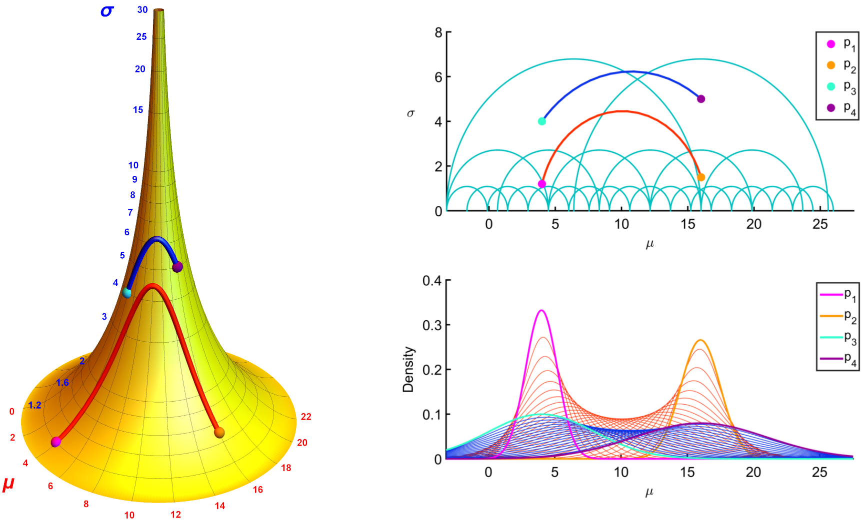
Original



Compressed

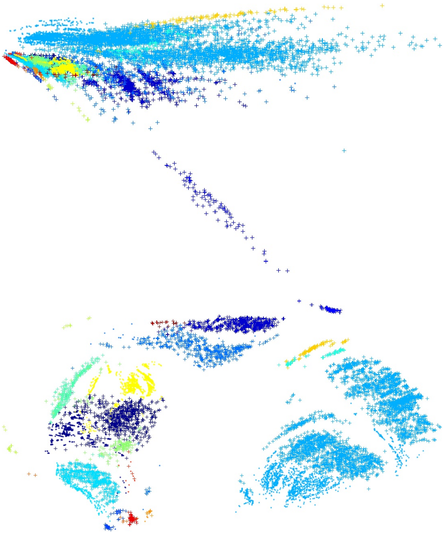


Geodesics on the Gaussian manifold



Data visualization with uncertainty

Plasma energy confinement time w.r.t. global plasma parameters

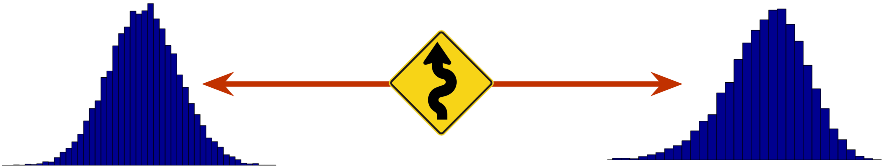


Overview

Challenges in regression analysis

- Data uncertainty: measurement error, fluctuations, ...
- Model uncertainty: missing variables, linear vs. nonlinear, Gaussian vs. non-Gaussian, ...
- Heterogeneous data and error bars
- Uncertainty on response (y) *and* predictor (x_j) variables
- Atypical observations (outliers)
- Near-collinearity of predictor variables
- Data transformations, e.g.

$$\ln(y) = \ln(\beta_0) + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \ldots + \beta_p \ln(x_p)$$



Least squares and maximum a posteriori

- Workhorse: ordinary least squares (OLS)
- Maximum likelihood (ML) / maximum *a posteriori* (MAP):

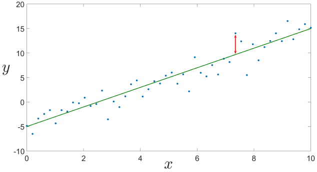
$$p(y_i|x_i,\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i \overset{\leftrightarrow}{-} \mu_i}{\sigma}\right)^2\right]$$

$$\mu_i = f_i(x_i,\boldsymbol{\theta}) \overset{\text{e.g.}}{=} \beta_0 + \beta_1 x_i$$

Michigan, circa 1890s.

- Need ***flexible*** and ***robust*** regression
- Parameter estimation $\rightarrow$ distance minimization:

Expected $\leftrightarrow$ Measured



The minimum distance approach

- ***Minimum distance estimation*** (Wolfowitz, 1952):

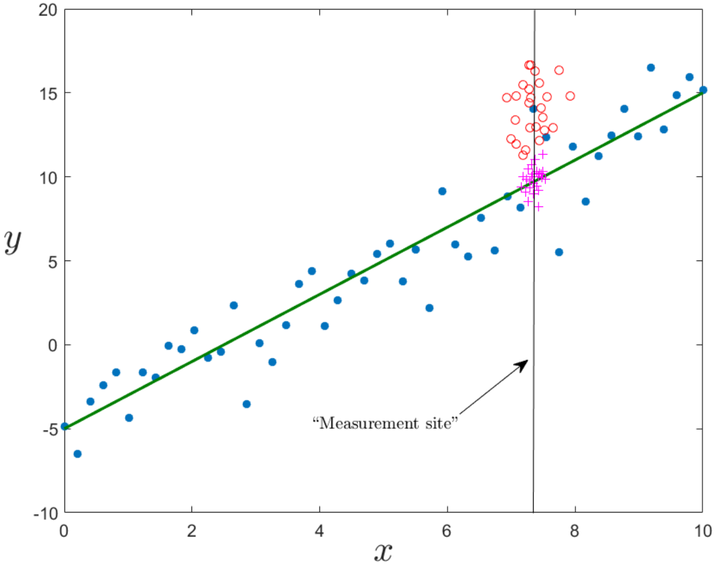
  Which distribution does the model predict?

  vs.

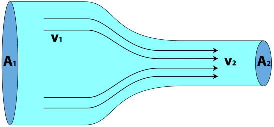
  Which distribution do you observe?

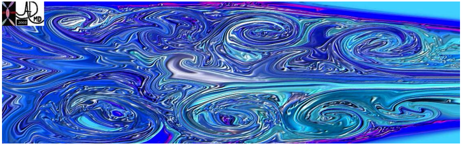
- Gaussian case: different means *and* standard deviations
- Hellinger divergence (Beran, 1977)
- Empirical distribution: kernel density estimate

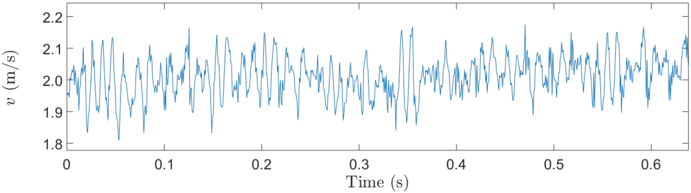
Modeled and *observed* distribution



Example: fluid turbulence







Geodesic least squares

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

Modeled distribution

σ_{mod}^2

Rao GD

Observed distribution

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

- Model-based approach: regression on probabilistic manifold
- To be estimated: $\sigma_{\text{obs}}, \beta_0, \beta_1, \ldots, \beta_m$
- iid data: minimize sum of squared GDs

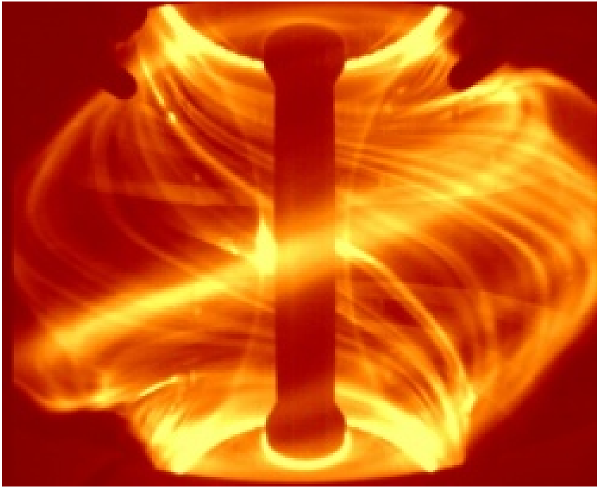
  $\Longrightarrow$ ***geodesic least squares (GLS)*** regression
- If $\sigma_{\text{mod}} = \sigma_{\text{obs}} \to$ Mahalanobis distance

G. Verdoolaege *et al.*, Entropy **17**, 4602, 2015

Overview

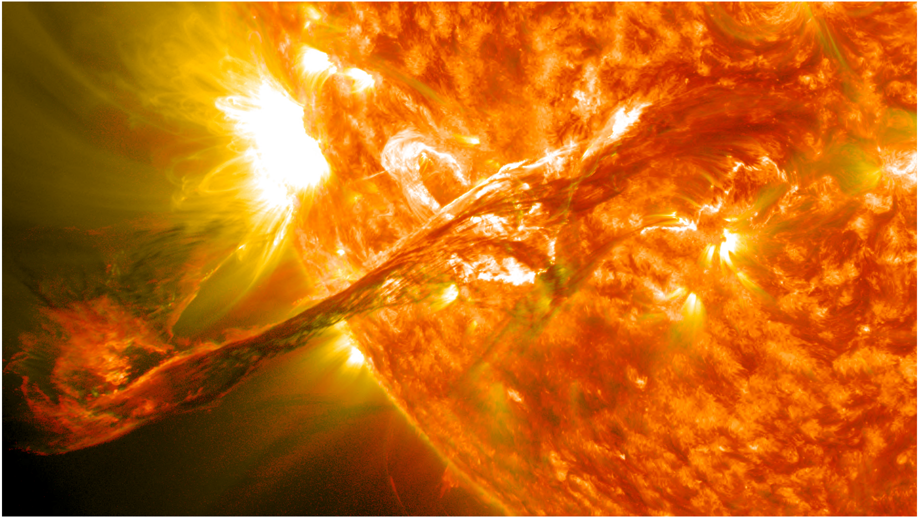
Edge-localized modes (ELMs)

- Repetitive instabilities in plasma edge
- Magnetohydrodynamic origin



MAST, Culham Centre for Fusion Energy, UK

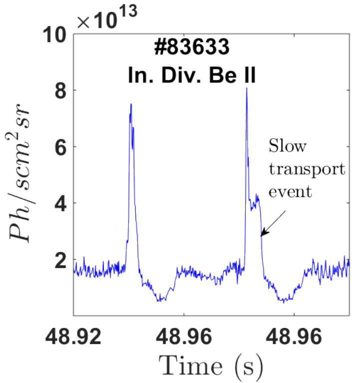
Analogy 1: Solar flares



Analogy 2: Cooking pot

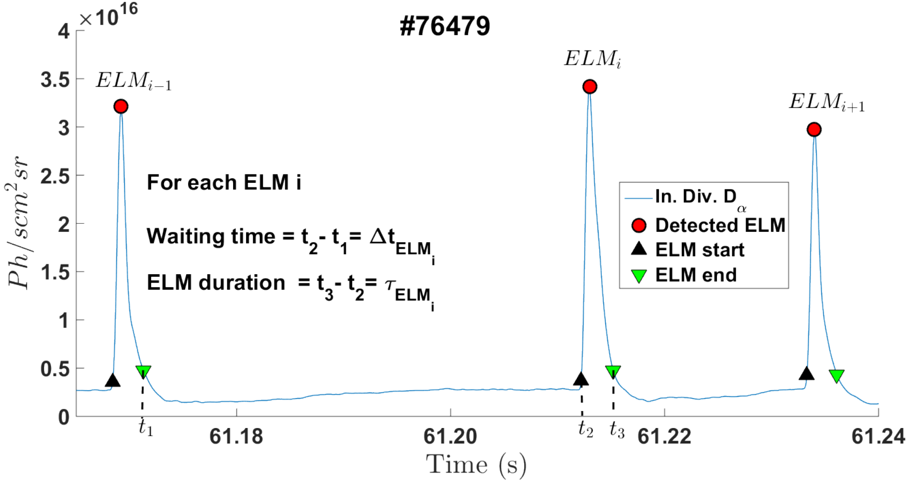
Importance of ELMs

- Confinement loss
- Potential damaging effects
- Impurity outflux
- $\rightarrow$ ELM control/mitigation
- Energy $\propto$ (frequency)$^{-1}$



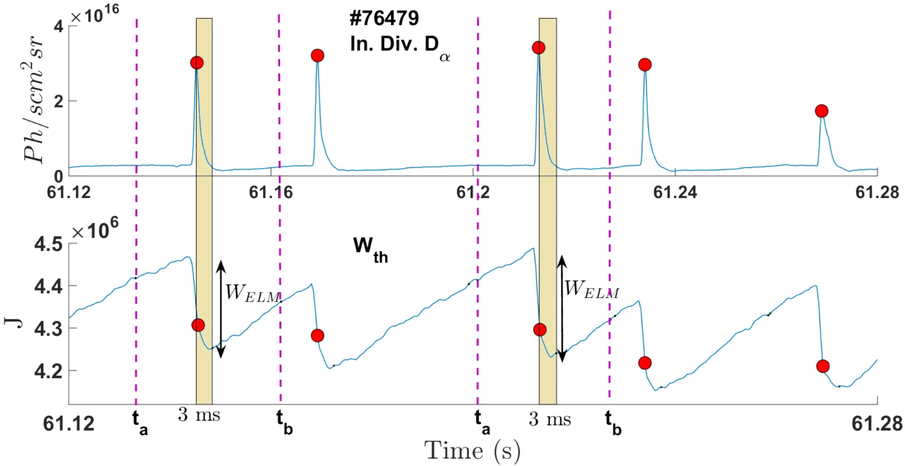
Data extraction: waiting times

- 32 recent JET discharges
- Waiting time: time before ELM burst

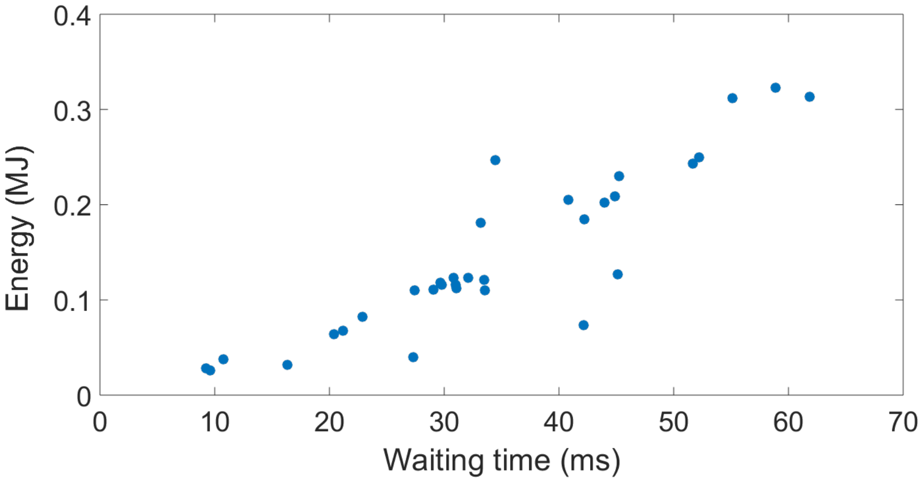


Data extraction: energies

- Energy carried from the plasma by an ELM

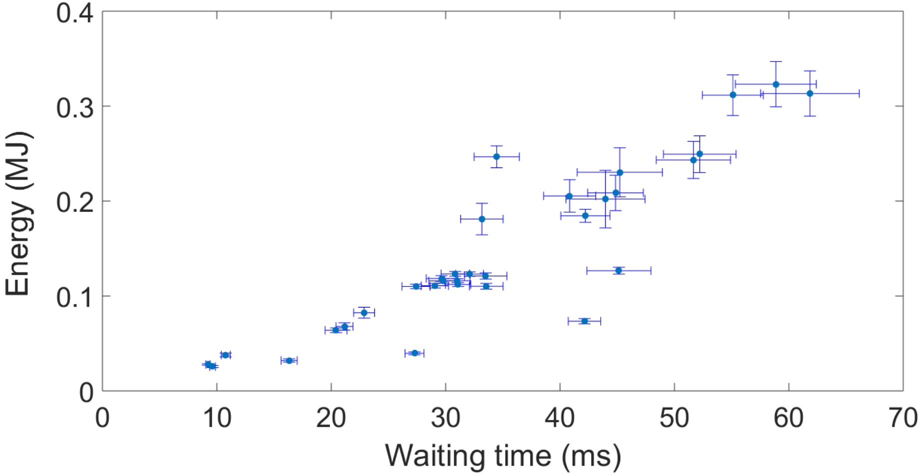


Average waiting times and energies



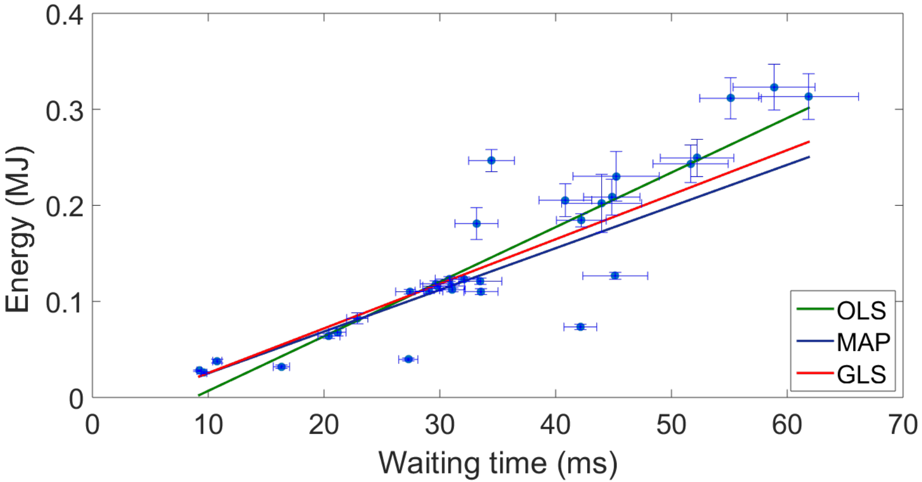
Error bars on averages

- Standard deviation / $\sqrt{n} \rightarrow$ error bars

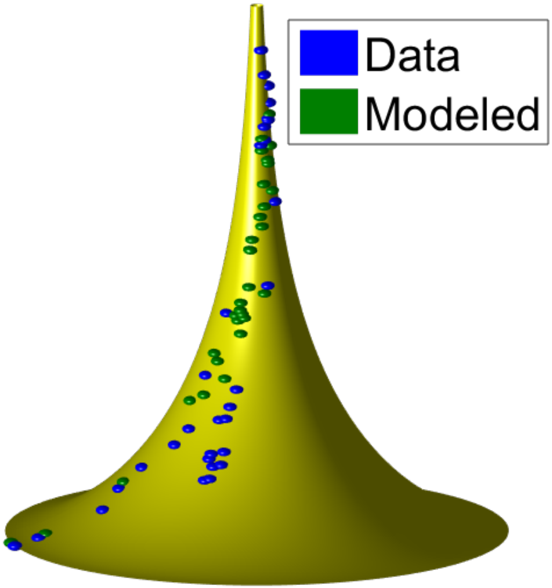


Regression on averages

$$E_{\mathrm{ELM}} = \beta_0 + \beta_1 \Delta t_{\mathrm{ELM}}, \qquad \sigma_{E,\mathrm{obs}} \propto \mu_{E,\mathrm{obs}}$$

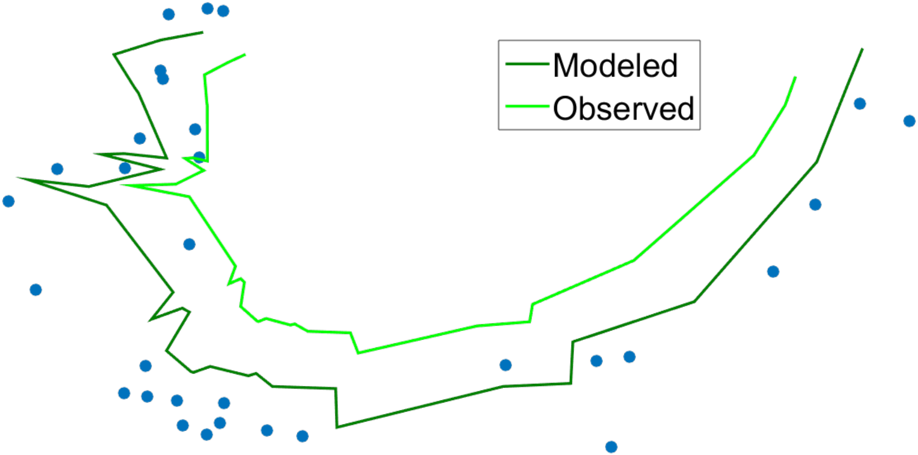


Regression results on pseudosphere



Projected regression results

Multidimensional scaling:



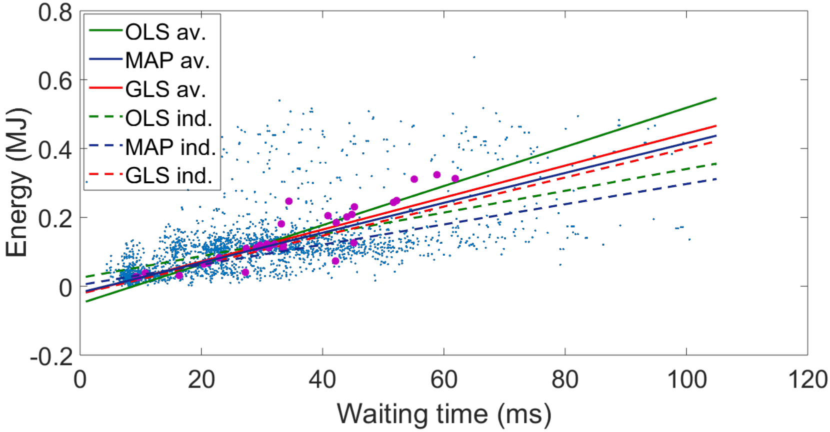
Average vs. collective trend

Average

Method	β_0 (MJ)	β_1 (MJ/s)
OLS	-0.050	5.7
GLS	-0.021	4.6

Individual

Method	β_0 (MJ)	β_1 (MJ/s)
OLS	0.024	3.2
GLS	-0.022	4.2



Overview

Baryonic Tully-Fisher Relation (BTFR)

- Simple, tight relation for disk galaxies:

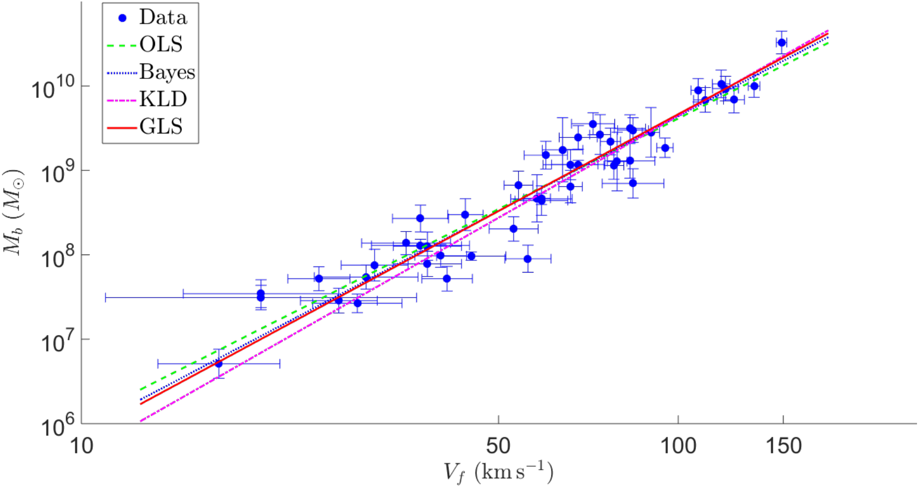
$$M_b = \beta_0 V_f^{\beta_1} \qquad \begin{cases} M_b = \text{total (stellar + gaseous) baryonic mass } (M_\odot) \\ V_f = \text{rotational velocity (km s}^{-1}) \end{cases}$$

- Various purposes:
  - Distance indicator
  - Constraints on galaxy formation models
  - Test for alternatives to ΛCDM cosmological model (slope and scatter)

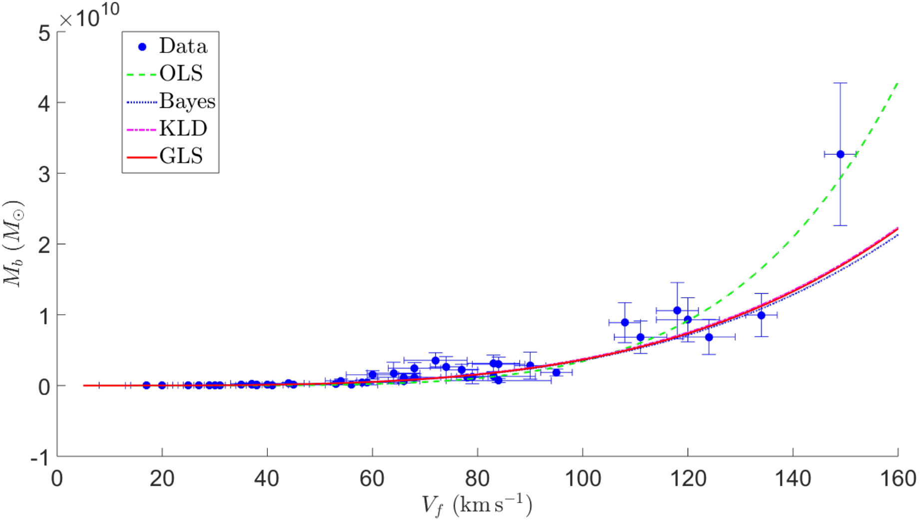
Experiments

- 47 gas-rich galaxies (McGaugh, Astron. J. **143**, 40, 2012)
- Loglinear ($\sigma_{\text{obs},i} \equiv s_{\text{obs}}$) and nonlinear ($\sigma_{\text{obs},i} = r_{\text{obs}}\, M_b$)
- Benchmarking:
  - Ordinary least squares (OLS)
  - Bayesian: errors in all variables, marginalized standard deviations (Bayes)
  - Geodesic least squares (GLS)
  - Kullback-Leibler least squares (KLS)

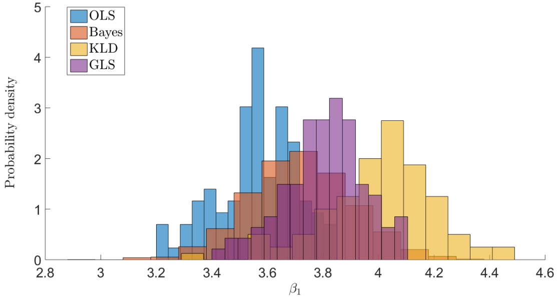
Loglinear regression

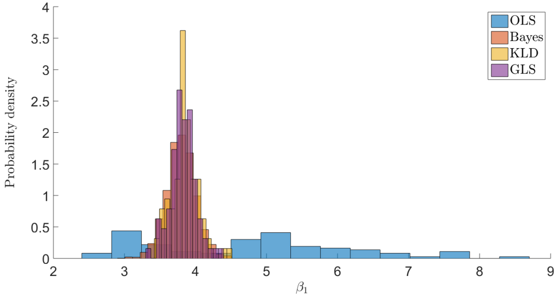


Nonlinear regression

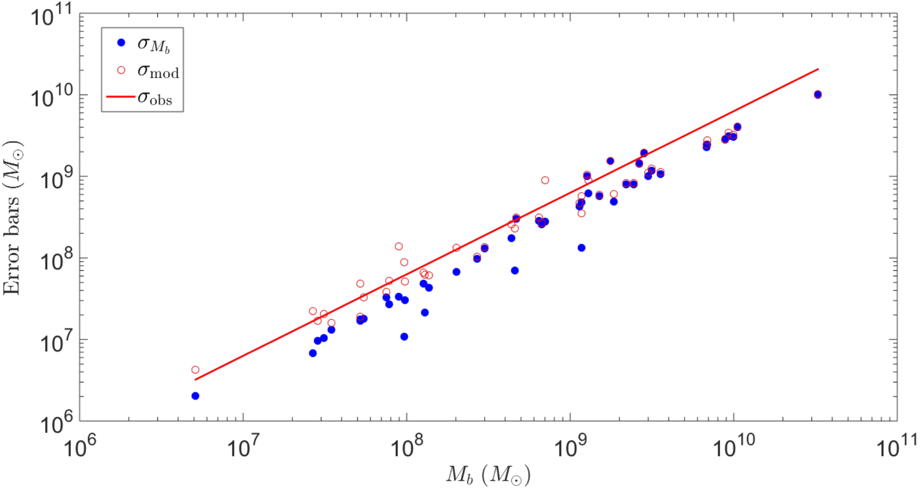


Parameter distributions



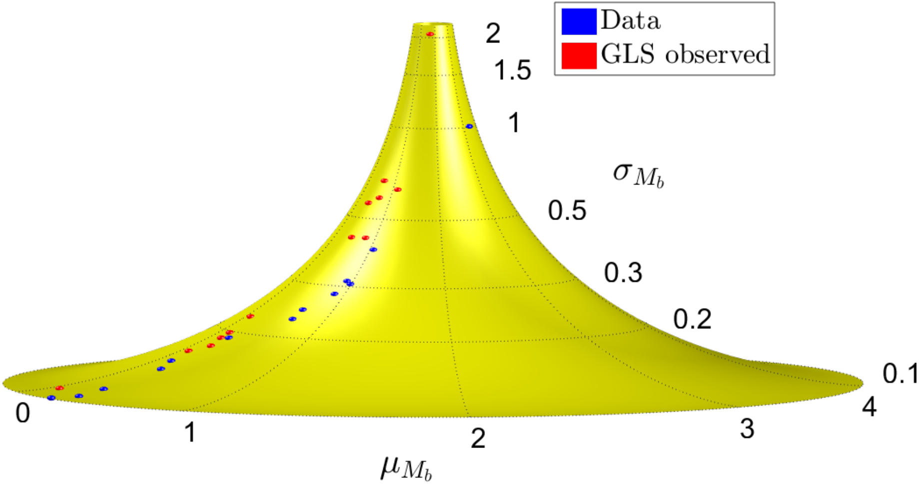


GLS uncertainty estimates



$r_{M_b} \approx 38\%$, $r_{\mathrm{obs}} \approx 63\%$

Interpretation on pseudosphere



Overview

Conclusions

- Probabilistic modeling of stochastic system properties
- Information geometry: distance measure, geometrical intuition
- Pattern recognition in probabilistic spaces
- More information, more flexibility
- Geodesic least squares regression: ***flexible*** and ***robust***
- ***Easy*** to use, ***fast*** optimization