UNIVERSITEIT GENT

FACULTY OF **ENGINEERING AND ARCHITECTURE**

Modeling Fusion Data in Probabilistic Metric Spaces for the Identification of Confinement Regimes and Scaling Laws

Geert Verdoolaege and Guido Van Oost

Department of Applied Physics, Ghent University, Ghent, Belgium

IAEA FEC 2012
San Diego, CA, October 10, 2012

Abstract

Pattern recognition plays an important role and has great potential in fusion data analysis. However, a drawback is that individual measurements are usually represented as unstructured points in a Euclidean data space. We argue that a fundamentally probabilistic approach offers significant advantages. It allows representing the data in a non-Euclidean probabilistic space, wherein the patterns of interest are much more distinct, simply because they are based on more information. In this work, we address the identification of confinement regimes and the establishment of a scaling law for the energy confinement time, using data from the International Global H-mode Confinement Database. We propose a single-level and a Bayesian multilevel model for capturing the statistical data uncertainty. We then show that pattern recognition operations working in the associated probability space are considerably more powerful than their counterparts in a Euclidean data space. This opens up new possibilities for analyzing confinement data and for fusion data processing in general.

Pattern recognition opportunities

- Dimensionality reduction: data visualization
- Clustering/classification: grouping of data points
- Regression: (nonlinear) deterministic relation between variables

Objectives

1. Physics from information: contribute to physics studies by extracting patterns, structure and relations from data
2. Contribute to plasma control through real-time data interpretation

Probability is fundamental

- Measurements are uncertain due to a lack of information:
  - Systematic: estimated through cross-validation
  - Stochastic: needs probability theory
- Traditional measurement: value + error bar
- Measurement = sample from underlying (non-Gaussian?) probability distribution
- Goal of measurement = probing the underlying distribution
- Probability density function (PDF) contains all information about measurement
- PDF is fundamental object resulting from measurement

Problem statement

- Pattern recognition usually operates in Euclidean data spaces with structureless data points
- This neglects additional information in the PDF!
- A huge potential remains unexplored: use additional information on the data to determine patterns:
  - Data probability distribution
  - Established theoretical models
  - Previous experiments

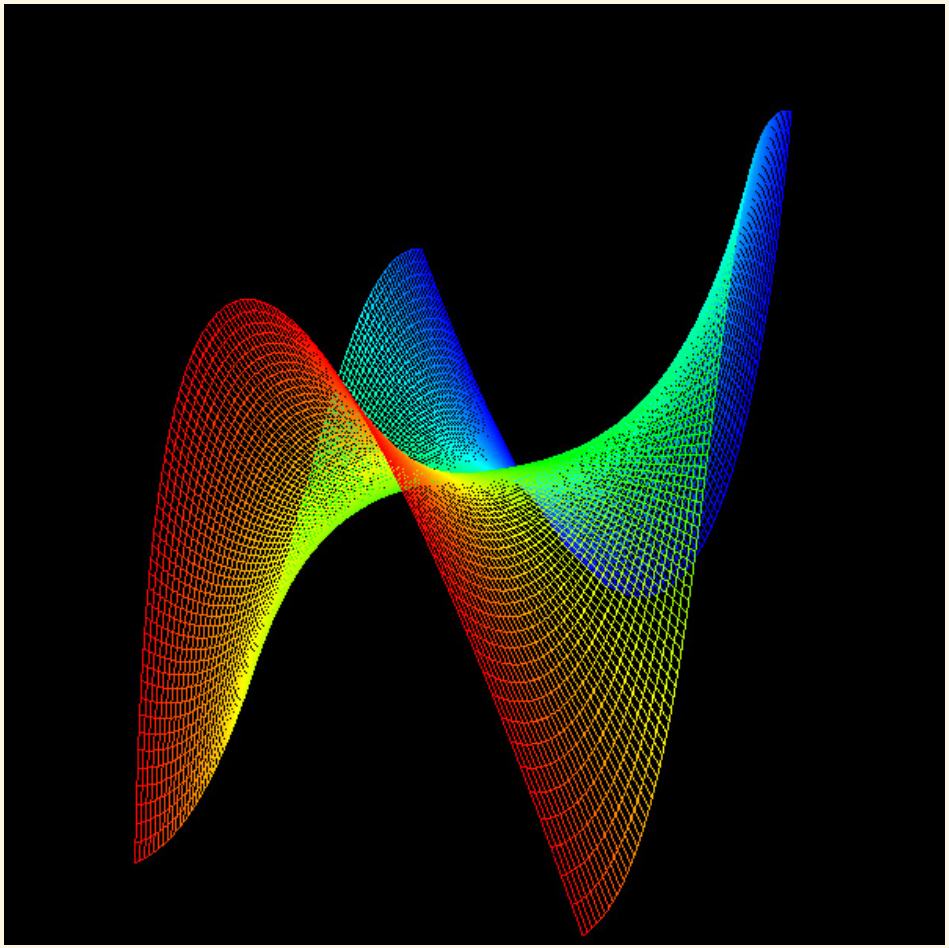
Challenge

To construct a probabilistic pattern recognition framework that exploits all available information for fast and efficient pattern recognition.

$\rightarrow$ Pattern recognition based on non-Euclidean geometry in probability spaces

Probability + geometry: a happy marriage

- Probabilistic manifold:
  - PDF = point on manifold
  - Coordinates = PDF parameters
- Distance between PDFs?
- Information geometry



Information geometry

- Riemannian differential geometry
- Fisher information = unique metric tensor:

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

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

$\vec{\theta}$ = N-dimensional parameter vector

- Line element:

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

- Minimum-length curve: geodesic
- Geodesic distance (GD)
- Natural and theoretically well motivated distance between PDFs

Univariate Gaussian distribution

- PDF:

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

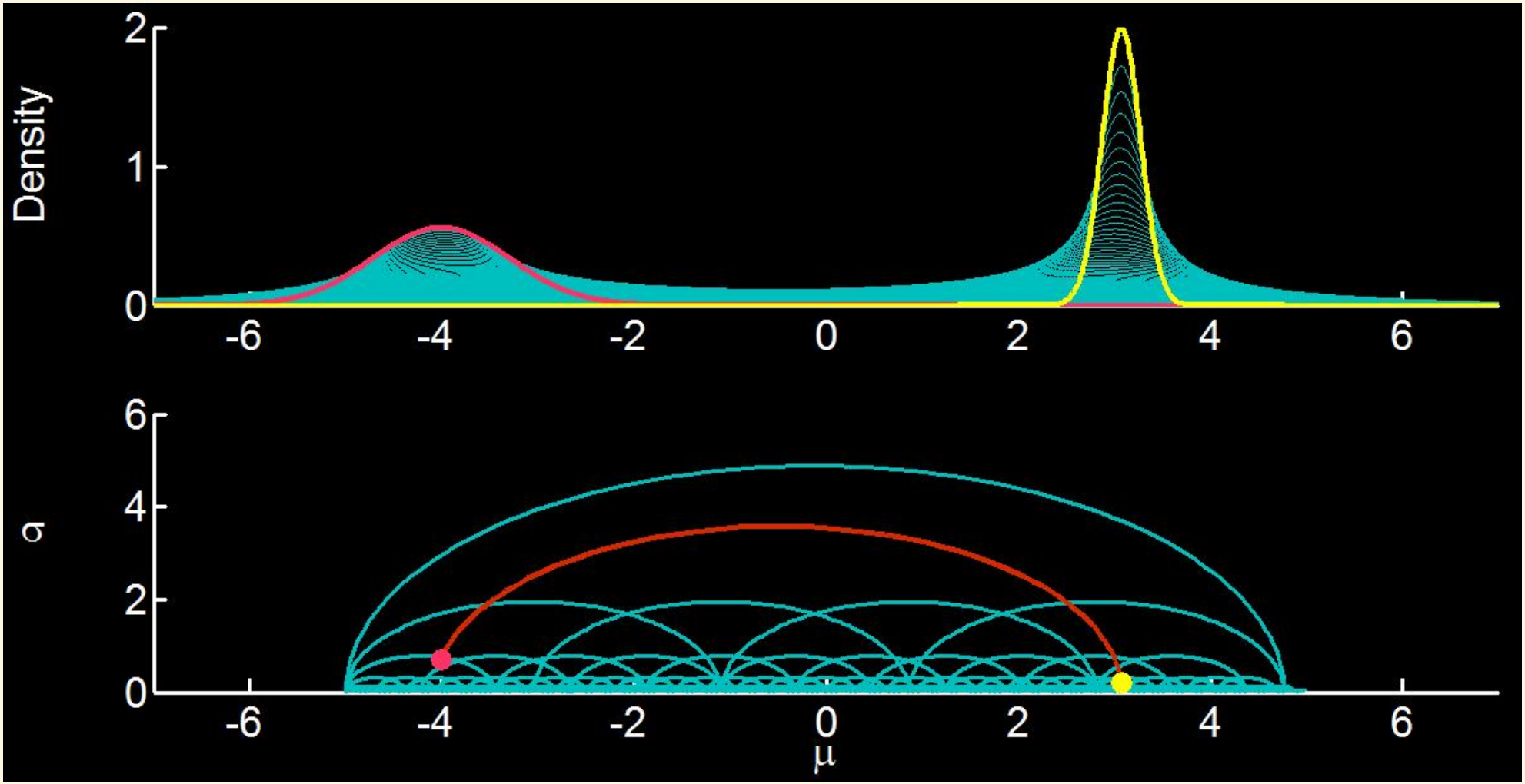
- Line element:

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

- Hyperbolic geometry: Poincaré half-plane model

Poincaré half-plane

p_1: $\mu_1 = -4$, $\sigma_1 = 0.7$; p_2: $\mu_2 = 3$, $\sigma_2 = 0.2$



ITPA Confinement Database

- ITPA Global H Mode Confinement Database (DB3)
  ITER H-Mode Database Working Group
  D.C. McDonald *et al.*, Nucl. Fusion **47**, pp. 147–174, 2007
  http://efdasql.ipp.mpg.de/hmodepublic

- $\sim 10^4$ entries from 19 tokamaks

- Approximate error estimates: limited information on PDF!

- Assume standard deviations $\rightarrow$ Gaussian PDFs (maximum entropy)

- Different machines $\rightarrow$ different error estimates: difficult to handle using classic approach!

Confinement regime classification

- Distinguish between L- and H-mode: 3845 L and 6207 H
- 8 global engineering variables: I_p, B_t, $\bar{n}_\mathrm{e}$, P_loss, R, a, M_eff, κ
- Variables statistically independent $\rightarrow$ product of Gaussians

  Note: this does not exclude the variables to be related through a deterministic relation!

Gaussian product manifold

- Plasma and machine variables: $x_\lambda \to \vec{x}$, $\lambda = 1, \ldots, 8$
- Distribution parameters: μ_λ, σ_λ
- Gaussian product distribution:

$$p\left(\vec{x}|\mu_1, \ldots, \mu_8, \sigma_1, \ldots, \sigma_8\right) = \prod_{\lambda=1}^{8} \mathcal{N}\left(x_\lambda|\mu_\lambda, \sigma_\lambda\right)$$

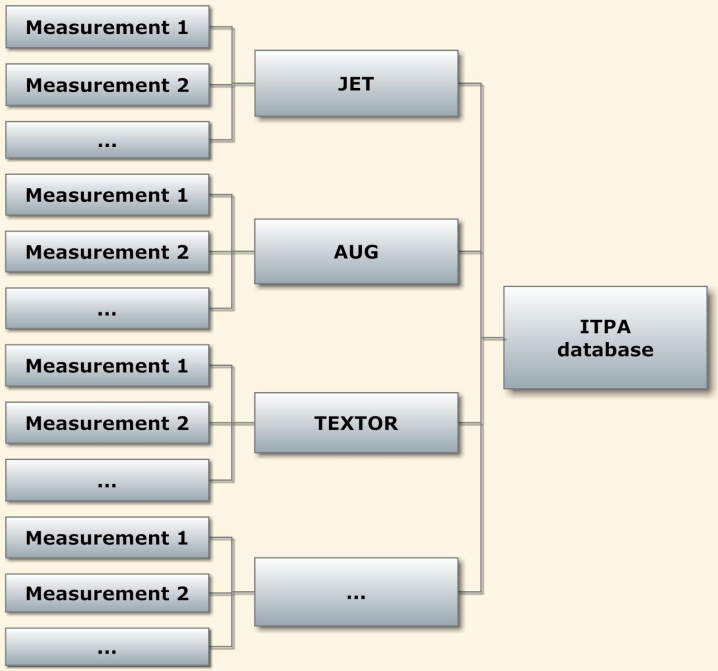
- Measurements A and B: $\vec{\mu}^{A,B} = \left(\mu_1^{A,B}, \ldots, \mu_8^{A,B}\right)$, $\vec{\sigma}^{A,B} = \left(\sigma_1^{A,B}, \ldots, \sigma_8^{A,B}\right)$
- GD in closed form:

$$\mathrm{GD}\left(\vec{\mu}^{\mathrm{A}}, \vec{\sigma}^{\mathrm{A}} || \vec{\mu}^{\mathrm{B}}, \vec{\sigma}^{\mathrm{B}}\right) = \sqrt{2}\left[\sum_{\lambda=1}^{8} \ln^2\left(\frac{1+\delta_\lambda^{\mathrm{AB}}}{1-\delta_\lambda^{\mathrm{AB}}}\right)\right]^{1/2},$$

$$\delta_\lambda^{\mathrm{AB}} = \left[\frac{\left(\mu_\lambda^{\mathrm{A}} - \mu_\lambda^{\mathrm{B}}\right)^2 + 2\left(\sigma_\lambda^{\mathrm{A}} - \sigma_\lambda^{\mathrm{B}}\right)^2}{\left(\mu_\lambda^{\mathrm{A}} - \mu_\lambda^{\mathrm{B}}\right)^2 + 2\left(\sigma_\lambda^{\mathrm{A}} + \sigma_\lambda^{\mathrm{B}}\right)^2}\right]^{1/2}$$

Bayesian multilevel model

- Introduce additional information on tokamak and database means and standard deviations for every variable
- At the minimum: from which tokamak were data obtained? $\rightarrow$ Expected range of variables
- Bayesian hierarchical/multilevel model
- Conjugate prior distributions for means, maximum-likelihood estimates for standard deviations



Level	Model
1	$x_{\lambda,ijk} \sim \mathcal{N}(x_{\lambda,ijk} \mid \mu_{\lambda,jk}, \sigma_{\lambda,jk})$
2	$\mu_{\lambda,jk} \sim \mathcal{N}(\mu_{\lambda,jk} \mid \mu_{\lambda,k}, \sigma_{\lambda,k})$
3	$\mu_{\lambda,k} \sim \mathcal{N}(\mu_{\lambda,k} \mid \mu_{\lambda,0}, \sigma_{\lambda,0})$
4	$\mu_{\lambda,0} \sim \mathcal{N}(\mu_{\lambda,0} \mid \phi_\lambda, \tau_\lambda)$
5	$\phi_\lambda \sim \mathcal{U}(-\infty, +\infty), \quad \tau_\lambda = 0.1\phi_\lambda$

Multilevel posterior distributions

- Bayes' rule:

$$p(\mu_{\lambda,jk}, \mu_{\lambda,k}, \mu_{\lambda,0}, \phi_\lambda | x_{\lambda,ijk}, \forall j, k)$$
$$\sim \prod_{j,k} \mathcal{N}(x_{\lambda,ijk} | \mu_{\lambda,jk}, \sigma_{\lambda,jk}) \mathcal{N}(\mu_{\lambda,jk} | \mu_{\lambda,k}, \sigma_{\lambda,k}) \mathcal{N}(\mu_{\lambda,k} | \mu_{\lambda,0}, \sigma_{\lambda,0})$$
$$\times \mathcal{N}(\mu_{\lambda,0} | \phi_\lambda, \tau_\lambda)$$

- Conditional posteriors for all parameters are also Gaussian, e.g.:

$$\mu_{\lambda,jk} | \mu_{\lambda,k}, \mu_{\lambda,0}, \phi_\lambda, \vec{x}_\lambda \sim \mathcal{N}(\hat{\mu}_{\lambda,jk}, \hat{\sigma}_{\lambda,jk})$$

$$\hat{\mu}_{\lambda,jk} = \frac{\frac{\mu_{\lambda,k}}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}} \bar{x}_{\lambda,\cdot jk}}{\frac{1}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}}} \qquad \hat{\sigma}^2_{\lambda,jk} = \frac{1}{\frac{1}{\sigma^2_{\lambda,k}} + \frac{n_{jk}}{\sigma^2_{\lambda,jk}}}$$

- Estimate parameters via (Gibbs) sampling

Dimensionality reduction for visualization

Step 1. Calculate all pairs of GDs

$\rightarrow$ proximity matrix $[D_{ij}]$

Step 2. Plot points arbitrarily in 2D Euclidean space

Step 3. Calculate Euclidean proximity matrix $[E_{ij}]$

Step 4. Minimize $\sum_{i,j}(D_{ij} - E_{ij})^2$

} Multi-dimensional scaling

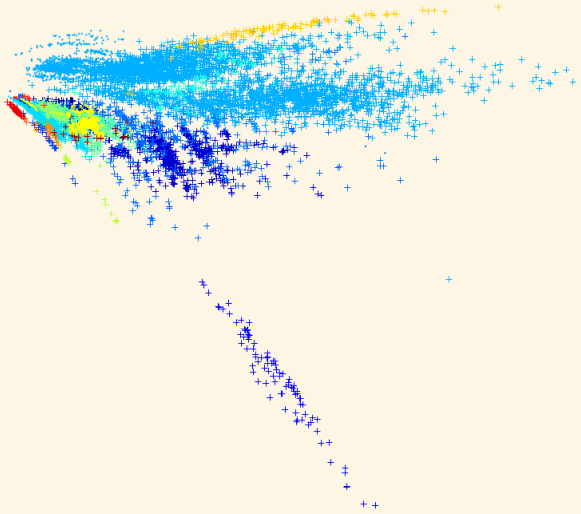
Step 5. Plot final configuration

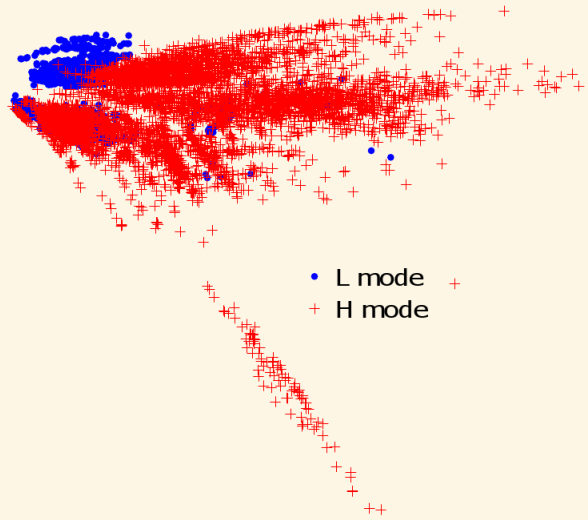
Confinement visualization

Tokamaks

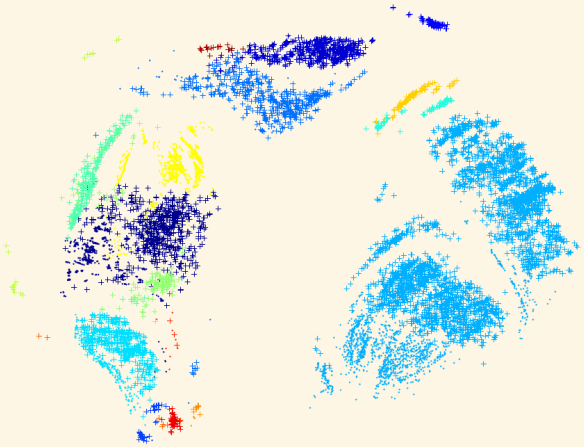
Confinement regime

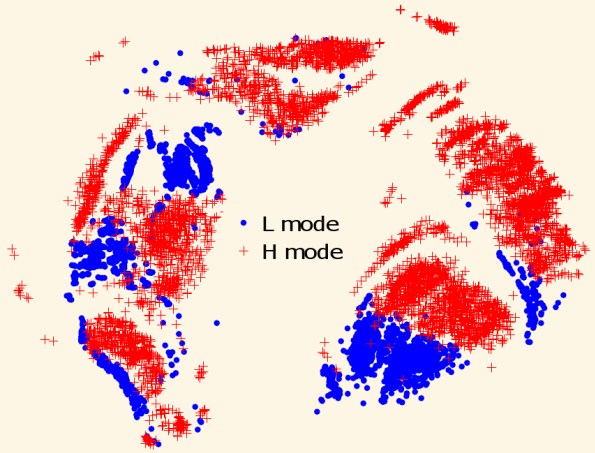
Euclidean
no errors





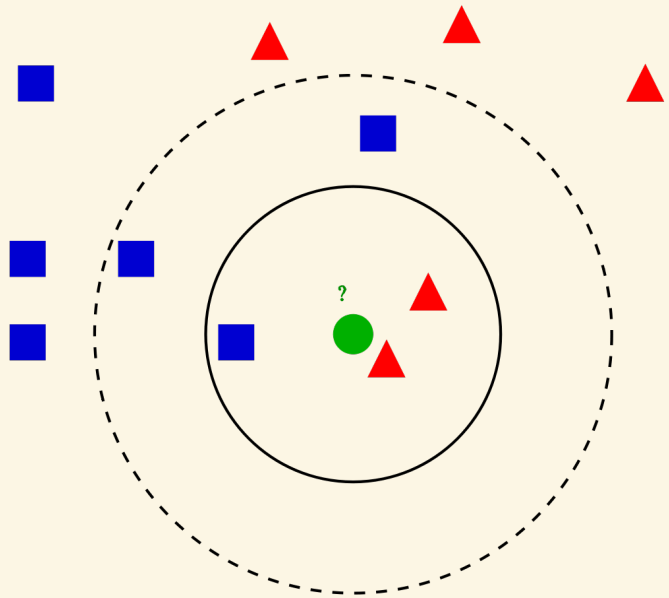
GD
with errors





k-nearest neighbor classification

- Confinement mode identification
- Training: 1%, testing: 99%
- $k = 1$: nearest neighbor
- Correct classification rates (%)



Mode	Euclidean w/o errors	Euclidean with errors	GD single-level	GD multilevel
L	89.7	91.2	91.9	93.2
H	89.1	90.5	93.3	94.3

Energy confinement scaling

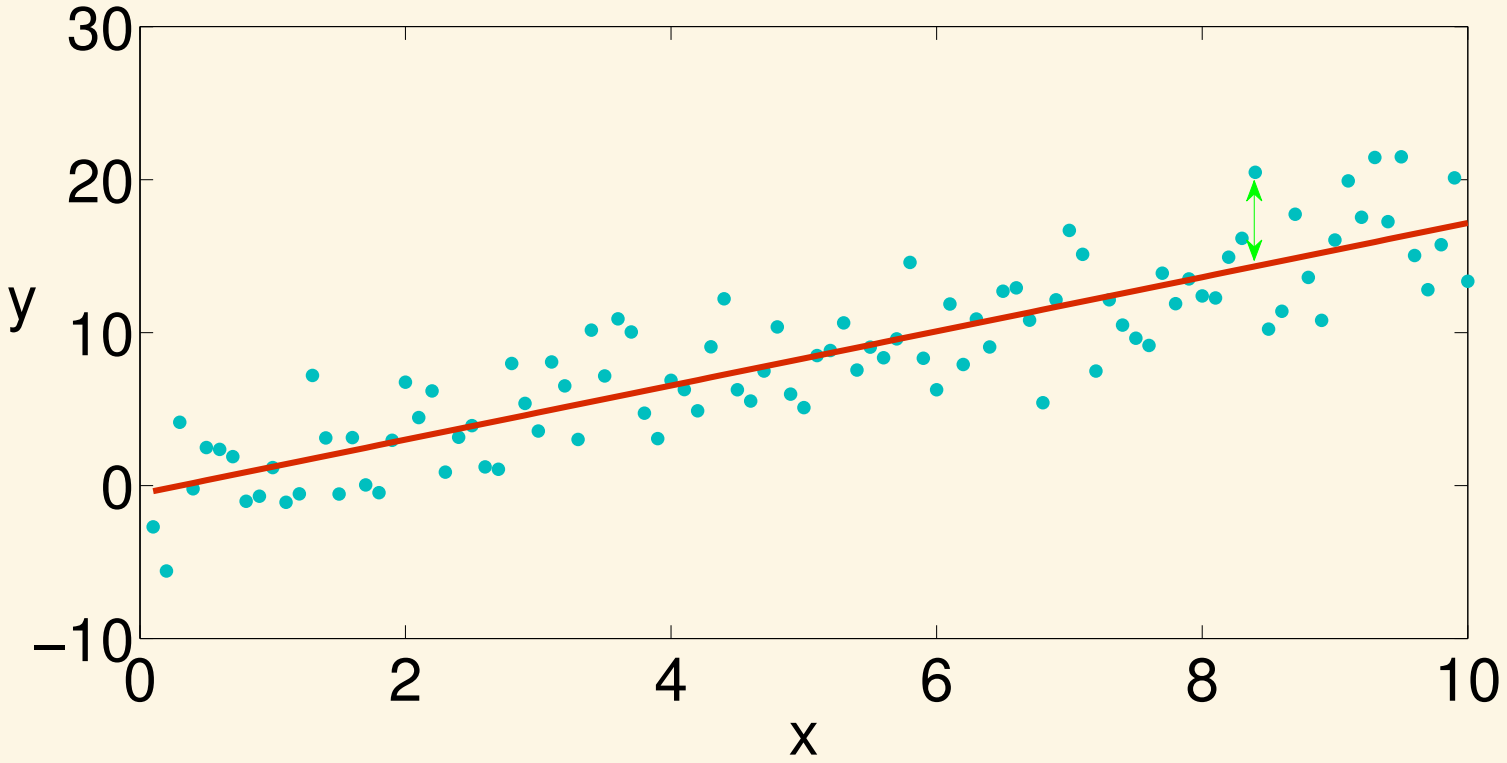
- Energy confinement time τ_{E}:

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

$$\begin{aligned} \Longrightarrow \ln(\tau_{\mathrm{E}}) = \ln(\beta_0) + \beta_1 \ln(I_{\mathrm{p}}) + \beta_2 \ln(B_{\mathrm{t}}) + \beta_3 \ln(\bar{n}_{\mathrm{e}}) + \beta_4 \ln(P_{\mathrm{loss}}) \\ + \beta_5 \ln(R) + \beta_6 \ln(\epsilon) + \beta_7 \ln(M_{\mathrm{eff}}) + \beta_8 \ln(\kappa) \end{aligned}$$

- Question: are all assumptions fulfilled?
  See e.g. D.C. Mc Donald *et al.*, PPCF **48**, pp. A439–A447, 2006
- Several methods:
  - Simple linear regression: ordinary least squares (OLS)
  - Errors-in-variables (EIV): total least squares
  - Geodesic regression (GR)
  - ...

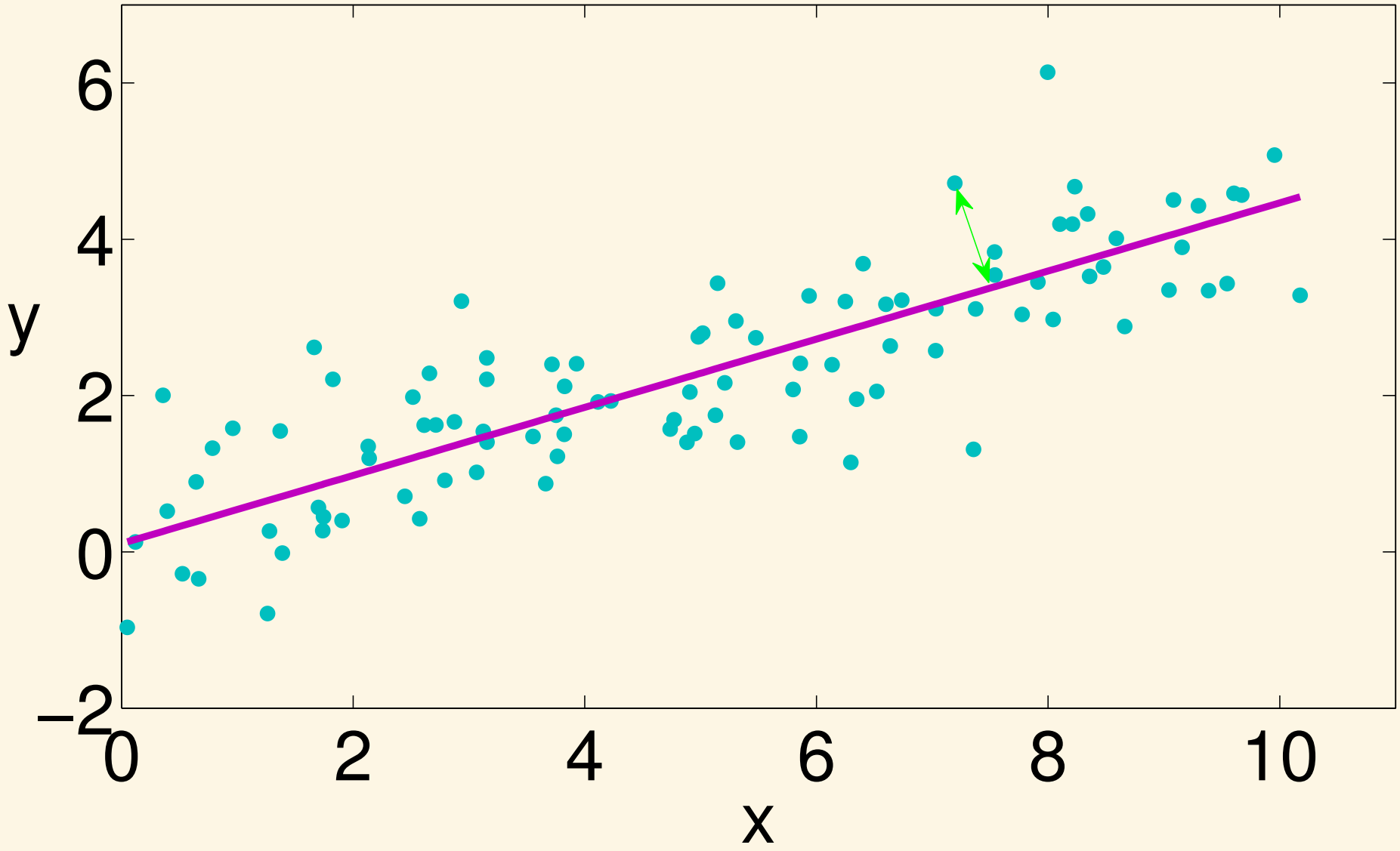
Ordinary least squares



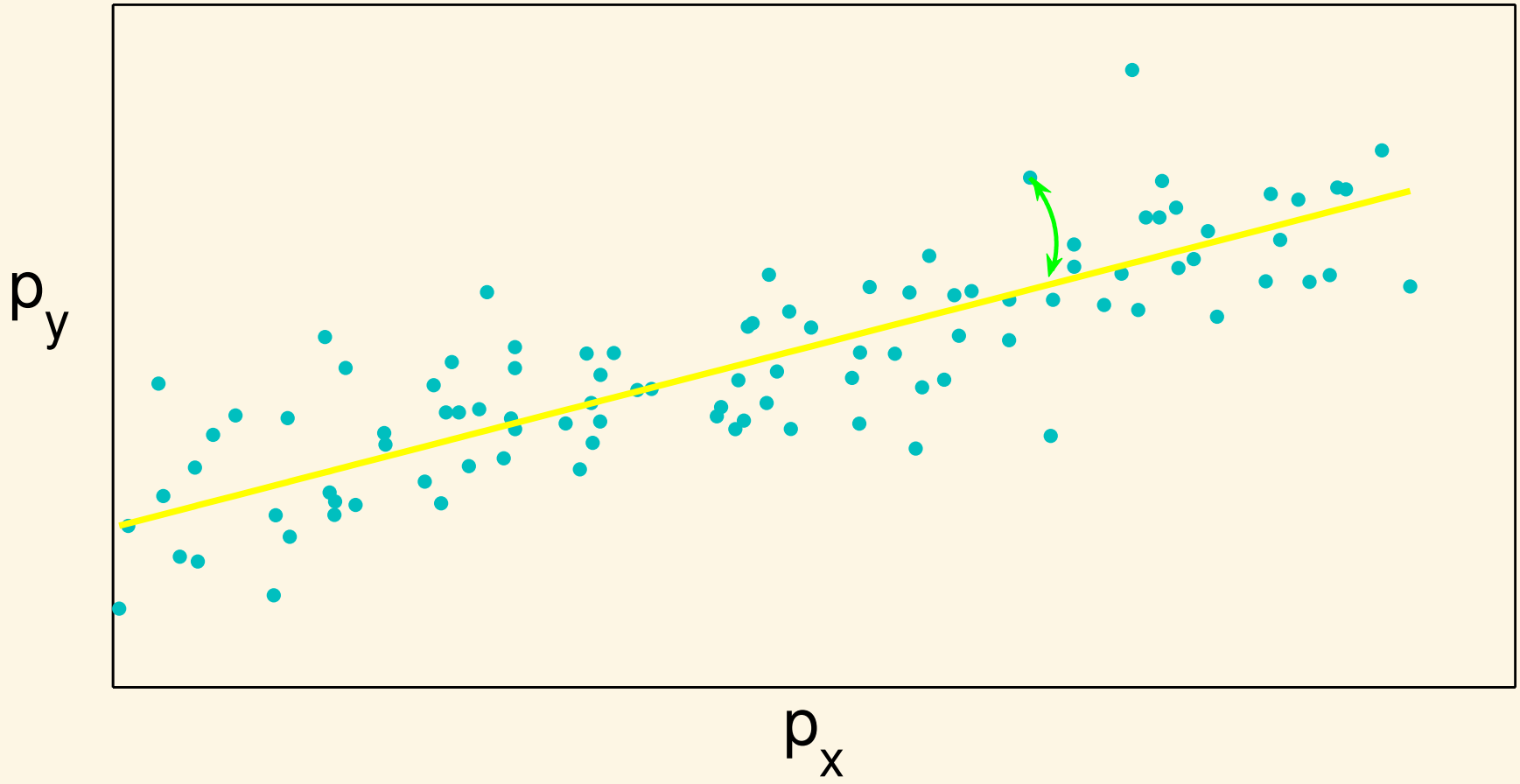
- Minimize $\mathrm{SS}_{\mathrm{err}} = \sum_i (y_i - \hat{y}_i)^2$
- Coefficient of determination:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{err}}}{\mathrm{SS}_{\mathrm{tot}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Errors-in-variables



Geodesic regression (GR)

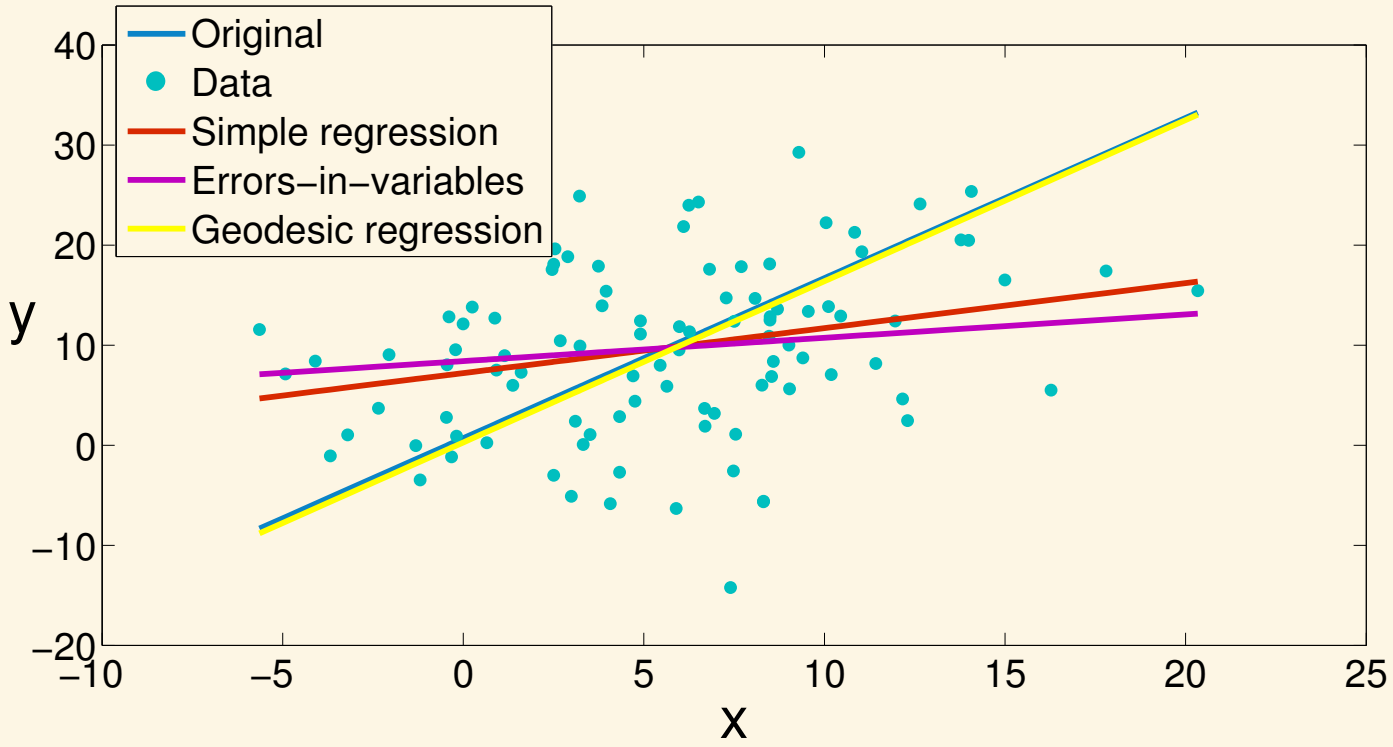


- Minimize sum of squared GD
- Use geodesic centroid

Synthetic data

$$\begin{cases} y^* = 0.7 + 1.6x^* \\ x = x^* + \eta, \quad \eta \sim \mathcal{N}(0, 4) \\ y = y^* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 8) \end{cases}$$

	Original	OLS	EIV	GR
β_0	0.70	7.21	8.41	0.28
β_1	1.60	0.45	0.23	1.61



Data and model

- DB3 standard data set
- Multiplicative Gaussian noise, both dependent and independent
- Logarithmic transformation $\Longrightarrow$ additive non-Gaussian noise
- Approximate with Gaussian distributions

GR empirical and model distribution

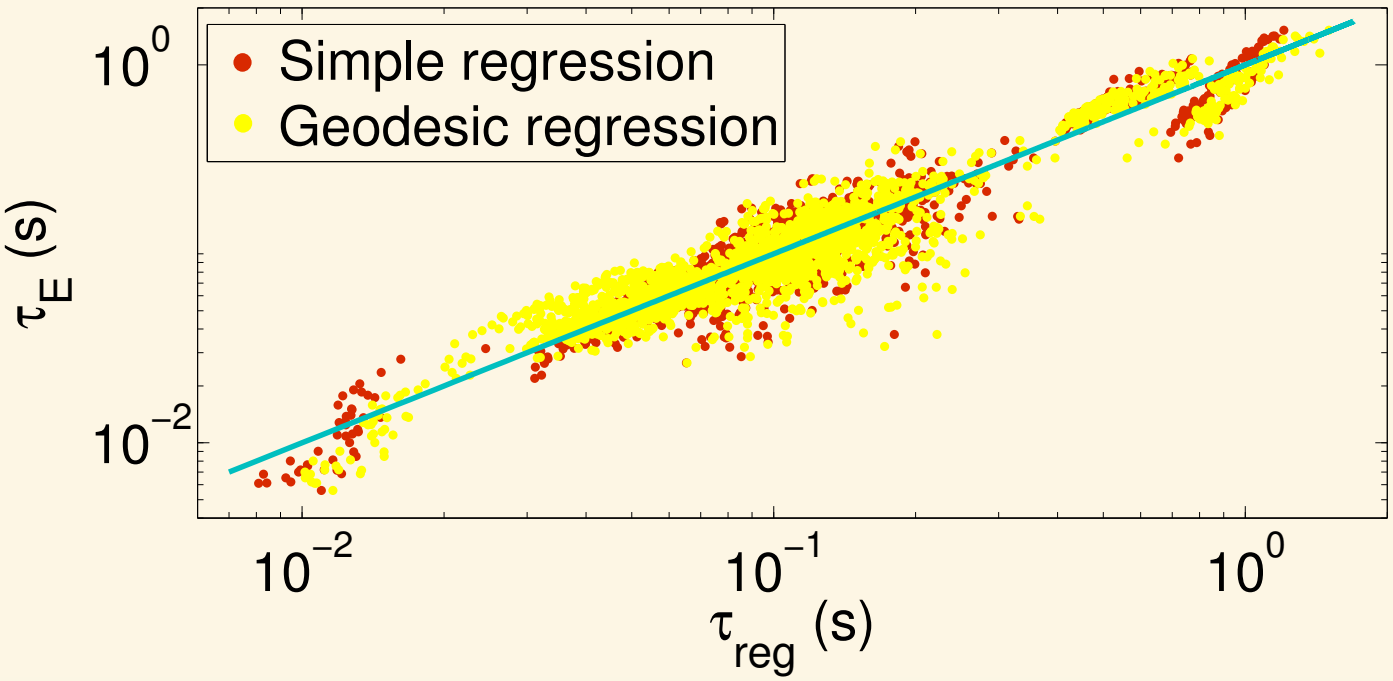
Empirical:

$$\ln(\tau_{\mathrm{E}}) \sim \mathcal{N}(\mu_{\ln \tau_{\mathrm{E}}}, \sigma_{\ln \tau_{\mathrm{E}}})$$

Model:

$$\begin{aligned}
\ln(\tau_{\mathrm{E}}^*) &= \ln(\beta_0) + \beta_1 \ln(I_{\mathrm{p}}^*) + \beta_2 \ln(B_{\mathrm{t}}^*) + \beta_3 \ln(\bar{n}_{\mathrm{e}}^*) + \beta_4 \ln(P_{\mathrm{loss}}^*) \\
&\quad + \beta_5 \ln(R^*) + \beta_6 \ln(\epsilon^*) + \beta_7 \ln(M_{\mathrm{eff}}^*) + \beta_8 \ln(\kappa^*) \\
\mu_{\ln \tau_{\mathrm{E}}^*} &= \ln(\beta_0) + \beta_1 \mu_{\ln I_{\mathrm{p}}^*} + \beta_2 \mu_{\ln B_{\mathrm{t}}^*} + \beta_3 \mu_{\ln n_{\mathrm{e}}^*} + \beta_4 \mu_{\ln P_{\mathrm{loss}}^*} + \beta_5 \mu_{\ln R^*} \\
&\quad + \beta_6 \mu_{\ln \epsilon^*} + \beta_7 \mu_{\ln M_{\mathrm{eff}}^*} + \beta_8 \mu_{\ln \kappa^*} \\
\sigma^2_{\ln \tau_{\mathrm{E}}^*} &= \beta_1^2 \sigma^2_{\ln I_{\mathrm{p}}^*} + \beta_2^2 \sigma^2_{\ln B_{\mathrm{t}}^*} + \beta_3^2 \sigma^2_{\ln n_{\mathrm{e}}^*} + \beta_4^2 \sigma^2_{\ln P_{\mathrm{loss}}^*} + \beta_5^2 \sigma^2_{\ln R^*} \\
&\quad + \beta_6^2 \sigma^2_{\ln \epsilon^*} + \beta_7^2 \sigma^2_{\ln M_{\mathrm{eff}}^*} + \beta_8^2 \sigma^2_{\ln \kappa^*} \\
\mu_{\ln I_{\mathrm{p}}^*} &\approx \ln(\mu_{I_{\mathrm{p}}^*}), \quad \sigma_{\ln I_{\mathrm{p}}^*} \approx \frac{\sigma_{I_{\mathrm{p}}^*}}{\mu_{I_{\mathrm{p}}^*}} \\
&\vdots \\
\mu_{\ln \kappa^*} &\approx \ln(\mu_{\kappa^*}), \quad \sigma_{\ln \kappa^*} \approx \frac{\sigma_{\kappa^*}}{\mu_{\kappa^*}}
\end{aligned}$$

Regression results



R^2 values:

OLS	EIV	GR single	GR multi	OLS in GR single	EIV in GR single
0.94	0.97	0.71	0.78	0.44	0.52

- GR yields full probability distributions
- Precise error estimates are not required, but may improve estimates

Conclusion

- Huge potential for pattern recognition in fusion
  - Physics studies
  - Plasma control
- Probability distributions are maximally informative
- Full probability structure actively determines patterns
- Geodesic distance is natural
- Even approximate probabilities are useful
- Very flexible to other probability models

Probability is fundamental

Probability distributions contain useful information for pattern recognition. Any useful information can be incorporated, including trustable theoretical models!