Geodesic Least Squares Regression on the Gaussian Manifold: Baryonic Tully-Fisher Scaling in Disk Galaxies

Geert Verdoolaege
Department of Applied Physics, Ghent University, Ghent, Belgium
Laboratory for Plasma Physics, Royal Military Academy (LPP-ERM/KMS), Brussels, Belgium

GSI2017, November 7, 2017
Overview

1. Regression analysis and minimum distance estimation
2. Geodesic least squares regression
3. Application in astronomy: Tully-Fisher Scaling
4. Conclusion
Overview

1. Regression analysis and minimum distance estimation
2. Geodesic least squares regression
3. Application in astronomy: Tully-Fisher Scaling
4. Conclusion
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):

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

\[
\mu_i = f_i(x_i, \theta) \quad \text{e.g.} \quad \beta_0 + \beta_1 x_i
\]

- 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
Regression analysis and minimum distance estimation

Geodesic least squares regression

Application in astronomy: Tully-Fisher Scaling

Conclusion
Difference/distance between points
Example: fluid turbulence
Modeled and observed distribution
Mahalanobis distance

$$\sum_i \left( \frac{y_i - \mu_i}{\sigma} \right)^2$$
Rao geodesic distance

- Based on Fisher information
- Observation/prediction (structureless number) $\rightarrow$ distribution (structured object)
- More flexibility, more information
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)
Geodesics on the Gaussian manifold
Geodesic least squares

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

Modeled distribution

Rao GD

Observed distribution

- 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
  \[ \implies \text{geodesic least squares (GLS) regression} \]
- If \( \sigma_{\text{mod}} = \sigma_{\text{obs}} \) \( \rightarrow \) Mahalanobis distance

G. Verduolaeghe et al., Entropy 17, 4602, 2015
Regression analysis and minimum distance estimation

Geodesic least squares regression

Application in astronomy: Tully-Fisher Scaling

Conclusion
Baryonic Tully-Fisher Relation (BTFR)

- Simple, tight relation for disk galaxies:
  \[ M_b = \beta_0 V_f^{\beta_1} \]
  \[ \begin{align*}
  M_b &= \text{total (stellar + gaseous) baryonic mass (}M_\odot) \\
  V_f &= \text{rotational velocity (km s}^{-1})
  \end{align*} \]

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


- 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 estimates

100 bootstrap samples

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loglinear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>340 ± 200</td>
<td>3.58 ± 0.15</td>
</tr>
<tr>
<td>Bayes</td>
<td>150 ± 200</td>
<td>3.74 ± 0.19</td>
</tr>
<tr>
<td>KLD</td>
<td>74 ± 87</td>
<td>3.98 ± 0.21</td>
</tr>
<tr>
<td>GLS</td>
<td>140 ± 83</td>
<td>3.81 ± 0.14</td>
</tr>
<tr>
<td>Nonlinear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>(1.0 ± 2.3)\times10^3</td>
<td>4.94 ± 1.40</td>
</tr>
<tr>
<td>Bayes</td>
<td>88 ± 140</td>
<td>3.81 ± 0.20</td>
</tr>
<tr>
<td>KLD</td>
<td>120 ± 100</td>
<td>3.91 ± 0.76</td>
</tr>
<tr>
<td>GLS</td>
<td>130 ± 130</td>
<td>3.79 ± 0.21</td>
</tr>
</tbody>
</table>
Parameter distributions

![Density plots for different methods](image)
GLS uncertainty estimates

$r_{M_b} \approx 38\%, \ r_{\text{obs}} \approx 63\%$
Interpretation on pseudosphere
Overview

1. Regression analysis and minimum distance estimation
2. Geodesic least squares regression
3. Application in astronomy: Tully-Fisher Scaling

4. Conclusion
Conclusions

- Geodesic least squares regression: *flexible* and *robust*
- Unified solution to various issues in regression analysis
- Works for *linear and nonlinear* relations and *any distribution* model
- Probability distributions more informative for regression
- Geometrical intuition
- *Easy* to use, *fast* optimization