View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Ghent University Academic Bibliography

Geodesic Least Squares Regression on the Gaussian Manifold with an Application in Astrophysics

Geert Verdoolaege[1,2]

[1] Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium
[2] Laboratory for Plasma Physics – Royal Military Academy (LPP–ERM/KMS), B-1000 Brussels, Belgium
geert.verdoolaege@ugent.be

Abstract. We present a new regression method called *geodesic least squares* (GLS), which is particularly robust against data and model uncertainty. It is based on minimization of the Rao geodesic distance on a probabilistic manifold. We apply GLS to Tully-Fisher scaling of the total baryonic mass vs. the rotation velocity in disk galaxies and we show the excellent robustness properties of GLS for estimating the coefficients and the tightness of the scaling.

Keywords: robust regression, geodesic least squares, Rao geodesic distance, Tully-Fisher scaling

1 Introduction

Many natural phenomena can be described by means of scaling laws, often in the form of a power law, e.g. in astrophysics, fluid and plasma dynamics, biology, geology, climatology and finance. However, in many application fields relatively simple or outdated statistical techniques are frequently used to estimate power laws. In the vast majority of cases, ordinary least squares (OLS) is applied to estimate the exponents (coefficients) of the power law on a logarithmic scale, despite its often poor performance in all but the simplest regression problems. Indeed, in more realistic settings, particularly when the goal is extrapolation of the scaling law, robustness is at least as important a quality compared to goodness-of-fit. This can become an issue in the presence of model uncertainty, heterogeneous data, atypical measurements (outliers) and skewed likelihoods [1].

Astrophysical data are often relatively complex from the statistical perspective and it has long been recognized that various assumptions of ordinary least squares regression are not valid in many applications in the field. Accordingly, several techniques from the domains of frequentist statistics and Bayesian probability theory have been applied to address the shortcomings of OLS. However, presently most techniques are designed to address one or a few shortcomings of OLS, but not all. In addition, judicious application of these techniques may require considerable expertise from the practitioner in statistics or probability theory, which can be an issue in various physics-centered application fields.

Presently, in many application domains there is a need for a robust general-purpose regression technique for estimating scaling laws.

For these reasons we have developed a new, robust regression method that is simple to implement, called *geodesic least squares regression* (GLS). It is based on minimization of the Rao geodesic distance between, on the one hand, the probability distribution of the response variable predicted by the regression model, and, on the other hand, a more data-driven distribution model of the response variable. GLS has recently been tested and applied in the field of magnetic confinement fusion [2, 3], showing its enhanced robustness over various traditional methods.

In this contribution, we apply GLS regression to estimate a key scaling law in astrophysics: the baryonic Tully-Fisher relation. This is a remarkably tight relation between the total baryonic mass of disk galaxies and their rotational velocity, of great practical and theoretical significance in astrophysics and cosmology.

2 Geodesic least squares regression

2.1 Principles of GLS

We here provide a brief overview of the GLS regression method. A more detailed description can be found in [1]. Implicitly, GLS performs regression on a probabilistic manifold characterized by the Fisher information. However, it is not directly based on a manifold regression technique like geodesic regression [4], where the relation between a manifold-valued response variable and a scalar predictor variable is modeled as a geodesic curve on the manifold. Rather, the idea behind GLS is to consider two different proposals for the distribution of a real-valued response variable y, conditional on the real-valued predictor variables, all of which can be affected by uncertainty. On the one hand, there is the distribution that one would expect if all assumptions were correct regarding the deterministic component of the regression model (regression function) and the stochastic component. We call this the *modeled distribution*. On the other hand, we try to capture the conditional distribution of y by relying as little as possible on the model assumptions, and much more on the actual measurements of y. For this we will use the term *observed distribution*. In this sense, GLS is similar to minimum distance estimation (MDE), where the Hellinger distance is a popular similarity measure [5], but there are several differences. First and foremost, GLS calculates the geodesic distance between each *individual* pair of modeled and observed distributions of the response variable. This often corresponds to an individual measurement point, together with an estimate of its error bar, provided by the experimentalist. The error bar estimate may have been obtained from previous experiments, or from a time series obtained at fixed (or stationary) values of the predictor variables. As such, each single data point is replaced by a probability density function describing the distribution of the response variable under fixed measured values of the predictor variables. In contrast, MDE

usually considers a distance between a kernel density estimate of the distribution of residuals on the one hand, and the parametric model on the other hand, but based on the entire data sample. Secondly, we explicitly model all parameters of the modeled distribution, similar to the idea behind the link function in the generalized linear model. In the present work this will be accomplished by explicitly modeling both the mean and standard deviation of the Gaussian modeled distribution. A final difference is that we use the Rao geodesic distance as a similarity measure.

2.2 The GLS algorithm

We start from a parametric multiple regression model between m predictor variables ξ_j $(j = 1, \ldots, m)$ and a single response variable η, all assumed to be infinitely precise. For n realizations of these variables, the regression model can be written as follows:

$$\eta_i = f(\xi_{i1}, \ldots, \xi_{im}, \beta_1, \ldots, \beta_p) \equiv f(\{\xi_{ij}\}, \{\beta_k\}), \qquad \forall i = 1, \ldots, n. \tag{1}$$

Here, f is the regression model function, in general nonlinear and characterized by p parameters β_k $(k = 1, \ldots, p)$. In regression analysis within the astronomy community, it is customary to add a noise variable to the idealized relation (1). This so-called *intrinsic scatter* serves to model the intrinsic uncertainty on the theoretical relation, i.e. uncertainty not related to the measurement process. We take another route for capturing model uncertainty, however.

In any realistic situation, we have no access to the quantities η_i and ξ_{ij}. Instead, a series of noisy measurements x_{ij}, resp. y_i is acquired for the predictor and response variables:

$$\begin{aligned} y_i &= \eta_i + \epsilon_{y,i}, & \epsilon_{y,i} &\sim \mathcal{N}\left(0, \sigma_{y,i}^2\right), \\ x_{ij} &= \xi_{ij} + \epsilon_{x,ij}, & \epsilon_{x,ij} &\sim \mathcal{N}\left(0, \sigma_{x,ij}^2\right). \end{aligned}$$

We have assumed independent Gaussian noise, but this can be generalized to any distribution. Also, in general the standard deviations are different for each point. For instance, in many real-world situations, such as the one discussed in this paper, there is a constant relative error on the measurements, so the standard deviation can be modeled to be proportional to the measurement itself.

Under this model, the distribution of the variable y, conditional on measured values x_{ij} of the m predictor variables (fixed i), as well as the parameters β_k, is given by

$$p_{\mathrm{mod}}(y|\{x_{ij}\}, \{\beta_k\}) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{mod},i}} \exp\left\{-\frac{1}{2}\frac{\left[y_i - f\left(\{x_{ij}\}, \{\beta_k\}\right)\right]^2}{\sigma_{\mathrm{mod},i}^2}\right\}. \tag{2}$$

This is the modeled distribution, where we suppose that estimates of the standard deviations $\sigma_{x,ij}$ and $\sigma_{y,i}$ are available. The uncertainty on the predictor

variables propagates through the function f and adds to the conditional uncertainty on the response variable, determined by $\sigma_{\mathrm{mod},i}$. We use standard Gaussian error propagation theory as a practical solution for this purpose. For example, referring to $f\big(\{x_{ij}\},\{\beta_k\}\big)$ as the modeled mean $\mu_{\mathrm{mod},i}$, for a linear model we have (with relabeled β_k):

$$\begin{aligned}\mu_{\mathrm{mod},i} &\equiv \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im},\\ \sigma^2_{\mathrm{mod},i} &\equiv \sigma^2_{y,i} + \beta_1^2\sigma^2_{x,i1} + \ldots + \beta_m^2\sigma^2_{x,im}.\end{aligned}$$

Relying on the maximum likelihood method, one would proceed to estimate the parameters β_k by maximizing (2), or, under the assumption of symmetry of the likelihood distribution and homoscedasticity, by minimizing the sum of squared differences (Euclidean distances) between each measured y_i and predicted $\mu_{\mathrm{mod},i}$. However, this assumes that the model is exact, specifically that $\sigma_{\mathrm{mod},i}$ is the only source of data variability. In order to take into account additional uncertainty sources, in particular model uncertainty, we therefore also consider the observed distribution of y, relying on as few assumptions as possible regarding the regression model. Specifically, we replace each data point y_i by a distribution $p_{\mathrm{obs}}(y|y_i)$. In the context of the GLM, this is known as the *saturated model*. In the present application, we choose again the normal distribution, but centered on each data point: $\mathcal{N}\big(y_i,\sigma^2_{\mathrm{obs},i}\big)$, where $\sigma_{\mathrm{obs},i}$ is to be estimated from the data. The extra parameters $\sigma_{\mathrm{obs},i}$ give the method added flexibility, since they are not *a priori* required to equal $\sigma_{\mathrm{mod},i}$. As a result, GLS is less sensitive to incorrect model assumptions. Choosing a Gaussian form for both the modeled and observed distribution offers a computational advantage, since the corresponding expression for the GD has a closed form [6]. Also, in principle, $\sigma_{\mathrm{obs},i}$ can be different for each point, although in practice it is clear that we will need to introduce some sort of regularization to render the model identifiable. In this paper we either assume $\sigma_{\mathrm{obs},i}$ a constant s_{obs}, or proportional to the response variable, $\sigma_{\mathrm{obs},i} = r_{\mathrm{obs}}|\bar{y}_i|$. The parameters s_{obs} or r_{obs} have to be estimated from the data. More complicated (parametrized) relations between $\sigma_{\mathrm{obs},i}$ and the response variable or other data would be possible too, but one should be careful not to put too many restrictions on p_{obs}, thereby defeating its purpose.

GLS now proceeds by minimizing the total GD between, on the one hand, the joint observed distribution of the n realizations of the variable y and, on the other hand, the joint modeled distribution. Owing to the independence assumption in this example, we can write this in terms of products of the corresponding marginal distributions (including all dependencies and with γ_{obs} either s_{obs} or r_{obs})):

$$\left\{\hat{\beta}_k, \hat{\gamma}_{\rm obs}\right\}$$
$$= \underset{\beta_k, \gamma_{\rm obs}\in\mathbb{R}}{\rm argmin}\ {\rm GD}^2\left[\prod_{i=1}^{n} p_{\rm obs}\left(y|y_i,\gamma_{\rm obs}\right), \prod_{i=1}^{n} p_{\rm mod}\left(y|\{x_{ij}\},\{\beta_k\},\sigma_{y_i},\{\sigma_{x_{ij}}\}\right)\right]$$
$$= \underset{\beta_k, \gamma_{\rm obs}\in\mathbb{R}}{\rm argmin} \sum_{n=1}^{n} {\rm GD}^2\Big[p_{\rm obs}\left(y|y_i,\gamma_{\rm obs}\right), p_{\rm mod}\left(y|\{x_{ij}\},\{\beta_k\},\sigma_{y_i},\{\sigma_{x_{ij}}\}\right)\Big]. \quad (3)$$

Note that the parameters β_k occur both in the mean and the variance of the modeled distribution. The last equality in (3) entails a considerable simplification, owing to the property that the squared GD between products of distributions can be written as the sum of squared GDs between the corresponding factors [6]. Hence, the optimization procedure involves, on the level of each measurement, matching not only y_i with $\mu_{{\rm mod},i}$, but also $\sigma_{{\rm obs},i}$ with $\sigma_{{\rm mod},i}$, in a way dictated by the geometry of the likelihood distribution. As will be shown in the experiments, the result is that GLS is relatively insensitive to uncertainties in both the stochastic and deterministic components of the regression model. The same quality renders the method also robust against outliers.

In the experiments below, we employed a classic active-set algorithm to carry out the optimization [7]. Furthermore, presently the GLS method does not directly offer confidence (or credible) intervals on the estimated quantities. Future work will address this issue in more detail, but for now error estimates were derived by a bootstrap procedure. The bootstrapping involved creating, from the measured data set, 100 artificial data sets of the same size, by resampling with replacement. The regression analysis was then carried out on each of the data sets and the mean and standard deviation, over all data sets, of each estimated regression parameter and of the predicted quantities were used as estimates of the parameter or prediction value and its error bar, respectively. This scheme typically results in rather conservative error bars, which could possibly be narrowed down using more sophisticated methods.

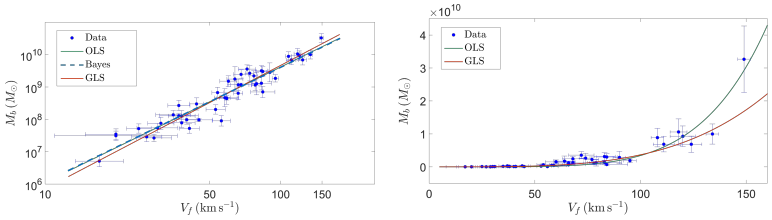
Incidentally, forcing $\sigma_{{\rm obs},i} \equiv \sigma_{{\rm mod},i}$ in (3), $\forall\, i$, would take us back to standard maximum likelihood estimation, since the Rao GD between two Gaussians p_1 and p_2 with means y_i, resp. $f\big(\{x_{ij}\},\{\beta_k\}\big)$, but with identical standard deviations σ_i (fixed along the geodesic path), is precisely the Mahalanobis distance [8]:

$${\rm GD}(p_1,p_2) = \frac{\left|y_i - f\big(\{x_{ij}\},\{\beta_k\}\big)\right|}{\sigma_i}.$$

3 Application of GLS to Tully-Fisher scaling

3.1 The baryonic Tully-Fisher relation

The baryonic Tully-Fisher relation (BTFR) between the total (stellar + gaseous) baryonic mass M_b of disk galaxies and their rotational velocity V_f is of fundamental importance in astrophysics and cosmology [9]. It is a remarkably simple



(a) (b)

Fig. 1: Baryonic mass M_b vs. rotation velocity V_f for 47 gas-rich galaxies and the fitted BTFR using various methods. (a) On the logarithmic scale and (b) on the original scale.

and tight empirical relation of the form

$$M_b = \beta_0 V_f^{\beta_1}. \tag{4}$$

The BTFR serves as a tool for determining cosmic distances, provides constraints on galaxy formation and evolution models, and serves as a test for the Lambda cold dark matter paradigm (ΛCDM) in cosmology. In this scaling problem, we use data from 47 gas-rich galaxies, as detailed in [9]. The data also contain estimates of the observational errors, which we treat here as a single standard deviation. Figure 2 shows a scatter plot of $\sigma_{\mathrm{mod},i}$, which is almost entirely determined by σ_{M_b}, vs. M_b for the 47 galaxies in the database. This suggests a measurement error on the response variable proportional to M_b, about 38%, i.e. a constant error bar on the logarithmic scale.

Table 1: Regression estimates for the BTFR parameters using loglinear and nonlinear OLS and GLS, and a robust Bayesian method in the loglinear case.

Loglinear	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	310	3.56
Bayes	160	3.72
GLS	110	3.81
Nonlinear	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	0.063	5.37
Bayes	91	3.80
GLS	79	3.83

3.2 Regression analysis

Owing to the power law character of most scaling laws, they are often estimated by linear regression on a logarithmic scale. However, it is known that this may lead to unreliable estimates, as the logarithm (heavily) distorts the distribution of the data [1]. This is in particular the case if the estimation is done using simple OLS or if there are outliers in the data. In contrast, we will show that GLS regression produces consistent results on both the logarithmic and original scales, demonstrating its robustness.

In view of the proportional error on M_b, the observed standard deviation in GLS is modeled here as $\sigma_{\mathrm{obs},i} = r_{\mathrm{obs}} M_b$, with r_{obs} an unknown scale factor to be estimated from the data using the optimization routine.

We compare the results of GLS regression with OLS and a standard Bayesian method. For the latter we choose the likelihood given in (2), but we use an unknown standard deviation σ_{u} instead of σ_{mod}, again assumed to be proportional to M_b through a scale factor r_{u}, to be estimated from the data. In addition, we use uninformative prior distributions for the regression parameters and a Jeffreys prior for r_{u}. This factor is then marginalized out of the posterior, which comes down to fitting a t-distribution to the data (shifted to sample mean zero). The t-distribution has heavier tails than a Gaussian, hence accommodating outliers.

The scalings obtained using OLS and GLS, are shown in Figure 1a for the case of linear regression on the logarithmic scale, and in Figure 1b for power-law regression on the original scale. In the loglinear case the result from the Bayesian analysis is also added, although it is very similar to the result of OLS. The coefficient estimates are given in Table 1. It is clear that GLS yields estimates that are much more consistent compared to OLS. In particular, whereas the data point corresponding to the largest V_f and M_b does not have the characteristics of an outlier on the logarithmic scale, it may be considered as such on the original scale. The nonlinear OLS estimate for the exponent β_1 is heavily influenced by this point, causing the discrepancy with the estimate on the logarithmic scale.

Next, 100 bootstrap samples were created from the data, yielding average parameter estimates and 95% confidence intervals on the basis of the OLS and GLS results, shown in Table 2. Again, the enhanced robustness of GLS compared to OLS stands out.

Finally, Figure 2 shows the plot of $r_{\mathrm{obs}} M_b$, with the scale factor r_{obs} (observed relative error) amounting to 63%. This is considerably larger than the value of 38% predicted by the model, possibly indicating that the scatter on the scaling law is not due to measurement error alone. This will be an important area of further investigation, as it may provide evidence for the ΛCDM vs. MOND cosmological models.

4 Conclusion

We have introduced geodesic least squares, a versatile and robust regression method based on regression between probability distributions. Part of the strength

Table 2: Average regression estimates and 95% confidence intervals for the BTFR using loglinear and nonlinear OLS and GLS, obtained from 100 bootstrap samples.

Loglinear	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	360 ± 220	3.57 ± 0.15
Bayes	220 ± 220	3.72 ± 0.19
GLS	140 ± 82	3.80 ± 0.16
Nonlinear	$\hat{\beta}_0$	$\hat{\beta}_1$
$(3.6 \pm 6.2) \times 10^3$	4.56 ± 1.19	
Bayes	130 ± 160	3.80 ± 0.21
KLD	560 ± 470	3.78 ± 0.19
GLS	390 ± 280	3.85 ± 0.18

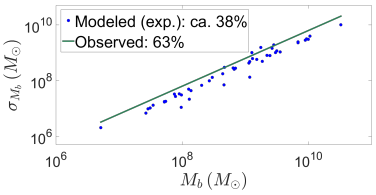


Fig. 2: Plot of σ_{M_b} ($\approx \sigma_{\text{mod}}$) and $r_{\text{obs}} M_b$ ($= \sigma_{\text{obs}}$) vs. M_b, as estimated by GLS.

of the method is its simplicity, allowing straightforward application by users in various application fields, without the need for parameter tuning. We have applied GLS to baryonic Tully-Fisher scaling, thereby demonstrating the robustness of the method and providing an alternative means for testing cosmological models based on the estimated intrinsic scatter.

References

1. G. Verdoolaege. A new robust regression method based on minimization of geodesic distances on a probabilistic manifold: Application to power laws. *Entropy*, 17(7):4602–4626, 2015.
2. G. Verdoolaege and J.-M. Noterdaeme. Robust scaling in fusion science: case study for the L-H power threshold. *Nucl. Fusion*, 55(11):113019 (19 pp.), 2015.
3. G. Verdoolaege, A. Shabbir, and G. Hornung. Robust analysis of trends in noisy tokamak confinement data using geodesic least squares regression. *Rev. Sci. Instrum.*, 87(11):11D422 (3 pp.), 2016.
4. P.T. Fletcher. Geodesic regression and the theory of least squares on riemannian manifolds. *Int. J. Comput. Vis.*, 105(2):171–185, 2013.
5. R.J. Pak. Minimum Hellinger distance estimation in simple regression models; distribution and efficiency. *Stat. Probab. Lett.*, 26(3):263–269, 1996.
6. J. Burbea and C.R. Rao. Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multivariate Anal.*, 12(4):575–596, 1982.
7. P.E. Gill, W. Murray, and M.H. Wright. *Numerical linear algebra and optimization, Vol. 1.* Addison Wesley, Boston, MA, 1991.
8. C.R. Rao. Differential metrics in probability spaces. In *Differential Geometry in Statistical Inference*. Institute of Mathematical Statistics, Hayward, CA, 1987.
9. S.S. McGaugh. The baryonic Tully-Fisher relation of gas-rich galaxies as a test of ΛCDM and MOND. *Astron. J.*, 143(2):40 (15 pp.), 2012.