Bayesian Model Fitting to Data With Both Intrinsic and Extrinsic Uncertainties in Two Dimensions

Nicholas C. Konz†, Adam S. Trotter†, and Daniel E. Reichart†

†Department of Physics and Astronomy, the University of North Carolina at Chapel Hill

Abstract and Introduction

Rigorously fitting a two dimensional statistical model to data that has intrinsic uncertainties (error bars) in both the independent variable and the dependent variable is a daunting task, especially if the data also has extrinsic uncertainty that cannot be fully accounted for by the error bars. While there are a few prescriptions that tackle this problem, they each have their downfalls. Here, we present a new statistic (described as the Trotter, Reichart, Konz statistic, or TRK) derived from basic principles that is advantageous towards model-fitting in this "worst-case data" scenario, especially when compared to other methods. The TRK statistic is fully invertible, but not scalable. However, an iterative algorithm is presented to obtain the optimal scale for the given data set which will give the best fit, using a nested downhill simplex method. To fit the model to the data, a Markov Chain Monte Carlo (MCMC) method is used to generate not just the best-fit values of the model parameters, but the full posterior probability distributions of them as well, including extrinsic scatter. The statistic is applicable to practically any data-driven field, for any custom model function, and can even be generalized to an arbitrary number of independent variables within the model.

A Quick Overview of Bayesian Statistics

The central principle of the Bayesian approach to statistics and probability (as opposed to the frequentist approach) is Bayes' theorem

$$p(H|DI) \propto \mathcal{L}(D|H)p(H), \tag{1}$$

where (excluding a normalization constant) $p(H|DI)$ is known as the *posterior* probability density function, $\mathcal{L}(D|HI)$ is called the *likelihood* function, and $p(H|I)$ is known as the *prior* probability density function. Given a set of observed data D and a set of parameters describing a hypothetical model H, the likelihood function $\mathcal{L}$ describes the conditional probability of obtaining the observed D given some model parameters H and the prior describes how any pre-existing information about the model (before data collection) affects or constrains the values of the model parameters.

Fitting a Model To "Worst Case" Data

In the most general case, a set of N two-dimensional datapoints $\{x_n, y_n\}$ (with $n = 1, 2 \dots N$) can have both intrinsic uncertainties in both directions (i.e. error bars) for each datapoint $\{\sigma_{x,n}, \sigma_{y,n}\}$, and extrinsic scatter, or "slop" for the data set $\{\sigma_x, \sigma_y\}$ (that must be parameterized and fit to as part of the model). In order to fit a model to such a dataset, we need to quantify the goodness of fit of some model to a dataset, including uncertainties. Such a model is not just a curve, but a probability distribution: technically a relative probability distribution $g(x, y)$ along the model curve $y_c(x; \vartheta_m)$, convolved with a two-dimensional Gaussian probability distribution $G(x, y)$ that is parameterized with the slop parameters σ_x and σ_y, where ϑ_m is the set of M model parameters. For each datapoint we also construct a *convolved error ellipse* with axes defined by the parameters $\Sigma_{x,n} = \sqrt{\sigma_{x,n}^2 + \sigma_x^2}$ and $\Sigma_{y,n} = \sqrt{\sigma_{y,n}^2 + \sigma_y^2}$ (see Fig.1, left).

In order to begin constructing the likelihood function, we must first determine the (joint) posterior distribution for a single datapoint, $p_n(\vartheta_m, \sigma_x, \sigma_y | x_n, y_n, \sigma_{x,n}, \sigma_{y,n})$. Each p_n is found by convolving the intrinsic and observed model distributions for the n^{th} datapoint, the latter of which requires convolving $g(x, y)$ and $G(x, y)$ with the model distribution. The model curve itself can be represented as a one-dimensional Dirac delta function in some arbitrary rotated coordinate system with axes (u_n, v_n), which can be different for each datapoint. In total, the expression for p_n includes four complicated, usually non-analytic integrals (see [5] for the explicit details). As such, three different, reasonable approximations must be made to make this computationally feasible, one of them being that the model curve y_c is approximately linear on the scale of the convolved error ellipse of the datapoint. We center this linear approximation of y_c at the point $(x_{t,n}, y_{t,n})$ where the ellipse is tangent to the model curve, with slope $m_{t,n}$ so that $y_c(x) \approx y_{t,n} + m_{t,n}(x - x_{t,n})$ (see Fig. 1 left). Following this, the tangent point can be found implicitly with

$$(y_c(x) - y_n)\frac{\mathrm{d}y_c(x; \vartheta_m)}{\mathrm{d}x}\Sigma_{x,n}^2 + (x - x_n)\Sigma_{y,n}^2 = 0, \tag{2}$$

(in practice with a two-point Newton-Raphson root finder, see [4]). Taking into account the approximations, we finally arrive at a computationally reasonable expression for p_n,

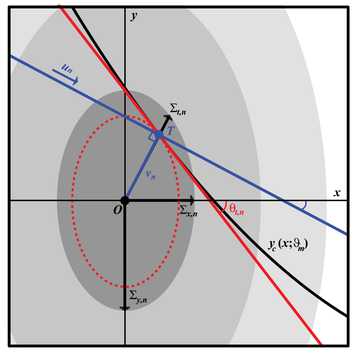
$$p_n(\vartheta_m, \sigma_x, \sigma_y | x_n, y_n, \sigma_{x,n}, \sigma_{y,n}) \approx f(x_n, y_n)g(x_n, y_n)\frac{\mathrm{d}u_n}{\mathrm{d}x}G_n\left(y_n\right), \tag{3}$$

where $G_n(y_n)$ is a Gaussian distribution of y_n with mean $y_{t,n} + m_{t,n}(x_n - x_{t,n})$ and deviation $\sqrt{m_{t,n}^2\Sigma_{x,n}^2 + \Sigma_{y,n}^2}$, and $f(x_n, y_n)$ represents the efficiency at which the data sampes the model distribution (see [5]).

The TRK Statistic

Recall from the previous section that the definition of p_n (3) utilizes some chosen rotated coordinate system (u_n, v_n) to define the one-dimensional model curve. We define the TRK statistic such that for some n^{th} datapoint, the u_n axis is perpendicular to the line segment connecting the datapoint (x_n, y_n) (see Fig. 1 left). This choice ends up being analogous to a one-dimension χ^2 statistic along the direction of v_n [5]. The TRK likelihood function—the central part of this statistic—which is the product of all N posterior distribution functions p_n, can then be defined as

$$\mathcal{L}^{\text{TRK}} \propto \prod_{n=1}^{N} \sqrt{\frac{m_{t,n}^2\Sigma_{x,n}^2 + \Sigma_{y,n}^2}{m_{t,n}^2\Sigma_{x,n}^4 + \Sigma_{y,n}^4}} \exp\left\{-\frac{1}{2}\frac{\left[y_n - y_{t,n} - m_{t,n}(x_n - x_{t,n})\right]^2}{m_{t,n}^2\Sigma_{x,n}^2 + \Sigma_{y,n}^2}\right\}. \tag{4}$$



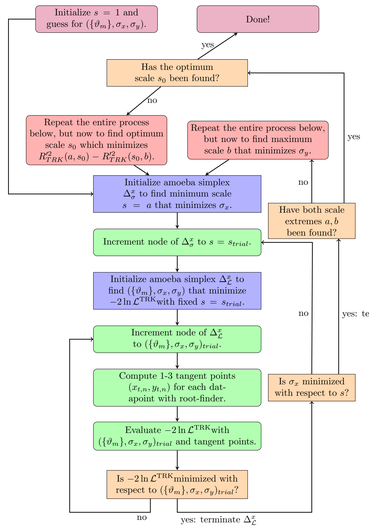


Fig. 1: Left: Illustration of the geometry of the TRK statistic. Data point centroid is at (x_n, y_n) (point O), with convolved widths $(\Sigma_{x,n}, \Sigma_{y,n})$. Model curve $y_c(x; \vartheta_m)$ is tangent to the convolved error ellipse at tangent point $(x_{t,n}, y_{t,n})$ (point T). Red line is the linear approximation of the model curve, with slope $m_{t,n} = \tan\theta_{t,n}$. Blue line indicates the rotated coordinate axis u_n for the TRK statistic, perpendicular to the v_n axis. Right: Flowchart of the scale optimization algorithm.

Invertibility and Scalability of the TRK Statistic

A desirable property for any 2D statistic is *invertibility*, i.e. that if fitting y vs. x to a model yields a curve $y_c(x)$, then fitting x vs y yields the inverse curve $x_c(y) = y_c^{-1}(x)$. In the Bayesian formalism, a statistic is invertible if fitting y vs x and x vs y gives the same likelihood [5]. Now, consider the well-known Pearson Correlation Coefficient R^2. One usage of R^2 is to quantify the invertibility of a statistic [5]. In the case of a linear model with slope m, for example, you have that $R^2 \equiv m_{yx}m_{xy}$, where m_{yx} is the slope of the model fit with y vs. x, and m_{xy} is the same but for x vs y. As we show in [4], it turns out that *the TRK statistic is completely invertible*; it follows that for fully invertible statistics like TRK, $R^2 = 1$.

Another important behavior to have in a statistic is *scalability*, which means that multiplying all y data values by some positive s should give the same fit for any s, i.e. the likelihood should be unchanged. As we show in [4], the $\mathcal{L}^{\text{TRK}}$ is not invariant to a change of s. However, we have created an algorithm that determines the *optimum scale* s_0 that yields the best fit, effectively mitigating the non-scalability of $\mathcal{L}^{\text{TRK}}$. First, we need some sort of new correlation coefficient R^2_{TRK} to compare TRK fits on different s for a given model and dataset, such that the closer the fits are to one another, the higher this R^2_{TRK} will be. R^2_{TRK} is a *measure of the variance of the TRK statistic's predictions under a change of s*; in the linear case, this means that $R^2_{\text{TRK}} = 1$ if the two fits have identical slope, and $R^2_{\text{TRK}} = 0$ if the two fits have orthogonal slope. So, in the linear case, R^2_{TRK} is a function of the difference in said slopes; this generalizes nicely to the nonlinear case using the aforementioned tangent slopes $m_{t,n}$ [5], which we use to define

$$R^2_{\text{TRK, linear}} \equiv \tan^2\left(\frac{\pi}{4} - \frac{|\theta_c - \theta_d|}{2}\right) \qquad R^2_{\text{TRK, nonlinear}} \equiv \frac{1}{N}\sum_{n=1}^{N}\tan^2\left(\frac{\pi}{4} - \frac{\left|\theta_{t,n;c} - \theta_{t,n;d}\right|}{2}\right) \tag{5}$$

where given two scales c and d, $\theta_c = \tan^{-1} m_c$ (the linear fit slope angle from scale c), $\theta_{t,n;c} = \tan^{-1} m_{t,n;c}$ (the nonlinear fit tangent point slope angles from scale c), and θ_d & $\theta_{t\ n;d}$ are defined similarly (see Fig. 1 left).

Scale Optimization and Model Distribution Fitting Algorithms

R^2_{TRK} has been defined as a way to compare fits done at different scales; now, we present an algorithm to determine the optimum scale s_0 for fitting. Consider the limiting behavior of TRK fits as $s \to 0$ and $s \to \infty$, which correspond to "extreme" scales a and b (respectively) that indicate the minimum and maximum scales such that the TRK fit remains physically meaningful. This is because at $s = a$, the TRK fit will force $\sigma_x \to 0$, and at $s = b$, $\sigma_y \to 0$, as $\sigma_x, \sigma_y < 0$ would be unphysical [5]. From this, we find that s_0 will then satisfy

$$R^2_{\text{TRK}}(a, s_0) = R^2_{\text{TRK}}(s_0, b) \equiv R^2_{\text{TRK}} \qquad \Rightarrow \qquad R^2_{\text{TRK}}(a, s_0) - R^2_{\text{TRK}}(s_0, b) = 0, \tag{6}$$

where the arguments of R^2_{TRK} are the two scales being evaluated (c and d in equation (5)).

In practice, we obtain a by stepping σ_x vs s until $\sigma_x = 0$ is reached using the well known "amoeba" simplex minimization algorithm of Nelder and Mead, with b obtaining analogously by minimizing σ_y. Then, we minimize (6), right, with respect to s_0 to obtain s_0 with the same method. To obtain σ_x or σ_y at some s, we use another nested simplex to minimize $-2\ln\mathcal{L}^{\text{TRK}}$ (analogous to minimizing the "regular" χ^2 value) with respect to $(\vartheta_m, \sigma_x, \sigma_y)$ (see Fig. 1, right).

Given some optimum fitting scale s_0, in order to perform the fit and generate probability distributions of the model parameters, we use a Markov Chain Monte Carlo method to properly explore the parameter space of the model. Specifically, we sample the joint posterior distribution of $(\vartheta_m, \sigma_x, \sigma_y)$ using the likelihood $\mathcal{L}^{\text{TRK}}$ (4) and any priors (see (1)), and use this generated histogram to obtain the marginalized probability distribution of each model parameter.

Preliminary Results

Shown in Fig. 2 are preliminary fits done on relationships between parameters describing empirical fits [1][2] to the observed spectral extinction by dust of stars in the Milky Way and Magellanic Clouds.

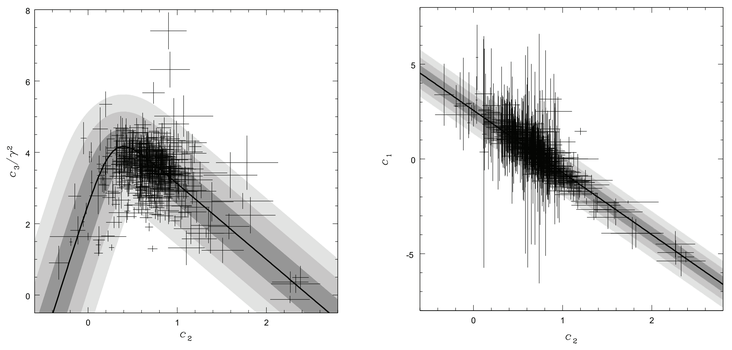


Fig. 2: Left: observed BH vs. c_2 data [3][6] fitted to a smoothly broken linear model distribution and Right: observed c_1 vs. c_2 data [3][6] fitted to a linear model distribution [5][1][2]. Shaded regions indicate the 1-, 2- and 3-σ slop envelopes of the model distribution.

Acknowledgements

This endeavor would've been impossible without the helpful guidance and wealth of knowledge of my advisor, Prof. Daniel Reichart, and our colleague Adam Trotter. This project was first introduced as Adam's PhD thesis (advised by Dr. Reichart), and I have worked to implement their statistic into a generalized computer algorithm.

References

[1] Jason A Cardelli, Geoffrey C Clayton, and John S Mathis. In: *The Astrophysical Journal* 345 (1989), pp. 245–256.

[2] Edward L Fitzpatrick and Derck Massa. In: *The Astrophysical Journal* 328 (1988), pp. 734–746.

[3] Karl D Gordon et al. In: *The Astrophysical Journal* 594.1 (2003), p. 279.

[4] A. S. Trotter, D. E. Reichart, and N. C. Konz. In: (2019). Publication in preparation.

[5] Adam S. Trotter. PhD thesis. The University of North Carolina at Chapel Hill, 2011.

[6] Lynne A Valencic, Geoffrey C Clayton, and Karl D Gordon. In: *The Astrophysical Journal* 616.2 (2004), p. 912.