

Enlightening the Relationship Between Distribution and Regression Fitting

Hang Yu, Qian Yin^(✉), and Ping Guo

Image Processing and Pattern Recognition Laboratory, Beijing Normal University,
Beijing 100875, China
yuhang@mail.bnu.edu.cn, yinqian@bnu.edu.cn, pguo@ieee.org

Abstract. Statistical distribution fitting and regression fitting are both classic methods to model data. There are slight connections and differences between them, as a result they outperform each other in different cases. A analysis model for processing natural data, say astronomical pulsar data in this paper, is proposed to improve data fitting method performance. Then the insight behind the comprehensive fitting model is given and discussed.

Keywords: Statistical distribution · Regression · Model selection · Goodness of fit

1 Introduction

Statistics is a subject aiming to model the world by rather collecting and analyzing data than to inference and prove precisely. Lots of natural data have been fitted into statistical distributions with significant performances. And with these statistical distributions scientists can utilize their off-the-shelf distribution properties. While unfortunately the diversity of our world is far more complex than the level those statistical distributions can totally model. Alternatively, polynomial regression is also an approach to automatically find empirical laws from data [1]. It is proposed to fit various patterns of data, and has already shown its powerful fitting ability. Researchers who use polynomial regression are conveniently not required to preliminarily estimate data distribution [2]. Regrettably, by implementing polynomial regression we eventually get a polynomial expression with less prior properties than traditional statistical distributions have.

Given that features of natural scientific data are relatively fixed, and they will still be studied in the future, it is significant to determine the best fitting method for certain scientific data feature. And we creatively name it fitting label. As a solution to the dilemma, we propose a discriminative fitting model to help scientists automatically fitting scientific data into the optimal expression. Once scientists get the fitting label of a certain feature of natural scientific data, those later scientists are able to directly update the fitting expression parameters on new data's arrival. We operated experiments on an astronomical data set, ANTF

Pulsar Catalogue [3]. Besides, given the reliability of P value is still controversial [4], and mean squared error and KL divergence are trustworthy measures for both fields of statistics and machine learning, we complementally took them in consideration.

The structure of this paper is shown as follows: the comprehensive fitting model is proposed in Sect. 2; and with it its corresponding experiments are showed and discussed in Sects. 3 and 4; the section of conclusion comes as the end.

2 Methodology

2.1 Comprehensive Fitting Model

As is shown in Algorithm 1, the selection model consists of mainly 3 steps. Step 1 is to determine whether the data is subject to a certain natural distribution or not through significance testing. Step 2, fitting data into significant distributions and operating regression, then calculating their respective similarities. Step 3 is to classify the data’s fitting label with distance-based classifier according to the similarities. Consequently, scientists are able to learn as more properties as possible of the data.

Algorithm 1. Comprehensive model

Input: Scientific Data \mathcal{D} ;
Candidate Statistical Distribution Families $\mathbf{\Pi} = \{\pi_1, \dots, \pi_i, \dots, \pi_m\}$;
Selected Significance Testing Methods $\mathbf{\Gamma} = \{\Gamma_1, \dots, \Gamma_j, \dots, \Gamma_n\}$;
Selected Similarity Measures $\mathbf{\Sigma} = \{\sigma_1, \dots, \sigma_j, \dots, \sigma_n\}$;
Polynomial Regression Order e ;
Cross-Validation Times t and Fold k .

Output: Fitting Labels π^* .

- 1: $\hat{\mathbf{\Pi}} \leftarrow \text{Algorithm2}(\mathcal{D}, \mathbf{\Pi}, \mathbf{\Gamma}, \Omega)$
- 2: $\hat{\mathbf{\Sigma}} \leftarrow \text{Algorithm3}(\mathcal{D}, \mathbf{\Sigma}, \hat{\mathbf{\Pi}})$
- 3: Fitting Labels $\pi^* \leftarrow kNN(\hat{\mathbf{\Sigma}})$

return π^*

2.2 Statistical Significance Testing

As is shown in Algorithm 2, firstly we need human experts, astronomers for example, to give us candidate statistical distribution families for each feature. These distribution families with empirical parameters are then be tested one by one for statistical significance under selected statistical significance measures, usually P value. Concretely with specified natural science feature data, we perform statistical significance test to screen out distribution families that have statistically large significance levels for the feature. Candidate statistical families and selected significance measures are given into the testing process, If the output P value reaches the threshold, a significance level of 0.05 by convention, the distribution will be recorded as a significant distribution.

Algorithm 2. Statistical significance testing

Input: Scientific Data \mathcal{D} ;Candidate Statistical Distribution Families $\mathbf{\Pi} = \{\pi_1, \dots, \pi_m\}$;Selected Significance Testing Methods $\mathbf{\Gamma} = \{\Gamma_1, \dots, \Gamma_n\}$;Testing Methods Weights $\mathbf{\Omega} = \{\omega_1, \dots, \omega_n\}$;Significance Level τ .**Output:** Significant Statistical Distribution Families $\hat{\mathbf{\Pi}} = \{\hat{\pi}_1, \dots, \hat{\pi}_l\}$.

```

1: for each  $\pi_i \in \mathbf{\Pi}$  do
2:   for each  $\Gamma_j \in \mathbf{\Gamma}$  do
3:      $\hat{\psi}_j^i \leftarrow \Gamma_j(\mathcal{D}, \pi_i)$ ;
4:      $\hat{\Psi}^i \leftarrow \hat{\Psi}^i \cup \hat{\psi}_j^i$ ;
5:   end for
6:    $\psi^i \leftarrow \mathbf{\Omega}^\top \times \hat{\Psi}^i$ ;
7:   if  $\psi^i \geq \tau$  then
8:      $\hat{\mathbf{\Pi}} \leftarrow \hat{\mathbf{\Pi}} \cup \pi_i$ ;
9:   end if
10: end for
    return  $\hat{\mathbf{\Pi}}$ 

```

2.3 Similarity Evaluation

For this step, we use similarity measures to detect how well the expression fits with the data, which is shown in Algorithm 3. Significant statistical distributions

Algorithm 3. Similarity evaluation

Input: Pulsar Feature Data \mathcal{D} ;Significant Statistical Distribution Families $\hat{\mathbf{\Pi}} = \{\hat{\pi}_1, \dots, \hat{\pi}_l\}$;Selected Similarity Measures $\mathbf{\Sigma} = \{\sigma_1, \dots, \sigma_n\}$;Similarity Measure Weights $\mathbf{\Theta} = \{\theta_1, \dots, \theta_n\}$;Polynomial Regression Highest Order e ;Cross-Validation Times t and Fold k .**Output:** Similarity/Goodness of Fit $\tilde{\mathbf{\Sigma}} = \{\tilde{\sigma}^1, \dots, \tilde{\sigma}^{l+1}\}$.

```

1: for each  $\hat{\pi}_i \in \hat{\mathbf{\Pi}}$  do
2:   for each  $\sigma_j \in \mathbf{\Sigma}$  do
3:      $\hat{\sigma}_j^i \leftarrow CV(t, k, \mathcal{D}, \hat{\pi}_i, \sigma_j)$ ;
4:      $\hat{\sigma}^i \leftarrow \hat{\sigma}^i \cup \hat{\sigma}_j^i$ ;
5:   end for
6:    $\tilde{\sigma}^i \leftarrow \mathbf{\Theta}^\top \times \hat{\sigma}^i$ ;
7:    $\tilde{\mathbf{\Sigma}} \leftarrow \tilde{\mathbf{\Sigma}} \cup \tilde{\sigma}^i$ ;
8: end for
9: for each  $\sigma_j \in \mathbf{\Sigma}$  do
10:   $\hat{\sigma}_j^0 \leftarrow CV(t, k, \mathcal{D}, Polyfit(e), \sigma_j)$ ;
11:   $\hat{\sigma}^0 \leftarrow \hat{\sigma}^0 \cup \hat{\sigma}_j^0$ ;
12: end for
13:  $\tilde{\sigma}^0 \leftarrow \mathbf{\Theta}^\top \times \hat{\sigma}^0$ ;
14:  $\tilde{\mathbf{\Sigma}} \leftarrow \tilde{\mathbf{\Sigma}} \cup \tilde{\sigma}^0$ 
    return  $\tilde{\mathbf{\Sigma}}$ 

```

obtained from last step are introduced here, and with it selected similarity measures. We perform t times k -fold cross-validation test to acquire the average similarity between the distribution and scientific data.

2.4 Distance-Based Classification

It is the final outcome, the fitting label for the feature data, producing step. We can take those similarity measures as distances, and run a distance-based classifier to eventually select the expression with minimal distance as the optimal fitting label. Here we bring a classic classifier k -Nearest Neighbours (abbreviated as kNN) with its parameter k chosen 1.

3 Experiments

A logical starting point is to test a distribution’s normality for its popularity in natural world [5]. While for features whose scatter diagram is shown in Fig. 1, they seem obviously subject to lognormal distribution [6,7]. We took the logarithm of that data first, then tested the resulting data’s normality, which is shown in Fig. 2. As we know, logarithm of scale parameter and shape parameter of lognormal distribution are respectively expectation and standard deviation of normal distribution.

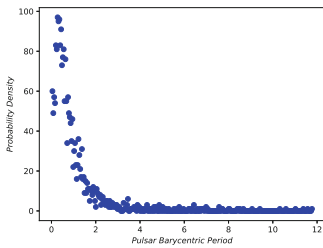


Fig. 1. Original plot

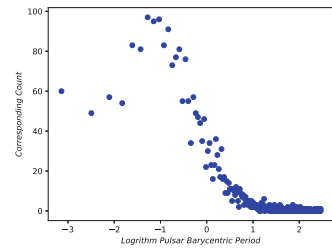


Fig. 2. Logarithm plot

Significance results are shown in Table 1. For those classic statistical testing methods, the values recorded are their P values. Results shows that lognormal has the highest significance among expert-selecting distributions.

Table 1. Significance testing of P0

P0	Normal	Lognormal	Power-law
KS test	4.21349250398968713e-61	0.487092039062517	0.0
Shaprio-Wilk test	0.0	0.155729278922081	-
Kurtosis test	9.7779914011373847e-154	0.10520924193801827	-
Skewness test	0.0	0.97680081829094689	-

Similarity evaluation results are shown in Table 2. The polynomial order is empirically chosen 2. Thus lognormal distribution has the lowest relative entropy and mean squared error. So it is the optimal fitting label for pulsar period feature. The temporary parameters of its corresponding normal distribution are mean -0.53455956344351729 and standard deviation 0.99686283436845236 .

Table 2. Similarity Evaluation of P0

P0	Lognormal	Polynomial
KL divergence	0.0188633190407	0.565469835801
MSE	2.49154711009	115.375894222

4 Discussions

As we know, the insight is a thought on inductive bias of Occam's razor and the principle of maximum entropy. Occam's razor has been the default heuristic technique in natural science for hundreds of years. It can be interpreted as stating that the one among competing hypotheses with the fewest assumptions should be selected [8,9]. The preference for simplicity also comes to machine learning's avoiding overfitting, a low-order fitting curve is almost always better than a high-order one. On the contrary, the second law of thermodynamics states that the total entropy of an isolated system can only increase over time. It is seen as an axiom of statistical thermodynamics. Here comes the principle of maximum entropy which states that the probability distribution that best represents the current state of knowledge is the one with largest entropy. And it also emphasizes a natural correspondence between statistical mechanics and information theory [10]. Numbers of statistical distributions widely used by natural scientists to model scientific data can be derived under the principle of maximum entropy. For any certain mean μ and variance σ^2 , the maximum entropy distribution $p(x|\mu, \sigma^2)$ is normal distribution $N(\mu, \sigma^2)$ [11].

5 Conclusions

For scientists we propose a discriminative fitting model through which diverse scientific data will be fitted into the optimal expression. And with our model as much as possible sample information can be used in future study. For data that is subject to a common statistical distribution, they then will be fitted into that distribution with parameters tuned by numerical optimization methods; comparatively for those who are not we model them by polynomial regression. Among main fields of natural science are still features whose fitting label remains unfixed, so it will be a great work to build a fitting label database with our model for scientists.

Acknowledgments. The research work in this paper was supported by the grants from National Natural Science Foundation of China (61472043, 61375045) and the Joint Research Fund in Astronomy (U1531242) under cooperative agreement between the NSFC and CAS, Beijing Natural Science Foundation (4142030). Prof. Qian Yin is the author to whom all the correspondence should be addressed.

References

1. Edwards, J.R., Parry, M.E.: On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Acad. Manag. J.* **36**(6), 1577–1613 (1993)
2. Theil, H.: A rank-invariant method of linear and polynomial regression analysis, 3; confidence regions for the parameters of polynomial regression equations. *Stichting Mathematisch Centrum. Statistische Afdeling (SP 5a/50/R)*, 1–16 (1950)
3. Manchester, R.N., Hobbs, G.B., Teoh, A., et al.: The Australia telescope national facility pulsar catalogue. *Astron. J.* **129**(4), 1993 (2005)
4. Nuzzo, R.: Statistical errors. *Nature* **506**(7487), 150 (2014)
5. Faucher-Giguere, C.A., Kaspi, V.M.: Birth and evolution of isolated radio pulsars. *Astrophys. J.* **643**(1), 332 (2006)
6. Lorimer, D.R., Faulkner, A.J., Lyne, A.G., et al.: The Parkes multibeam pulsar SurveyCVI. Discovery and timing of 142 pulsars and a Galactic population analysis. *Mon. Not. R. Astron. Soc.* **372**(2), 777–800 (2006)
7. Bates, S.D., Lorimer, D.R., Rane, A., Swiggum, J.: PsrPopPy: an open-source package for pulsar population simulations. *Monthly Not. Roy. Astrono. Soc.* **439**(3), 2893–2902 (2014). <https://academic.oup.com/mnras/article/439/3/2893/1108811/PsrPopPy-an-open-source-package-for-pulsar>
8. Gauch, H.G.: *Scientific Method in Practice*. Cambridge University Press, Cambridge (2003)
9. Hoffmann, R., Minkin, V.I., Carpenter, B.K.: Ockham’s razor and chemistry. *Bulletin de la Socit chimique de France* **133**(2), 117–130 (1996)
10. Jaynes, E.T.: Information theory and statistical mechanics. *Phys. Rev.* **106**(4), 620–630 (1957)
11. Jaynes, E.T.: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge (2003)