Submitted to the Annals of Applied Statistics

INSIGHTS INTO COORDINATE RELATING TO RADIAL VELOCITY MEASUREMENT OF NGC 7531

BY ZIYE LIN (20825543)

University of Waterloo, z294lin@uwaterloo.ca

This project aims to construct a suitable model to describe the association between the measured radial velocity of NGC 7531 and the coordinates of positions in the sky from which it is measured. We assume a Gaussian error with homoscedasticity on the measurements and constructed models using ordinary least squares, regularized least squares with ℓ^2 norm penalty, cubic polynomials, and cubic B-splines additive model in this project. After assessing their performances using the AIC, training error, and estimated mean prediction squared error via cross-validation, we select the cubic B-splines additive model as the final model to make inferences about the association between the measured radial velocity of NGC 7531 and its coordinates.

1. Introduction. Professor Ronald J. Buta (1987) from the University of Alabama performs a series of analyses to explore the photometric and kinematic properties of NGC 7531, "a nonbarred (SA) spiral possessing a very bright inner ring" (pg. 1). One of the analyses includes studying the radial velocity of NGC 7531 by which measurements were taken from different positions of the sky that NGC 7531 covers. It turns out that the measured values of the radial velocity appear to be inconsistent when measuring from different positions of NGC 7531. According to the COSMOS Encyclopedia of Astronomy by SUT (n.d.) (Swinburne University of Technology), "radial velocities can be determined by examining the redshift of spectral lines in a star or galaxy's spectrum. This allows astronomers to compute the distance to galaxies using the Hubble expansion law and also study the orbits of stars in binaries". This explains why Buta includes the study of the radial velocity of NGC 7531 in his paper.

This project aims to construct a suitable model to describe how the measured radial velocity of NGC 7531 changes as we measure from different positions. We make homoscedastic Gaussian assumptions on the errors of measurements and consider the linear model, the polynomial model, and the smoothing spline additive model, and use the Akaike information criterion and cross-validation to access their fits and generalization performances. Our goal is to get better insight into the variation of the measured radial velocity of NGC 7531 at different positions, which does not include a discussion of why such a phenomenon exists.

2. Data. According to Buta (1987), a two-dimensional reference frame is set up on the area of the sky covered by NGC 7531. The $x-$ axis represents the east-west coordinates of NGC 7531 with the east being negative and the west being positive The $y-$ axis represents the north-south coordinates of NGC 7531 with the south being negative and the north being positive. The measurements are taken using the "TAURUS Fabry-Perot interferometer (Taylor and Atherton 1980) on the AAT in 1982 November" which includes 323 observations in total (Buta, pg. 8). The unit for the radial velocity is kilometer per second. There is no unit for the coordinates.

Figures 1 and 2 provide visualizations of the data from two different angles. Both figures suggest a potential non-linearity association between the radial velocity and the coordinates of NGC 7531.

Keywords and phrases: Galaxy, NGC 7531, Radial Velocity.

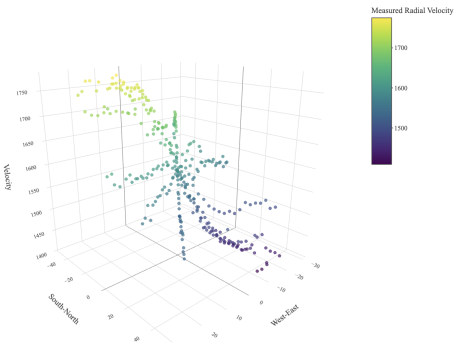

FIGURE 1. *Radial Velocity of NGC 7531 (1)*

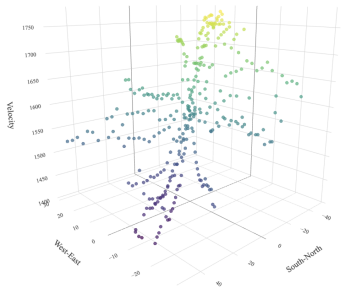

FIGURE 2. *Radial Velocity of NGC 7531 (2)*

3. Methods. The response is the measured radial velocity of NGC 7531 which we denote as y_i , $i = 1, \ldots, n$, where n denotes the total sample size of the data. The covariates are the east-west coordinates which we denote as x_{i1} , and the north-south coordinates which we denote as x_{i2} , $i = 1, \ldots, n$. We use $\boldsymbol{y} = \begin{pmatrix} y_1 \; y_2 \; \ldots \; y_n \end{pmatrix}^T$ to denote the vector of responses and $\boldsymbol{\beta} = \begin{pmatrix} \alpha \; \beta_1 \; \beta_2 \end{pmatrix}^T$ to denote the vector containing the intercept and the regression coefficients.

We assume an independent, homoscedastic Gaussian error on the response for all our models:

$$\epsilon_i \sim^{ind} \mathrm{N}(0, \sigma^2) \; , \; i = 1, \ldots, n \; .$$

We use the squared loss throughout this project:

$$Loss\left(y_i, x_{i1}, x_{i2}\right) = \left(y_i - \hat{f}(x_{i1}, x_{i2})\right)^2 \; , \; i = 1, \ldots, n \; ,$$

where $\hat{f}$ is the fitted model.

We consider the following regression techniques and evaluate their performances based on Akaike Information Criterion (AIC), training errors, and estimated mean prediction squared errors obtained via cross-validation:

3.1. *Ordinary Least Squares.* Assuming a simple linear association between the response and the covariates, this model is written as

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \; , \; i = 1, \ldots, n \tag{3.1}$$

This yields the ordinary least squares estimates for α, β_1, β_2:

$$\left[\hat{\alpha} \; \hat{\beta}_1 \; \hat{\beta}_2\right]^T = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \; ,$$

where $\boldsymbol{X}$ is the corresponding design matrix for the covariates.

3.2. *Regularized Least Squares - ℓ^2 Norm Penalty.* Assuming a simple linear association between the response and the covariates:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \; , \; i = 1, \ldots, n \; , \tag{3.2}$$

we add the ℓ^2 norm penalty to the loss function:

$$Loss(\boldsymbol{y}, \boldsymbol{X}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \hat{\lambda}\|\boldsymbol{\beta}\|^2 \; ,$$

where $\hat{\lambda} = \arg\min_\lambda \; \mathrm{AIC}(\hat{f}, \lambda)$ is chosen to minimize the AIC value of this model.

This yields the estimates for α, β_1, β_2:

$$\left[\hat{\alpha}\ \hat{\beta}_1\ \hat{\beta}_2\right]^T = \left(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}\,\mathbf{I}_{n+1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\,,$$

where $\boldsymbol{X}$ is the corresponding design matrix for the covariates.

3.3. *Cubic Polynomial.* We perform a cubic polynomial regression:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1}^3 + \beta_6 x_{i2}^3 + \epsilon_i\ ,\ i = 1, \ldots, n\,. \tag{3.3}$$

This yields the estimates for the regression coefficients:

$$\left[\hat{\alpha}\ \hat{\beta}_1\ \hat{\beta}_2\ \hat{\beta}_3\ \hat{\beta}_4\ \hat{\beta}_5\ \hat{\beta}_6\right]^T = \left(\boldsymbol{X}_c^T\boldsymbol{X}_c\right)^{-1}\boldsymbol{X}_c^T\boldsymbol{y}\,,$$

where $\boldsymbol{X}_c$ is the corresponding design matrix for the covariates.

3.4. *Additive Cubic B-Splines.* We consider an additive model using the cubic B-splines basis functions:

$$\begin{gathered} y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \epsilon_i\ ,\ i = 1, \ldots, n\,, \\ f_1(x_{i1}) = \sum_j B_{j1}(x_{i1})\beta_{j1}\ ,\ f_2(x_{i2}) = \sum_j B_{j2}(x_{i2})\beta_{j2}\,, \end{gathered} \tag{3.4}$$

where B_{j1}, B_{j2} are the cubic B-splines basis functions.

The intercept α estimated by $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^n y_i$ which is the sample mean of the responses. We also decide to use the point constraint $f_1(0) = 0\ ,\ f_2(0) = 0$ for interpretation purposes which will be discussed later in this paper. The position of knots for the B-splines basis functions is chosen by the $MGCV$ package in R.

We estimate the regression coefficients by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix} = \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^n [y_i - \hat{\alpha} - (f_1(x_{i1}) + f_2(x_{i2}))]^2) + \hat{\lambda}\left(\int \left(f_1^{''}(x)^2 + f_2^{''}(x)^2\right) dx\right)\,,$$

where $\hat{\lambda} = \arg\min_\lambda\ \text{AIC}(\hat{f}, \lambda)$ is chosen to minimize the AIC of this model, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the corresponding coefficient vectors for covariate x_1 and x_2.

3.5. *Model Selection Criteria.* Given a fitted model $\hat{f}$, we calculate the AIC of $\hat{f}$ by

$$\text{AIC}(\hat{f}) = -\frac{2}{n}\,\text{log-likelihood}(\boldsymbol{y} \mid \hat{f}) + \frac{2}{n}\text{Trace}(H_{\hat{f}})\,, \tag{3.5}$$

where $H_{\hat{f}} = \boldsymbol{X}_{\hat{f}}\left(\boldsymbol{X}_{\hat{f}}^T\boldsymbol{X}_{\hat{f}}\right)^{-1}\boldsymbol{X}_{\hat{f}}$ is the corresponding "Hat Matrix", $\boldsymbol{X}_{\hat{f}}$ is the corresponding design matrix of the covariates given the fitted model $\hat{f}$. A smaller AIC value indicates that the value of log-likelihood$(\boldsymbol{y} \mid \hat{f})$ is closer to 0, meaning that the fitted model is "more likely" to produce the given data and thus providing a better fit based on Frequentist-Inference framework.

Given a fitted model $\hat{f}$, we calculate the training error by

$$\text{Err}_{train}(\hat{f}) = \frac{1}{n}\sum_{i=1}^n Loss(y_i, \hat{f}(x_i)) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \hat{f}(x_i)\right)^2\,. \tag{3.6}$$

A smaller training error indicates that the fitted values are closer to the observed values in the given data set, meaning that the fitted model provides a good fit to the observed data.

We estimate the mean prediction squared error via both the K-fold cross-validation and the leave-one-out cross-validation. The estimate of the mean prediction squared error is computed by

$$\widehat{\text{MPSE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{-i}(x_i)\right)^2 , \tag{3.7}$$

where $\hat{f}^{-i}$ is the model constructed using data from all folds except the one containing y_i. We choose $K = 10$ so that each of the first 9 folds contains exactly 32 data points, and the last fold contains exactly 35 data points. This results in an approximately $90\%;$ training $- 10\%$ testing split of our data each time.

4. Results.

4.1. *Model Comparison and Selection.*

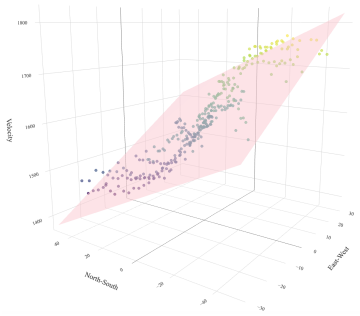

FIGURE 3. *Ordinary Least Squares*

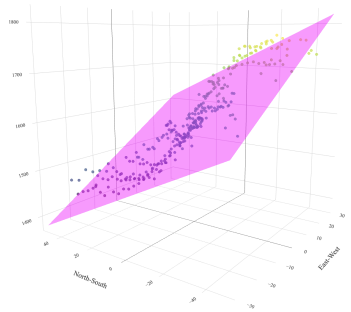

FIGURE 4. *Least Squares:* ℓ^2 *Norm Penalty*

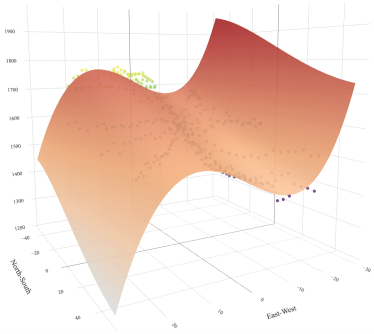

FIGURE 5. *Cubic Polynomial*

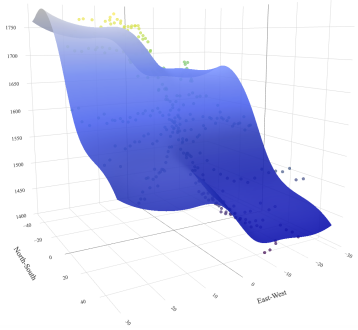

FIGURE 6. *Cubic B-Splines Additive Model*

Figures 3, 4, 5, and 6 provide visualizations of the fitted surfaces using ordinary least squares, regularized least squares with ℓ^2 norm penalty, cubic polynomial, and cubic B-splines additive model. The models in Figures 3 and 4 do not appear to be much different from each other, mainly because there is no multicollinearity between the two covariates as the coordinates are constructed artificially. The cubic polynomial model seems to fit the data well in the region around the origin $(0, 0)$, and fit the data very poorly as it moves further along the East-West direction. The cubic B-splines additive model has the greatest flexibility

TABLE 1
Model Performances

Model	AIC	Training Error	Est. MPSE (10-Fold CV)	Est. MPSE (Leave-One-Out CV)
Ordinary Least Squares	9.769	1004.587	1025.709	1027.596
Least Squares with ℓ^2 Norm Penalty	9.769	1004.588	1025.721	1027.598
Cubic Polynomial	9.011	459.228	494.062	493.24
Additive Cubic B-Splines	8.647	305.941	339.485	344.503

to capture the variation of measured radial velocity across different coordinates and seems to provide a good fit consistently over different regions.

Table 1 provides a summary of the AIC values, training errors, and estimated mean prediction squared errors for each of the four models. We select the cubic B-splines additive model as our final model to make inferences about the association between the measured radial velocity and the coordinates of NGC 7531 as it has the smallest AIC value, smallest training error, and smallest estimated mean prediction squared errors among the four models. We will simply refer to it as the additive model for the rest of this paper.

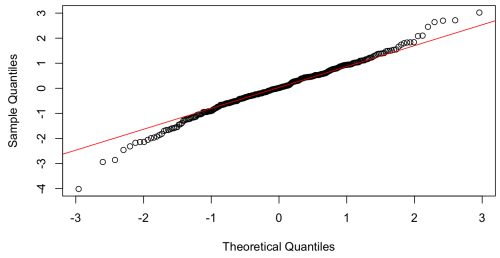

FIGURE 7. *Residuals Q-Q Plot*

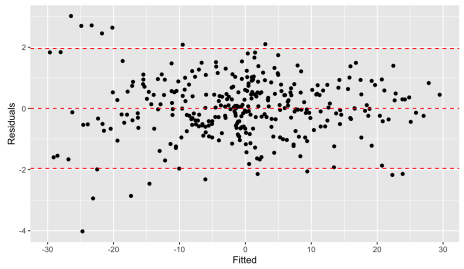

FIGURE 8. *Residuals Plot*

4.2. *Model Assumption Checking for Additive Model.* We perform a distributional check on the residuals computed using the additive model.

Based on our assumption of the errors, we expect the studentized residuals to independently follow a standard Gaussian distribution as well:

$$\hat{r}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}} \sim^{ind} \mathrm{N}(0,1)\ ,\ i = 1, \ldots, n\ ,$$

where $\hat{y}_i\ ,\ i = 1, \ldots, n$ are the fitted values under the additive model.

We estimate σ^2 using the mean squared error of the additive model:

$$\hat{\sigma}^2 = \frac{(y_i - \hat{y}_i)^2}{n - \mathrm{Trace}(H(\hat{\lambda}))}\ ,$$

where $H(\hat{\lambda})$ is the corresponding "Hat Matrix" for the additive model evaluated at $\hat{\lambda} = \arg\min_\lambda\ \mathrm{AIC}(\hat{f}, \lambda)$.

Figure 7 shows the Normal Q-Q Plot of the studentized residuals. Despite being slightly off at the tail regions, the majority of the studentized residuals lie reasonably well along the theoretical Normal Q-Q line in the middle region. Figure 8 shows a plot of the studentized residuals versus the fitted values. Almost 95% of the studentized residuals lie within the range $[-1.96, 1.96]$ and distribute randomly about 0. Both figures suggest that our Gaussian assumption of the error appears to be reasonable in this case.

5. Conclusions.

5.1. *Interpretation of the Final Model.* We use contrast with a reference point at the origin $(0,0)$ to provide an intuitive interpretation of the final result. Based on our Gaussian assumption on the errors, this contrast represents the difference in the expected radial velocity measured at coordinate (x_1, x_2) versus at the origin $(0,0)$:

$$\mathbb{E}[y \mid (x_1,x_2)] - \mathbb{E}[y \mid (0,0)] = \hat{f}_1(x_1) + \hat{f}_2(x_2) - \left(\hat{f}_1(0) + \hat{f}_2(0)\right) . \tag{5.1}$$

Figure 9 provides a three-dimensional visualization of contrast (5.1). Figure 10 provides a top view of the contrast surface shown in Figure 9 with contours showing levels of equal radial velocity measurements. Based on the two figures, the measured radial velocity tends to decrease when we measure from the eastern area of the galaxy and tend to increase when we measure from the western area of the galaxy. The steepest change occurs along the Northeast-Southwest diagonal as shown by the red dashed line in Figure 10.

5.2. *Discussion and Limitation.* Our model building and selection process rely solely on the data from the paper by Buta (1987), which is the only data set we have. This limits the generalization performance of our final model to predict radial velocity at coordinates that are not included in our data set. Since there is no natural way to anticipate how the radial velocity of a celestial object would change as we measure from different positions of the object, the only way we can infer whether overfitting or underfitting occurs is by looking at the difference between training error and the estimated mean prediction squared error. We conclude based on Table 1 that we may have avoided overfitting in our final model since the difference between the training error and the estimated mean prediction error are not large. However, we cannot provide a more rigorous evaluation of whether overfitting or underfitting occurs in our model. Also, because the diameter of NGC 7531 is about 92,120.90 light-years according to UniverseGuide (2023), a sample size of 323 may not constitute a reasonably large sample under this context. Therefore, it is worth obtaining a new, larger set of observations from NGC 7531 in the future to provide a more thorough assessment of our final model if possible. It should be noted that our result cannot be generalized to data that do not come from NGC 7531 as there is no guarantee that other galaxies would exhibit similar kinematics properties as NGC 7531.

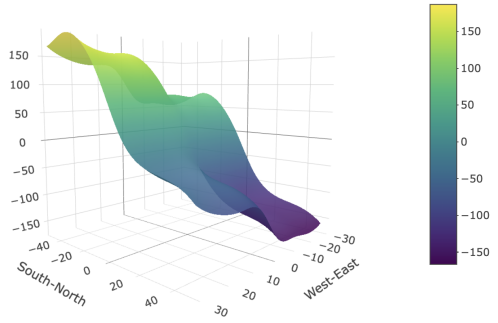

FIGURE 9. *Contrast Surface: Reference Point at (0,0)*

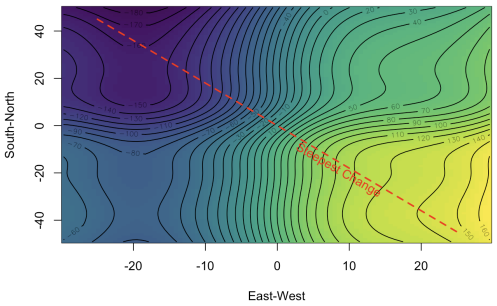

FIGURE 10. *Top View of Figure 9*

REFERENCES

BUTA, R. (1987). The Structure and Dynamics of Ringed Galaxies, III: Surface Photometry and Kinematics of the Ringed Nonbarred Spiral NGC7531. *The Astrophysical J. Supplement Ser.* **64** 1-37.

SUT (n.d.). Radial Velocity, Cosmos, https://astronomy.swin.edu.au/cosmos/r/Radial+velocity.

UNIVERSEGUIDE (2023). NGC 7531 Galaxy Facts. *https://www.universeguide.com/galaxy/ngc7531*.