Multivariate Regression Estimation of Temperature Signals in the Active Reflector Health Monitoring System for the FAST Project (Postprint)

Authors: Sun Xiao, Wang Qingmei, Li Zhenwei, Qiao Feng, Chu Jingjing

Date: 2022-01-14T15:19:05+00:00

Abstract

In large-scale structural health monitoring systems, temperature sensor failures can pose safety hazards. Based on the active reflector health monitoring system of the FAST project, we extracted and analyzed the linear correlations among data from nine measurement points, grouped candidate variables and selected the optimal regression subset, established a multiple linear regression model, and fused data from normal sensors to estimate faulty measurement points. To address multicollinearity among variables, ridge regression was further applied with the ridge parameter selected as 6. The model' s significance and validity were tested using F-test and goodness-of-fit test, and the estimation accuracy was verified using data from different dates. Results demonstrate that the multiple linear regression model achieves higher goodness-of-fit and accuracy than univariate approaches, with a root mean square error of 0.475°C, while the ridge regression method exhibits higher stability with a root mean square error of only 0.435°C.

Full Text

Multivariate Regression Estimation of Temperature Signals in the FAST Project Active Reflector Health Monitoring System

Sun Xiao[1], Wang Qingmei[2], Li Zhenwei[1], Qiao Feng[1], Chu Jingjing[1]
[1]College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China
[2]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Abstract: Temperature measurement point failures in large-scale structural health monitoring systems can create safety hazards. Based on the FAST project

active reflector health monitoring system, this study extracts and analyzes the linear correlations among data from nine measurement points. Candidate variables are grouped and the optimal regression subset is selected to establish a multiple linear regression model that fuses normal measurement point data to estimate faulty points. To address multicollinearity among variables, ridge regression is further applied with the ridge parameter selected as 6. The F-test and goodness-of-fit test verify model significance and validity, while data from different days validate estimation accuracy. Results demonstrate that the multiple linear regression model achieves higher fitting degree and precision than univariate models, with a root mean square error (RMSE) of 0.475°C. The ridge regression method provides greater stability, with an RMSE of only 0.435°C.

Keywords: Five-hundred-meter Aperture Spherical Telescope (FAST); temperature sensor; structural health monitoring; multiple linear regression; ridge regression

0 Introduction

Structural health monitoring is a technology that evaluates the current state of structures by deploying numerous sensors, with widespread applications in engineering safety monitoring. The Five-hundred-meter Aperture Spherical Telescope (FAST) is a large steel-structure project with an active reflector [**?**]. Literature [**?**] discusses the complexity of structural forces in FAST. Due to its special structure, large spatial span, and numerous components, the active reflector undergoes controlled deformation during observation, resulting in complex structural forces. To ensure safe operation, an active reflector health monitoring system has been constructed to assess the health status of FAST by monitoring stress and environmental information at key structural locations.

Literature [**?**] indicates that during cable net construction, the monitored real-time stress in the ring beam reached approximately 60 MPa, consistently below the design safety value of 201.5 MPa, remaining within safe limits, while emphasizing the importance of temperature signals for stress analysis. During FAST operation, active deformation of the cable net makes structural forces even more complex. Taking July 2019 monitoring data as an example, the maximum stress value at all ring beam lattice column measurement points was 137.37 MPa, with a minimum of -135.88 MPa, still within safe ranges. However, some measurement points exhibited large stress variations, with the largest range reaching 167.17 MPa at a single point. Stress monitoring remains crucial for real-time health assessment and long-term fatigue damage evaluation in subsequent telescope operations.

Large-scale engineering structures have substantial spatial dimensions, with different temperature rises across various locations. Temperature differences and thermal effects cause significant structural stress changes. Literature [**?**] notes that in FAST structures, temperature loads control the stiffness of the ring beam,

making distributed temperature measurement points critical for structural state assessment [**?**]. FAST employs 416 Fiber Bragg Grating (FBG) strain sensors [**?**] to monitor stress, including 100 measurement points on the ring beam and lattice columns and 316 on main cables. Although measurement points are distributed differently, all require temperature compensation for strain sensors based on actual temperatures at each point. Literature [**?**] further discusses the necessity of separating temperature effects from structural stress for state assessment, as missing temperature information directly compromises strain measurement reliability.

Actual engineering environments are complex, and numerous sensors operating long-term experience certain failure rates. Maintenance has revealed that individual FBG demodulators in FAST' s active reflector health monitoring system have malfunctioned, with several sensors producing abnormal data. The proportion of confirmed abnormal measurement points is approximately 5%, preventing normal acquisition of some monitoring data. However, due to site constraints and special installation locations, faulty sensors often cannot be repaired immediately. Data loss degrades health monitoring system performance and creates safety hazards.

Using adjacent measurement points to replace faulty temperature information presents challenges. For FBG sensors, which are often deployed in series, damage to a data channel can disable an entire group of adjacent sensors. Using distant sensor values is problematic due to large distances, varying solar exposure, and significant temperature differences, which is precisely why temperature sensors are deployed at each point. Accurate estimation of temperature signals at faulty measurement points is therefore essential for improving health monitoring system reliability.

---

1 Sensor Deployment and Data Extraction Analysis

The FAST active reflector health monitoring system uses FBG temperature sensors to monitor temperature information at key strain measurement points on the ring beam lattice columns. The sensor principle relies on temperature-induced changes in grating period and effective refractive index, causing reflected wavelength shifts. Temperature is obtained by measuring the center wavelength variation.

To effectively investigate correlations among temperature measurement points deployed over long distances, temperature data from corresponding measurement points on FAST' s edge ring beam support lattice columns were extracted. Each monitored lattice column provides one measurement point. FAST has 50 ring beam lattice columns total, with 10 equipped with measurement points. The numbers and distribution of monitored lattice columns are shown in Figure 1, with sensors installed in the green horizontal tie rods on the inner side of

ring beam supports as illustrated in Figure 2. The maximum distance between measurement points is 500 m.

The monitoring system samples data at 1 Hz. To reduce data volume, filter high-frequency interference, and avoid analysis difficulties from non-uniform data lengths, raw data are averaged using a 10-minute window. The average temperature value for each 10-minute period serves as the temperature data for that moment, yielding 144 samples per measurement point per day. July 2019 monitoring data were extracted, with the measurement point at lattice column #31 having no data due to acquisition channel maintenance, resulting in nine measurement points total. Since environmental temperature cycles typically follow diurnal patterns, data were grouped by day. Data for July 16 are shown in Figure 3.

Figure 3 shows significant temperature differences between measurement points, with maximum differences exceeding 5°C and varying fluctuation patterns, though overall trends are consistent. Two data samples are denoted as $X = [x_1, x_2, ..., x_n]$ and $Y = [y_1, y_2, ..., y_n]$. Linear correlation between different samples is calculated using the correlation coefficient formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is sample length and $\bar{x}, \bar{y}$ are respective means. The resulting correlation coefficient matrix for different measurement point temperature data is shown in Table 1, revealing high linear correlation between points, with the lowest correlation coefficient being 0.959.

2.1 Model Introduction

When variables are highly linearly correlated, linear regression methods are commonly used for modeling. With one independent variable denoted as x and dependent variable as y, the general model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0, β_1 are model coefficients and ε is random error. In health monitoring temperature estimation applications, univariate regression suffers from poor prediction accuracy at certain times due to non-uniform solar exposure [**?**].

Introducing more variables and integrating temperature effects from different measurement points yields better prediction results. With multiple independent variables, this is called multiple linear regression [**?**], with the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

where $\beta_0, \beta_1, ..., \beta_p$ are coefficients and p is the number of variables. With n data groups, the model can be simplified to matrix form:

$$Y = XB + E$$

where Y is the dependent variable vector, X is the independent variable matrix, B is the coefficient vector, and E is the random error vector. Random error terms follow a normal distribution $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, ..., n$. Neglecting error terms, least squares method solves for model coefficients:

$$\hat{B} = (X^T X)^{-1} X^T Y$$

Using the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, the empirical regression equation is established:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

where $x_1, x_2, ..., x_p$ are independent variable inputs and $\hat{y}$ is the model output estimate.

For the extracted nine measurement point datasets, taking one point as the dependent variable (the point to be estimated) and the remainder as independent variable inputs allows training a multiple linear regression model. When data from the estimated point are missing, its signal can be estimated using the independent variable inputs.

2.2 Optimal Independent Variable Selection

Using lattice column #1 measurement point as the dependent variable (point to be estimated) and the remaining eight measurement points' temperature information as candidate variables, there are $2^8 - 1 = 255$ possible variable subset combinations. Using July 16 data to train the model, variables are sorted by correlation in descending order and added incrementally, divided into eight groups.

The C_p statistic and adjusted coefficient of determination R_a^2 are combined as selection criteria to determine the optimal variable subset:

$$C_p = \frac{RSS_p}{RSS_m/(n-m)} - (n - 2p)$$

where $n = 144$ is sample length, $m = 8$ is the total number of candidate variables, p is the number of selected variables, RSS_m is the residual sum of squares for modeling with all variables, and RSS_p is the model residual sum of squares.

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

where R^2 is the coefficient of determination (model goodness-of-fit):

$$R^2 = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squares for dependent variable $y = [y_1, y_2, ..., y_n]$:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where y_i are actual values, $\hat{y}_i$ are model predictions, and $\bar{y}$ is the mean of actual values.

Variable grouping and corresponding model C_p and R_a^2 values are shown in Table 2. The optimal variable subset is selected by minimizing C_p and maximizing R_a^2, revealing that using all eight variables yields the best model fit.

---

3 Multicollinearity Problem and Ridge Regression Model

Table 1 shows extremely strong correlations between all measurement point pairs. For multiple linear regression models, such multicollinearity causes model stability issues. There may exist a set of numbers $k_0, k_1, ..., k_p$, not all zero, such that:

$$k_0 + k_1 x_{i1} + k_2 x_{i2} + ... + k_p x_{ip} \approx 0 \quad (i = 1, 2, ..., n)$$

At this point, the design matrix rank $rank(X) < p + 1$, meaning $|X^T X| \approx 0$ in formula (7). The coefficient vector estimate matrix $\hat{B}$ has variance matrix $D(\hat{B}) = \sigma^2 (X^T X)^{-1}$, whose diagonal elements become excessively large, reducing estimation precision for coefficient vector B. The model becomes overly sensitive to small input data changes, increasing prediction uncertainty and reducing model stability and accuracy.

Ridge regression optimizes the multicollinearity problem in multiple linear regression models [**?**]. The ridge estimate of the coefficient vector is defined as:

$$\hat{B}(\lambda) = (X^T X + \lambda I)^{-1} X^T Y$$

where $\lambda > 0$ is called the ridge parameter. Adding a positive constant matrix I makes $X^T X + \lambda I$ full-rank and invertible, reducing near-singularity compared to the original matrix.

The L_2 norm penalty term increases with λ, reducing model variance, but the coefficient estimate $\hat{B}(\lambda)$ deviates from the original value, increasing model bias. When $\lambda = 0$, the model reduces to ordinary multiple linear regression. Therefore, ridge parameter selection requires careful consideration.

Applying the ridge trace method, the ridge traces for eight independent variable regression coefficients are plotted in Figure 4. Following the principle of selecting the smallest λ value that stabilizes all coefficient vectors, $\lambda = 6$ is determined. Figure 4 shows that variables T_{26} and T_6 coefficients approach zero, indicating that ridge regression performs variable selection during training.

4.1 Testing and Evaluation Metrics

After establishing the multiple linear regression model, reliability must be assessed. The F-test determines regression equation significance, with the statistic:

$$F = \frac{ESS/p}{RSS/(n-p-1)}$$

where RSS is model residual sum of squares, ESS is regression sum of squares, $n = 144$ is sample length, and $p = 8$ is the number of selected variables:

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

where $\hat{y}_i$ are model predictions and $\bar{y}$ is the mean of actual values.

Model goodness-of-fit R^2 is tested using formula (11), with values closer to 1 indicating better fit. Using T_1 data as the point to be estimated and other measurement points as input variables, models are trained using July 16 data and tested on July 17 and 18 data. Estimation performance is evaluated using root mean square error (RMSE), where smaller values indicate better prediction:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

4.2 Test Results

Corresponding to the aforementioned groups, univariate through eight-variable linear regression models and ridge regression models were established and evaluated. Test and model estimation results are shown in Table 3. The lowest F-test statistic is 4972.29, far exceeding the critical value $F_{\alpha}(1, 135) = 3.911$, so all models pass significance tests. Goodness-of-fit values are all above 0.99, meeting requirements. As the number of fused variables increases, model testing RMSE decreases correspondingly.

Table 3 shows that the univariate, eight-variable linear regression, and ridge regression models all track the true values closely, with prediction curves nearly overlapping actual values. All models estimate signals effectively, with ridge regression achieving higher precision.

Temperature monitoring data from nine measurement points spanning 500 m in FAST' s active reflector health monitoring system were extracted. Analysis revealed extremely strong linear correlations between point pairs. Using measurement point #1 as the point to be estimated, methods for fusing data from other points were investigated, and model estimation performance was compared across different variable sets.

Multiple linear regression demonstrates superior performance over univariate models in both fitting degree and practical estimation applications. When input variables increase to three, testing RMSE falls below 0.5°C. Compared with original multiple linear regression, ridge regression slightly reduces model fitting degree but effectively avoids multicollinearity issues. Testing on July 17 and 18 data shows ridge regression achieves better prediction accuracy than multiple regression, with stronger anti-interference capability and stability. The average RMSE over two days is only 0.435°C.

These results indicate that when a measurement point fails, ridge regression can be used to train and establish a multiple linear regression model using recent historical data. Normal measurement point temperature information serves as input to estimate output values for faulty points. This method achieves high precision and can supplement missing node temperature information in health monitoring systems, maintaining normal operation during sensor repair periods. However, this approach uses recent data for fitting and prediction, serving only as a short-term replacement solution after point failure. If a point remains faulty for extended periods with significant time differences in available data, estimation accuracy will decrease due to variations in solar exposure and weather conditions causing changes in equipment temperature field patterns. Long-term data missing estimation methods require further research.

References

[?] R. D. Nan, "Five Hundred Meter Aperture Spherical Radio Telescope (FAST)," *Science in China: Series G: Physics Mechanics & Astronomy*, vol. 49, no. 2, pp. 129-148, 2006.

[?] X. L. Wang, D. P. Li, M. Zhu, et al., "Design and application of health monitoring system of main reflector system for FAST," *Journal of Guangxi University of Science and Technology*, vol. 29, no. 4, pp. 84-91, 2018.

[?] Y. T. Wang, Q. M. Wang, M. Zhu, et al., "Application of Fiber Grating Sensor in Construction Stage of Cable Net for FAST," *China Mechanical Engineering*, vol. 27, no. 20, pp. 2760-2764, 2016.

[?] B. Q. Zhao, Q. M. Wang, Z. H. Li, et al., "Performance research of FAST ring beam by theoretical and experimental," *Engineering Mechanics*, vol. 35, no. S1, pp. 200-211, 2018.

[?] H. Wang, Y. W. Ning, and H. Yan, "Temperature Distribution and Deformation Impact Analysis of 26m Antenna Frame," *Astronomical Research and Technology*, vol. 15, no. 02, pp. 208-215, 2018.

[?] X. Sun, Q. M. Wang, M. Zhu, et al., "Application of optical fiber Bragg grating strain gauge to cable force monitoring of FAST," *Optics and Precision Engineering*, vol. 23, no. 4, pp. 919-925, 2015.

[?] Q. M. Wang, M. Zhu, Q. M. Wang, et al., "Research and application of Data Processing Method of the Health Monitoring System for the Main Active Reflector of FAST," *Astronomical Research and Technology*, vol. 14, no. 2, pp. 164-171, 2017.

[?] Z. C. Guo, P. Sun, Z. C. Li, et al., "Research on the Influence of Prediction Model and Modeling Sequence Length on Short-term Prediction Accuracy of Clock Bias," *Astronomical Research and Technology*, vol. 17, no. 03, pp. 299-307, 2020.

[?] X. Xiao and X. G. Wu, "Geometric Interpolation of Multicollinearity in Linear Regression," *Statistics & Decision*, no. 21, pp. 46-51, 2021.

[?] A. E. Hoerl and R. W. Kennard, "Ridge regression: application for non-orthogonal problems," *Technometrics*, vol. 12, pp. 69-72, 1970.

Figures

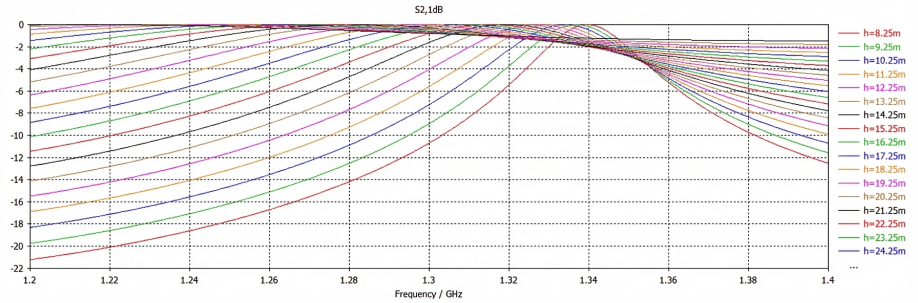

Figure 1: Figure 4