TYPE-1 FUZZY LASSO REGRESSION FUNCTIONS

Nihat Tak – Aylin Ucan – Dogan Yildiz

Abstract

Fuzzy Functions (FFs), originally proposed by Turksen, represent a non-rule-based inference method. In this approach, data are clustered using the Fuzzy C-Means (FCM) algorithm; each observation is assigned to all clusters with specific membership degrees, and these degrees are incorporated into the input matrix as additional variables. Turksen suggested that this expanded input matrix could enhance the model's predictive performance. Moreover, including functions of the membership degrees in the model has been shown to further improve predictive accuracy. However, such transformations can introduce a large number of correlated variables into the input matrix, potentially leading to multicollinearity issues. In this study, it is aimed to use Lasso regression with Type-1 Fuzzy Functions to address this issue and improve model performance. Lasso performs variable selection by excluding ineffective or less significant variables, thereby increasing the model's generalizability and reducing the risk of overfitting. The method has been evaluated on three real-world datasets, and the results show that it outperforms the other methods included in the study.

Key words: Type-1 Fuzzy Functions, Lasso Regression, Feature Selection, Forecasting,

JEL Code: C45, C53

1 Introduction

Forecasting is essential for effective decision-making, as it enables the prediction of future events or trends by examining past data. This structured process relies on mathematical modeling, statistical analysis, and a variety of techniques to estimate what lies ahead. Across fields such as business, economics, and weather forecasting, it provides a reliable basis for planning in uncertain conditions. By interpreting patterns in historical data, forecasting supports data-driven and strategic predictions.

Forecasting approaches are generally classified into two primary categories: statistical and non-statistical (alternative) techniques. Statistical approaches, which rely on deterministic structures and relatively simple models, tend to perform effectively when the underlying data satisfies certain predefined assumptions. Nonetheless, their effectiveness significantly

diminishes in the presence of real-world data that exhibit complexity, non-linearity, and unpredictability.

Contemporary forecasting research has increasingly focused on alternative methodologies, including artificial intelligence (AI) and fuzzy logic, which offer enhanced flexibility and accuracy compared to conventional techniques. AI, particularly through advanced machine learning models such as artificial neural networks (McCulloch & Pitts, 1943), is proficient in uncovering complex data structures and modeling nonlinear associations. This strength enables forecasters to effectively process and interpret intricate and evolving datasets, thereby facilitating a more comprehensive analysis of underlying patterns and behaviors. On the other hand, fuzzy logic is especially advantageous in contexts involving vagueness and uncertainty, as it accommodates gradations of truth and linguistic variables, offering a robust solution where traditional models fall short in managing ambiguity.

Fuzzy Inference Systems (FIS), originally proposed by Zadeh (1973), serve as robust artificial intelligence frameworks for modeling uncertainty and handling imprecise information. At the heart of these systems lies a rule base composed of expert-defined fuzzy rules, which articulate the relationships between inputs and outputs through "IF-THEN" constructs. These rules apply fuzzy logic to input variables to infer corresponding outputs. Consequently, the effectiveness of an FIS is heavily dependent on the quality and comprehensiveness of the expert knowledge embedded in its rule set. However, this dependence poses a significant limitation, particularly in dynamic environments, as traditional FIS lack the capacity to autonomously adapt or learn from new data. To address this challenge, Turksen (2008) introduced type-1 fuzzy function-based methodologies that facilitate automatic rule generation, thereby enhancing the adaptability and learning ability of fuzzy systems.

Type-1 Fuzzy Functions (T1FFs) rely on a regression approach that utilizes both membership degrees and their transformations as predictors. However, this structure may lead to multicollinearity, a common issue in high-dimensional datasets, which violates the core assumptions of Ordinary Least Squares (OLS) regression. In regression analysis, multicollinearity increases the variance of parameter estimates, causing the model to overfit the training data and thereby reducing its generalizability. (Bas et al., 2019-2020), (Kizilaslan et al., 2020) and (Tak and Inan, 2022) addressed the multicollinearity problem in their studies by using Ridge Regression. Although ridge regression can resolve multicollinearity, it does not address the issue of overfitting. Lasso regression (Least Absolute Shrinkage and Selection Operator) is an effective method for reducing multicollinearity and selecting important

variables. With the L1 penalty term, it shrinks some coefficients to zero, which simplifies the model and makes it easier to interpret. This helps prevent overfitting (Tibshirani, 1996). In high-dimensional datasets, Lasso also improves model accuracy and generalizability by identifying the most relevant variables (Zou & Hastie, 2005; James et al., 2013). The positive effects of Lasso on multicollinearity and overfitting have been supported by various studies in the literature. For example, Kökçü and Gençtürk (2021) demonstrated the effectiveness of Lasso in variable selection and overfitting control in their study on environmental datasets. Similarly, a study by Demir and Özkan (2018) reported that Lasso yielded lower error rates compared to classical methods and provided more stable results in the presence of multicollinearity.

T1FFs use the Fuzzy C-Means (FCM) algorithm to determine the membership degrees included in the input matrix. The clustering process is critically important for identifying accurate patterns, and therefore, extensive research has been conducted on clustering algorithms in the literature. Specifically, the FCM algorithm, originally proposed by Dunn (1973) and later developed by Bezdek (1984), is among the most widely used fuzzy clustering methods. The obtained membership degrees and functions are incorporated into the model by adding them to the input matrix. In this context, Celikyılmaz and Türksen (2009) stated that applying mathematical transformations such as exponential and logarithmic to the membership degrees could improve the prediction performance of T1FFs. These transformations make the model more flexible and generalizable.

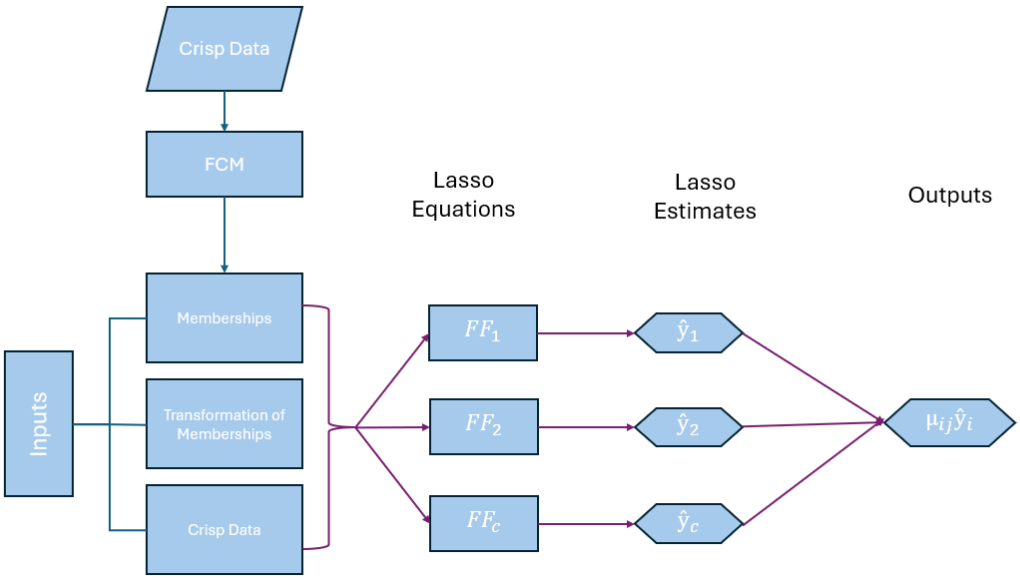
However, this extended structure may lead to high correlations among explanatory variables, causing modeling issues such as multicollinearity and overfitting. In this study, to mitigate these problems and balance model complexity through variable selection, the Type-1 Fuzzy Lasso Regression (T1FLR) method was employed, which integrates the Lasso regression technique within the framework of Type-1 Fuzzy Functions. In this way, it is aimed to overcome both multicollinearity and overfitting issues.

2 Method

The primary objective of clustering methods is to group observations with similar characteristics while separating those that differ significantly. In this process, the accurate determination of cluster centers plays a critical role. In the Type-1 Fuzzy Functions (T1FFs) framework, data are clustered using the Fuzzy C-Means (FCM) algorithm, where each observation is assigned to all clusters with specific degrees of membership. These membership

degrees and corresponding functions are then incorporated into the input matrix and included in the model. However, this extended structure may introduce high correlations among explanatory variables, leading to issues such as multicollinearity and overfitting. To mitigate these problems, the Lasso regression technique is employed within the T1FF framework in this study. By performing variable selection, Lasso retains only the most relevant predictors in the model, thereby enhancing both its generalizability and interpretability. The detailed steps and overall structure of the method are presented below.

Fig. 1: Architecture of Fuzzy Functions with Lasso



Algorithm 1:

Step 1: The fuzziness parameter m and the number of clusters c are selected.

Step 2: The input matrix $X \in R^{nxp}$ is clustered using FCM, and the membership degrees and cluster centers are obtained.

Step 2.1: Calculate the membership value using the formula in Equation (1):

$$\mu_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d(x_k,v_i)}{d(x_k,v_j)}\right)^{\frac{2}{m_i-1}}\right]^{-1} \quad i=1,2,\ldots,c;\ k=1,2,\ldots,n \tag{1}$$

Here, x denotes the input matrix, v represents the cluster centers, $d(\cdot)$ is the Euclidean distance function, c is the number of clusters, and m is the fuzziness parameter, as shown in Equations (1) and (2).

Step 2.2 Update the cluster centers using the formula in Equation (2):

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^{m_i} x_k}{\sum_{k=1}^{n} (\mu_{ik})^{m_i}} \quad (2)$$

Step 2.3 Repeat Steps 2.1 and 2.2 until the difference between cluster centers in two consecutive iterations falls below a predefined threshold, or until the maximum number of iterations is reached.

Step 3: Adding the membership degrees and their functions to the input matrix, the X and Y matrices corresponding to the i-th cluster are obtained as follows.:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix}, X = \begin{bmatrix} \mu_{i1} & \mu_{i1}^2 & ln_{(\mu_{i1})} & e^{\mu_{i1}} & x_{11} \ldots & x_{p1} \\ \mu_{i2} & \mu_{i2}^2 & ln_{(\mu_{i2})} & e^{\mu_{i2}} & x_{12} \ldots & x_{p2} \\ . & . & . & . & . \ldots & . \\ \mu_{in} & \mu_{in}^2 & ln_{(\mu_{in})} & e^{\mu_{in}} & x_{1n} \ldots & x_{pn} \end{bmatrix}, i = 1, 2, \ldots, c$$

Step 4: For each cluster, a regression model is fitted using Lasso to perform parameter estimation and feature selection simultaneously.

$$\arg min_B \left\{ \sum_{i=1}^{c} \left(y^{(i)} - (X^i)^T B^{(i)} \right)^2 + \lambda \left\| B^i \right\|_1 \right\} \text{ for } i = 1, 2, ..., c \quad (3)$$

Step5: The predictions obtained from the Lasso regression model for each cluster are calculated using Equation (4).

$$\hat{y}_i = \frac{\sum_{i=1}^{c} \hat{y}_{ik} \mu_{ik}}{\sum_{i=1}^{c} \hat{y}_{ik}} \quad k = 1, 2, \ldots, n \quad (4)$$

3 Applications

In this study, three different regression datasets from various domains were utilized. The first dataset, "Concrete Compressive Strength" (Data 1), was obtained from the UCI Machine Learning Repository and contains various features related to concrete mixtures. The second dataset, "Steel Fatigue Strength Prediction" (Data 2), was retrieved from the Kaggle platform and includes measurements related to materials science. The third dataset consists of Near-Infrared (NIR) Spectroscopy data (Data 3) obtained from the “chemometrics” package in R, and is used for chemical content prediction. All datasets were randomly split into training (80%), validation (10%), and test (10%) subsets. The number of clusters (c) was determined by searching within the range of 2 to 5. For the Lasso regression model, the regularization parameter λ was optimized using cross-validation via the “cv.glmnet” function in R. All analyses were conducted using the R programming language. The method was evaluated in comparison with multiple linear regression (MLR), Ridge regression, and classical Lasso regression methods. Summary information and the optimal hyperparameter values selected for each dataset are presented in Table 1.

Table 1 Hyper parameter detections and datasets information

Data No	Methods	n	p	c	m	α	λ
Data 1	MLR	1030	8	-	-	-	-
Data 1	T1FLR	1030	8	2	2	1	0.0075
Data 1	Lasso	1030	8	-	-	1	0.0075
Data 1	Ridge	1030	8	-	-	0	0.8048
Data 2	MLR	437	25	-	-	-	-
Data 2	T1FLR	437	25	5	2.6	1	0.0826
Data 2	Lasso	437	25	-	-	1	0.0826
Data 2	Ridge	437	25	-	-	0	15.4752
Data 3	MLR	166	235	-	-	-	-
Data 3	T1FLR	166	235	4	3	1	0.0855
Data 3	Lasso	166	235	-	-	1	0.0855
Data 3	Ridge	166	235	-	-	0	85.4942

The model selection process was carried out based on the RMSE values calculated from the validation datasets. At this stage, the best-performing methods were identified, and the validation results for each dataset are presented in detail in Table 2.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2} \tag{5}$$

Table 2 RMSE values of the methods

Methods	Data 1	Data 2	Data 3
MLR	10.0861	36.3706	131.7285
T1FLR	**9.6189**	**33.6250**	**5.3599**
Lasso	10.0895	36.8116	5.7110
Ridge	10.3412	40.1546	10.8752

According to Table 2, the T1FLR method achieved the lowest RMSE value on the validation dataset for each data set. Therefore, performance evaluation in the testing phase was

conducted based on this method. Tables 3, 5, and 7 present the performance results of the T1FLR method on the test datasets for Data 1, 2, and 3, respectively.

In addition to the evaluation metrics of the T1FLR method, hypothesis testing was conducted based on the predicted and actual values to assess whether the differences between T1FLR and the other methods were statistically significant. For all datasets, the null (H_0) and alternative (H_1) hypotheses are defined as below, and the comparison results are presented in Tables 4, 6, and 8. According to Table 4 and Table 6, the differences between T1FLR and the Ridge method are statistically significant at the 95% confidence level. Similarly, based on Table 8, the differences between T1FLR and both the Ridge and MLR methods are also statistically significant.

H_0: There is no statistically significant difference in test error values between the T1FLR method and the compared methods (Lasso, Ridge, MLR)

H_1: There is a statistically significant difference in test error values between the T1FLR method and the compared methods.

Table 3 Evaluation metrics of the test dataset of Data 1

Methods	RMSE	MAE	MAPE
T1FLR	12.2986	9.9901	0.3619

Table 4 Pairs Method Comparison of Data 1

Methods	Test Type	p-value
T1FLR vs Lasso	Wilcoxon	0.2229
T1FLR vs Ridge	Wilcoxon	0.0464
T1FLR vs MLR	Wilcoxon	0.2382

Table 5 Evaluation metrics of the test dataset of Data 2

Methods	RMSE	MAE	MAPE
T1FLR	33.4378	22.6183	0.0400

Table 6 Pairs Method Comparison of Data 2

Methods	Test Type	p-value
T1FLR vs Lasso	Wilcoxon	0.6281
T1FLR vs Ridge	Paired t-test	0.0093
T1FLR vs MLR	Wilcoxon	0.4543

Table 7 Evaluation metrics of the test dataset of Data 3

Methods	RMSE	MAE	MAPE
T1FLR	4.6868	3.7433	0.4795

Table 8 Pairs Method Comparison of Data 3

Methods	Test Type	p-value
T1FLR vs Lasso	Wilcoxon	0.463700
T1FLR vs Ridge	Paired t-test	0.004400
T1FLR vs MLR	Wilcoxon	0.000031

Conclusion

In this study, the performance of the regression method integrating Lasso regularization into Type-1 Fuzzy Functions is investigated. The method was evaluated on three real-world datasets from different domains (concrete compressive strength, material fatigue strength, and chemical content prediction). Hyperparameters were optimized by minimizing the RMSE value on the validation dataset, and the proposed method was compared with classical regression techniques including MLR, Lasso, and Ridge. The application results showed that the T1FLR method achieved the lowest RMSE values during the validation phase across all datasets. Evaluations conducted on independent test datasets further confirmed the superior performance of T1FLR in terms of RMSE, MAE, and MAPE criteria. In future studies, it is planned to test the method with different feature selection techniques on various linear and nonlinear datasets, and to expand the hyperparameter search space.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) (Grant 123F266).

References

Bas, E., Egrioglu, E., Yolcu, U., & Grosan, C. (2019). Type 1 fuzzy function approach based on ridge regression for forecasting. *Granular Computing,* 4, 629-637.

Bas, E., Yolcu, U., & Egrioglu, E. (2020). Picture fuzzy regression functions approach for financial time series based on ridge regression and genetic algorithm. *Journal of Computational and Applied Mathematics*, 370, 112656.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & geosciences, 10(2-3), 191-203.

Celikyilmaz A, I. B. Turksen, Modeling uncertainty with fuzzy logic, *Studies in fuzziness and soft computing* 240 (1) (2009) 149–215

Demir, M., & Özkan, S. (2018). A comparison of ridge, lasso and elastic net methods under multicollinearity. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, 67(2), 1184–1197.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kizilaslan, B., Egrioglu, E., & Evren, A. A. (2020). Intuitionistic fuzzy ridge regression functions. *Communications in Statistics-Simulation and Computation,* 49(3), 699-708.

Kökçü, Ö., & Gençtürk, M. (2021). Comparison of variable selection methods in environmental data: A case study. *Journal of Environmental Statistics*, 13(2), 45–58.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5 (4), 115–133. doi: 10. 1007/BF02478259

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58(1), 267–288.

Tak, N., & İnan, D. (2022). Type-1 fuzzy forecasting functions with elastic net regularization. *Expert Systems with Applications,* 199, 116916.

Türkşen, I. B. (2008). Fuzzy functions with LSE. *Applied Soft Computing,* 8(3), 1178-1188.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67(2), 301–320.

Contact

Nihat Tak
Marmara University, Goztepe Campus, Faculty of Science, Department of Statistics,
Yerleşkesi 34722 Kadıköy - Istanbul
nihat.tak@marmara.edu.tr

Aylin Ucan
Marmara University, Goztepe Campus, Faculty of Science, Department of Statistics,
Yerleşkesi 34722 Kadıköy - Istanbul
aylin.ucan@std.yildiz.edu.tr

Dogan Yildiz
Department of Statistics, Yıldız Technical University,
Istanbul, 34220, Turkey
dyildiz@yildiz.edu.tr