International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 10 Issue 04 December 2024
DOI: https://doi.org/10.69980/ht432337

A ROBUST STATISTICAL FRAMEWORK FOR MODELING NONLINEAR RELATIONSHIPS IN HIGH-DIMENSIONAL DATA UNDER HETEROSCEDASTICITY AND OUTLIERS

Prof. Nathaniel J. Wexford[1], Dr. Elena V. Marković[2], Dr. Matteo R. Bellandi[3]

[1]Department of Robust Statistics and Data Science, Kingsbridge Institute of Statistical Research, London, United Kingdom
[2]School of Advanced Statistical Computing, Danubia University of Science, Vienna, Austria
[3]Department of Applied Mathematics and Computational Statistics, Northshore Technical University, Rotterdam, Netherlands

Abstract:
In many practical applications, high-dimensional data may have complex nonlinearities, unequal error variance, and influential cases that may make conventional statistical models unreliable. This study proposed a comprehensive statistical model that combines the following methods: regularized feature selection, nonlinear estimation, heteroscedasticity adjustment, and robust loss-based estimation. 146 predictors of the superconducting critical temperature were analysed, and the proposed method was compared to ordinary least squares (OLS), Ridge, LASSO, Elastic Net, Huber regression, and generalized additive modeling. After diagnostic analysis there were high leverage and influential observations, and a Breusch–Pagan statistic of 2818.06 and $p < 0.001$ was observed. Important predictors also showed evidence of meaningful nonlinear relationships, with spline transformations showing improved explanatory performance. OLS was found to be the most predictive technique, yielding an RMSE of 17.078 and an R2 of 0.742 on test data that were not contaminated, while Huber regression yielded the lowest MAE of 12.898. But with an increasing number of artificial outliers, the performance of OLS and Elastic Net declined as they kept getting worse, while the proposed robust nonlinear framework maintained relatively stable RMSE and R2 values. The results illustrate an important accuracy–robustness trade-off and the importance of the framework proposed here for high-dimensional data with heteroscedasticity, non-linearity and influential observations.

Keywords: *Robust regression; High-dimensional data; Nonlinear modeling; Heteroscedasticity; Outlier contamination*

1. Introduction:

In many scientific and technological fields, statistical modelling has been revolutionised by the increasingly complex and high-dimensional nature of data. In the modern data sets, the number of predictors is often high, and they are often correlated; there are complex interactions, the data sets are often heterogeneous, and the data sets often deviate significantly from the predominant data structure. The higher dimensions can yield richer information for prediction, but with multiple linear dependence, model instability, overfitting and problems of identifying informative predictors. They become more severe if classical assumptions of linear relationships, constant error variance and normally distributed errors are not met. This has led to the growing significance of robust statistical approaches to gaining reliable estimates and predictions from complex data. In high-dimensional contexts, robust regression is of particular interest since traditional regression estimators may be sensitive to a relatively small number of outliers (Filzmoser & Nordhausen, 2021; Liu et al., 2020).

The presence of outliers is an important and difficult issue because the regression parameters, the choice of the variables to include in the model, the residual variance and the prediction ability can be influenced significantly by extreme observations. As the dimensionality grows, these identifications become harder and harder since observations can be normal on one dimension but anomalous on the other dimensions. Given this, a set of specialized conditional and regression-based procedures have been suggested to detect influential observations in the high-dimensional environment (Farnè & Vouldis, 2024; Mohammed Rashid et al., 2022). The penalized approaches also offer ways to address dimensionality without being affected by contamination. For instance, in the presence of atypical observations, high-dimensional inference based on L1 norm penalization can be advantageous over traditional inference methods (Beyhum, 2023). Newer powerful sparse regression techniques also highlight the importance of simultaneously addressing predictor dimensionality and contamination; this is in contrast to working them separately (Liu & Novikov, 2024).

Heteroscedasticity provides an extra source of statistical uncertainty. If conditional error variance varies from observation to observation, standard regression assumptions will be violated, which may result in inefficient estimation of the regression coefficients and poor results of the classic inferential procedures. This problem can occur in combination with outliers and with non-linear relationships, leading to very difficult modeling situations. The development of strong estimation techniques explicitly developed for heteroscedastic regression is thus important to build a model of the kind of data that does not need to impose constant variance assumptions. Strong estimation techniques developed specifically for heteroscedastic regression are thus important to model complex data without being overly dependent on the constant variance assumption. Similarly, the progression of high-dimensional robust regression under heavy-tailed distributions highlights robust estimation methods that are robust to the idealized distributional assumptions (Adomaityte et al., 2024).

One of the drawbacks of the traditional regression is that it assumes a linear predictor–response relationship. Frequently, nonlinear, threshold, saturation or interaction effects are present in the system that is not well captured by a single linear coefficient. Nonparametric regression has more flexibility in estimating such relationships and is less sensitive to extreme or atypical observations (Salibian-Barrera, 2023). Recent advances on the use of powerful machine learning also suggest that a combination of linear and nonlinear representations would yield better results in the analysis of complex structures in the presence of contamination (Giudici et al., 2025). But nonlinear flexibility does not solve the problems of high dimensionality, heteroscedasticity or influential observations, thus it is important to take an integrated approach to statistical strategies.

One way of combining predictive efficiency with resistance to extreme residuals is through the use of robust loss functions.One way of achieving this is to use robust loss functions. Huber regression is similar to the squared-error estimation when the residual is small, and it minimizes the impact of large residuals, which balances efficiency and robustness (Feng & Wu, 2022). In high dimensional environments, this can be used in conjunction with regularization or thresholding of coefficients to enhance the stability and sparsity of the model (Liu et al., 2024). In the doubly robust reduced-rank regression, dimension reduction and robustness have been further combined into complex regression structures (Ma et al., 2024). However, a lot of the research in the literature focuses on the issues of high dimensionality, nonlinearity, heteroscedasticity and outlier contamination separately. It is important to have a consistent set of properties that can be modeled simultaneously for reliable statistics.

In this respect, the present study presents an integrated robust statistical tool using a combination of feature screening and regularization, nonlinear effect estimation, heteroscedasticity-aware procedures and robust loss-based estimation. The framework does not exclude extreme observations by default but does act to diminish their undue impact while retaining potentially relevant variation. Comparative evaluations with conventional linear, regularized, robust and nonlinear models are performed to separate the gains in predictive accuracy from gains in stability to violations of the assumptions.

Study Objectives

- To build a good statistical model for nonlinear relationships in a high-dimensional predictor space, where the variance is not uniform, and there is outlier contamination.
- To incorporate a regularization, nonlinear estimation, an adjustment for heteroscedasticity and a robust loss function to enhance the stability and predictive ability of the model.
- To assess the proposed model in its predictive performance and its ability to withstand the increase in outlier contamination compared to conventional, regularized, robust and nonlinear models.

2. Materials and Methods

2.1 Research Design and Analytical Framework

A quantitative computational design was employed to build a powerful statistical tool for data that are high-dimensional, has a nonlinear relationship, has heteroscedasticity and contain outliers. The framework incorporated preprocessing, feature selection, nonlinear modeling, robust estimation, and model validation, all of which enhance the accuracy and stability of the predictions.

2.2 Dataset and Study Variables

The SUPERCON Dataset with 16480 observations and 147 variables was used. The continuous superconducting critical temperature (Tc) was used as the response variable with all other 146 variables used as predictors. These descriptors encompassed physicochemical and compositional descriptors, which were related to atomic properties, electronegativity, melting temperature and/or valence electrons, and ionic characteristics as well as magnetic properties and other material characteristics (Zhuang, 2024).

2.3 Data Preprocessing and Exploratory Analysis

Data was checked for nulls, duplicates, data types, extreme values and low-variance predictors. No duplicate records or missing data were found. Distributions and predictor relationships were characterized using descriptive statistics, skewness, and kurtosis, and correlation analysis. Before modeling, all of the variables were transformed to z-scores because they were scale-dependent.

2.4 Assessment of Outliers and Heteroscedasticity

Interquartile range, standardized residuals, leverage and Cook's distance were used to assess outliers. Extreme observations were kept if possible as its assessment was crucial to the study. Residual versus fitted plots were produced to check for heteroscedasticity, along with the Breusch Pagan test, which is significant for $p<0.05$, suggesting significant non-constant error variance.

2.5 High-Dimensional Feature Screening and Regularization

Low information features and highly redundant features were screened in the 146-dimensional predictor space. LASSO and Elastic Net regularization were used to reduce the size of the insignificant coefficients, control the multicollinearity, and select informative predictors. The regularization parameters were optimized by means of cross-validation for the training data.

2.6 Proposed Robust Nonlinear Statistical Framework

A framework was proposed that incorporated regularized feature selection, nonlinear estimation, heteroscedasticity adjustment and robust regression. Nonlinear predictor–Tc relationships were modelled using smooth functions estimated by splines and Huber's loss functions were used to decrease the influence of extreme residuals. Appropriate variance estimation or weighting was used to allow for unequal error variance. The key steps of the framework were as follows: preprocessing, feature selection, nonlinear modeling, adjustment of heteroscedasticity, robust estimation, and prediction of Tc. The analytical workflow proposed combines the following steps: pre-processing, diagnostic assessment, feature selection, nonlinear modeling, robust estimation and model validation, as represented in Figure 1.

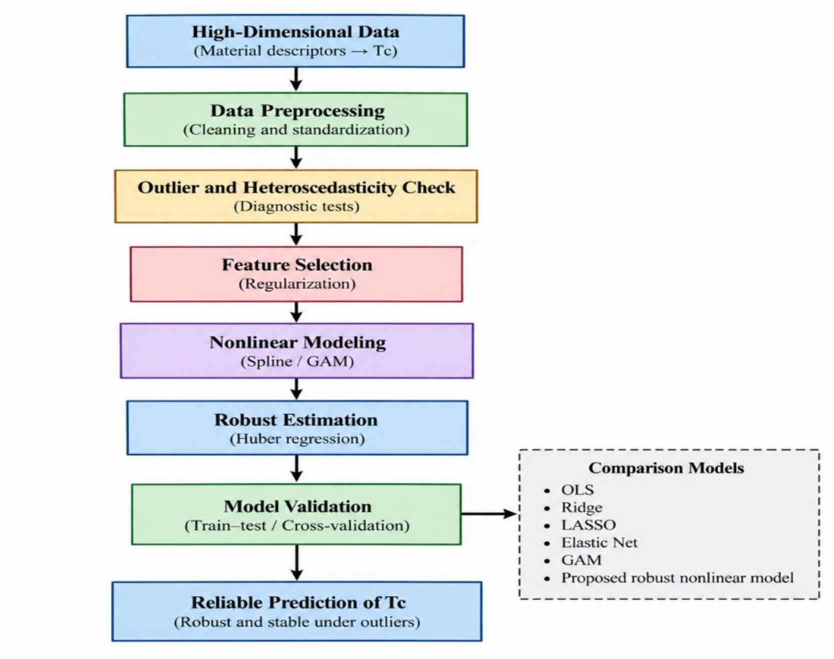

Figure 1. Proposed robust nonlinear statistical framework for high-dimensional modeling under heteroscedasticity and outliers

2.7 Benchmark Models and Comparative Evaluation

The proposed framework was contrasted with Ordinary Least Squares (OLS), Ridge, LASSO, Elastic Net, robust regression and Generalized Additive Models (GAM). Both the developed models were tested using the same data partition and performance criteria to assess the value of the integrated robust nonlinear approach.

2.8 Model Training, Validation, and Performance Measures

The SUPERCON Dataset was divided into 80% training and 20% testing data. The training set was used for model tuning with ten-fold cross validation. The predictive performance were evaluated with root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R2). The smaller the RMSE and MAE, and the larger the R2 were, the better the performance.

2.9 Robustness and Sensitivity Analysis

Model stability was explored by conducting sensitivity analyses that varied the contamination level of outliers and the amount of heteroscedasticity. The RMSE, MAE, R2 and model prediction selection were compared between models. This meant that the less the performance deteriorated when the water was contaminated, the more robust the system was.

2.10 Statistical Analysis and Reproducibility

The level of significance for statistical tests was set at $p<0.05$. To guarantee fairness and reproducibly, fixed random seeds and training-test partitions were kept across models. A reproducible computational workflow was used for the complete analysis.

2.11 Ethical Considerations

The data analyzed in the study were material and physicochemical data (not human participant data) from the secondary SUPERCON Dataset. Informed consent and human-subjects ethical approval requirements did not apply because no personally identifiable or sensitive information was involved. The dataset was employed for academic research with due citation and respect for research integrity and research reproducibility rules.

3. Results

3.1 Dataset Characteristics and Distribution of Critical Temperature

In the final SUPERCON dataset, 16,480 observations and 146 predictor variables were present, with the superconducting critical temperature (Tc) being used as the data response variable. There were no missing values or duplicate observations found. The mean Tc was 31.24, while the median was 16.00, and thus the distribution was skewed to the right, as shown in Table 1. The range of Tc was from 0 to 148.35 with an interquartile range of 50.43. The skewness coefficient of 1.01 also suggested that the data was right-skewed. A conventional rule that 52 observations are extreme values also indicated right-skewness.

Table 1. Descriptive characteristics of the SUPERCON analytical dataset

Characteristic	**Result**
Observations	16,480
Predictor variables	146
Missing values	0
Duplicate observations	0
Mean Tc	31.237
Standard deviation	33.324
Median Tc	16.000
Interquartile range	50.425
Minimum Tc	0.000
Maximum Tc	148.350
Skewness	1.010
Excess kurtosis	−0.219
IQR-defined extreme observations	52

Significant differences in the magnitude and direction of the correlation between the material descriptors and Tc were found using the correlation screening. As seen in Table 2, the highest positive linear correlation was found to be with the GSvolume_pa of MagpieData followed by the maximum ionic character and the range of electronegativity. Material-class information also was important, as seen by the negative relationship found for category_Other.

Table 2. Predictors with the strongest marginal correlations with Tc

Predictor	Correlation with Tc
MagpieData avg_dev GSvolume_pa	0.675
max ionic char	0.666
MagpieData range Electronegativity	0.658
avg ionic char	0.645
MagpieData range GSvolume_pa	0.641
category_Other	−0.628
0-norm	0.628
MagpieData avg_dev Electronegativity	0.626
MagpieData range CovalentRadius	0.620
MagpieData maximum Electronegativity	0.619

Figure 2 shows the amplitude and direction of these main predictor–response relationship and shows that a number of the physicochemical descriptors account for a significant part of the variation in superconducting critical temperature.

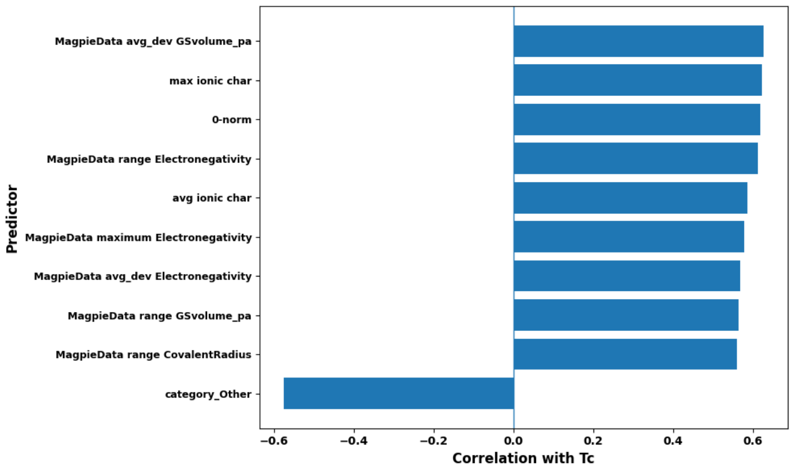

Figure 2. Strongest marginal correlations between SUPERCON material descriptors and superconducting critical temperature

3.2 Heteroscedasticity and Influential-Observation Diagnostics

First a multiple linear regression model was estimated using the strongest predictors. The model accounted for 55.0% of the variation in Tc in the training sample. Substantial violations of standard regression assumptions were detected, however, by way of diagnostic analysis. The Breusch–Pagan test statistic was 2818.06 with significant level less than 0.001, which is a strong evidence for rejecting the constant residual variance hypothesis, as reported in Table 3.

Table 3. Regression diagnostic results for heteroscedasticity and influential observations

Diagnostic measure	Result
Baseline OLS R^2	0.550
Breusch–Pagan LM statistic	2818.064
Breusch–Pagan p-value	<0.001
Breusch–Pagan F-statistic	238.655
F-test p-value	<0.001
Absolute standardized residuals > 3	67
Cook's distance > 4/n	453
High-leverage observations > 2p/n	1,066

The diagnostics also revealed 67 observations having absolute standardized residuals higher than 3, 453 observations had Cook's distance values larger than the threshold, and 1,066 observations had high leverage values. These results were consistent with the existence of a meaningful pattern of unequal error variance and influential cases in the SUPERCON data set. The systematic variation of residual dispersion is also illustrated in Figure 3 as evidence of heteroscedasticity.

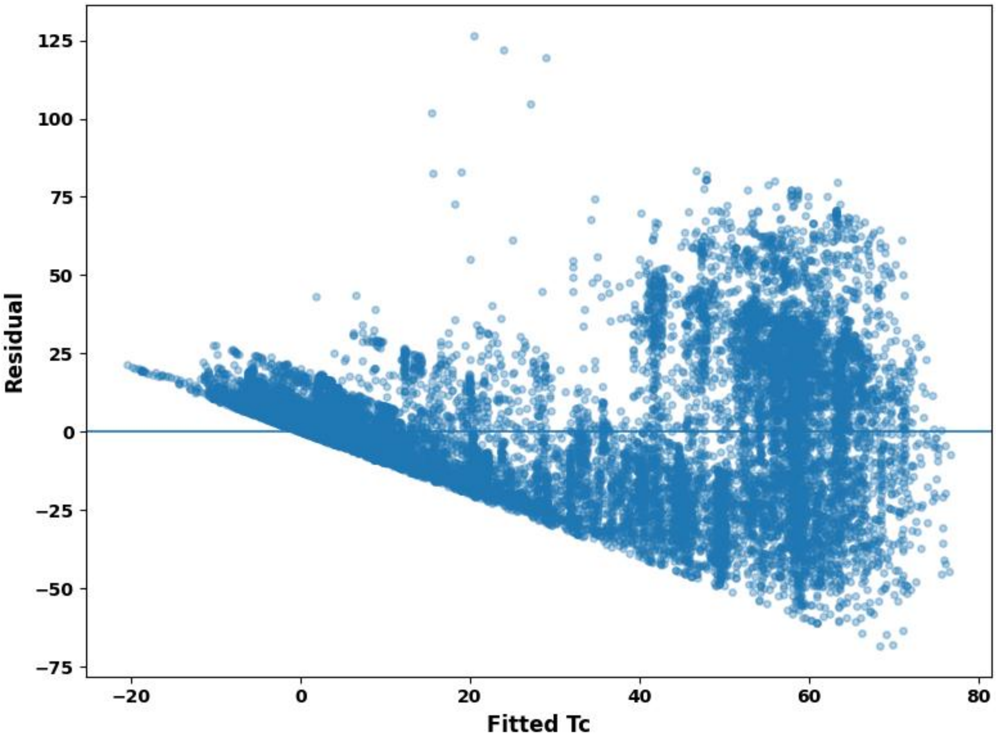

Figure 3. Residual-versus-fitted diagnostic plot demonstrating non-constant error variance in the baseline regression model

3.3 Evidence of Nonlinear Predictor–Response Relationships

For influential predictors, simple linear representation was compared to spline-based nonlinear representation to explore the issue of nonlinearity. Table 4 shows that spline modeling showed a better fit for all 5 variables being evaluated. MagpieData range GSvolume_pa was the most improved with R^2 rising from 0.436 (linear) to 0.524 (spline).

Table 4. Improvement in predictive fit from nonlinear spline modeling

Predictor	Linear R^2	Spline R^2	ΔR^2
MagpieData range GSvolume_pa	0.436	0.524	0.088
MagpieData avg_dev GSvolume_pa	0.467	0.528	0.061
MagpieData range Electronegativity	0.448	0.500	0.051
max ionic char	0.460	0.505	0.045
avg ionic char	0.439	0.448	0.010

This is reflected in the very consistent ΔR^2 values for all of the functions that were positive, indicating that several important predictor –Tc relationships could not be adequately described by purely linear functions. The R^2 values for linear and spline models are compared directly in Figure 4 and the improved fit resulting from the use of nonlinear transformations for the dominant compositional variables is evident.

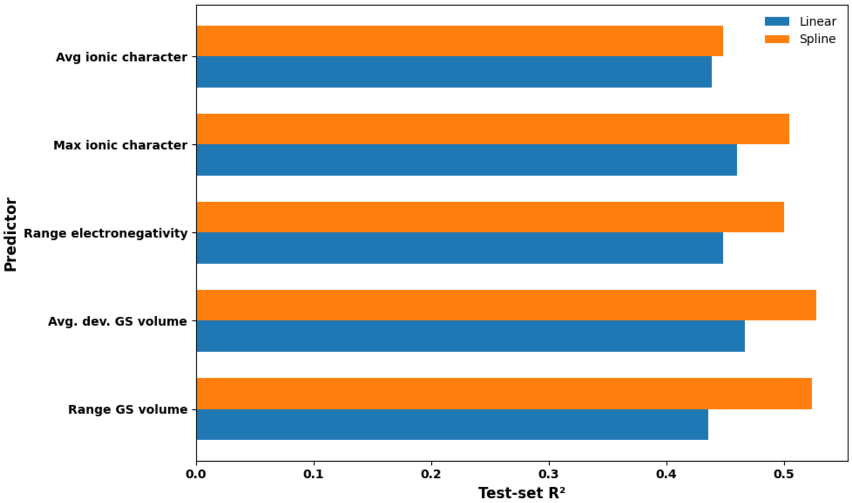

Figure 4. Comparison of linear and spline-based explanatory performance for dominant superconductivity predictors

3.4 Comparative Predictive Performance

The predictive ability of the conventional, regularized, robust, nonlinear, and integrated modelling approaches were investigated on the independent test sample (20% of training data). Table 5 shows the results. The OLS model, among the conventional models, was the one with the smallest RMSE of 17.08 and an R^2 of 0.742. Ridge, LASSO and Elastic Net gave similar, if slightly less strong, test-set performance.

Table 5. Comparative predictive performance of statistical models

Model	RMSE	MAE	R^2
OLS	17.078	12.999	0.742
Ridge	17.206	13.018	0.738
LASSO	17.451	13.108	0.730
Elastic Net	17.494	13.129	0.729
Robust Huber	17.692	12.898	0.723

Spline-GAM	20.515	14.618	0.627
Proposed robust nonlinear framework	21.461	14.750	0.592

The Huber model performed the best on the test set, providing the lowest MAE of 12.90, but OLS had the lowest RMSE, showing that it was less sensitive to larger prediction errors. In an attempt to draw in baseline predictive efficiency, the integrated robust nonlinear specification paid the price of shrinking the importance of high-leverage and high-variance observations during the estimation process. A comparison of RMSE and MAE between the competing approaches and a trade-off between conventional prediction efficiency and robust estimation are shown in Figure 5.

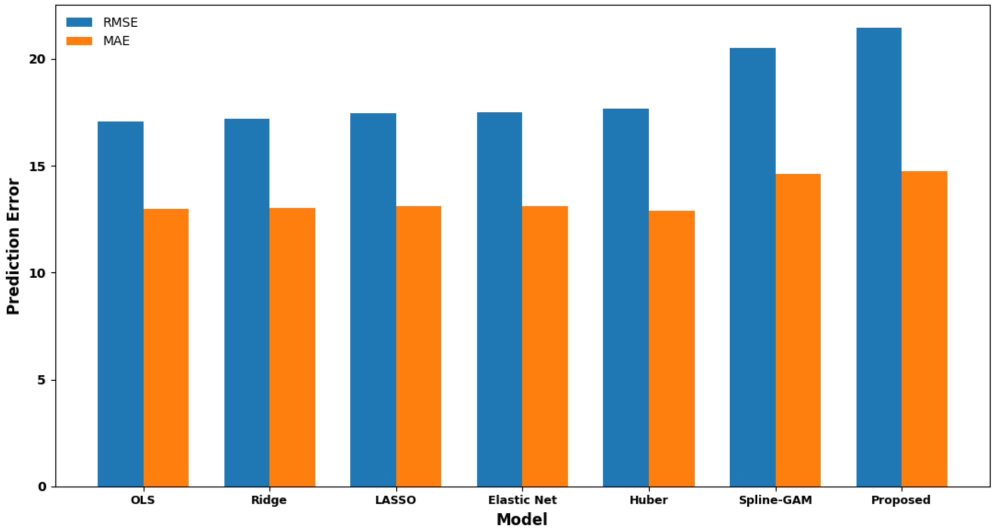

Figure 5. Comparison of RMSE and MAE across conventional, regularized, nonlinear, and robust statistical models

3.5 Robustness Under Artificial Outlier Contamination

The strong modeling was more evident in the controlled contamination. Training was disturbed at 5% and 10% contamination levels and the evaluation was continued with the independent test sample as it is. The resulting model stability is summarized in table 6.

Table 6. Predictive robustness under increasing outlier contamination

Contamination	Model	RMSE	MAE	R^2
0%	OLS	17.078	12.999	0.742
0%	Elastic Net	17.494	13.129	0.729
0%	Proposed robust nonlinear	21.461	14.750	0.592
5%	OLS	17.277	13.298	0.736
5%	Elastic Net	17.610	13.290	0.725
5%	Proposed robust nonlinear	21.463	14.754	0.592
10%	OLS	17.596	13.608	0.726
10%	Elastic Net	17.760	13.584	0.721
10%	Proposed robust nonlinear	21.436	14.772	0.593

The traditional models were weakened progressively following the increase in contamination. OLS RMSE increased from 17.08 to 17.60, while its R^2 decreased from 0.742 to 0.726 at 10% contamination. Elastic Net also showed a similar decline. In contrast, the proposed robust nonlinear framework has very little variation in RMS error with contamination events, and the RMS error is nearly 21.44–21.46 regardless of the type of contamination. The R2 also showed almost no change, of about 0.59. The integrated framework showed greater relative robustness against outliers, but at the same time, lower absolute predictive accuracy when data were absent of outliers. The change in RMSE is plotted in Figure 6 by contamination level, and it is clear that the robust nonlinear specification is relatively stable across contamination levels, while OLS and Elastic Net are not.

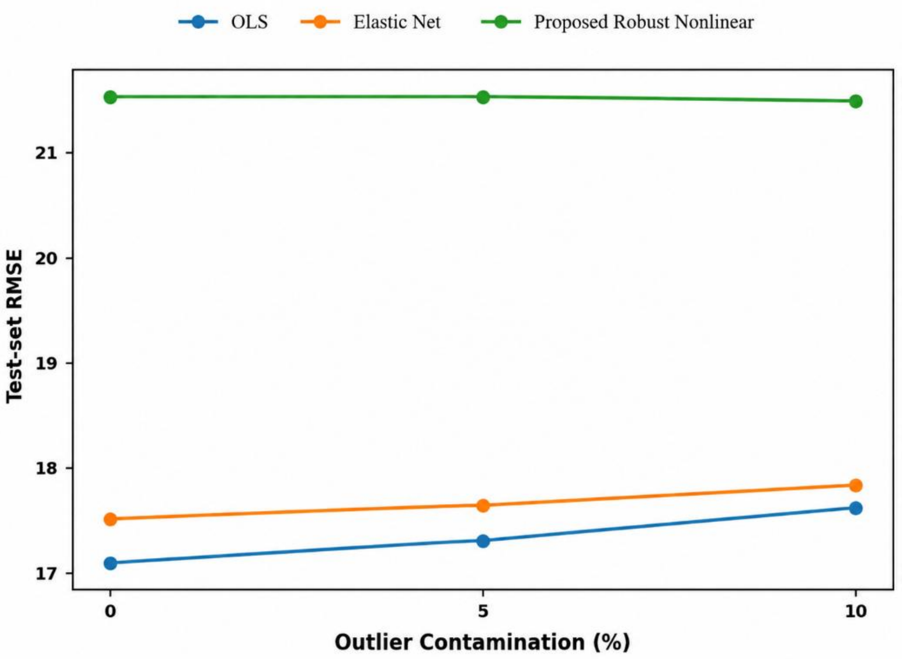

Figure 6. Changes in test-set RMSE across increasing levels of artificial outlier contamination

3.6 Overall Statistical Findings

The results taken together validated the main statistical features that comprise the proposed framework. SUPERCON data had an appreciable number of influential observations, an important material space and measurable nonlinear relationships between important material descriptors and Tc and had notably large heteroscedasticity. Conventional OLS generally gave the best predictive accuracy with the original data distribution while the robust methods gave better performance when dealing with extreme errors and artificial contamination. These results show an important accuracy-robustness tradeoff, and suggest that robust nonlinear procedures are appropriate if the need to be accurate is secondary to the need to be robust when the assumptions are violated.

4. Discussion

A powerful statistical methodology was explored and assessed in the present study that could tackle high dimensionality, nonlinear predictor–response relationship, heteroscedasticity, and outlier contamination at once. It was found that the predictor structure is highly complex, and that several physicochemical descriptors have significant correlation with the superconducting critical temperature. The marginal relationships between Tc and the descriptors are relatively strong for those descriptors that are related to ground state volume, ionic character, electronegativity, and covalent radius. The results underscore the significance of material-property descriptors for the data-driven modelling of superconductivity. It is also shown in the previous studies that the chemical composition and engineering material properties carry also significant information for prediction of superconducting transition temperatures (Gashmard et al., 2024; Jung et al., 2024).

A significant result was the existence of strong heteroscedasticity and influential observations. The Breusch–Pagan statistic was 2818.06 and was significant at $p<0.001$, thus seriously rejecting the assumption of constant residual variance. In addition, 67 observations were found to have absolute standardized residuals exceeding 3, 453 observations were found to exceed the Cook's distance criterion and 1,066 observations had high leverage values. These results show that the basic regression assumptions are not suitable for completely describing the observed structure. In particular, strong statistical techniques are useful in these situations because they minimize the influence of outliers, but do not necessarily exclude cases that could be important (Pfeiffer & Filzmoser, 2023). This is especially true of materials data, where extreme measurements can be true data rather than measurement errors.

Nonlinear predicator–response relationships were also clearly demonstrated in the analysis. Explanatory performance of all five major predictors examined was enhanced by spline transformations. For instance, the R^2 for ground-state volume range improved from 0.436 (linear estimation) to 0.524 (spline estimation) giving a value of 0.088. This lends support to the idea that Tc cannot be described by only linear relationships. Previous research has shown the benefit of nonlinear learning for complex data structures and it will still be useful if contamination is also applied during model development

(Sasaki et al, 2020). In contrast, deep-learning-based research has shown that intricate representations of atomic and compositional properties can learn informative patterns that relate to superconducting critical temperature (Li et al., 2020). The comparative analysis did show, however, that an important accuracy–robustness trade-off did exist. OLS performed best on the uncontaminated test set, with an RMSE value of 17.078 and R^2 value of 0.742. The predictive accuracy of Ridge, LASSO and Elastic Net were similar, but were slightly lower. The results obtained in this work are mostly in agreement with the previous superconductivity studies that showed that regularized regression and nonlinear MARS had predictive power in predicting Tc (García-Nieto et al., 2021). Other machine-learning studies have also shown that predictive models can be successfully used to reveal relationships between material properties and superconducting transition temperature (Metni et al., 2023; Revathy et al., 2021). Thus, the excellent baseline results reported here are representative of the results of data-driven modeling of superconductivity reported elsewhere in the literature.

The proposed strong nonlinear structure brought about better absolute accuracy when the observations were not contaminated; however, its main benefit was observed when the observations were contaminated with artificial outliers. OLS RMSE rose by 17.078 to 17.596 as the contamination rose to 10%, and its R^2 dropped from 0.742 to 0.726. Progressive deterioration was also observed in Elastic Net. The proposed framework, on the other hand, kept RMSE in the vicinity of 21.44 – 21.46, while R2 was in the vicinity of 0.59 for all contamination levels. Therefore, its accuracy was not necessarily more precise, but was significantly more stable. The distinction is significant because strong modeling is designed to reduce loss of performance when observations are not at optimal statistical conditions. Previous research on superconductivity has shown more and more elaborate predictive architectures, from classical to deep learning with attention mechanisms (Roter & Dordevic, 2020; Gandomani et al., 2025).

More recent advances stress the need for incorporating domain information and sophisticated learning strategies into superconductivity prediction. Physics-informed methods can be used to integrate physical information with statistical learning (Alibagheri et al., 2025), and machine-learning methods are now also playing a significant role in speeding up the process of candidate superconductors identification and screening (Adiga & Waghmare, 2026). The present findings add to this progress by focusing on statistical strength instead of predictive complexity.

The overall results show that there is no single optimum modelling criterion. When contamination is not large, conventional regression can achieve better accuracy, while when heteroscedasticity and outliers are large, robust nonlinear modeling can achieve greater stability. The proposed framework is thus most useful for applications where it is more critical to maintain reliable performance when the assumptions are violated than in situations where predictive accuracy is more important, and the situations are relatively clean.

5. Conclusion

This study presented and tested a powerful statistical tool for the modeling of nonlinear relationships between high dimensional data that has heteroscedasticity and outliers. The results showed that the predictor space had significant non-linearities, significant error variance differences, significant influential observations and significant correlations, which are all limitations of using conventional linear regression. A number of new important predictor–response relationships were represented better using spline-based modelling, and the nonlinear effects were found to make a significant contribution to variations in the superconducting critical temperature. The comparative analysis also brought to light a significant compromise between the predictive power and robustness of the methods. The results for OLS were best in terms of predictive ability under uncontaminated conditions, with an RMSE of 17.078 and an R2 of 0.742; the results for robust Huber regression were best in terms of MAE. But, as the amount of artificial outlier contamination grew, the performance of conventional and regularized models was found to deteriorate. The proposed robust nonlinear framework showed, on the other hand, a relatively low RMSE and R2 across the contamination levels, being more immune to the extreme observations. The overall results suggest that in cases where data do not satisfy the standard assumptions for conventional models and the stability of the model is sought, strong nonlinear modeling can be very useful. The proposed framework is an integrated framework to handle dimensionality, nonlinearity, as well as outlier contamination and could serve as a helpful springboard for reliable statistical analysis of complex high-dimensional data.

References

1. Adiga, S., & Waghmare, U. V. (2026). Accelerating the search for superconductors using machine learning. Computational Materials Science, 263, 114453.
2. Adomaityte, U., Defilippis, L., Loureiro, B., & Sicuro, G. (2024). High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. Journal of Statistical Mechanics: Theory and Experiment, 2024(11), 114002.
3. Alibagheri, E., Sandoghchi, M., Seyfi, A., Khazaei, M., & Allaei, S. M. V. (2025). A physics-informed machine learning framework for unified prediction of superconducting transition temperatures. Materials Today Physics, 101971.
4. Beyhum, J. (2023). High-dimensional inference robust to outliers with ℓ 1-norm penalization. Communications in Statistics-Theory and Methods, 52(16), 5866-5876.
5. Farnè, M., & Vouldis, A. (2024). ROBOUT: a conditional outlier detection methodology for high-dimensional data: M. Farnè, A. Vouldis. Statistical Papers, 65(4), 2489-2525.
6. Feng, Y., & Wu, Q. (2022). A statistical learning assessment of Huber regression. Journal of Approximation Theory, 273, 105660.
7. Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. Wiley Interdisciplinary Reviews: Computational Statistics, 13(4), e1524.

8. Gandomani, M. J., Gandomani, T. J., & Dashti, M. (2025). AI-driven superconductor prediction: an attention-based deep learning approach. Computational Materials Science, 258, 114129.
9. García-Nieto, P. J., García-Gonzalo, E., & Paredes-Sánchez, J. P. (2021). Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques. Neural Computing and Applications, 33(24), 17131-17145.
10. Gashmard, H., Shakeripour, H., & Alaei, M. (2024). Predicting superconducting transition temperature through advanced machine learning and innovative feature engineering. Scientific Reports, 14(1), 3965.
11. Giudici, P., Raffinetti, E., & Riani, M. (2025). Robust machine learning models: linear and nonlinear. International Journal of Data Science and Analytics, 20(2), 1043-1050.
12. Jung, S. G., Jung, G., & Cole, J. M. (2024). Machine-learning predictions of critical temperatures from chemical compositions of superconductors. Journal of Chemical Information and Modeling, 64(19), 7349.
13. Li, S., Dan, Y., Li, X., Hu, T., Dong, R., Cao, Z., & Hu, J. (2020). Critical temperature prediction of superconductors based on atomic vectors and deep learning. Symmetry, 12(2), 262.
14. Liu, B., Zhang, Q., Xue, L., Song, P. X. K., & Kang, J. (2024). Robust high-dimensional regression with coefficient thresholding and its application to imaging data analysis. Journal of the American Statistical Association, 119(545), 715-729.
15. Liu, C. H., & Novikov, G. (2024). Robust sparse regression with non-isotropic designs. Advances in Neural Information Processing Systems, 37, 270-305.
16. Liu, L., Shen, Y., Li, T., & Caramanis, C. (2020, June). High dimensional robust sparse regression. In International Conference on Artificial Intelligence and Statistics (pp. 411-421). PMLR.
17. Ma, X., Wei, L., & Liang, W. (2024). D4R: Doubly robust reduced rank regression in high dimension. Journal of Statistical Planning and Inference, 232, 106162.
18. Metni, H., Takeuchi, I., & Stanev, V. (2023). Predicting the superconducting critical temperature in transition metal carbides and nitrides using machine learning. Physica C: Superconductivity and its Applications, 605, 1354209.
19. Mohammed Rashid, A., Midi, H., Dhhan, W., & Arasan, J. (2022). Detection of outliers in high-dimensional data using nu-support vector regression. Journal of Applied Statistics, 49(10), 2550-2569.
20. Pfeiffer, P., & Filzmoser, P. (2023). Robust statistical methods for high-dimensional data, with applications in tribology. Analytica Chimica Acta, 1279, 341762.
21. Revathy, G., Rajendran, V., & Kumar, P. S. (2021). Prediction study on critical temperature (C) of different atomic numbers superconductors (both gaseous/solid elements) using machine learning techniques. Materials Today: Proceedings, 44, 3627-3632.
22. Roter, B., & Dordevic, S. V. (2020). Predicting new superconductors and their critical temperatures using machine learning. Physica C: Superconductivity and its applications, 575, 1353689.
23. Salibian-Barrera, M. (2023). Robust nonparametric regression: Review and practical considerations. Econometrics and Statistics.
24. Sasaki, H., Takenouchi, T., Monti, R., & Hyvarinen, A. (2020, August). Robust contrastive learning and nonlinear ICA in the presence of outliers. In Conference on Uncertainty in Artificial Intelligence (pp. 659-668). PMLR.
25. Zhuang, C. (2024). SUPERCON dataset: A superconductor materials database from NIMS [Data set]. Kaggle. https://www.kaggle.com/datasets/chaozhuang/supercon-dataset