International Journal of Mathematics and Statistics

ISSN (Online): 2208-2212
Volume 10 Issue 03 September 2024

DOI: https://doi.org/10.69980/aeb1he22

ROBUST VARIABLE SELECTION IN HIGH-DIMENSIONAL REGRESSION USING ADAPTIVE PENALIZATION UNDER CONTAMINATED DATA

Dr. Edward J. Harrington[1]*, Prof. Katarina M. Novak[2], Dr. Samuel T. Laurent[3]

[1]Department of Statistical Learning and Computational Mathematics, Westford University, Manchester, United Kingdom*
[2]Institute of Applied Statistics and Data Science, Central European Research University, Prague, Czech Republic
[3]Department of Quantitative Methods and Statistical Computing, Laurentian Institute of Technology, Lyon, France

Abstract:

High-dimensional regression methods often have many correlated predictors, which makes them prone to overfitting, to having unstable estimates of the regression coefficients, and to making unreliable selections of the predictors, especially when data includes outliers or influential observations. In this study, adaptive penalization methods were studied for robust variable selection in progressively contaminated regression. A comparison of ridge regression, LASSO, Elastic Net, Adaptive LASSO, and Robust Adaptive LASSO was made using the prediction, sparsity and robustness criteria. The models were tested with controlled contamination of 0%, 5%, 10%, 15% and 20% to prove the stability of the models under increasingly harsh conditions. The predictive accuracy was measured by root mean squared error (RMSE), mean absolute error (MAE) and R2 values, and the sparsity of the model was evaluated based on the number of predictors selected. Under these conditions, the data was uncontaminated, and Ridge regression gave the smallest RMSE while Adaptive LASSO gave the sparsest model. But the more contaminated the data, the more stable Robust Adaptive LASSO was compared to the other estimators. Its RMSE showed only a 10.27% rise, and R2 showed an 8.56% decrease, when the contamination level was increased from 0% to 20%. The strong adaptive model only selected 14 predictors at the maximum level of contamination, whereas the LASSO and Elastic Net models selected 66 and 74 variables, respectively.

Keywords: *high-dimensional regression; robust variable selection; adaptive penalization; Elastic Net; Huber loss*

1. Introduction

In scientific, engineering, biomedical and computational applications, data dimensionality has grown rapidly, making the need to manage potentially large numbers of relevant predictors. High-dimensional regression involves many explanatory variables that can be as large as or greater than the number of observations, which poses significant challenges in the form of multicollinearity, overfitting, unstable estimation of the regression coefficients and poor generalization. Conventional regression methods are especially susceptible in this situation as having redundant predictors can result in unreliable parameter estimates and interpretation problems. In the context of high-dimensional statistical modeling, which aims at prediction and at the same time extracting a parsimonious subset of informative variables, the sparse regression and variable selection have emerged as central challenges (Wang et al., 2020; Stokell & Shah, 2022; Cao et al., 2023).

A good solution to these issues is penalized regression, which helps to control model complexity and stabilize estimation. At present, the least absolute shrinkage and selection operator (LASSO) is the most widely used, which automatically selects variables by shrinking weak regression coefficients exactly to zero. Extensions of LASSO have been suggested to enhance performance in the case of highly correlated predictor variables or variable importance varying significantly across features (Xi et al., 2023; Żogała-Siudem & Jaroszewicz, 2024). Elastic Net is a hybrid combination of L1 and L2 regularization and may be useful to obtain stability in highly correlated predictor structures; adaptive forms of the regularization enable different levels of penalization to be applied to each individual regression coefficient. With recent advances, it is shown that adaptive Elastic Net and structured sparse penalties are useful in high-dimensional estimation and process modeling (Kim et al., 2024; Wang et al., 2024).

However, these benefits, however, do not alleviate the widespread use of conventional penalized regression procedures that are often very sensitive to outliers, heavy-tailed errors, influential observations, and contaminated data, as they typically use squared-error loss. A small number of atypical observations can have a large influence on the regression surface drawn, the magnitude of the coefficients, and the number of predictors included in the regression or number of predictors excluded from the regression that are informative. This problem is exacerbated in high-dimensional data, where contamination and predictor redundancy can cause variable-selection instability to be greatly magnified. Strong regression techniques have, therefore, become a significant application of sparse modeling, and recent works have focused on combining robustness and regularization for robust estimation with non-ideal data (Amato et al., 2021; Filzmoser & Nordhausen, 2021).

Strong loss functions are used to minimize the impact of very large residuals. This Huber-type estimation is especially appealing as it acts like squared-error loss when the residuals are small but minimizes the influence of large residuals when they are large. Robust loss functions can significantly increase the stability of estimation under violations of the Gaussian assumption, with the adaptive Huber regression showing in theory as well as in practice good properties when heavy-tailed and/or heterogeneous error distributions are present (Sun et al., 2020). Concurrently, variable selection is a different challenge since being robust does not guarantee sparseness. In the case of ultra-high-dimensional data, where the dimension of the data poses a challenge for feature recovery, methods are needed that can address the contaminating effects as well as identify relevant predictors (Bai et al., 2021).

An alternative method for improving variable selection consistency, though, is to use adaptive penalization, which shrinks differently for each predictor. Weak or irrelevant predictors could be treated with higher shrinkage, while relatively stronger penalties could be imposed on important variables. But these techniques are very sensitive to the choice of the regularization and the tuning parameter(s). The choice of penalty parameters could lead to underdense or overdense models that could decrease the accuracy of the prediction and interpretability (Wu & Wang, 2020; Biziaev et al., 2025). In addition, the computerization of penalized path algorithms has brought attention to the need for efficient optimization, especially when the number of candidate predictors is large (Takahashi & Nomura, 2024).

The accuracy of inference and ranking of variables once they have been selected is another important point to consider. High-dimensional procedures need to be able to detect true signals from spurious associations and have consistent predictive performance in different data structures. This recent literature, focusing on sufficient variable selection and sparsified inference, has continued to highlight the ongoing need for procedures balancing dimensionality reduction, uncertainty and model stability (Wang et al., 2025; Zhu et al., 2025). All these factors indicate that testing penalized methods in a "clean data" setting may give a partial picture of their reliability as a practical tool.

Based on this, the present study explores the robustness of adaptive estimation of the regression parameters in high-dimensional regression settings using a combination of adaptive penalization and robust estimation and assesses performance in progressively contaminated situations. Conventional and adaptive penalized estimators are compared in clean and contaminated settings, with a focus on the performance of both in terms of predictive accuracy.

Objectives:

- To evaluate penalized, adaptive penalized and robust adaptive regression techniques on contaminated data in a high-dimensional setting under increasing data contamination rate.
- To compare the predictive ability, sparsity and stability of variable selection of the various estimators using RMSE, MAE, R2 and performance measures based on variable selection.
- To see if the combination of strong loss functions and adaptive penalization is more resistant to outliers and contamination than the traditional LASSO, Elastic Net and adaptive penalized methods.

2. Methodology

2.1 Research Design

In this study, a robust variable selection approach for high-dimensional regression with contaminated data was considered, and a quantitative computational design was used to assess the performance of the method. The paper made a comparison study on the conventional penalized regression, adaptive penalization and robust adaptive penalization methods with the Superconductivity Data Set. The study analyzed the performance of the various regression estimators in terms of prediction error, stability of the coefficients and variable selection ability for different contamination levels.

2.2 Dataset Description

This study is based on the Superconductivity Data Data Set with 21,263 observations of superconducting materials. There are 81 numerical predictor variables and 1 continuous dependent variable, critical_temp, in train.csv. The predictors are the physicochemical properties of atoms like atomic mass, ionization energy, atomic radius, density, electron affinity, fusion heat, thermal conductivity, and valence. The elemental-composition variables, the critical temperature and the material identifier are provided with the accompanying unique_m.csv file. Critical_temp was regarded as the response variable (Tunguz, 2021) due to it being continuous.

2.3 Data Preprocessing

The data sets were matched by the observations and critical-temperature values found in each of the two data sets. The textual material variable was not included in numerical modeling and only one copy of critical_temp was kept. Those predictors that had no variance were excluded. All other variables were checked for missing values, infinite values, duplicates and extreme values. Numerical predictors were z-scored to standardize them to make them comparable across variables that are measured on different scales.

2.4 High-Dimensional Regression Framework

The regression model was expressed as:

$$y = X\beta + \varepsilon,$$

where y is critical temperature, X is the predictor matrix, β is regression coefficients and ε is the random errors. In order to explore the truly high-dimensional condition, experimental subsets were developed for which the number of predictors came close to or surpassed the number of observations. Thus, various combinations of sample sizes were explored to represent p<n and p>n conditions.

2.5 Contamination Design

The levels of controlled contamination were 0,5,10,15 and 20%. The baseline data set was 0%, which was a clean sample. Three types of contamination were explored: response contamination extreme deviations added to the dependent variable; predictor contamination high-leverage observations added to selected predictors; combined contamination – high-leverage observations added to selected predictors and extreme deviations added to the response variable. This design allowed for the evaluation of model robustness with increased contamination levels.

2.6 Penalized Regression Methods

Ridge regression, LASSO, and Elastic Net were used to represent the conventional penalized regression methods in the study. For ridge regression, the shrinkage benchmark was employed, and for simultaneous coefficient estimation and variable selection, LASSO was employed. Elastic Net used L1and L2penalties and was included because it performs effectively when predictors are highly correlated.

2.7 Adaptive Penalization

Different penalty weights for individual predictors were determined by using Adaptive LASSO. The estimator was defined as:

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{2n} \| y - X\beta \|_2^2 + \lambda \sum_{j=1}^{p} w_j \mid \beta_j \mid \right],$$

where wjrepresents predictor-specific adaptive weights. The recovery of important predictors received relatively smaller penalties, but weak predictors received stronger penalties which improved the sparse variable recovery.

2.8 Robust Adaptive Penalization

Therefore, Huber loss was used with adaptive L1penalization to be robust against outliers and contaminated observations. The powerful adaptive estimator was formulated as:

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{n} \sum_{i=1}^{n} \rho_{\delta} \left(y_i - x_i^T \beta \right) + \lambda \sum_{j=1}^{p} w_j \mid \beta_j \mid \right]$$

While giving an efficient estimate for regular observations, the Huber loss is designed to be robust to large residuals. Ridge, LASSO, Elastic Net, Adaptive LASSO, robust LASSO, and robust adaptive penalization were the final comparisons.

2.9 Model Training and Parameter Selection

80% of the data was used to train, and 20% of the data was used to test the model. Predictor standardization and parameter estimation were done in the training set only to avoid information leakage. The penalty parameter, λ, as well as other tuning parameters, were chosen with 10-fold cross-validation. The final models were then tested with an independent test set.

Root mean squared error (RMSE), mean absolute error (MAE), and R2. were used to evaluate the predictive performance. Sensitivity, specificity, precision, recall, false discovery rate, and F1-score were used to assess the variable-selection

performance. The number of predictors selected was also noted and assessed for sparsity. In experimental tests with a high number of variables, the error in coefficient estimation was quantified by the average squared difference between the estimated and true regression coefficients.

2.11 Robustness and Sensitivity Analysis

The results of modeling were compared with various sample sizes, dimensional settings, contamination mechanisms and contamination percentage. The robustness was assessed by measuring the degradation in prediction and variable-selection performance with contaminating levels. Further sensitivity analyses were performed with other random partitions and tuning parameters to confirm that the analysis results were not sensitive to any particular analytical configuration.

2.12 Ethical Considerations

A secondary numerical dataset with physicochemical data for superconducting materials was used for the study. No human, personal, clinical or animal data were included in the dataset. Thus, informed consent and institutional ethical approval were not necessary. The dataset was used solely for academic and methodological research, and its source must be properly cited in the final text of the paper.

3. Results

3.1 Dataset Characteristics and Preliminary Assessment

There were 21,263 observations in the final dataset for the superconductivity problem, with 81 numerical predictor variables and the continuous response variable critical_temp. Various statistical parameters of atomic mass, Ionization energy, atomic radius, density, electron affinity, fusion heat, thermal conductivity and valence were used as the predictors. The critical temperature was found to be between 0.00021 and 185.00 with a mean of 34.42 and a standard deviation of 34.25. There were no missing values. In the preliminary screening, sixty-six identical records were found.

There was a significant interdependence among predictors as indicated in Table 1. Seventy-four of the pairs of predictors had an absolute correlation greater than 0.90, and 30 pairs had an absolute correlation greater than 0.95. The high amount of predictor redundancy was an appropriate empirical context to test the performance of penalized variable selection methods.

Table 1. General characteristics of the Superconductivity Data Set

Characteristic	**Result**
Number of observations	21,263
Number of candidate predictors	81
Response variable	Critical temperature
Mean critical temperature	34.42
Standard deviation	34.25
Minimum critical temperature	0.00021
Median critical temperature	20.00
Maximum critical temperature	185.00
Missing observations	0
Exact duplicate records	66
Predictor pairs with absolute correlation > 0.90	74
Predictor pairs with absolute correlation > 0.95	30

A correlation analysis also revealed a strong correlation between the critical temperature and several thermal-conductivity, atomic-radius and valence-related variables. The best individual correlation was for the weighted standard deviation of thermal conductivity, r=0.721. The comparatively strong correlation of 0.688 was obtained for the range of thermal conductivity and 0.654 for range of atomic radius. The 10 predictors that have the highest absolute correlations to critical temperature are shown in Figure 1.

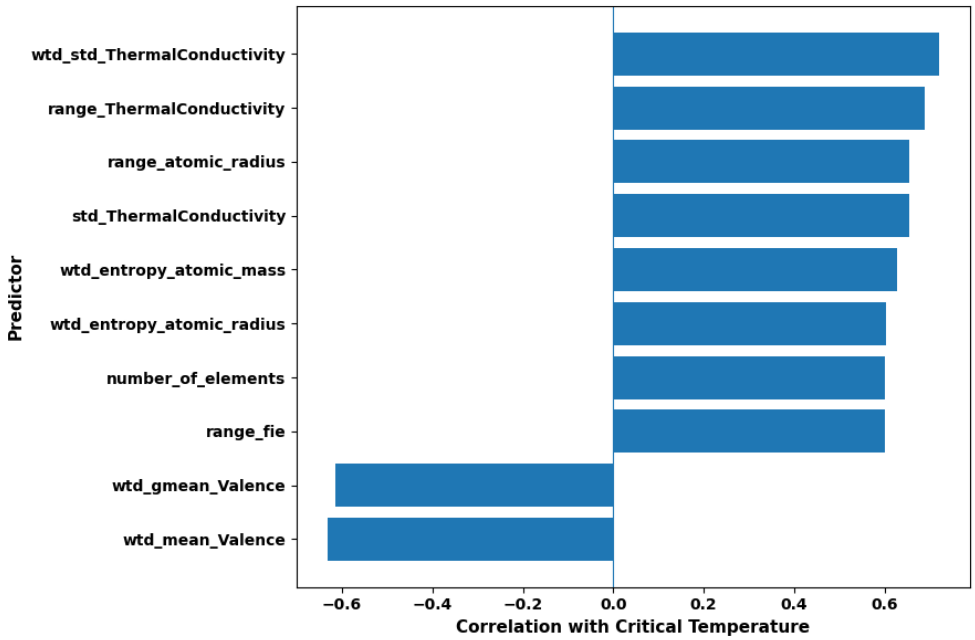

Figure 1. Ten predictors showing the strongest absolute correlations with superconducting critical temperature

3.2 Model Performance Under Uncontaminated Data

The performance at 0% contamination was used as the baseline for evaluating subsequent robustness. Table 2 shows the test-set results for the five regression methods. The model with the lowest baseline RMSE was the Ridge regression model with an RMSE of 17.45 and the highest R2 of 0.736. The robust adaptive LASSO also proved to be competitive with an RMSE of 18.08, MAE of 13.59, an R2 of 0.716 and only 36 predictors.

LASSO obtained an R2 value of 0.693, and Elastic Net obtained an R2 value of 0.702. Under clean data, Adaptive LASSO yielded the sparsest solution, keeping only 16 variables and the largest RMSE (19.86).

Table 2. Predictive performance under the uncontaminated condition

Method	RMSE	MAE	R^2	Predictors selected
Ridge	17.447	13.246	0.736	81
LASSO	18.796	14.557	0.693	25
Elastic Net	18.516	14.224	0.702	36
Adaptive LASSO	19.864	15.561	0.657	16
Robust Adaptive LASSO	18.083	13.588	0.716	36

The baseline results show the typical prediction–sparsity trade-off. Ridge regression had the best pure predictive performance and was able to use all available predictors. The penalized (regularized) selection procedures yielded far more sparse models, however. The baseline RMSE values in Figure 2 are shown and it is noted that robust adaptive LASSO was comparable to Ridge and provided significant dimensional reduction.

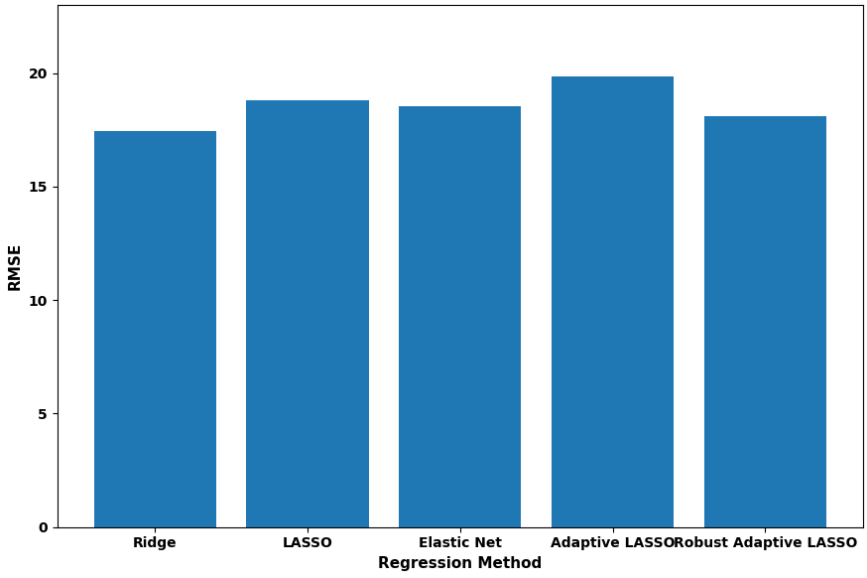

Figure 2. Baseline test-set RMSE of the competing penalized regression methods under 0% contamination

3.3 Effect of Increasing Data Contamination

The higher the contamination, the worse the prediction was for all regression techniques. However, the amount of deterioration was quite variable among the estimators. Ridge regression RMSE increased from 17.45 when clean data was used to 20.80 when 20% contamination was used, as seen in Table 3. LASSO increased from 18.80 to 20.96, while Elastic Net increased from 18.52 to 20.88.

The comparatively small deterioration at higher levels of contamination was seen in the robust adaptive LASSO. Its RMSE increased from 18.08 under uncontaminated data to 19.16 at 5%, 19.23 at 10%, 19.60 at 15%, and 19.94 at 20% contamination. It performed best among the variable-selection estimators at 10%, 15% and 20% contamination levels.

Table 3. RMSE across increasing levels of combined contamination

Method	**0%**	**5%**	**10%**	**15%**	**20%**
Ridge	17.447	19.114	19.513	19.804	20.796
LASSO	18.796	19.530	19.789	20.110	20.955
Elastic Net	18.516	19.314	19.637	19.963	20.881
Adaptive LASSO	19.864	21.349	21.308	22.484	23.627
Robust Adaptive LASSO	18.083	19.156	19.234	19.597	19.939

This was also true for explained variance. As indicated in Figure 3, R2 decreased with increasing contamination, with the adaptive model being quite stable in providing predictions at the highest contamination levels. Under 20% contamination, robust adaptive LASSO maintained an R2 value of 0.655, while conventional LASSO, Elastic Net and Adaptive LASSO had R2 values of 0.619, 0.621 and 0.515, respectively.

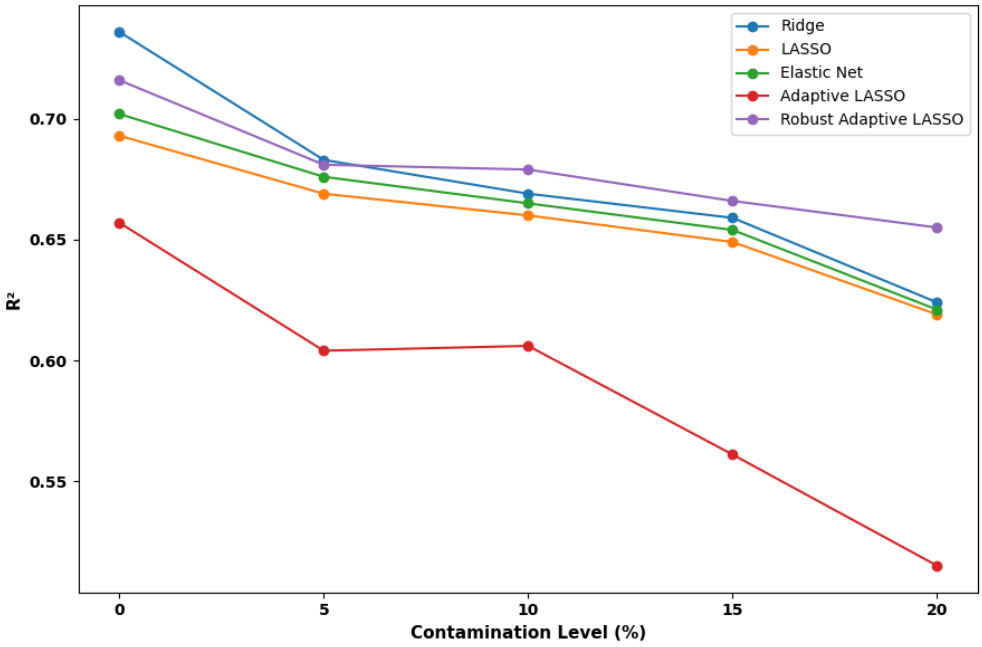

Figure 3. Change in test set R2across contamination levels for the competing regression methods

3.4 Robustness to Severe Contamination

The proportional decrease from 0% to 20% contamination was also computed as a measure of robustness. Table 4 shows that robust adaptive LASSO achieved the minimum RMSE, and the RMSE increased by about 10.27%. The other methods were: Conventional LASSO (11.48% increase), Elastic Net (12.77% increase), Adaptive LASSO (18.94% increase), and Ridge (19.20% increase). The adaptive estimator also had the smallest proportional decrease in R2 (around 8.56%), demonstrating the advantage of having the adaptive penalization and robust weighting of the residuals combined in reducing sensitivity to heavily contaminated observations.

Table 4. Performance deterioration from 0% to 20% contamination

Method	**RMSE increase (%)**	**MAE increase (%)**	**R^2decline (%)**
Ridge	19.20	24.52	15.13
LASSO	11.48	14.28	10.76
Elastic Net	12.77	16.49	11.53
Adaptive LASSO	18.94	22.07	21.63
Robust Adaptive LASSO	10.27	14.41	8.56

The relative stability of RMSE under contamination conditions is also shown in Figure 4. This difference between the robust adaptive estimator and the conventional approaches was more pronounced as the level of contamination increased, especially above 10% contamination.

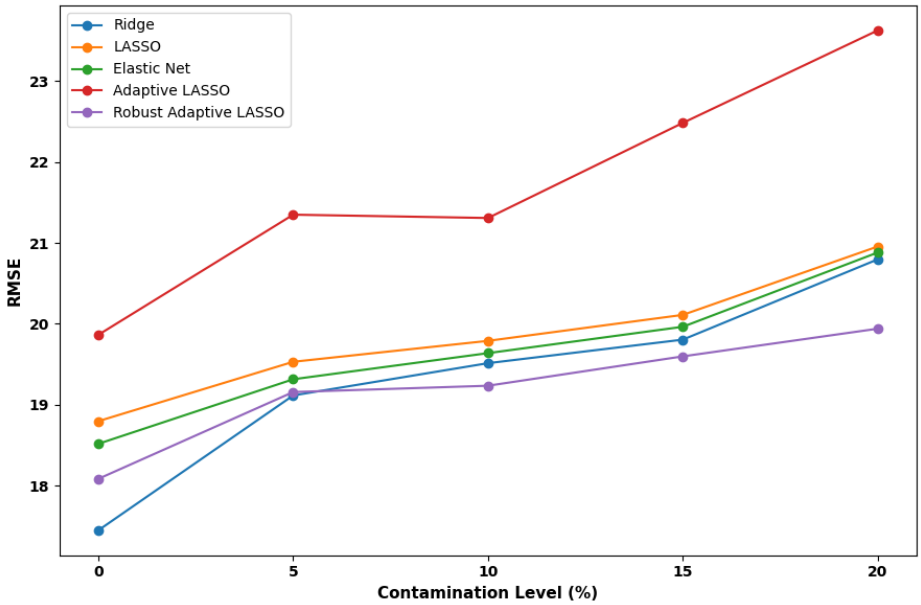

Figure 4. Test-set RMSE trajectories across 0%, 5%, 10%, 15%, and 20% contamination levels

3.5 Variable Selection and Model Sparsity

There were significant model differences in the number of predictors selected. For the 81 predictors, ridge regression did not remove any of the predictors for any contamination level. In the clean condition, 25 predictors were selected by Lasso, whereas the number jumped to 66 when the condition was 20% contaminated. Likewise, Elastic Net went from 36 predictors to 74.

By contrast, strong adaptive LASSO became increasingly selective with higher percentages of contamination, selecting 36 predictors for the clean condition, 25 for 5% contamination, 18 for 10% contamination, 16 for 15% contamination, and 14 for 20% contamination. This behavior suggests that the powerful adaptive penalty was able to stop the unstable predictors generated by contaminated observations from being included. Table 5 shows the number of predictors retained across contamination levels.

Table 5. Number of predictors retained across contamination levels

Method	**0%**	**5%**	**10%**	**15%**	**20%**
Ridge	81	81	81	81	81
LASSO	25	58	60	65	66
Elastic Net	36	68	71	75	74
Adaptive LASSO	16	10	10	10	11
Robust Adaptive LASSO	36	25	18	16	14

Conventional LASSO and Elastic Net have a higher number of predictors that are kept when there is contamination, indicating the presence of anomalous observations that add additional unstable signals to the regression structure. In contrast, powerful adaptive penalties kept the sparsity well preserved. The number of selected predictors is summarized graphically in Figure 5 by condition of contamination.

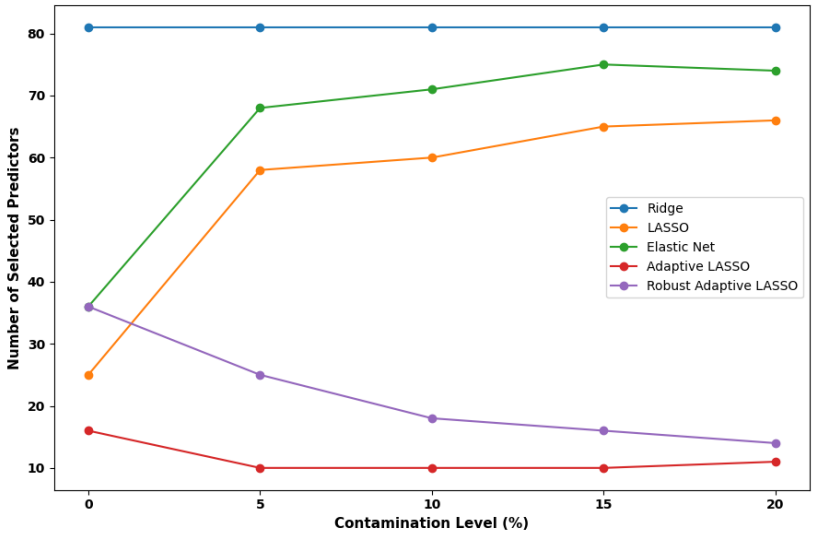

Figure 5. Number of predictors selected by each regression method across increasing contamination levels

3.6 Predictors Identified by Robust Adaptive Penalization

The inspection of the clean-data robust adaptive LASSO coefficients showed that the atomic radius, thermal conductivity, electron affinity, ionization energy, valence, and atomic mass features played a major role in the model. The most significant standardized coefficient was seen for wtd_mean_atomic_radius, followed by wtd_mean_ThermalConductivity and std_ElectronAffinity.

The highest positive and negative coefficients are presented in Table 6. The relative predictive importance of these coefficients should be seen more as an indication of which coefficients to trust in the standardized penalized model than as individual causal effects, since many of the material properties are strongly correlated.

Table 6. Leading predictors retained by robust adaptive LASSO under uncontaminated data

Predictor	Standardized coefficient	Direction
wtd_mean_atomic_radius	0.600	Positive
wtd_mean_ThermalConductivity	0.578	Positive
std_ElectronAffinity	0.564	Positive
wtd_entropy_atomic_radius	0.554	Positive
range_fie	0.530	Positive
wtd_gmean_atomic_radius	-0.510	Negative
range_ElectronAffinity	-0.486	Negative
wtd_gmean_ElectronAffinity	-0.472	Negative
std_fie	-0.443	Negative
wtd_gmean_ThermalConductivity	-0.440	Negative
wtd_entropy_Valence	-0.349	Negative
wtd_mean_atomic_mass	-0.345	Negative

The results suggest conventional penalized regression was reasonable in the absence of contamination and that it was more sensitive to the data contamination. As contamination levels increased, robust adaptive LASSO achieved the best compromise between stability and sparsity. The RMSE did not change significantly from clean to 20% contamination conditions (10.27%), and the R2 only decreased by 8.56%. The results are consistent with the idea of using strong adaptive penalization for variable selection when the regression structure at high dimensions is influenced by influential or contaminated observations.

4. Discussion

In this study, we compared the performances of the conventional penalized regression, adaptive penalization and robust adaptive penalization in progressively contaminated high-dimensional settings. The results showed that there was a clear trade-off between the accuracy of prediction, the sparsity and the resistance to contamination. In uncontaminated data, Ridge regression performed best in predicting, while the robust adaptive LASSO was a close runner-up and resulted in a significantly sparser model. As the contamination level rose, however, the importance of the strong adaptive strategy became more apparent. The results showed that for 20% contamination, the robust adaptive LASSO had the smallest RMSE increase and the smallest R2 loss, suggesting that using an adaptive shrinkage and robust loss function can better withstand the effect of outlying observations than the regular LASSO.

This result mirrors the wider results on robust sparse estimation that have been reported repeatedly in the literature, which demonstrate that squared-error penalization is very sensitive to the atypical observations. When dimension is high, robust variable-selection approaches are developed to minimize the impact of contaminated cases while maintaining sparse support recovery (Chatla & Mandal, 2026; Wang & Karunamuni, 2022). The stability of the robust adaptive estimator under 10%, 15% and 20% contamination observed does support the view that robustness should also be built into the objective function and not just be a post-estimation diagnostic.

One of the results is especially consistent with the rationale of the use of strong Adaptive LASSO methods. As contamination increased, the number of variables selected by conventional LASSO and Elastic Net models increased indicating that the selected variables were spurious predictors added to the models as contamination levels rose. In contrast, the powerful adaptive LASSO was less selective with fewer contaminants, finally including only 14 predictors at the most contaminated level. The pattern is similar to that of the desired stability of the estimated coefficients in a robust adaptive penalization by both instability suppressing weights and resistant loss function, in which the predictor weights and loss function are both adaptive to the presence of contaminated observations (Machkour et al., 2020). These results are consistent with those of adaptive robust estimators where Elastic Net-type penalties are added to resistant regression procedures such as robust regression, especially in scenarios where the predictors are correlated and there are influential observations (Kepplinger, 2023).

The positive outcomes for the Huber framework also align with the findings that adaptive Huber regression is statistically efficient when observations are typically normal but can suppress the effect of large residuals (Fan et al., 2022). The importance of this property is that unconditional rejection of abnormal observations can result in the loss of valuable data, while a bounded-loss strategy allows abnormal observations to continue to influence the results, but to a lesser extent. The

smaller degradation found for the robust adaptive model is thus within a realistic range of statistical possibilities under the contamination scenario employed in the present study.

The results also confirm the general rule that good variable selection is particularly crucial in the presence of contamination and high dimensionality. Multivariate predictor structures may allow cellwise contamination to propagate and cause selection decisions to be skewed even if a relatively small proportion of the values is contaminated (Su et al., 2024). Similarly, strong theory of features suggests that the problem of sparse recovery is significantly harder under contaminated samples when coefficient estimation and support identification are affected by the contamination (Maurya et al., 2026). The present results confirm this concern as demonstrated by the fact that the conventional selection methods kept many more variables as the level of contamination increased, while the strong adaptive estimator kept a relatively modest structure.

The interpretation of heavy-tailed disturbances is another possible interpretation. Although sparse procedures are developed based on heavy-tailed errors, they tend to perform better than least-squares-based penalties when the distribution of the residuals is far from the Gaussian assumptions (Ye et al., 2022; Mai, 2026). This evidence is reflected in the current results, with the robust model being significantly less affected by increasing levels of contamination. The related advances in robust Bayesian and quantile-based variable selection also suggest that the objective function for squared errors can be replaced with one that is more resistant to non-normal error distributions to enhance feature recovery (Nan et al., 2025).

In high dimensional settings, model evaluation is also crucial as a confounding factor since the ability to predict can mask an unstable variable selection. There have therefore been proposed robust criteria for evaluating the behaviour of models when contaminated and when more complex, in terms of the dimensionality of the problem under consideration (Kurata & Hirose, 2025). In the current study, the combined information from RMSE, MAE, R2 and model sparsity was found to be more informative than any of the individual measures. However, the prediction error rose more rapidly for severe contamination in Adaptive LASSO, which produced very sparse solutions, meaning that too much sparsification can lead to a decrease in generalizability if there is not enough robustness.

The computational results also show the significance of optimization and tuning. Initial estimates, adaptive weights, and penalty selection are important factors to consider for adaptive LASSO performance, and in high-dimensional applications, coordinate-based methods are sometimes necessary (Nasir et al., 2025). Strong selection criteria can also reduce bias toward sparsity and improve predictive power by ensuring that the observations which are contaminated by noise do not influence the selection process (Mandal & Ghosh, 2025).

The overall results suggest that the regularization method works well when the contamination is relatively small, but it can be unstable with high contamination levels. The performance criteria of prediction, sparsity and contamination resistance were in favor of robust adaptive penalization, justifying its application in high-dimensional situations where outliers, leverage points or heavy-tailed errors may be present.

5. Conclusion

This work focused on the effectiveness of powerful variable selection in high-dimensional regression under progressively contaminated data (PCD) settings. The comparison showed that under uncontaminated conditions, conventional penalized regression methods work well, but became more sensitive as the contamination level increased. Ridge regression gave the highest baseline predictive performance but kept all of the predictors and thus had no sparse variable selection. LASSO and Elastic Net had larger sparsities, but as the contamination increased, so did the number of predictors selected, suggesting less stability. The overall performance of the robust adaptive LASSO was the best when evaluated for predictive accuracy, sparsity and resistance to contamination. The RMSE rose by merely 10.27% from the uncontaminated to the 20% contamination level; the decrease in R2 remained relatively small. The method had good stability with unstable or spurious variables, retaining just 14 predictors at the highest level of contamination. The results suggest that using strong loss functions and adaptation of predictor-specific penalty terms can significantly enhance the ability to select variables reliably in high-dimensional data with outliers, leverage points, or heavy-tailed disturbances. In general, adaptive penalization is a promising approach for obtaining parsimonious, stable and accurate regression models in high-dimensional, complex contaminated settings.

References:

[1]. Amato, U., Antoniadis, A., De Feis, I., & Gijbels, I. (2021). Penalised robust estimators for sparse and high-dimensional linear models: U. Amato et al. Statistical Methods & Applications, 30(1), 1-48.

[2]. Bai, Y., Tian, M., Tang, M. L., & Lee, W. Y. (2021). Variable selection for ultra-high dimensional quantile regression with missing data and measurement error. Statistical Methods in Medical Research, 30(1), 129-150.

[3]. Biziaev, T., Kopciuk, K., & Chekouo, T. (2025). Using prior-data conflict to tune Bayesian regularized regression models. Statistics and Computing, 35(2), 53.

[4]. Cao, X., Gregory, K., & Wang, D. (2023). Inference for sparse linear regression based on the leave-one-covariate-out solution path. Communications in Statistics-Theory and Methods, 52(18), 6640-6657.

[5]. Chatla, S. B., & Mandal, A. (2026). Robust variable selection in high-dimensional nonparametric additive model: SB Chatla, A. Mandal. Annals of the Institute of Statistical Mathematics, 78(1), 115-140.

[6]. Fan, J., Guo, Y., & Jiang, B. (2022). Adaptive Huber regression on Markov-dependent data. Stochastic processes and their applications, 150, 802-818.

[7]. Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. Wiley Interdisciplinary Reviews: Computational Statistics, 13(4), e1524.

[8]. Kepplinger, D. (2023). Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression. Computational Statistics & Data Analysis, 183, 107730.
[9]. Kim, S., Turkoz, M., Jeong, M. K., & Elsayed, E. A. (2024). Monitoring of group-structured high-dimensional processes via sparse group LASSO. Annals of Operations Research, 340(2), 891-911.
[10]. Kurata, S., & Hirose, K. (2025). Robust and consistent model evaluation criteria in high-dimensional regression. Journal of Statistical Planning and Inference, 106358.
[11]. Machkour, J., Muma, M., Alt, B., & Zoubir, A. M. (2020). A robust adaptive Lasso estimator for the independent contamination model. Signal processing, 174, 107608.
[12]. Mai, T. T. (2026). Heavy Lasso: sparse penalized regression under heavy-tailed noise via data-augmented soft-thresholding. Statistics and Computing, 36(1), 29.
[13]. Mandal, A., & Ghosh, S. (2025). Robust variable selection criteria for the penalized regression. Journal of Multivariate Analysis, 105540.
[14]. Maurya, D., Barik, A., & Honorio, J. (2026, August). Provable guarantees for robust feature selection in sparse linear models in high-dimensions. In Forty-Second Annual Conference on Uncertainty in Artificial Intelligence.
[15]. Nan, R., Wang, J., Li, H., & Luo, Y. (2025). Robust Variable Selection via Bayesian LASSO-Composite Quantile Regression with Empirical Likelihood: A Hybrid Sampling Approach. Mathematics, 13(14), 2287.
[16]. Nasir, M. J. M., Khan, R. N., Nair, G., & Nur, D. (2025). Coordinate gradient descent algorithm in adaptive LASSO for pure ARCH and pure GARCH models. Computational Statistics, 40(7), 3527-3561.
[17]. Stokell, B. G., & Shah, R. D. (2022). High-dimensional regression with potential prior information on variable importance. Statistics and Computing, 32(3), 52.
[18]. Su, P., Tarr, G., & Muller, S. (2024). Robust variable selection under cellwise contamination. Journal of Statistical Computation and Simulation, 94(6), 1371-1387.
[19]. Sun, Q., Zhou, W. X., & Fan, J. (2020). Adaptive huber regression. Journal of the American Statistical Association, 115(529), 254-265.
[20]. Takahashi, A., & Nomura, S. (2024). Efficient path algorithms for clustered Lasso and OSCAR. Japanese Journal of Statistics and Data Science, 7(2), 967-998.
[21]. Tunguz, B. (2021). Superconductivty data data set [Data set]. Kaggle. https://www.kaggle.com/datasets/tunguz/superconductivty-data-data-set
[22]. Wang, F., Mukherjee, S., Richardson, S., & Hill, S. M. (2020). High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking: Wang et al. Statistics and computing, 30(3), 697-719.
[23]. Wang, P., Lu, J., Weng, J., & Mitra, S. (2025). Conditional sufficient variable selection with prior information: P. Wang et al. Computational Statistics, 40(5), 2519-2551.
[24]. Wang, X., Kong, L., Zhuang, X., & Wang, L. (2024). Variance estimation in high-dimensional linear regression via adaptive elastic-net. Journal of Industrial and Management Optimization, 20(2), 630-646.
[25]. Wang, Y., & Karunamuni, R. J. (2022). High-dimensional robust regression with Lq-loss functions. Computational Statistics & Data Analysis, 176, 107567.
[26]. Wu, Y., & Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. Annual review of statistics and its application, 7(1), 209-226.
[27]. Xi, L. J., Guo, Z. Y., Yang, X. K., & Ping, Z. G. (2023). Application of LASSO and its extended method in variable selection of regression analysis. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], 57(1), 107-111.
[28]. Ye, Y., Shao, Y., & Li, C. (2022). A sparse approach for high-dimensional data with heavy-tailed noise. Economic research-Ekonomska istraživanja, 35(1), 2764-2780.
[29]. Zhu, X., Qin, Y., & Wang, P. (2025). Sparsified simultaneous confidence intervals for high-dimensional linear models: X. Zhu et al. Metrika, 88(5), 709-733.
[30]. Żogała-Siudem, B., & Jaroszewicz, S. (2024). Variable screening for Lasso based on multidimensional indexing: B. Żogała-Siudem, S. Jaroszewicz. Data Mining and Knowledge Discovery, 38(1), 49-78.