Dhaka Univ. J. Sci. 73(2): 143-150, 2025 (July)

ISSN 1022 - 2502, E-ISSN 2408-8528
DOI: https://doi.org/10.3329.dujs.v73i2.82773

Robust Variable Selection in High-Dimensional Data: Mitigating Cellwise Contamination Through Comparative Analysis

Nadia Mehjabeen Oyshi, Md. Tuhin Rana, and Md. Jamil Hasan Karami*

Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh

(Received: 13 December 2024; Accepted: 19 May 2025)

Abstract: The proliferation of high-dimensional data has heightened challenges posed by cellwise outliers, where contamination in individual cells distorts analyses more pervasively than traditional rowwise outliers. This study conducts a comprehensive comparison of robust variable selection methods under cellwise contamination, evaluating four rank-based techniques (ALGR, ALRP, LGR, LRP) against traditional approaches (Lasso, Adaptive Lasso, sLTS). Simulations under varying correlation structures, contamination rates (2%, 5%, 10%), and outlier magnitudes (γ = 2, 6, 10) demonstrate that Gaussian Rank correlation-based methods (ALGR, LGR) achieve superior F1 scores, balancing high true positives and low false positives. Real-data applications on life expectancy and crime datasets corroborate these findings, with ALGR and LGR maintaining robustness in low- and high-dimensional settings. Results emphasize the critical need for methods resilient to cellwise contamination in fields reliant on accurate high-dimensional data analysis, such as healthcare and genomics.

Keywords: Cellwise contamination, Robust variable selection, Gaussian Rank correlation, High-dimensional regression, independent contamination model, Sparse robust regression.

I. Introduction

The rise of Big Data brings challenges like contamination from outliers—data points that skew traditional analyses. These outliers distort statistical models, risking severe consequences: flawed financial risk assessments, compromised healthcare diagnoses, and misleading genomic patterns. Neglecting them leads to inaccurate predictions, misinformed decisions, and hindered progress. Robust solutions include outlier detection algorithms to identify anomalies and statistical methods resistant to contamination. Addressing outliers ensures reliable models, accurate insights, and better outcomes in fields like finance, healthcare, and bioinformatics. Proactive data cleaning and robust modeling are vital for trustworthy analyses and informed decision-making in the era of Big Data. It is widely acknowledged that raw datasets often include about 1% to 10% outliers, as noted by Hampel[1]. Typically, when we refer to "outliers," we are discussing rowwise outliers. These outliers are characterized by entire rows in a dataset being flagged as anomalous, as depicted in the left panel of Figure 1. The foundational principles of classical robust statistics are rooted in the concept of rowwise contamination, where each observation is considered either entirely uncontaminated or wholly outlying. By focusing only on the uncontaminated observations, such as by down-weighting rows identified as outliers, robust estimates can be derived that withstand the influence of these anomalies.

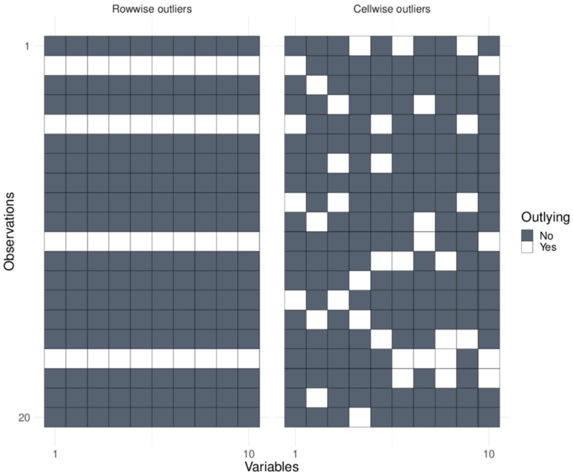

Fig. 1. Rowwise and cellwise outliers: The outlying cells are rendered in white, and the uncontaminated cells are shown in gray. For both panels, 20% of the cells are contaminated. However, the left panel has 4 out of 20 rows outlying while the right panel has 18 rows outlying.

The Tukey-Huber contamination model (THCM), proposed by Tukey and Huber[2,3], is frequently used to model rowwise outliers. In this model, we observe n independent observations $\mathbf{x}_i \in \mathbb{R}^p$ for $i = 1, \dots, n$, which may include outliers. This model can be expressed as:

$$\mathbf{x}_i = (1 - b_i)\mathbf{x}_i^{\text{clean}} + b_i\mathbf{x}_i^{\text{contam}},$$

where b_i follows a Bernoulli distribution with contamination probability e, $\mathbf{x}_i^{\text{clean}} \in \mathbb{R}^p$ represents a clean

[*] Author for correspondence. e-mail: karami.stat@du.ac.bd

observation following a multivariate distribution F, and $\mathbf{x}_i^{\text{contam}} \in \mathbb{R}^p$ represents a contaminated observation following a multivariate distribution H. The indicator variable b_i is independent of both $\mathbf{x}_i^{\text{clean}}$ and $\mathbf{x}_i^{\text{contam}}$.

However, this traditional paradigm of rowwise outliers is increasingly viewed as overly restrictive[4]. Recent research has shifted towards the concept of cellwise outliers, which focuses on individual cells within an observation. Unlike rowwise outliers, where an entire row is contaminated, cellwise outliers may affect only a small subset of cells within a row or column.

This distinction is significant because it provides a more nuanced understanding of outliers and their impact on the dataset. As shown in the right panel of Figure 1, cellwise outliers illustrate that for a given observation, only specific cells may be contaminated, while the rest remain clean and informative.

Cellwise outliers can be modeled using the independent contamination model (ICM) as described by Alqallaf[5]. In this model, we observe independent observations of a random vector $\mathbf{x}_i$ given by:

$$\mathbf{x}_i = \left(\mathbf{I} - \text{Diag}(\mathbf{b}_i)\right)\mathbf{x}_i^{\text{clean}} + \text{Diag}(\mathbf{b}_i)\mathbf{x}_i^{\text{contam}},$$

where $\mathbf{b}_i = \left(b_{i1}, \ldots, b_{ip}\right)^\top$ and $b_{i1}, \ldots, b_{ip}$ are independently drawn from a Bernoulli distribution.

Cellwise outliers pose a significant challenge compared to the traditional approach of identifying and down-weighting outlying rows. Even a small percentage of contaminated cells can affect many rows, dramatically altering the overall dataset. Moreover, down weighting such observations may lead to the loss of valuable information in the uncontaminated cells.

To better comprehend this phenomenon, Alqallaf[5] describe the propagation of cellwise outliers under the independent contamination model. For a contamination rate e of cells, the expected proportion of contaminated observation rows is $1 - (1 - e)^p$. This proportion rapidly exceeds 50% as the dimension p increases, highlighting the significant impact of cellwise contamination on high-dimensional data.

Traditional methods often focus on detecting outliers at the row level, identifying entire observations that deviate from expected patterns. The Detect Deviating Cells (DDC) technique[6] identifies cellwise outliers in multivariate data through robust standardization of variables, univariate outlier flagging using z-score thresholds, and analysis of bivariate relationships to compute correlations and regression slopes. Predicted cell values are derived from connected variables, adjusted via de-shrinkage to avoid underestimation, and compared to observed values through residual analysis to flag outliers. Rows with excessive flagged cells are marked as rowwise outliers, and detected anomalies are replaced with imputed predictions, ensuring data integrity. Traditional robust regression estimators, such as the M-estimator[3] and the MM-estimator[7], are designed to mitigate the impact of these outliers by down-weighting the observations that deviate substantially from the model.

the M-estimator of β is obtained by solving:

$$\hat{\beta}_M = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^\top \beta}{\hat{\sigma}_\epsilon}\right),$$

where $\rho(\cdot)$ is a loss function that grows slower than the quadratic function used in ordinary least squares (OLS), making the estimator less sensitive to large residuals. A common choice for ρ is Huber's loss function.

The Least Absolute Shrinkage and Selection Operator (LASSO), a regression method proposed by Tibshirani[8], is designed to improve the predictive power and interpretability of statistical models by concurrently conducting variable selection and regularization. This constraint leads to the shrinkage of some coefficients to exactly zero, effectively eliminating the corresponding variables from the model and thereby achieving variable selection. The Lasso estimator is defined as the solution to the following optimization problem:

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2n} \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1\right),$$

where $\parallel \beta \parallel_1 = \sum_{j=1}^{p} \left|\beta_j\right|$ is the $L1$ norm of β, and $\lambda \geq 0$ is a tuning parameter that controls the amount of regularization. The term $\lambda \parallel \beta \parallel_1$ acts as a penalty, shrinking the coefficients towards zero. When λ is large, more coefficients are set to zero, resulting in a simpler model with fewer predictors.

The Adaptive LASSO, introduced by Zou[9], is an enhancement of the standard LASSO method designed to improve variable selection consistency and predictive performance. The Adaptive LASSO modifies the penalty applied to the coefficients by incorporating adaptive weights, thereby addressing some of the limitations of the traditional LASSO, particularly its tendency to select only one predictor from a group of highly correlated predictors.

The Adaptive Lasso estimator is defined by modifying Lasso's objective function to include adaptive weights ω_j for each coefficient β_j:

$$\hat{\beta}_{ALasso} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2n} \parallel y - X\beta \parallel_2^2 + \lambda \sum_{j=1}^{p} \omega_j \left|\beta_j\right|\right),$$

where the weights $\omega_j = \frac{1}{\left|\hat{\beta}_j^{(init)}\right|^\gamma}$ are derived from initial estimates $\hat{\beta}_j^{(init)}$ of the coefficients, and $\gamma > 0$ is a tuning parameter. These weights allow the penalty to adapt based on the initial estimates, reducing the bias for larger coefficients and improving the selection of relevant variables. The initial estimates $\hat{\beta}_j^{(init)}$ can be obtained using various methods, such as the ordinary least squares (OLS) if $p < n$, the Ridge regression, or even the standard Lasso.

The Sparse Least Trimmed Squares (sLTS) estimator, introduced by Alfons[10], is a robust regression technique designed to handle high-dimensional data with outliers. sLTS combines the principles of robust regression with sparsity-inducing penalties to provide reliable parameter estimates in the presence of contamination, while simultaneously performing variable selection. The sLTS estimator aims to minimize the sum of the smallest squared residuals while imposing an $L1$ penalty to induce sparsity:

$$\hat{\beta}_{sLTS} = \underset{\beta}{\operatorname{argmin}} \sum_{i \in H} (y_i - x_i^\top \beta)^2 + \lambda \parallel \beta \parallel_1,$$

where H is a subset of indices corresponding to the h smallest residuals, and λ is a tuning parameter that controls the amount of regularization.

We aim to conduct a comprehensive comparative analysis of robust variable selection techniques in the presence of cellwise contaminated data, with a particular focus on contrasting rank-based and pairwise methods against traditional approaches. The main objective is to assess the effectiveness of the two-rank based and pairwise methods Adaptive Lasso estimator based on Gaussian rank correlation (ALGR) and Adaptive Lasso estimator based on pairwise correlation (ALRP)[11], we compare its performance against several established methods: Lasso[8] adaptive Lasso (ALasso)[9] and sparse least trimmed squares (sLTS)[10].

II. Methodology

A robust estimator is the Gaussian Rank (GR) correlation, introduced by Boudt[12].

For a data matrix $Z = (y, X)$, given a pair of observation vectors for variables z_j and z_k, where $z_j, z_k \in \mathbb{R}^n$ and $1 \leq j, k \leq (p+1)$, an efficient and robust correlation estimator is the Gaussian rank (GR) correlation. The GR correlation is defined as the sample correlation estimated from the normal scores of the data. For z_{ij}, the i-th observation of the j-th variable, we compute:

$$\tilde{z}_{ij} = \Phi^{-1}\left(\frac{\operatorname{Rank}(z_{ij})}{n+1}\right).$$

After constructing a pseudo dataset $\tilde{Z} = (\tilde{z}_{ij})_{n \times (p+1)}$, we derive the GR correlation matrix by calculating the Pearson correlation matrix of $\tilde{Z}$. This estimated correlation matrix is necessarily positive semi-definite, even in high-dimensional settings, ensuring that the subsequent regression optimization process remains convex.

Once robust estimates of the scale parameters are obtained using Qn estimators[13] and the correlation matrix $\hat{R}$, the robust empirical covariance matrix $\hat{\Sigma}$ is computed as:

$$\hat{\Sigma} = \hat{S}\hat{R}\hat{S},$$

where $\hat{S} = \operatorname{Diag}\left(\hat{\sigma}_{z_1}, \ldots, \hat{\sigma}_{z_{p+1}}\right)$ is a diagonal matrix consisting of the robustly estimated scale parameters using Qn estimators. Compared to other nonparametric correlation estimators, the GR correlation demonstrates strong robustness, consistency, and significant efficiency[12,14]. Owing to these excellent properties, the GR correlation has been considered a reliable plug-in estimator for covariance matrix estimation[15].

In the context of robust variable selection under cellwise contamination, we also explore various pairwise techniques to estimate correlations. These techniques are known for their robustness but often sacrifice efficiency to some extent. One such method is the Gnanadesikan-Kettenring (GK) pairwise estimator[16], which lean on the identity:

$$\hat{r}_{GK}(x^j, x^k) = \frac{\hat{\sigma}^2(\tilde{x}^j + \tilde{x}^k) - \hat{\sigma}^2(\tilde{x}^j - \tilde{x}^k)}{\hat{\sigma}^2(\tilde{x}^j + \tilde{x}^k) + \hat{\sigma}^2(\tilde{x}^j - \tilde{x}^k)},$$

where $\hat{\sigma}(\cdot)$ is a robust estimator of scale, $\tilde{x}^j = \frac{x^j}{\hat{\sigma}_j}$, and $\hat{\sigma}_j = \hat{\sigma}(x^j)$. This estimator is utilized by Raymaekers and Rousseeuw[4] for quickly obtaining robust correlations. Once the correlation coefficients are estimated, they are assembled into an empirical correlation matrix $\hat{R} = \left(\hat{r}(x^j, x^k)\right)_{p \times p}$, following the approach used by Tarr[17].

In line with the work of Loh and Wainwright[18], the objective loss function for a linear regression model is given by:

$$\hat{\beta}_{LS} = \underset{\beta}{\operatorname{argmin}}\left\{\hat{\Sigma}_{yy} + \beta^\top \hat{\Sigma}_{xx} \beta - 2\beta^\top \hat{\Sigma}_{xy}\right\},$$

where $\hat{\Sigma}_{yy}$ denotes the estimated variance of y, $\hat{\Sigma}_{xy}$ represents the estimated covariance matrix between x and y, and $\hat{\Sigma}_{xx}$ is the estimated covariance matrix of x. These components form the estimated covariance matrix $\hat{\Sigma}$ among the predictors and the response variable:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{yy} & \hat{\Sigma}_{xy}^\top \\ \hat{\Sigma}_{xy} & \hat{\Sigma}_{xx} \end{pmatrix}.$$

The solution to the above optimization problem can be expressed as $\hat{\beta}_{LS} = \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xy}$.

The objective loss function can be rephrased in a more elegant form. Given that $\hat{\Sigma}$ is positive semi-definite, we define $(v, W) = \hat{\Sigma}^{1/2}$ as the square root of $\hat{\Sigma}$, where v is the first column of $\hat{\Sigma}^{1/2}$ and W is a matrix composed of the remaining columns. Considering the relationships $v^\top v =$

$\hat{\Sigma}_{yy}$, $W^{\top}v = \hat{\Sigma}_{xy}$, and $W^{\top}W = \hat{\Sigma}_{xx}$, we rewrite the objective loss as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{\hat{v} - \widehat{W}\beta \|_2^2\},$$

which is a classic quadratic optimization problem.

To ensure effective variable selection, one can integrate the adaptive Lasso penalty, leading to a regularized objective loss function. Given the adaptive Lasso's consistency, utilizing its penalty allows for the formulation of this regularized objective loss function[11].

$$\hat{\beta}_{RLS-\text{ALasso}} = \underset{\beta}{\operatorname{argmin}}\left\{\| \hat{v} - \widehat{W}\beta \|_2^2 + \lambda \sum_{j=1}^{p} \widehat{\omega_j}\, |\beta_j| \right\},$$

where λ is a tuning parameter, $\omega_j = \frac{1}{\tilde{\beta}_j}$, and $\tilde{\beta}_j$ is an initial robust estimate of β_j. While there are many criteria that could be applied to determine the tuning parameter λ, we have selected the BIC because of its ease of implementation and good performance.

III. Simulation Study

Data is generated from a linear regression model with number of observations n = 100, number of variables p = 20 (low-dimensional) and p = 200 (high-dimensional), where predictors follow a correlation structure $\Sigma_{ij} = r^{|i-j|}$ with varying r over $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Regression Coefficients are set as $\beta = (1,2,1,2,1, 0_{15}^{\top})^{\top}$.To introduce outliers, we consider contamination proportions e of 2%, 5%, and 10% for all predictors separately. The outlying cells, x_{ij}^{contam}, are generated randomly from $0.5\mathcal{N}(\gamma, 1) + 0.5\mathcal{N}(-\gamma, 1)$, where $\gamma \in \{2, 6, 10\}$ simulates outliers of varying magnitudes. These represent small, medium and large magnitude of outliers respectively. Each scenario is repeated 200 times. For every iteration, the entire dataset, including the design matrix, is regenerated. The performance of each method in variable selection is evaluated using the true positive rate (TPR), the false positive rate (FPR).

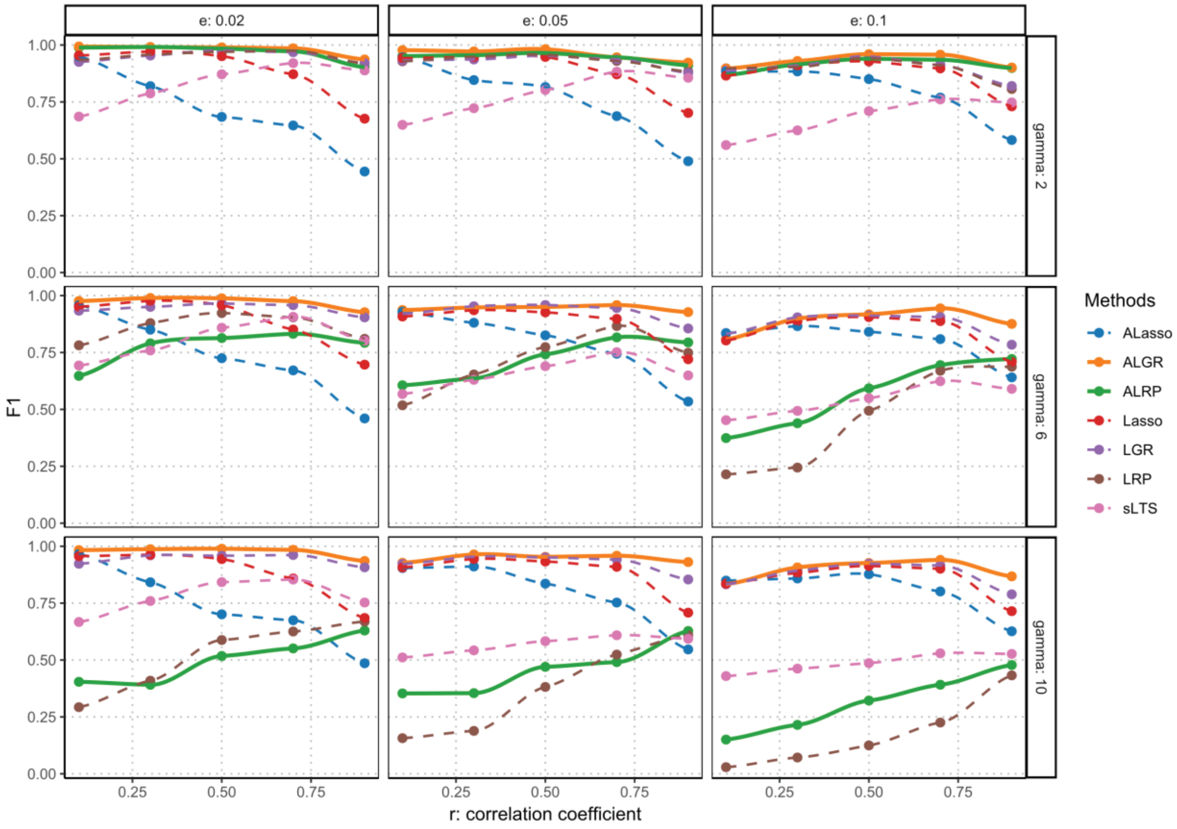

Fig. 2. Selection results as summarized by the F1 score over 200 simulation runs when p = 20.

The F1 Score is calculated as:

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Where TP denotes the true positive number, FP the false positive number and FN the false negative number.

Figure 2 shows the F1 score of low dimensional setting for various robust variable selection methods under different conditions of cellwise contamination**.** X-axis represents the correlation coefficient (r), which indicates the degree of correlation among variables, ranging from 0.1 to 0.9. Y-axis represents the F1 score, ranging from 0 to 1. Columns represent different contamination rates (e = 0.02, 0.05, 0.1). Rows represent different magnitudes of outliers (gamma = 2, 6, 10).

ALGR remains the most reliable method in maintaining high F1 scores across varying contamination rates, outlier magnitudes, and correlation coefficients. sLTS performs well at lower gamma values but is sensitive to higher outlier magnitudes and contamination rates. ALRP is effective in specific scenarios with lower contamination and better handling of higher correlations. ALasso, Lasso, and LRP are less robust, showing higher sensitivity to challenging conditions.

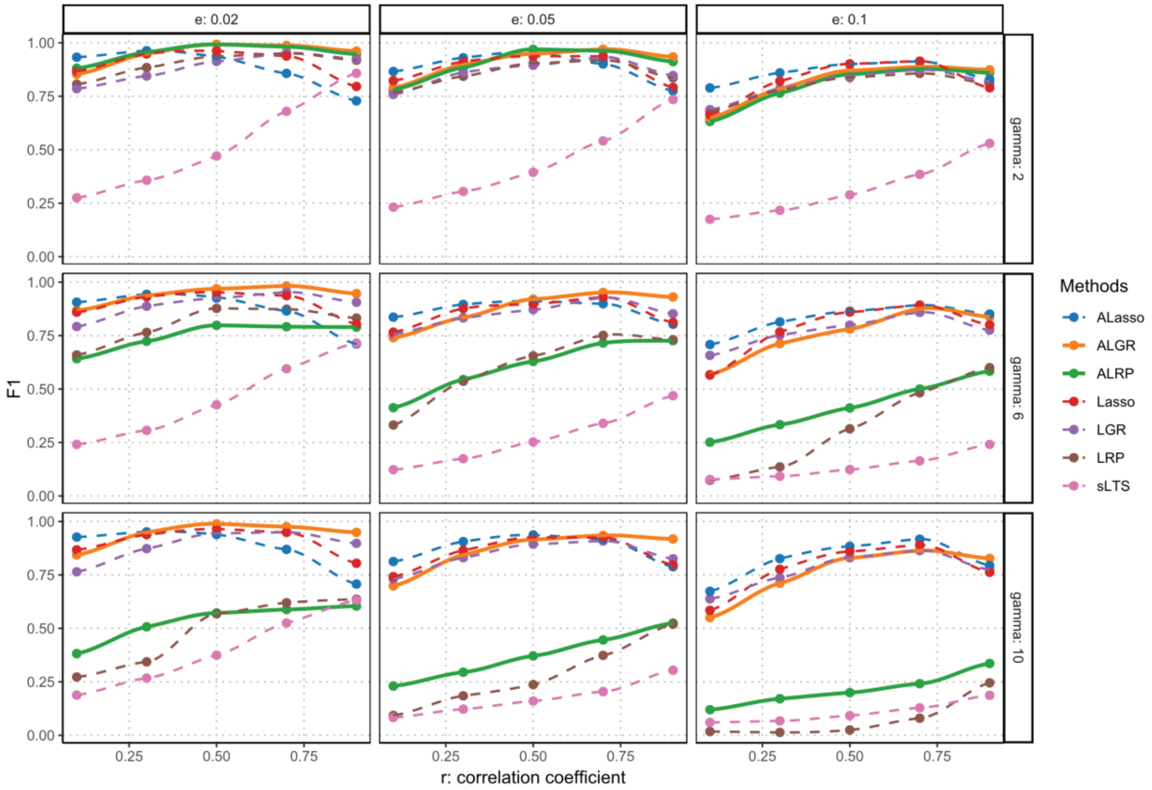

Fig. 3. Selection results as summarized by the F1 score over 200 simulation runs for various contamination rates when p = 200.

Figure 3 shows the F1 score for various robust variable selection methods under different conditions of cellwise contamination in a high-dimensional setting. ALGR and also LGR remain really reliable methods in maintaining high F1 scores across varying contamination rates, outlier magnitudes, and correlation coefficients. ALRP performs well in specific scenarios with lower contamination and better handling of higher correlations. ALasso, Lasso, and LRP are less robust, showing higher sensitivity to challenging conditions. sLTS is less robust, showing a noticeable decline in F1 score with increasing outlier magnitudes and contamination rates, though it performs better with increasing correlation.

IV. Real Data Application

For the low-dimensional scenario, we use a dataset concerning life expectancy available on *https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who*.

Low-Dimensional Setting Using life expectancy Data

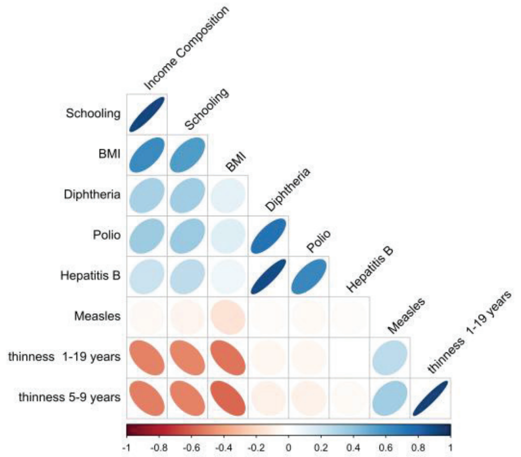

Fig. 4. Correlation map of variables in the life expectancy dataset.

Figure 4 illustrates the correlation between various predictors of life expectancy. Dark blue indicates strong positive correlations, such as between ‘Income Composition of Resources’ and ‘Schooling’, and between ‘BMI’ and ‘Diphtheria’ immunization rates. Measles cases show weak negative correlations with other predictors, indicating that higher immunization rates correspond to fewer Measles cases. Overall, the plot highlights how some predictors are closely related, while others contribute unique information, aiding in robust model building and variable selection for predicting life expectancy.

To rigorously evaluate the robustness of the compared variable selection methods in a real-world context, we intentionally introduce two common challenges: redundant predictors and cellwise contamination. This approach serves several key purposes:

Assessing Specificity and Resilience to Irrelevant Information: In real datasets, it's common to have many potential predictors, not all of which are truly influential. By adding known redundant predictors (variables that, by design, should not be related to the outcome), we test the ability of each method to correctly identify and discard these irrelevant variables, thus minimizing false positives. A truly robust method should maintain its focus on the genuine predictors even in the presence of such noise.

Evaluating Performance Under Data Imperfection: Real-world data is rarely perfect and often suffers from various forms of errors or outliers. Introducing cellwise contamination simulates this scenario, allowing us to observe how each method's variable selection stability and parameter estimation are affected by corrupted individual data points. This directly tests the core claim of robustness. The ability to provide reliable results despite deviations from ideal data conditions.

By systematically manipulating the real-data environment in these ways, we can more comprehensively assess how well each method performs under conditions that mimic the complexities and imperfections often encountered in practical data analysis. This provides a more nuanced understanding of their true robustness beyond what can be observed from the original dataset alone, thereby strengthening the justification for their application in challenging scenarios.

To assess the robustness of these methods, we introduce 10 additional random variables as redundant predictors. These variables are generated from a multivariate normal distribution with a correlation structure $\Sigma_{ij} = 0.5^{|i-j|}$. The 19 predictors are then standardized using robust estimators of location (median) and scale (Qn). Subsequently, 10% of the cells in these predictors are replaced by cellwise outliers, generated from a mixture of two normal distributions, 0.5N(10,1) + 0.5N(-10,1). As a comparison, we will also run simulations without any contamination to investigate how stable the various methods are when known outliers are present in the data. We repeat this process of adding ten redundant variables followed by generating 10% of outliers in the 19 explanatory variables 1,000 times and then compute the selection rate of each variable.

Table 1. Variable selection rates for life expectancy data over 1,000 simulation runs.

Method	Contamination (e)	Income Composition of Resources	Hepatitis B	Measles	BMI	Polio	Diphtheria	Thinness 1-19 Years	Thinness 5-9 Years	Schooling	FPR
ALRP	0.0	1.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.001875
ALRP	0.1	0.479	0.118	0.001	0.299	0.382	0.275	0.298	0.299	0.399	0.016500
LRP	0.0	1.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.002625
LRP	0.1	0.433	0.124	0.000	0.297	0.333	0.253	0.287	0.269	0.369	0.023875
ALGR	0.0	1.000	0.000	0.000	0.020	0.000	0.000	0.000	0.025	0.000	0.002500
ALGR	0.1	0.993	0.051	0.009	0.313	0.418	0.318	0.535	0.676	0.952	0.004875
LGR	0.0	1.000	0.001	0.000	0.292	0.000	0.000	0.986	0.000	0.000	0.005000
LGR	0.1	0.997	0.156	0.037	0.496	0.629	0.568	0.720	0.857	0.985	0.018500
Lasso	0.0	1.000	0.000	0.000	0.017	0.018	0.000	0.000	0.000	0.007	0.000000
Lasso	0.1	0.980	0.052	0.013	0.221	0.412	0.369	0.436	0.611	0.899	0.003625
ALasso	0.0	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000375
ALasso	0.1	0.901	0.017	0.002	0.055	0.171	0.124	0.172	0.271	0.647	0.002375
sLTS	0.0	1.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.178000
sLTS	0.1	1.000	0.349	0.027	0.371	0.326	0.357	0.315	0.328	0.327	0.371000

The sparse regression model used for variable selection is:

$$\begin{aligned}\text{Life Expectancy} = \beta_0 &+ \beta_1(\text{Income Composition of Resources}) \\ &+ \beta_2(\text{Hepatitis B}) + \beta_3(\text{Measles}) \\ &+ \beta_4(\text{BMI}) + \beta_5(\text{Polio}) + \beta_6(\text{Diphtheria}) \\ &+ \beta_7(\text{Thinness 1-19 Years}) \\ &+ \beta_8(\text{Thinness 5-9 Years}) + \beta_9(\text{Schooling}) \\ &+ \epsilon\end{aligned}$$

The results of the variable selection methods, both with and without cellwise contamination, are summarized in Table 1.

The results indicate that the performance of variable selection methods varies significantly with the presence of cellwise contamination. We assess each method's consistency based on its ability to maintain high selection rates for relevant predictors while minimizing false positives, particularly in the presence of contamination. Among the methods evaluated, ALGR stands out as the most consistent and robust in the presence of cellwise contamination. They maintain high selection rates for relevant predictors and low false positive rates, making suitable choices for real-world applications involving contaminated data

High-Dimensional Setting Using Communities and Crime Unnormalized Dataset

For the high-dimensional setting, we utilize the Communities and Crime Unnormalized dataset available on:

https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized

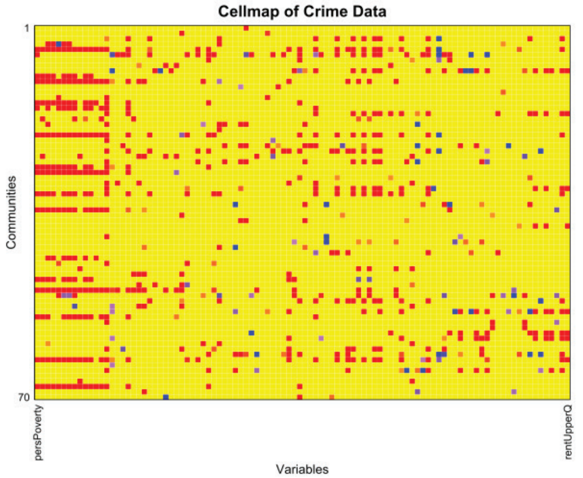

Fig. 5. Application of the Detect Deviating Cells (DDC) method on Crime dataset to identify cellwise outliers. Outlier cell map for 100 selected variables on 70 patients from the Communities and Crime Unnormalized Dataset. Most cells are yellow, showing they are not detected as outliers. A red cell means the observed value is larger than the predicted value and a blue cell means the observed value is smaller than the predicted value significantly.

Table 2. Model sizes, RMSPEs, MAPEs and RTMSPEs of the compared methods for the Communities and Crime Unnormalized Dataset with post-MM-estimator and Leave-One-Out Cross-Validation.

Method	Size	RMSPE	MAPE	RTMSPE
ALRP	4	5.743212	1.978014	1.828408
LRP	10	6.219900	2.298627	2.205932
ALGR	5	1.897773	0.864743	0.824961
LGR	6	1.681322	0.739802	0.651681
ALasso	3	2.696067	1.168824	1.079631
Lasso	4	1.756885	0.883119	0.900543
sLTS	20	58.012661	18.361718	14.349794

The results for each method are summarized in Table 2. The table presents the size of the model (number of selected variables) and the three-evaluation metrics for each method.

Here, the Gaussian Rank correlation (LGR) method proves to be the best and most consistent approach for robust variable selection.

V. Conclusion

Both in simulated and real-world data analyses, methods employing Gaussian Rank Correlation (ALGR and LGR) consistently demonstrated superior performance across varying levels of contamination, outlier magnitudes, and correlation coefficients. These methods maintained high True Positive Rates and low False Positive Rates, high F1 score indicating their robustness in variable selection under challenging conditions.

References

1. Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, & W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. (Wiley, 2005).

2. Tukey, J. W. The Future of Data Analysis. *The Annals of Mathematical Statistics* **33**, 167 (1962).

3. Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35**, 73101 (1964).

4. Raymaekers, J. & Rousseeuw, P. J. Challenges of cellwise outliers. (2023) doi:10.48550/ARXIV.2302.02156.

5. Alqallaf, F., Van Aelst, S., Yohai, V. J. & Zamar, R. H. Propagation of outliers in multivariate data. *The Annals of Statistics* **37**, (2009).

6. Rousseeuw, P. J. & Bossche, W. Van Den. Detecting Deviating Data Cells. *Technometrics* **60**, 135145 (2017).

7. Yohai, V. J. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* **15**, (1987).

8. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Series B Stat Methodol* **58**, 267288 (1996).

9. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J Am Stat Assoc* **101**, 14181429 (2006).

10. Alfons, A., Croux, C. & S. Gelper, Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat***7**, (2013).

11. Su, P., Tarr, G. & Muller, S. Robust Variable Selection under Cellwise Contamination. (2021) doi:10.48550/ARXIV.2110.12406.

12. Boudt, K., J. Cornelissen, & C. Croux, The Gaussian rank correlation estimator: robustness properties. *Stat Comput* **22**, 471483 (2011).

13. Rousseeuw, P. J. & Croux, C. Alternatives to the Median Absolute Deviation. *J Am Stat Assoc* **88**, 12731283 (1993).

14. Amengual, D., E. Sentana, & Z. Tian, Gaussian Rank Correlation and Regression. in 269–306 (Emerald Publishing Limited, 2022). doi:10.1108/s0731-90532021000043b012.

15. Öllerer, V. & C. Croux, Robust high-dimensional precision matrix estimation (2015) doi:10.48550/ARXIV.1501.01219.

16. Gnanadesikan, R. & J. R. Kettenring, Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics* **28**, 81 (1972).

17. Tarr, G., Müller, S. & N. C. Weber, Robust estimation of precision matrices under cellwise contamination. *Comput Stat Data Anal* **93**, 404420 (2016).

18. Loh, P.-L. & M. J. Wainwright, High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. (2011) doi:10.48550/ARXIV.1109.3714