2023-08-01

Robust Penalized Density Power Divergence Regression With Scad Penalty For High Dimensional Data Analysis

Maxwell Kwesi Mac-Ocloo
University of Texas at El Paso

ROBUST PENALIZED DENSITY POWER DIVERGENCE REGRESSION WITH SCAD PENALTY FOR HIGH DIMENSIONAL DATA ANALYSIS

MAXWELL KWESI MAC-OCLOO

Master's Program in Statistics and Data Science

APPROVED:

Abhijit Mandal, Ph.D, Chair

Suneel Babu Chatla, Ph.D.

Sourav Roy, Ph.D.

Stephen Crites, Ph.D
Dean of the Graduate School

to my

MOTHER, Victoria Ameley Mac-Ocloo

with much love

ROBUST PENALIZED DENSITY POWER DIVERGENCE REGRESSION WITH SCAD PENALTY FOR HIGH DIMENSIONAL DATA ANALYSIS

by

MAXWELL KWESI MAC-OCLOO

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2023

Acknowledgement

I extend my heartfelt gratitude to the Almighty Lord for guiding me on this successful journey. Special thanks to my thesis advisor, Dr. Abhijit Mandal, for his invaluable guidance and support throughout this project. I am also grateful to Dr. Suneel Baabu Chatla and Dr. Sourav Roy for their mentorship and valuable insights. Their contributions enriched this work, and I am sincerely appreciative of their involvement. Without their support, this project would not have been possible. I am truly thankful for the opportunity to learn and grow under their guidance.

Abstract

Amidst the exponential surge in big data, managing high-dimensional datasets across diverse fields and industries has emerged as a significant challenge. Conventional statistical methods struggle to handle their complexity, making analysis intricate. In response, we've formulated a robust estimator tailored to counter outliers and heavy-tailed errors. Our approach integrates the SCAD penalty into the Density Power Divergence method, effectively reducing insignificant coefficients to zero. This enhances analysis precision and result reliability. We benchmark our robust and penalized model against existing techniques like Huber, Tukey, LASSO, LAD, and LAD-LASSO. Employing both simulated and UCI machine learning repository datasets, we assess method performance using RMPE, Sensitivity, Specificity, and Mean Dimension reduction. In simulations, BIC(DPD) and EBIC(DPD) consistently yielded the lowest RMPE values for outlier proportions (0%, 5%, 10%) and signal-to-noise ratios (0.5, 1, 5), with sample size increasing from 100 to 500. Cp(DPD) exhibited strong sensitivity. Our model, Cp(DPD), surpassed LASSO and LAD-LASSO in achieving dimension reduction within high-dimensional data. While constrained by computational complexity, our model's predictor inclusion was limited. Future research should expand this aspect, validating established methods against our innovation, the Robust Penalized Density Power Divergence Regression with SCAD penalty.

Keywords: Penalized regression, high-dimensional data, outliers, density power divergence, SCAD.

Contents

List of Tables

Chapter 1

Introduction

The rapid growth of big data presents a significant challenge in analyzing high-dimensional datasets across various scientific domains and industries. These datasets often have more variables than observations, leading to issues like multicollinearity, overfitting, and limited interpretability. High-dimensional data analysis has become prominent in diverse fields such as genomics, finance, and machine learning. However, traditional statistical methods often face difficulties in handling the complexity and noise inherent in such datasets. Hence, innovative and efficient approaches are needed to tackle these challenges effectively. The current usage of statistical theory and methods may involve massive datasets, which may contain vast amounts of observations for each of a limited number of experimental units (Johnstone & Titterington 2009). We define a multivariate linear model as a straightforward yet valuable approach for high-dimensional data analysis. This model is given as

$$Y_i = \beta_0 + \sum_{i=1}^{p-1} \beta_k X_{ik} + \varepsilon_i, \tag{1.1}$$

for $i = 1, 2, \ldots, n$, and β_0 as the unknown intercept. Parameter vector, β_k's are intuitively clear that, they can reasonably be estimated well based on n observations if β is sparse in some sense.

High-dimensional data analysis is a challenging task because many traditional statistical methods are not suitable for such datasets. For example, methods that rely on low-dimensional linear models or assumptions of normality may not be appropriate for high-dimensional data. Therefore, new methods and techniques have been developed specifically for high-dimensional data analysis, including machine learning algorithms, regularization

methods, and dimensionality reduction techniques.

The prediction ability of high-dimensional data analysis faces numerous challenges, among which is the curse of dimensionality, first coined by Richard E. Bellman. According to Venkat (2018), it refers to the explosive growth of spatial dimensions and its consequences, such as the exponential increase in computational effort, inefficient use of space, and poor visualization capabilities. These challenges can significantly impede the accurate prediction of certain quantities since the predictive power of the model reduces exponentially with each added variable. Consequently, even a slight increase in dimensionality can necessitate a considerable expansion in the volume of data to maintain comparable levels of task performance.

High-dimensional data analysis presents several challenges, including the requirement for complex algorithms that can handle large datasets. However, these algorithms can be computationally intensive, leading to difficulties in analyzing data in a timely manner. Moreover, high-dimensional data analysis is susceptible to overfitting, where the model fits noise rather than underlying signals, leading to poor predictive performance and inaccurate results. Outliers are also common in large datasets, and traditional statistical methods may not be robust enough to handle them, leading to inaccuracies in results.

Handling outliers is a crucial element of predictive analysis, particularly in the realm of high-dimensional data analysis. To tackle the challenges posed by high-dimensional data, such as the curse of dimensionality, overfitting, and outliers, robust methods are necessary. These methods are designed to provide reliable and accurate results despite these challenges. Among the robust methods suitable for high-dimensional data analysis are L1-regularization for variable selection and the Huber loss function, which is a robust alternative to the mean squared error loss function used in traditional regression. In scenarios with outliers, the Tukey M-estimator, a robust regression method, is frequently employed. It is less sensitive to outliers and can produce more accurate results than traditional regression methods. In this research, the Tukey M-estimator will be utilized as the robust method for high-dimensional data analysis in the presence of these challenges.

Nahar & Purwani (2017) explores the application of robust M-estimator regression in handling data outliers. The authors explain that data outliers can significantly affect the accuracy of regression analysis and lead to incorrect conclusions. Therefore, robust regression methods are needed to provide accurate results despite the presence of outliers. They then described the M-estimator regression method and how it can be used to identify and handle outliers. They also provided a case study on a real dataset to demonstrate the effectiveness of the method in handling outliers. Overall, the paper highlights the importance of robust regression methods, specifically the M-estimator, in handling outliers in regression analysis.

The paper by Elsaied & Fried (2016) emphasized the use of Tukey's M-estimator for the estimation of Poisson parameter, especially for small means. It highlights the problems associated with the Maximum Likelihood Estimator (MLE) for Poisson parameter estimation, which can be biased and have high variance. The authors demonstrated the superiority of Tukey's M-estimator over MLE in terms of robustness to outliers and better performance for small sample sizes. The findings indicate that Tukey's M-estimator is a dependable alternative to MLE for estimating the Poisson parameter, particularly in cases with small means and outliers.

The importance of robust methods in statistical inference is widely recognized, particularly in high-dimensional settings where data often exhibit irregularities such as data contamination or heavy-tailed errors. Traditional statistical methods that assume a Gaussian or normal distribution may not be suitable in such scenarios, as they can lead to biased results and unreliable inferences. In contrast, robust methods are specifically designed to handle these challenges and provide more reliable statistical inference even in the presence of outliers or non-normal data. These methods aim to minimize the impact of extreme observations or deviations from the assumed distribution, enabling more robust and accurate estimation of model parameters (Luo 2020).

This research aims to develop robust statistical methodologies specialized for high-dimensional data analysis and outliers, achieved by integrating density power divergence

and the SCAD penalty concepts into a regression framework. The primary objective is to overcome the constraints of traditional methods, offering a more precise, interpretable, and resilient solution for high-dimensional datasets. The presence of outliers and the curse of dimensionality present challenges, resulting in biased and inefficient estimates, especially in real-world datasets with numerous variables. Consequently, the goal is to construct a regression model capable of effectively handling high-dimensional data and outliers, encouraging sparse representations while preserving predictive accuracy.

The primary focus is on exploring a robust penalized regression framework that leverages density power divergence to capture complex relationships and non-linear associations among variables, enhancing modeling flexibility. Robust methods are incorporated to mitigate the impact of outliers and heavy-tailed errors, ensuring more reliable conclusions and inferences, even in the presence of data irregularities. This approach enhances the reliability of statistical analysis, making it particularly valuable in high-dimensional settings.

The SCAD penalty promotes sparsity by encouraging certain model coefficients to be exactly zero, facilitating variable selection and enhancing model interpretability. This combination of techniques offers precise and robust estimates, even in the presence of outliers and influential data points. By addressing challenges related to high-dimensional datasets, this study aims to improve the accuracy and reliability of statistical analyses across diverse fields, ultimately advancing the understanding of complex phenomena through data-driven approaches (Fan & Li 2001, Zou 2006).

Chapter 2

Literature Review

This section gives a brief description of different existing methods of high-dimensional data analysis, their advantages and disadvantages.

2.1 Ordinary Least Squares (OLS)

The ordinary least squares (OLS) method is frequently utilized for estimating the parameters of various functional relationships. It minimizes the sum of squared differences between the observed dependent variable and the linear function of the independent variable. OLS encompasses both simple and multiple linear regression, aiming to find the best-fitting line or hyperplane for the data.

2.1.1 Linear Model

Suppose we have a dataset with n observations, where each observation i includes a scalar response y_i and a column vector x_i with p parameters (regressors). The linear regression model can be expressed as:

$$Y_i = \beta_0 + \sum_{i=1}^{p-1} \beta_k X_{ik} + \varepsilon_i, \tag{2.1}$$

for $i = 1, 2, \ldots, n$ and where the dependent variable is represented Y, X_i's are the independent variables, with β_0 being the intercept term. The β_k's are the slope coefficients, and ε_i is the error term and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and p is the number of β's in the model.

The model above can also be represented in matrix notation as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \tag{2.2}$$

where $\mathbf{Y}$ and ε are $n \times 1$ vectors of the response variables and the errors of the n observations, and $\mathbf{X}$ is an $n \times p$ matrix of regressors, often referred to as the design matrix with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

When holding all other independent variables constant, a change in the beta coefficient corresponding to an independent variable results in an increase or decrease in the response variable. In other words, for every unit change in the beta coefficient of the independent variable, there is a corresponding change in the dependent variable. In general, the multiple linear regression model is widely utilized in various fields, such as economics, finance, engineering, and social sciences. In economics, multiple linear regression is frequently utilized to establish connections between various economic variables. For instance, a regression model may be utilized to evaluate the association between income, education, and health outcomes. Similarly, in finance, multiple linear regression is widely employed to model the relationships between financial variables. For example, a regression model may be used to analyze the correlation between stock prices, interest rates, and other financial indicators. Multiple linear regression is widely used in health sciences to study connections between health variables, like diet, exercise, and disease risk. In conclusion, multiple linear regression is a versatile technique with diverse applications. It enables modeling relationships between variables and predicting their impacts on each other.

In multiple linear regression, coefficients are determined using the least squares method to minimize the sum of squared differences between observed and predicted response variable values. The residuals vector indicates the discrepancies between the actual and estimated values and this is given by

$$\varepsilon = Y - X\beta,$$

with the residuals sum of squares being represented as:

$$\varepsilon^{'}\varepsilon = (Y - X\beta)^{'}(Y - X\beta),$$

$$\varepsilon^{'}\varepsilon = Y^{'}Y - 2\beta^{'}X^{'}Y + \beta^{'}X^{'}X\beta.$$

Take the partial derivative by differentiating with respect to β

$$\frac{\partial\varepsilon^{'}\varepsilon}{\partial\beta} = -2X^{'}Y + 2X^{'}X\beta.$$

Now, by setting the partial derivative to zero and solving for β, we obtain:

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y. \tag{2.3}$$

The fitted values, $\hat{Y}$ can be represented in a vector form as:

$$\hat{Y} = X\hat{\beta} = X(X^{T}X)^{-1}X^{T}Y. \tag{2.4}$$

When the Gauss-Markov assumptions, also known as the OLS assumptions or assumptions of the Classical Linear Regression Model (CLRM), are satisfied, the OLS estimator is reliable and provides consistent results. The key requirements that must be fulfilled include:

- Linearity: In multiple linear regression, the association between the dependent variable and the independent variables should be constant and linear across different values of the independent variable.

- Independence: The observations in the dataset should be independent of each other. This means that the value of the dependent variable for one observation should not be influenced by the value of the dependent variable for any other observation.

- Homoscedasticity: In multiple linear regression, the errors (ε's) should exhibit constant variance across all independent variable values. This ensures that the errors are evenly distributed throughout the range of the independent variables.

- Normality: The errors should be normally distributed with a mean of zero. This means that the distribution of errors should be symmetrical around zero.

OLS (Ordinary Least Squares) for linear regression offers several advantages:

- Simplicity and Flexibility: Estimating linear regression coefficients using OLS is an uncomplicated and user-friendly technique. On the other hand, multiple linear regression is a versatile and flexible statistical approach capable of addressing various research inquiries and data formats.

- Interpretability: The coefficients in an OLS model have clear and easily interpretable meanings, representing the change in the response variable per unit change in the corresponding predictor variable.

- Efficiency and Applicability: If the OLS assumptions are fulfilled, this method is highly efficient and yields dependable and unbiased estimates. Moreover, it can be readily employed in diverse fields requiring linear regression modeling to examine variable relationships and forecast future outcomes based on observed data.

Despite its advantages, ordinary least squares (OLS) for linear regression also has some limitations:

- Limited applicability: It is only applicable to linear models and does not work well with nonlinear relationships.

- Sensitive to outliers: OLS is sensitive to outliers, which can significantly influence the estimates of the regression coefficients.

- Assumption-dependent: OLS relies on the assumption that the errors follow a normal distribution with constant variance. If these assumptions are not met, the accuracy and reliability of the regression results may be affected.
- Multicollinearity: When there is multicollinearity (high correlation) among the predictor variables, the OLS estimates may be unstable and difficult to interpret.

2.2 Least Absolute Shrinkage and Selection Operator

Lasso, also known as L_1 regularization, was introduced by Tibshirani (1996) as a method for variable selection and regularization of coefficients to prevent overfitting in high-dimensional data. The Lasso regularization method has gained popularity in high-dimensional estimation problems due to its statistical accuracy in prediction and variable selection, along with its computational feasibility (Bühlmann & Van De Geer 2011). This approach involves augmenting the traditional regression loss function by introducing a penalty term that encourages less important features to shrink towards zero, effectively excluding them from the final model. The degree of shrinkage is determined by a tuning parameter called lambda (λ), which can be optimized through cross-validation. Lasso is especially advantageous when working with datasets containing numerous predictors, while only a handful of these predictors are expected to be significant for the outcome variable.

The lasso model adds an L_1 regularization penalty term to the ordinary least squares (OLS) regression that minimizes:

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.5}$$

where y is the response variable, β_0 is the intercept, $X_i (i = 1, \ldots, p)$ are the predictor variables, β_i are their corresponding coefficients, p is the number of predictors, and ε is the error term. The λ is the shrinkage or penalty parameter which determines the degree of regularization applied to the model and $\sum_{j=1}^{p} |\beta_j|$ is the sum of the absolute values of the

regression coefficients.

L_1 regularization has several advantages in high-dimensional data analysis. Lasso's variable selection is designed to shrink the coefficients of the less important variables to zero, effectively excluding them from the model. This feature of Lasso makes it useful for variable selection in high-dimensional data where there are many predictors, and only a few are expected to be important for the outcome variable (Tibshirani 1996).

In addition, Lasso's computational feasibility has made it popular for solving high-dimensional estimation problems. Lasso adds a linear penalty term to the traditional regression loss function, making the optimization problem solvable using coordinate descent, which is a computationally efficient algorithm. Lasso mitigates overfitting without needing a separate validation dataset by performing variable selection and coefficient regularization. This results in an increased statistical accuracy in prediction (Ranstam & Cook 2018, Bühlmann & Van De Geer 2011). Lasso is known to be robust to outliers in the data, which can be beneficial in real-world applications. Overall, Lasso's advantages make it a popular tool for high-dimensional data analysis in various fields, including machine learning, genetics, and economics.

Although the Lasso method for linear regression offers advantages in terms of variable selection and regularization, it does have some drawbacks. Specifically, Lasso-based methods excel at selecting important predictors and applying regularization to enhance model performance. However, they may not consistently yield stable subset selection, especially when dealing with highly correlated predictors. In certain situations, this instability can lead to inconsistent results, thereby limiting the reliability and robustness of the Lasso-based approach (Signorino & Kirchner 2018, Zou 2006). Additionally, the LASSO estimator has limitations such as the maximum number of variables it can select when p>n and a lack of grouping property, resulting in selecting only one variable from highly correlated predictors. Furthermore, LASSO's prediction performance is often inferior to other regression models like ridge regression in scenarios where there are more observations than predictors and high correlations among predictors (Emmert-Streib & Dehmer 2019).

Lasso regression is potent but not a universal solution for overfitting and optimism bias. External validation remains necessary. Lasso sacrifices precision in parameter estimation to enhance overall prediction accuracy. This might limit the interpretability of regression coefficients as independent risk factors, as the emphasis is on combined prediction rather than precise estimation (Ranstam & Cook 2018).

2.3 Least Absolute Deviation (LAD)

The Least Squares method is effective for estimating linear model parameters when residuals are normally distributed without large outliers. However, when residuals are non-normally distributed and contain significant outliers, estimates are affected in many applications across science and engineering. To mitigate the influence of outliers, robust regression methods have been developed. The LAD method, also known as the L_1-norm, is a statistical optimization approach that employs a statistical optimality criterion. It aims to minimize the sum of absolute deviations, residuals, or errors. By minimizing the sum of absolute errors, this estimator achieves greater efficiency than the OLS method. That is;

$$\min_{\hat{\beta}} |y_i - \hat{y}_i| \,.$$

The primary objective of this method is to reduce the L_1 norm of these absolute values. Unlike other methods that require a tuning mechanism, the LAD approach is considered the most straightforward method for robust regression. This is mainly because of its inherent robustness, as it does not rely on squared values like the Least Squares method, and thus, it is often preferred (Dasgupta & Mishra 2004, Li & Arce 2004).

The LAD regression technique, characterized by errors following a Laplacian distribution, lacks a readily available closed-form solution. Consequently, the approach necessitates the use of numerical and iterative algorithms to address the computational aspects (Li & Arce 2004). In accordance with Taylor's account in 1974, KF Gauss and PS Laplace can be

credited with developing and employing the method for solving an over-determined system of linear algebraic equations. They proposed and utilized the approach of Least Squares, and also practiced the method of Least Absolutes, which endeavors to minimize the sum of the absolute values of the residuals in the equations (Dasgupta & Mishra 2004).

Unlike the Least Squares (LS) method, the LAD method is less affected by outliers and provides robust estimates(Chen et al. 2008). Within the framework of the Minkowski norm, the mathematical expressions for the Least Absolute (L_1) and Least Squares (L_1) methods are as follows:

$$Min(S) = \min_{a} \left(\sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} a_j X_{ij} \right|^p \right)^{\frac{1}{p}}, \tag{2.6}$$

for $p = 2$ and $p = 1$ respectively. Some advantages of the Least Absolute Deviation (LAD) method include its robustness compared to the least squares method and its ability to handle outliers effectively. LAD exhibits robustness by being resistant to the presence of outliers in the data. It demonstrates a strong resistance to the influence of outliers or any other forms of data contamination within the dataset. This robustness makes LAD a widely used technique in various fields.

In heavy-tailed distributions where outliers are more common, the Least Absolute Deviation (LAD) method demonstrates higher efficiency compared to the Least Squares (LS) method. LAD achieves this increased relative efficiency by assigning lower weights to outliers, resulting in more accurate estimation. This characteristic allows LAD to provide improved performance and reliability when dealing with datasets that exhibit heavy-tailed distributions and a higher incidence of outliers. In order to achieve more efficient parameter estimation in robust regression, Thanoon (2015) illustrated that the LAD (L1-norm) method, in conjunction with the Iteratively Reweighted Least Squares (IRWLS) approach, exhibits improved efficiency when estimating model parameters across different error distribution scenarios. This advantage becomes particularly noticeable when comparing it to the performance of the LS method assuming a normal distribution, regardless of the sample

sizes.

Chen et al. (2008) highlight two main drawbacks of the Least Absolute Deviation (LAD) method when compared to the Least Squares (LS) method. Firstly, LAD lacks a convenient inference procedure, making statistical inference and drawing meaningful conclusions from estimated parameters more challenging. Secondly, the LAD method does not possess a valid analysis-of-variance approach, limiting its applicability in analyzing variance components within a model (Chen et al. 2008). These limitations restrict the usability and interpretability of LAD in certain statistical analyses.

The LAD method offers robustness to outliers, but it comes with some drawbacks. It is computationally more complex than the OLS method due to its iterative nature, requiring more time and resources. Moreover, LAD may produce multiple solutions for specific datasets, complicating the determination of the most appropriate one and posing challenges in result interpretation. Nevertheless, LAD remains a valuable tool in scenarios where outlier resistance is crucial, and careful consideration should be given to its computational demands and potential for multiple solutions.

2.4 LAD-LASSO

LAD-Lasso, or Least Absolute Deviations-Lasso, is a powerful statistical method that merges the principles of least absolute deviations (LAD) with Lasso regularization. This technique finds its primary application in feature selection and regression analysis. The lasso, introduced by Tibshirani (1996), is known as the "least absolute shrinkage and selection operator" and stands as a robust approach that excels in both variable selection and regression parameter estimation. Notably, the lasso has gained substantial recognition for its ability to effectively identify essential explanatory variables while ensuring accurate estimation of regression parameters.

LAD-Lasso offers the advantage of performing parameter estimation and variable selection simultaneously, setting it apart from LAD regression. Additionally, in comparison to

the traditional Lasso method, LAD-Lasso demonstrates robustness in handling heavy-tailed errors or outliers in the response variable. The core concept behind LAD-Lasso involves the integration of the conventional LAD criterion with a lasso-type penalty, resulting in the development of the LAD-Lasso method (Wang et al. 2007). Now, suppose we have a multivariate linear regression model defined as;

$$y_i = x_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{2.7}$$

for $i = 1, \ldots, n$ and where $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is the p-dimensional regression covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ are independent associated regression coefficients and ε_i are independent and identically distributed (iid) random errors with mean 0 and variance σ^2. The parameters of the model can be estimated by minimizing the Ordinary Least Squares (OLS) criterion, $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$. Moreover, in order to shrink unnecessary coefficients towards zero, Tibshirani (1996) introduced the lasso criterion, which can be expressed as follows:

$$LASSO = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + n\lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda > 0$ is the tuning parameter and due to the uniform application of tuning parameters for all regression coefficients in the lasso method, there is a potential for the resulting estimators to display notable bias. Hence, the modified lasso criterion by Fan & Li (2001) is given as:

$$LASSO* = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + n \sum_{j=1}^{p} \lambda_j |\beta_j|.$$

Wang et al. (2007) acknowledged the well-known fact that the OLS criterion utilized in the aforementioned equation is highly susceptible to the influence of outliers. To attain a more resilient lasso-type estimator, they proposed a modification to the lasso* objective

function, leading to the formulation of the LAD-lasso criterion as follows:

$$LADlasso = Q(\beta) = \sum_{i=1}^{n} |y_i - x_i^{'}\beta| + n\sum_{j=1}^{p} \lambda_j |\beta_j|. \tag{2.8}$$

The LADlasso criterion effectively combines the LAD criterion and the lasso penalty, resulting in an estimator that is robust against outliers and promotes sparsity in the representation (Wang et al. 2007).

LAD-Lasso offers several unique advantages. In high-dimensional datasets with outliers, LAD-Lasso tends to outperform Lasso by producing solutions with smaller standard errors. Its robustness against outliers enables LAD-Lasso to provide more accurate and precise estimations, resulting in reduced standard errors and increased reliability of the obtained results (Rahardiantoro & Kurnia 2015).

LAD-Lasso provides the unique capability of simultaneous variable selection and regression parameter estimation. By identifying the most important variables and estimating their coefficients, it constructs a concise and interpretable model. Moreover, users have the flexibility to choose the tuning parameter, allowing for a trade-off between model complexity and predictive performance based on individual needs and prior knowledge. This flexibility enhances the applicability and adaptability of LAD-Lasso in various data analysis scenarios.

Some drawbacks of LAD-LASSO are:

- Computational complexity: LAD-Lasso can be computationally intensive, especially for large-scale or high-dimensional datasets. The non-differentiable penalty function and the need to solve an optimization problem for each value of the tuning parameter can increase the computational burden.

- Lack of variable selection consistency:The LAD-Lasso method differs from Lasso in that it may not consistently select the correct relevant variables as the sample size increases. It has the potential to include irrelevant variables or exclude important

ones, which can lead to model misspecification.

- Another drawback of LAD-Lasso is its sensitivity to the selection of the tuning parameter (λ). Choosing the appropriate value for λ can be challenging, and an improper choice may result in biased coefficient estimates or suboptimal variable selection. Careful consideration is essential when determining the optimal λ value to ensure the accuracy and reliability of the LAD-Lasso model.

2.5 Huber Loss Function

Numerous challenges in the fields of learning, optimization, and statistics necessitate robustness, which implies that a model trained or optimized should be less influenced by outliers than by inliers, that is, the regular data (Gokcesu & Gokcesu 2021). Furthermore, the authors recommended that rather than employing outlier detection methods, it is crucial to develop loss functions that possess inherent robustness to outliers. The Huber loss is a loss function employed in robust regression within statistics, which is comparatively less responsive to outliers in data compared to the popular mean squared error (MSE) loss function. It is also occasionally utilized in a modified form for classification purposes and combines the advantages of both MSE and mean absolute error (MAE) by behaving like MSE for small errors and like MAE for large errors. Due to its strong convexity and ability to facilitate fast learning, the square loss is highly effective. On the other hand, the absolute loss is capable of withstanding arbitrary outliers, as their impact on the estimation is determined by their position in the data, rather than their actual values. Combining the benefits of these two loss functions is crucial in the development of algorithms that can exhibit robustness against outliers, while also achieving rapid convergence with minimal loss (Gokcesu & Gokcesu 2021).

Therefore, the Huber (1964) loss function is defined as

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise .} \end{cases}$$

For small values of a, the function behaves quadratically, but for large values, it behaves linearly. The two sections of the function have the same values and slopes at two points where $|a| = \delta$. Typically, the variable a represents residuals, which are the differences between observed and predicted values ($a = y - f(x)$). So the former can be expanded to

$$L_\delta(a) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta, \\ \delta(|y - f(x)| - \frac{1}{2}\delta), & \text{otherwise .} \end{cases}$$

Therefore, in order to develop a loss function that is both robust and converges quickly, it is necessary to merge the characteristics of the absolute and quadratic losses. The simplest method involves using a piecewise function to blend the quadratic and absolute losses in a manner that maximizes their effectiveness (Gokcesu & Gokcesu 2021).

One advantage of the Huber loss function is its wide application across various fields, including machine learning, control systems, and computer vision. It offers a significant benefit over the Mean Squared Error (MSE) by demonstrating reduced sensitivity to outliers. This characteristic is particularly valuable in real-world scenarios where data may contain unexpected or noisy values. By mitigating the influence of outliers, the Huber loss function can improve model accuracy and stability, resulting in more reliable and robust predictions.

The Huber loss function strikes a balance between the advantages of mean squared error (MSE) and mean absolute error (MAE). It is quadratic for small errors and linear for large errors, providing a trade-off between the two. As a result, it is less sensitive to outliers than MSE while still being differentiable everywhere like MSE. Huber loss is continuous and differentiable, which makes it amenable to optimization using gradient-based methods. This is

particularly significant in machine learning, where gradient-based optimization algorithms are frequently employed to minimize the loss function by adjusting the model parameters (Gokcesu & Gokcesu 2021). Due to its benefits, Huber loss is commonly used in regression problems, particularly in real-world scenarios where the data may have outliers or other sources of noise.

Despite its advantages, Huber loss has a major disadvantage that is associated with complexity. The loss function has a hyperparameter, δ, which determines the point at which the function changes from being quadratic to linear. Tuning the value of delta(δ) to suit the specific problem at hand can be a challenging and time-consuming process that requires expert knowledge. However, in general, the drawbacks of Huber loss are minor compared to its benefits. Nevertheless, when selecting a loss function, it is essential to consider the specific needs of the problem at hand and the characteristics of the data and model being used.

2.6 Tukey M-estimator

In 1964, Huber introduced the M-estimation method, which has since become a widely adopted and popular approach for robust regression. It is an extension of maximum likelihood estimation and is specifically designed to handle outliers in location models effectively. Huber's M-estimation method has gained significant popularity and is widely used in robust regression due to its ability to provide robust and reliable parameter estimates in the presence of outliers. It offers comparable efficiency to ordinary least squares (OLS) while using a different objective function. Instead of minimizing the sum of squared errors, the M-estimation method minimizes a residual function, which contributes to its robustness against outliers. In linear regression, the M-estimate objective function is given as:

$$\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{X_i}^\top \beta).$$

Filzmoser & Nordhausen (2021) highlight the importance of maintaining desired equivariance properties, such as robustness unaffected by rescaling the response variable. To achieve this, scaling the residuals becomes necessary. This leads to the development of the regression M-estimator, which incorporates scaled residuals to ensure robustness. The M-estimator is a widely used technique in robust regression, known for its effectiveness in estimating parameters that are influenced by outliers. In general, M-estimation involves minimizing an objective function to obtain robust regression estimates (Nahar & Purwani 2017). By utilizing scaled residuals, the robustness of the estimator remains consistent even when the response variable is rescaled.

John Tukey introduced the Tukey's M-estimator in 1960 as a robust regression technique. Like the Huber loss function, the Tukey M-estimator is specifically designed to be less sensitive to outliers compared to traditional regression methods. Its robustness allows it to generate more dependable parameter estimates, even when dealing with outliers, making it an invaluable tool for robust data analysis. To obtain a scale-invariant version of the Tukey M-estimator in regression, a common approach is to divide the residuals by a robust measure of scale. The regression Tukey M-estimator with scale adjustment is achieved by using the median absolute deviation (MAD) as the scale estimator. This approach aims to minimize the impact of outliers and influential data points in linear regression. The objective function for the regression Tukey M-estimator with scale adjustment can be defined as follows:

$$\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{X_i}^\top \beta}{s}\right), \tag{2.9}$$

where s is the scale of robust estimation. The S estimator that is often used is

$$s = \frac{median|e_i - median(e_i)|}{0.6745}.$$

- The inclusion of the constant 0.6745 in the calculation of S makes it an approximately unbiased estimate of σ under the conditions of a large sample size and a normal distribution.

- y_i's are the observed values of the response variable.
- $\mathbf{X_i}^\top$ is the vector of predictors for the ith observation.
- β is the vector of regression coefficients to be estimated and ρ measures the agreement between an observation, y_i and any possible value of the parameter of interest.

To minimize Equation (2.9), first take partial derivatives with respect to β and set them equal to zero. Then,

$$\sum_{i=1}^{n} X_i \psi \left(\frac{y_i - \mathbf{X_i}^\top \beta}{s} \right) = 0,$$

where $\psi = \rho^{'}$ is the influence function. From Equation (2.9), the matrix notation can be written as follows:

Therefore, the matrix notation based on the equation above can be written as follows:

$$\mathbf{X}^\top \mathbf{W} \mathbf{X} \beta = \mathbf{X}^\top \mathbf{W} \mathbf{Y}, \tag{2.10}$$

where $\mathbf{W}$ is $n \times n$ diagonal matrix of weights (that is, the weight function). X is the independent variable matrix size $(n \times (p+1))$ and Y is the dependent variable matrix size $(n \times n)$. Therefore, the robust regression Tukey M-estimator for β is:

$$\hat{\beta}_M = (X^\top W X)^{-1} (X^\top W Y). \tag{2.11}$$

The influence function $\psi(\cdot)$ quantifies the impact of individual data points on the estimation procedure, helping identify influential observations that can skew results or lead to biased estimates. It is a valuable tool for assessing robustness, detecting outliers, and understanding the behavior of estimators in the presence of influential data.

There are commonly used influence functions for M-estimators. For the Tukey M-

estimator, the influence function is given as;

$$\psi(z) = \begin{cases} z(1-(z/k)^2)^2 & \text{if } |z| \leq k, \\ 0 & \text{if } |z| > k. \end{cases}$$

The tuning constant k plays a crucial role in balancing the robustness and efficiency of the estimators. Opting for a larger k enhances efficiency but compromises robustness to outliers. In the case of the Tukey function, k values typically range from 3 to 5 to strike an optimal balance between efficiency and robustness (Elsaied & Fried 2016).

The Tukey M-estimator for linear regression offers numerous advantages:

- Robustness to outliers: Tukey M-estimators exhibit resilience against the impact of outliers, which can have a significant effect on conventional estimators like least squares. When confronted with substantial outliers, Elsaied & Fried (2016) suggested employing the Tukey function because of its redescending nature. This distinctive feature enables the Tukey M-estimator to effectively reduce the influence of extreme observations, resulting in improved robustness and dependability when dealing with data containing outliers.

- Flexibility in error distribution: Tukey M-estimators are advantageous in linear regression as they can handle non-normal error distributions, such as heavy-tailed or skewed distributions, which may deviate from the assumption of normality. This flexibility enhances their robustness and enables accurate capture of the underlying relationship between variables.

- Robustness customization: Tukey M-estimators provide researchers with the ability to customize the robustness level by manipulating a tuning constant. This allows them to adjust the estimator's robustness according to the unique attributes of their data, enabling them to strike an ideal trade-off between robustness and efficiency that aligns with their preferences and requirements.

Despite its advantages, the Tukey M-estimator for linear regression also presents some limitations and challenges. Some of these challenges are:

- Inconsistency: The Tukey M-estimator for linear regression may lack consistency in the presence of outliers or errors, leading to biased results. It is important to consider alternative robust estimation methods to obtain reliable estimates in such scenarios.

- Sensitivity to the choice of tuning constant: The effectiveness of Tukey M-estimators heavily relies on the choice of the tuning constant. Improper selection can lead to biased estimates and reduced efficiency. Finding the optimal tuning constant is challenging and often requires trial-and-error or data-driven techniques.

- Computational complexity: Tukey M-estimators can be computationally demanding, especially for large datasets or complex models. The iterative nature of the estimation procedure, where the tuning constant is updated in each iteration, can lead to increased computational time and resource requirements.

Chapter 3

Methodology

This chapter presents the methodology employed in this study, focusing on the Density Power Divergence (DPD), the SCAD (Smoothly Clipped Absolute Deviation) penalty, and the robust penalized regression model.

3.1 Density Power Divergence (DPD)

The density power divergence (DPD) measure, introduced by Basu et al. (1998), incorporates a tuning parameter $\alpha \geq 0$ in minimum density power divergence estimation. This estimation framework is widely used in robust statistics due to its flexibility in handling outliers and heavy-tailed distributions. The key step in this approach involves numerically minimizing the power divergence, which leads to the desired estimates. Riani et al. (2020) demonstrated that by choosing a suitable value of α and minimizing the power divergence, robust and reliable estimates can be obtained, even when dealing with challenging data characteristics.

Therefore, the divergence measure between the model density f_θ with parameter $\boldsymbol{\theta} \, \epsilon \, \boldsymbol{\Theta}$ and the true or data density g is defined as

$$d_\alpha(f_\theta, g) = \begin{cases} \int_z \left\{ f_\theta^{1+\alpha}(z) - (1 + \frac{1}{\alpha}) f_\theta^{\alpha}(z) g(z) + \frac{1}{\alpha} g^{1+\alpha}(z) \right\} dz, & \text{for } \alpha > 0, \\ \int_z g(z) \log \left(\frac{g(z)}{f_\theta(z)} \right) dz, & \text{for } \alpha = 0. \end{cases}$$

When $\alpha = 0$, the DPD becomes the Kullback-Leibler divergence as a limiting case when $\alpha \to 0^+$. The minimum density power divergence estimator (MDPDE) is obtained by

minimizing the density power divergence (DPD) measure with respect to the parameter $\boldsymbol{\theta}$ over its parametric space $\boldsymbol{\Theta}$. The tuning parameter plays a crucial role in the MDPDE as it controls the balance between efficiency and robustness for the power divergence estimator. By choosing an appropriate value of α, one can adjust the trade-off between these two characteristics, allowing for more tailored and robust estimation (Ghosh & Basu 2016, Riani et al. 2020). In general, it can be shown that as the tuning parameter α increases, the robustness of the Minimum Density Power Divergence estimator increases while its efficiency decreases (Basu et al. 1998).

Considering a linear regression model from Equation (2.7), we let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ be the parameter with the probability density function (pdf) of y_i, denoted by $f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$ or f_i follows a normal distribution which is given by

$$f_i \equiv f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2\right), \tag{3.1}$$

where f_i is then the estimate of the vector of the parameters according to the Maximum Likelihood (ML) criterion is

$$\hat{\beta}_{ML} = \arg\min_{\beta} \left[\frac{1}{(2\pi\sigma^2)^{(n/2)}} \left(\frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right)\right]. \tag{3.2}$$

This is equivalent to the solution given by ordinary least squares method. In Durio & Isaia (2011), we let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of size $n \geq 2$ from $\boldsymbol{X}$, the Minimum Density Power Divergence Estimator for $\boldsymbol{\theta}_0$ corresponding to the vector $\hat{\boldsymbol{\theta}}_\alpha$ by minimizing the divergence $d_\alpha(f_\theta, g)$. In Riani et al. (2020), since the third term of the divergence is independent of θ, the power divergence estimator of θ can be found by minimizing

$$\int f_\theta^{1+\alpha}(z)dz - \left(1 + \frac{1}{\alpha}\right)\frac{1}{n}\sum_{i=1}^{n} f_\theta^\alpha(y_i). \tag{3.3}$$

According to Durio & Isaia (2011), when $\alpha = 0$ the MDPDE reduces to the Maximum Likelihood estimator while for $\alpha = 1$, the divergence $d_1(f_\theta, g)$ yields the L_2 metric and the

estimator minimizes the L_2 distance between the densities. If we assume that the random variables $Y|\boldsymbol{x}$ are distributed as a $\mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma)$, then with reference to Equation (3.3), the estimate of the vector $\boldsymbol{\theta}_\alpha = [\beta_0, \ldots, \beta_p, \sigma]$, is given by

$$\hat{\boldsymbol{\theta}}_\alpha = \underset{\beta,\sigma}{\arg\min} \left[\frac{1}{\sigma^\alpha \sqrt{(2\pi)^\alpha (1+\alpha)}} - \frac{\alpha+1}{\alpha} \frac{1}{n} \sum_{i=1}^{n} f_i^\alpha (Y_i | \boldsymbol{X}_i^\top \boldsymbol{\beta}, \sigma) \right]. \tag{3.4}$$

Hence, from the above equation, the penalized MDPDE $(\tilde{\beta}, \tilde{\sigma})$ using iterative algorithm gives

$$\tilde{\beta}_{md} = (X^\top \hat{W}_\alpha X)^{-1} (X^\top \hat{W}_\alpha Y), \tag{3.5}$$

and

$$\tilde{\sigma}^2 = \frac{1}{trace(\hat{W}_\alpha)} \left[\frac{1}{n} (Y - X^\top \hat{\beta})^\top \hat{W}_\alpha (Y - X^\top \hat{\beta}) + \frac{\alpha}{(2\pi\hat{\sigma}^2)^{\alpha/2} (1+\alpha)^{3/2}} \right], \tag{3.6}$$

where $\hat{W}_\alpha$ is defined as a diagonal matrix with $f_1^\alpha, f_2^\alpha, \ldots, f_n^\alpha$ as diagonal elements and $Y = (Y_1, \ldots, Y_n)^\top$.

3.2 Robust Penalized Regression Method

Robust regression techniques have gained interest due to the limitations of classical methods in variable selection. The development of penalized robust regression approaches has been driven by the advancements in collecting and analyzing high-dimensional data. These methods combine robustness with variable selection, addressing the challenges of high-dimensional data. Penalized robust regression effectively handles outliers and non-normality while identifying relevant predictors by incorporating penalty terms in the objective function. The SCAD penalty, among others, strikes a balance between sparsity promotion and robustness, resulting in more accurate estimation and inference (Luo 2020).

In high-dimensional settings, the sensitivity of the quadratic loss function to heavy-tailed errors or outliers presents a challenge for linear regression models. To mitigate

this challenge, a robust penalized selection and estimation procedure can be utilized. This procedure replaces the conventional sum of squares loss function with a robust loss function that is more resilient to outliers or heavy-tailed errors. According to Luo (2020), the corresponding robust estimator $\hat{\beta}$ takes the following form

$$\hat{\beta} = \arg\min_{\beta} \left(\mathcal{L}(\boldsymbol{\beta}; Z_1^n) + \rho_\lambda(\boldsymbol{\beta}) \right), \tag{3.7}$$

where $\mathcal{L}(\boldsymbol{\beta}; Z_1^n)$ is the empirical loss function, $Z_1^n = (Z_1, Z_2, \ldots, Z_n)$ denote a collection of n samples and $Z_i = (\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$. It is important to note that a penalized robust procedure is defined by its loss function $\mathcal{L}(\boldsymbol{\beta}; Z_1^n)$ and the penalty function. The loss function is designed to handle outliers and heavy-tailed errors, while the penalty function promotes sparsity in the parameter vector β.

Therefore, by referring to the equation above, the robust estimator $\hat{\beta}$ for the penalized regression model can be expressed as follows:

$$\hat{\beta} = \arg\min_{\beta} \left(d_\alpha(f_i, g) + \rho_\lambda(\boldsymbol{\beta}) \right), \tag{3.8}$$

where

- $d_\alpha(f_i, g)$ is the density power divergence measure between the observed data g and the model density function f_i.
- λ is the tuning parameter that controls the strength of the penalty term
- $\rho_\lambda(\cdot)$ is the SCAD penalty function applied to the regression coefficients $\boldsymbol{\beta}$.
- $\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho_\lambda(|\beta_j|)$

The penalized DPD regression aims to achieve unbiased and selective coefficient estimates. It achieves this by minimizing the DPD measure while incorporating the SCAD penalty. This combination encourages sparsity in the parameter estimate, allowing for vari-

able selection in high-dimensional settings. By promoting sparsity and addressing outliers, this approach enhances the accuracy and robustness of the estimation process.

3.3 Smoothly Clipped Absolute Deviation (SCAD)

In high-dimensional statistical modeling, the process of variable selection is crucial. However, traditional methods like LASSO often suffer from bias in this process. To overcome this limitation, the smoothly clipped absolute deviation (SCAD) estimator was introduced. The SCAD estimator addresses bias while promoting sparsity in the selected variables through a continuous penalty. By striking a balance between bias reduction and sparsity, the SCAD estimator provides a robust and effective approach to variable selection in high-dimensional settings. The SCAD estimator, originally proposed by Fan & Li (2001), offers desirable properties such as continuity, sparsity (encouraging coefficient shrinkage towards zero), and unbiasedness.

In the article of Fan & Li (2001), it is discussed that the L_q penalty function and the hard thresholding penalty function do not meet the mathematical conditions required for achieving unbiasedness, sparsity, and continuity simultaneously. This finding highlights the limitations of these penalty functions and the challenges involved in incorporating all desired properties into a single penalty function. For this reason, Fan (1997) and Fan & Li (2001) proposed a non-concave penalty function referred to as the smoothly clipped absolute deviation (SCAD) which is given by

$$p_{\lambda}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq \lambda; \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\beta_j| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda, \end{cases}$$

where $a > 2$ is a fixed parameter. The function exhibits continuity and possesses a first derivative for certain values of $a > 2$ and $\beta > 0$. Specifically, the first derivative of the

function which is given by

$$p'_{\lambda}(\beta) = \lambda \left\{ I(\beta \leq \lambda + \frac{(a\lambda - \beta}{(a-1)\lambda} + I(\beta > \lambda) \right\}, \tag{3.9}$$

enhances the properties of the L_1 penalty and the hard thresholding penalty function. This equation represents a quadratic spline function with knots at λ and $a\lambda$. By introducing these knots, the penalty function avoids excessive penalization for large values of λ, while also ensuring the solution remains continuous. This improvement in the penalty function helps achieve desirable properties in variable selection and estimation (Fan & Li 2001, Fan 1997).

Now, analogous to the SCAD estimator when using the square-error loss function, we derive $\hat{\beta}_{SCAD}$, the penalized MDPDE of β using the SCAD penalty function. Now, suppose $\hat{\beta}_j$ is the unpenalized MDPDE of the j-th component of β, then, the penalized MDPDE can be written as

$$\hat{\beta}_{j,SCAD} = F^{SCAD}_{\lambda,\alpha}(\hat{\beta}_j). \tag{3.10}$$

Fan (1997) demonstrated that the SCAD penalty can yield sparse solutions and approximately unbiased coefficient estimates for large coefficients. Therefore, the solution to the SCAD penalty in the equation above can be given as

$$F^{SCAD}_{\lambda,\alpha}(\hat{\beta}_j) = \begin{cases} sign(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ & \text{if } |\hat{\beta}_j| < 2\lambda; \\ \{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda\}/(a-2) & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda; \\ \hat{\beta}_j & \text{if } |\hat{\beta}_j| > a\lambda, \end{cases} \tag{3.11}$$

where λ, α and a are tuning parameters. Here, our λ parameter is selected using information criteria (AIC, BIC, etc.) and the DPD parameter α is selected based on another information criterion, called H-score.

During it implementation in the Bayesian analysis, Fan & Li (2001) assumed that, for given a and λ, the prior distribution for β_j is a normal distribution with zero mean and variance $a\lambda$ were the Bayes was computed through numerical intergration. They further suggested that, for the universal thresholding, $\lambda = \sqrt{2\log(d)}$ (Donoho & Johnstone 1994). Fan & Li (2001) recommended choosing $a = 3.7$ based on Bayesian statistical considerations and simulation studies. They argued that this value provides good performance for various variable selection problems. Furthermore, they noted that data-driven methods for selecting the value of a do not significantly improve the performance of variable selection in practice.

Robust regression is a valuable tool for analyzing data with outliers, as it aims to produce stable results even in the presence of these outliers. Unlike traditional approaches, robust estimation methods are specifically designed to mitigate the influence of outliers during the estimation process, resulting in more reliable and resilient analyses. These methods ensure that outliers do not disproportionately impact the results, leading to more accurate and robust statistical inferences (Almetwally & Almongy 2018).

3.4 Model Selection Criteria

Model selection is crucial in high-dimensional settings like regression to prevent biases and errors. Reliable criteria identify relevant variables and improve parameter estimation and prediction. By employing robust model selection, researchers ensure unbiased results and enhance statistical analyses' validity.

3.4.1 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) is a widely used statistical measure for model evaluation and selection. It strikes a balance between the goodness of fit of a model and its complexity, addressing the trade-off between overfitting and underfitting. AIC quantifies the relative amount of information lost by each model and favors models with lower information loss.

By comparing the information loss of different models, AIC allows researchers to assess the quality of each model's approximation to the true underlying process. Therefore, suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model and $\hat{L}$ be the maximized value of the likelihood function for the model. The general AIC formula is given by

$$AIC = 2k - 2\ln(\hat{L}).$$

Therefore, our modified AIC formula under the DPD setting is

$$AIC(\lambda_n) = n\log(n\hat{\sigma}^2) + 2df. \tag{3.12}$$

The preferred model is chosen based on the minimum AIC value, which penalizes models with more parameters, discouraging overfitting. By selecting the model with the lowest AIC, researchers can choose a model that achieves a good fit without being overly complex (Akaike 1973, 1974).

3.4.2 Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) is a highly renowned and extensively employed tool for statistical model selection. Its widespread usage is attributed to its computational ease and its effectiveness in various modeling frameworks, even in cases where prior distributions are not readily available (Neath & Cavanaugh 2012).

The BIC introduces a penalty for model complexity, preventing overfitting by favoring simpler models. It strikes a balance between complexity and accuracy, aiding model selection by identifying the most appropriate model. Widely used in statistics, the BIC's Bayesian principles offer a principled approach for selecting models in diverse fields.

Mathematically, the Bayesian information criterion (BIC) is derived as an asymptotic result under certain assumptions. Specifically, it is based on the assumption that the data

distribution follows the exponential family. The BIC formula incorporates this assumption and provides an approximation of the model's information loss in relation to the true underlying data distribution. Therefore, the most widely used BIC formula is given as

$$BIC = k\ln(n) - 2\ln(\hat{L}).$$

In our method, the BIC formula to be used is expressed as

$$BIC(\lambda_n) = n\log(n\hat{\sigma}^2) + df. \tag{3.13}$$

The BIC is a useful model selection tool, favoring models with lower values, indicating better fit or fewer variables. It penalizes additional parameters more strongly than AIC, helping researchers balance model fit and complexity for appropriate selection.

3.4.3 Mallows's Cp

Mallows's Cp, named after Colin Lingwood Mallows, is a statistical measure used to evaluate the goodness of fit of a regression model estimated using ordinary least squares. It is particularly useful in the context of model selection, where the objective is to identify the best model that includes a subset of available predictor variables for predicting an outcome. The Cp value provides an indication of the precision of the model, with a smaller value suggesting a more precise fit to the data. Interestingly, Mallows's Cp has been found to be equivalent to the Akaike information criterion (AIC) in the special case of Gaussian linear regression (Mallows 1973, Gilmour 1996, Boisbunon et al. 2013).

In an article published by Hocking (1976), the technique involves comparing a full model, which includes all the parameters, with a smaller model that includes only a subset of the parameters. It assesses the amount of error that remains unexplained by the smaller model. This is done by estimating the standardized total mean square of estimation for

the partial model using the following formula :

$$C_p = \frac{SSE_p}{S^2} - N + 2(P+1),$$

where;

- $SSE_p = \sum_{i=1}^{N}(Y_i - Y_{pi})^2$ is the error sum of squares for the model with P predictor variables,
- Y_{pi} is the predicted value of the ith observation of Y from the P regressors,
- S^2 = the residual mean square for the model (estimated by MSE)
- N is the sample size

Therefore, the Cp formula used in our model selection is given as

$$Cp(\lambda_n) = \frac{n\hat{\sigma}^2}{\hat{\sigma}_u^2} - n + 2df. \tag{3.14}$$

Mallows's Cp addresses the issue of overfitting, in which model selection statistics such as the residual sum of squares always get smaller as more variables are added to a model. Different interpretations have been proposed for the Mallows's Cp statistic, but the consensus is that smaller values indicate better model fit. A smaller Cp value suggests a smaller amount of unexplained error in the model.

3.4.4 Extended Bayesian Information Criterion (EBIC)

In the context of high-dimensional data analysis, the Extended Bayesian Information Criterion (EBIC) is a statistical criterion for model selection. The penalty of selecting models with a lot of variables is incorporated into this version of the Bayesian Information Criterion (BIC).

The BIC, for example, tends to prefer models with lots of variables in high-dimensional settings, which is one of the limits of other criteria. The EBIC is intended to overcome

these issues. To encourage sparsity and the selection of models with fewer variables, the EBIC penalizes model complexity more severely.

The EBIC is defined as:

$$EBIC(\lambda) = -2\ln(L) + n\ln(\hat{\sigma}^2) + 2s(\lambda, n),$$

where L represents the maximum likelihood estimation of the model, $\hat{\sigma}^2$ is the estimate of the error variance, n is the sample size, and $s(\lambda, n)$ is a penalty term that depends on the tuning parameter λ and the sample size n.

The penalty term $s(\lambda, n)$ is calculated as:

$$s(\lambda, n) = \lambda \left(p - \frac{1}{2}\ln(n) \right),$$

where p is the number of variables in the model and λ is a positive tuning parameter that controls the amount of penalization.

Therefore, the EBIC formula in our method is given as

$$EBIC(\lambda_n) = n\log(n\hat{\sigma}^2) + (\log n + \log p)df. \tag{3.15}$$

The EBIC is a useful criterion for model selection in high-dimensional data analysis, as it promotes sparsity and parsimony while accounting for model fit.

For instance, Chen & Chen (2008) in their article established that the EBIC are extremely useful for variable selection in problems with a moderate sample size but with a huge number of covariates, especially in genome-wide association studies, which are now an active area in genetics research.

Where $\hat{\sigma}$ is the estimate of σ obtained from the sub-model and $\hat{\sigma}_u$ is an unbiased and robust estimator of σ from the full model. The df is the degrees of freedom of the sub-model, that is, the dimension of non-zero β coefficients obtained from the group SCAD estimator. Therefore, for each case, we select an optimum λ_n that minimizes each information criterion.

Chapter 4

Simulation and Real Data Analysis

The penalized density power divergence (DPD) regression model is a promising approach in statistical analysis, offering robustness and flexibility. This study aims to evaluate the performance of the penalized DPD model through a comprehensive simulation study. The simulation results provide valuable insights into the potential applications of the penalized DPD regression in practical data analysis settings.

4.1 Simulation Results

We conducted a comprehensive simulation experiment to assess and validate the performance of various models using metrics such as Root Mean Prediction Error (RMPE), mean of Sensitivity, and Specificity. To assess the performance of the robust penalized DPD model, a synthetic data was generated to mimic real-world scenarios with varying distributions, noise levels, and data irregularities. The simulations were conducted using a penalized regression model with different numbers of predictors, including cases with zero coefficients. The simulations were replicated 100 times to ensure robustness and were performed for sample sizes of 100 and 500.

This approach allowed us to evaluate the effectiveness of our model under different data sizes and signal to noise ratio (SNR) values. By analyzing the simulation results, we gained valuable insights into the performance and reliability of the method in high-dimensional settings.

The data is generated from a linear regression model

$$Y_i = \beta_0 + \sum_{i=1}^{p} \beta_k X_{ik} + \epsilon_i$$

In assessing the robustness of our model, we introduced different levels of contamination in the data by adding outliers. The contamination proportions were set to 0%, 5%, and 10% for samples of size 100 and 500. The total number of predictors, excluding the intercept, was chosen as 15 and 25, with 7 and 20 predictors having zero coefficients (p_0).

In the process of evaluating the performance of the robust model under various data sizes, we selected different signal-to-noise ratio (SNR) values: 0.5, 1, and 5 for samples of size 100 and 500, respectively. Each simulation was replicated 100 times to ensure reliable results. The models were evaluated based on their Root Mean Prediction Error (RMPE), as well as the sensitivity and specificity metrics.

By considering these factors, we were able to gain insights into the effectiveness and robustness of the penalized DPD regression model in different scenarios, providing a comprehensive evaluation of its performance in high-dimensional settings.

4.1.1 RMPE, Sensitivity and Specificity of Simulated data

In this section, we conducted an analysis of the Root Mean Prediction Error (RMPE) for various estimators. The focus of our investigation was on sample sizes of 100 and 500, with varying numbers of predictors (15 and 25, excluding the intercept) and predictors with zero coefficients (7 and 20). To understand the impact of outliers, we considered different proportions (0%, 5%, and 10%) and signal-to-noise ratios (0.5, 1, and 5).

Our objective was to examine how the RMPE of the models changed as the proportion of outliers was manipulated. By altering this proportion, we aimed to gain insights into the models' performance under different outlier scenarios. Additionally, we evaluated the average sensitivity and specificity of the simulated data, providing a comprehensive assessment of the models' predictive capabilities across multiple categories.

The Table 4.1 gives the results for RMPE, the mean of sensitivity and specificity based on the following settings: $n = 100$, $p = 15$, $p_0 = 7$, $\epsilon = 0$, and SNR $= 0.5$. The RMPE values for most estimators are generally low, indicating good predictive performance. However, among the estimators, **LASSO** estimator stands out with a lower RMPE of **1.1437** and the highest sensitivity value of **0.9563** indicating a remarkable accuracy of **95.63%** in correctly identifying positive instances. On the other hand, **LADlasso** obtained the highest specificity value of **0.9900**, demonstrating a strong ability to accurately identify negative instances at a rate of **99%**. Overall, LASSO demonstrated superior predictive performance and effectively captured positive cases, while LADlasso exhibited robustness in detecting negative cases.

Table 4.1: Results for n = 100, SNR = 0.5, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.1678	-	-
Huber	1.1766	-	-
Tukey	1.1786	-	-
LAD	1.2373	-	-
LASSO	**1.1437**	**0.9563**	0.4014
LADlasso	1.4893	0.1125	**0.9900**
AIC(DPD)	1.1795	0.8850	0.6514
BIC(DPD)	1.2299	0.7538	0.8071
EBIC(DPD)	1.3641	0.3838	0.9500
Cp(DPD)	1.1782	0.8875	0.6443

In Table 4.2, with an introduction of a 5% contamination, the **Tukey** estimator performs better in terms of the RMPE with a value of **1.1779**. The table again presents the sensitivity and specificity values for the other models where **Cp(DPD)** achieved the highest sensitivity value of **0.7738**, indicating a remarkable accuracy of **77.38%** in correctly identifying positive instances. On the other hand, **LADlasso** obtained the highest specificity value of **0.9886**, demonstrating a strong ability to accurately identify negative instances at a rate of **98.86%**.

Table 4.2: Results for n = 100, SNR = 0.5, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.8102	-	-
Huber	1.2711	-	-
Tukey	**1.1779**	-	-
LAD	1.2964	-	-
LASSO	1.5695	0.5700	0.6314
LADlasso	1.4697	0.1400	**0.9886**
AIC(DPD)	1.4485	0.7550	0.6886
BIC(DPD)	1.4324	0.5925	0.8514
EBIC(DPD)	1.4665	0.3238	0.9400
Cp(DPD)	1.4393	**0.7738**	0.6686

Table 4.3: Results for n = 100, SNR = 0.5, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	2.5979	-	-
Huber	1.4856	-	-
Tukey	**1.2202**	-	-
LAD	1.3878	-	-
LASSO	2.0429	0.4050	0.7757
LADlasso	1.4924	0.1475	**0.9871**
AIC(DPD)	2.0058	0.4538	0.7929
BIC(DPD)	1.8032	0.3113	0.8957
EBIC(DPD)	1.7771	0.2600	0.9386
Cp(DPD)	1.9969	**0.4988**	0.7643

Comparing the simulation results in Table 4.3 to those in Tables 4.1 and 4.2, we observe that the RMPE of the estimators increases as the percentage of outliers rises, leading to a significant decrease in sensitivity values. Among the estimators, the **Tukey** estimator achieved the lowest RMPE of **1.2202**, while the **Cp(DPD)** estimator demonstrated a higher sensitivity value of **0.4988**. Additionally, concerning specificity, **LADlasso** obtained the highest value of **0.9871**, indicating exceptional accuracy in its predictions

Table 4.4: Results for n = 100, SNR = 1, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.1814	-	-
Huber	1.1902	-	-
Tukey	1.1922	-	-
LAD	1.3213	-	-
LASSO	**1.1531**	**0.9963**	0.3500
LADlasso	1.9843	0.1238	**0.9971**
AIC(DPD)	1.1655	0.9825	0.6986
BIC(DPD)	1.1778	0.9588	0.8343
EBIC(DPD)	1.2379	0.8913	0.8929
Cp(DPD)	1.1634	0.9838	0.7000

Table 4.5: Results for n = 100, SNR = 1, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	2.4773	-	-
Huber	1.2970	-	-
Tukey	**1.1923**	-	-
LAD	1.3861	-	-
LASSO	2.0866	0.6413	0.6057
LADlasso	1.9206	0.1488	**0.9986**
AIC(DPD)	1.2810	0.9650	0.7557
BIC(DPD)	1.2881	0.9400	0.8571
EBIC(DPD)	1.3208	0.8913	0.8943
Cp(DPD)	1.2805	**0.9663**	0.7486

Table 4.6: Results for n = 100, SNR = 1, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	4.0343	-	-
Huber	1.5242	-	-
Tukey	**1.2258**	-	-
LAD	1.4906	-	-
LASSO	3.0783	0.4313	0.7471
LADlasso	1.9343	0.1700	**0.9943**
AIC(DPD)	2.1809	0.6763	0.8171
BIC(DPD)	2.0666	0.6000	0.9029
EBIC(DPD)	2.0738	0.5600	0.9229
Cp(DPD)	2.3841	**0.7663**	0.7143

In Tables 4.4 to 4.6, the estimators generally exhibit low RMPE values. However, the **LASSO** estimator outperforms others with a value of **1.1531** and a higher sensitivity of **0.9963** when the data is uncontaminated. In the presence of 5% and 10% data contamination, the **Tukey** estimator stands out with low RMPE values of **1.1923** and **1.2258**, respectively.

Regarding sensitivity and specificity values, the **Cp(DPD)** estimator achieves the highest sensitivity values of **0.9663** and **0.7663** in Tables 4.5 and 4.6, respectively, showcasing its remarkable accuracy in identifying positive instances. On the other hand, **LADlasso** consistently obtains higher specificity values, demonstrating its strong ability to accurately identify negative instances at rates of **99.71%**, **99.86%**, and **99.43%** across all tables.

Table 4.7: Results for n = 100, SNR = 5, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.2926	-	-
Huber	1.3013	-	-
Tukey	1.3036	-	-
LAD	1.6938	-	-
LASSO	1.2557	**1.0000**	0.3429
LADlasso	5.4415	0.1950	**1.0000**
AIC(DPD)	1.2501	**1.0000**	0.8114
BIC(DPD)	1.2212	**1.0000**	0.9586
EBIC(DPD)	**1.2118**	**1.0000**	0.9871
Cp(DPD)	1.2501	**1.0000**	0.8029

Table 4.8: Results for n = 100, SNR = 5, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	7.7919	-	-
Huber	1.4603	-	-
Tukey	1.3116	-	-
LAD	1.9059	-	-
LASSO	6.1379	0.6888	0.5671
LADlasso	5.4112	0.1788	**1.0000**
AIC(DPD)	1.2842	**1.0000**	0.8129
BIC(DPD)	1.2577	**1.0000**	0.9457
EBIC(DPD)	**1.2374**	**1.0000**	0.9929
Cp(DPD)	1.2856	**1.0000**	0.8000

Table 4.9: Results for n = 100, SNR = 5, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	15.4891	-	-
Huber	1.7410	-	-
Tukey	**1.3422**	-	-
LAD	2.0611	-	-
LASSO	2.0612	0.4588	0.7357
LADlasso	5.2521	0.2125	**0.9986**
AIC(DPD)	3.5122	0.8788	0.7729
BIC(DPD)	3.1631	0.8613	0.9600
EBIC(DPD)	3.0245	0.8450	0.9843
Cp(DPD)	4.1276	**0.9075**	0.7286

Tables 4.7 to 4.8 revealed that the robust estimator, **EBIC(DPD)**, outperformed other methods in terms of RMPE, achieving values of **1.2118** and **1.2374**, respectively. All DPD-based robust estimators and the LASSO estimator achieved **100%** accuracy in sensitivity, while the **LADlasso** showed **100%** accuracy in specificity. In Table 4.9, **Cp(DPD)** exhibited the highest success rate of **0.9075** in identifying positive cases, with **Tukey** having the lowest RMPE.

In this simulation study, we again evaluate the Root Mean Prediction Error (RMPE), the mean sensitivity, and the mean specificity for various parameter settings. The parameters used are as follows: sample size (n) = 500, total predictors $p = 25$ (excluding the intercept), predictors with zero coefficients $p_0 = 20$, different proportions of outliers (ϵ), and signal-to-noise ratios (SNR). By analyzing these simulations, we aim to gain insights into the performance of different models under varied conditions, providing a comprehensive understanding of their predictive capabilities.

Table 4.10: Results for n = 500, SNR = 0.5, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.0029	-	-
Huber	1.0063	-	-
Tukey	1.0064	-	-
LAD	1.0826	-	-
LASSO	0.9657	**1.0000**	0.6625
LADlasso	1.0036	0.8840	**0.9970**
AIC(DPD)	0.9757	**1.0000**	0.8020
BIC(DPD)	**0.9645**	**1.0000**	0.9595
EBIC(DPD)	0.9653	**1.0000**	0.9700
Cp(DPD)	0.9757	**1.0000**	0.8020

Table 4.11: Results for n = 500, SNR = 0.5, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.1874	-	-
Huber	1.0162	-	-
Tukey	1.0080	-	-
LAD	1.0957	-	-
LASSO	1.1002	0.9860	0.6765
LADlasso	1.0419	0.8120	**0.9880**
AIC(DPD)	1.0188	**0.9960**	0.8195
BIC(DPD)	**0.9983**	**0.9960**	0.9520
EBIC(DPD)	0.9986	0.9940	0.9690
Cp(DPD)	1.0218	**0.9960**	0.8015

Table 4.12: Results for n = 500, SNR = 0.5, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.5680	-	-
Huber	1.0546	-	-
Tukey	**1.0118**	-	-
LAD	1.1301	-	-
LASSO	1.4238	**0.9560**	0.6760
LADlasso	1.0970	0.746	**0.9840**
AIC(DPD)	1.1730	0.9380	0.8830
BIC(DPD)	1.1747	0.8780	0.9350
EBIC(DPD)	1.1841	0.8380	0.9545
Cp(DPD)	1.1733	0.9480	0.8380

In Tables 4.10 to 4.12, we present the results based on a new sample size. Our robust estimator with DPD (**BIC(DPD)**) demonstrated superior performance in terms of RMPE, achieving values of **0.9645** and **0.9983**, respectively. As for sensitivity, **AIC(DPD)**, **BIC(DPD)**, and **Cp(DPD)** all exhibited the same high value of **0.9960**, representing **99.60%** accuracy in identifying positive cases. On the other hand, the **LADlasso** estimator maintained its superiority in accurately identifying negative cases. These findings shed light on the robustness and reliability of the various estimators under different conditions.

Table 4.13: Results for n = 500, SNR = 1, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.0063	-	-
Huber	1.0098	-	-
Tukey	1.0099	-	-
LAD	1.1623	-	-
LASSO	0.9689	**1.0000**	0.6665
LADlasso	1.0148	0.9760	**0.9985**
AIC(DPD)	0.9777	**1.0000**	0.8165
BIC(DPD)	**0.9635**	**1.0000**	0.9870
EBIC(DPD)	0.9643	**1.0000**	0.9920
Cp(DPD)	0.9776	**1.0000**	0.8175

Table 4.14: Results for n = 500, SNR = 1, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.3775	-	-
Huber	1.0148	-	-
Tukey	1.0113	-	-
LAD	1.1791	-	-
LASSO	1.2382	0.9940	0.6790
LADlasso	1.0891	0.9300	0.9930
AIC(DPD)	1.0189	**1.0000**	0.7540
BIC(DPD)	0.9865	**1.0000**	0.9870
EBIC(DPD)	**0.9855**	**1.0000**	**0.9935**
Cp(DPD)	1.0199	**1.0000**	0.7475

Table 4.15: Results for n = 500, SNR = 1, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	2.1449	-	-
Huber	1.0480	-	-
Tukey	**1.0153**	-	-
LAD	1.2181	-	-
LASSO	1.8983	0.9640	0.6790
LADlasso	1.1996	0.8380	0.9825
AIC(DPD)	1.0674	**0.9980**	0.7435
BIC(DPD)	1.0331	**0.9980**	0.9805
EBIC(DPD)	1.0373	0.9900	**0.9915**
Cp(DPD)	1.0679	**0.9980**	0.7365

In Tables 4.13 to 4.14, we observed that the **BIC(DPD)** and **EBIC(DPD)** estimators achieved lower RMPE values of **0.9635** and **0.9855**, respectively, compared to other estimators when the data was both uncontaminated and contaminated at 5%. Furthermore, all **DPD** estimators demonstrated perfect sensitivity values, scoring **100%**. Notably, in Tables 4.14 and 4.15, **EBIC(DPD)** outperformed **LADlasso** in terms of specificity, achieving rates of **99.35%** and **99.15%**, respectively. These results highlight the superior perfor-

mance and robustness of the **BIC(DPD)** and **EBIC(DPD)** estimators under different data conditions.

Table 4.16: Results for n = 500, SNR = 5, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.0345	-	-
Huber	1.0382	-	-
Tukey	1.0382	-	-
LAD	1.8439	-	-
LASSO	0.9956	**1.0000**	0.6645
LADlasso	1.2049	**1.0000**	**1.0000**
AIC(DPD)	1.0056	**1.0000**	0.8175
BIC(DPD)	0.9900	**1.0000**	0.9960
EBIC(DPD)	**0.9897**	**1.0000**	0.9995
Cp(DPD)	1.0056	**1.0000**	0.8175

Table 4.17: Results for n = 500, SNR = 5, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	2.8938	-	-
Huber	1.0230	-	-
Tukey	1.0393	-	-
LAD	1.9024	-	-
LASSO	2.3577	**1.0000**	0.6855
LADlasso	1.5305	**1.0000**	0.9915
AIC(DPD)	1.0203	**1.0000**	0.7765
BIC(DPD)	0.9968	**1.0000**	0.9935
EBIC(DPD)	**0.9961**	**1.0000**	**0.9995**
Cp(DPD)	1.0202	**1.0000**	0.7790

Table 4.18: Results for n = 500, SNR = 5, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	6.7639	-	-
Huber	1.0339	-	-
Tukey	1.0446	-	-
LAD	1.9843	-	-
LASSO	5.6969	0.9780	0.6755
LADlasso	1.9802	0.9840	0.9670
AIC(DPD)	1.0383	**1.0000**	0.7840
BIC(DPD)	1.0145	**1.0000**	0.9980
EBIC(DPD)	**1.0141**	**1.0000**	**1.0000**
Cp(DPD)	1.0376	**1.0000**	0.7875

The **EBIC(DPD)** estimator consistently outperformed other estimators, achieving lower RMPE values (**0.9897**, **0.9961**, and **1.0141**) under uncontaminated and contaminated data scenarios at 5% and 10% levels. It demonstrated perfect sensitivity (**100%**) and superior specificity (**99.95%** and **100%**) compared to **LADlasso** in Tables 4.16 to 4.18. These findings highlight the robustness and superior performance of the **EBIC(DPD)** estimator across various data conditions.

Table 4.19: Results for n = 100, SNR = 0.5, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.1678	-	-
Huber	1.1766	-	-
Tukey	1.1786	-	-
LAD	1.2373	-	-
LASSO	**1.1437**	**0.9563**	0.4014
LADlasso	1.4893	0.1125	**0.9900**
AIC(DPD)	1.1795	0.8850	0.6514
BIC(DPD)	1.2299	0.7538	0.8071
EBIC(DPD)	1.3641	0.3838	0.9500
Cp(DPD)	1.1782	0.8875	0.6443

Table 4.20: Results for n = 100, SNR = 1, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.1814	-	-
Huber	1.1902	-	-
Tukey	1.1922	-	-
LAD	1.3213	-	-
LASSO	**1.1531**	**0.9963**	0.3500
LADlasso	1.9843	0.1238	**0.9971**
AIC(DPD)	1.1655	0.9825	0.6986
BIC(DPD)	1.1778	0.9588	0.8343
EBIC(DPD)	1.2379	0.8913	0.8929
Cp(DPD)	1.1634	0.9838	0.7000

Table 4.21: Results for n = 100, SNR = 5, $\epsilon = 0$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.2926	-	-
Huber	1.3013	-	-
Tukey	1.3036	-	-
LAD	1.6938	-	-
LASSO	1.2557	**1.0000**	0.3429
LADlasso	5.4415	0.1950	**1.0000**
AIC(DPD)	1.2501	**1.0000**	0.8114
BIC(DPD)	1.2212	**1.0000**	0.9586
EBIC(DPD)	**1.2118**	**1.0000**	0.9871
Cp(DPD)	1.2501	**1.0000**	0.8029

The Tables 4.19 to 4.21 above gives the results for different SNR values when data is uncontaminated with $p = 15$ and $p_0 = 7$. **EBIC(DPD)** performed better in terms of RMPE (**1.2118**) with **SNR = 5** with all the DPD methods perfectly predicting positive cases.

Table 4.22: Results for n = 100, SNR = 0.5, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.8102	-	-
Huber	1.2711	-	-
Tukey	**1.1779**	-	-
LAD	1.2964	-	-
LASSO	1.5695	0.5700	0.6314
LADlasso	1.4697	0.1400	**0.9886**
AIC(DPD)	1.4485	0.7550	0.6886
BIC(DPD)	1.4324	0.5925	0.8514
EBIC(DPD)	1.4665	0.3238	0.9400
Cp(DPD)	1.4393	**0.7738**	0.6686

Table 4.23: Results for n = 100, SNR = 1, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	2.4773	-	-
Huber	1.2970	-	-
Tukey	**1.1923**	-	-
LAD	1.3861	-	-
LASSO	2.0866	0.6413	0.6057
LADlasso	1.9206	0.1488	**0.9986**
AIC(DPD)	1.2810	0.9650	0.7557
BIC(DPD)	1.2881	0.9400	0.8571
EBIC(DPD)	1.3208	0.8913	0.8943
Cp(DPD)	1.2805	**0.9663**	0.7486

Table 4.24: Results for n = 100, SNR = 5, $\epsilon = 0.05$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	7.7919	-	-
Huber	1.4603	-	-
Tukey	1.3116	-	-
LAD	1.9059	-	-
LASSO	6.1379	0.6888	0.5671
LADlasso	5.4112	0.1788	**1.0000**
AIC(DPD)	1.2842	**1.0000**	0.8129
BIC(DPD)	1.2577	**1.0000**	0.9457
EBIC(DPD)	**1.2374**	**1.0000**	0.9929
Cp(DPD)	1.2856	**1.0000**	0.8000

The **EBIC(DPD)** estimator again had a lower RMPE value of **1.2374** in Table 4.24 and **Cp(DPD)** consistently predicting higher values of sensitivity in in all the three scenarios of the SNR.

Table 4.25: Results for n = 100, SNR = 0.5, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	2.5979	-	-
Huber	1.4856	-	-
Tukey	**1.2202**	-	-
LAD	1.3878	-	-
LASSO	2.0429	0.4050	0.7757
LADlasso	1.4924	0.1475	**0.9871**
AIC(DPD)	2.0058	0.4538	0.7929
BIC(DPD)	1.8032	0.3113	0.8957
EBIC(DPD)	1.7771	0.2600	0.9386
Cp(DPD)	1.9969	**0.4988**	0.7643

Table 4.26: Results for n = 100, SNR = 1, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	4.0343	-	-
Huber	1.5242	-	-
Tukey	**1.2258**	-	-
LAD	1.4906	-	-
LASSO	3.0783	0.4313	0.7471
LADlasso	1.9343	0.1700	**0.9943**
AIC(DPD)	2.1809	0.6763	0.8171
BIC(DPD)	2.0666	0.6000	0.9029
EBIC(DPD)	2.0738	0.5600	0.9229
Cp(DPD)	2.3841	**0.7663**	0.7143

Table 4.27: Results for n = 100, SNR = 5, $\epsilon = 0.1$, p = 15, $p_0 = 7$

Estimators	RMPE	Sensitivity	Specificity
OLS	15.4891	-	-
Huber	1.7410	-	-
Tukey	**1.3422**	-	-
LAD	2.0611	-	-
LASSO	2.0612	0.4588	0.7357
LADlasso	5.2521	0.2125	**0.9986**
AIC(DPD)	3.5122	0.8788	0.7729
BIC(DPD)	3.1631	0.8613	0.9600
EBIC(DPD)	3.0245	0.8450	0.9843
Cp(DPD)	4.1276	**0.9075**	0.7286

Table 4.28: Results for n = 500, SNR = 0.5, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.0029	-	-
Huber	1.0063	-	-
Tukey	1.0064	-	-
LAD	1.0826	-	-
LASSO	0.9657	**1.0000**	0.6625
LADlasso	1.0036	0.8840	**0.9970**
AIC(DPD)	0.9757	**1.0000**	0.8020
BIC(DPD)	**0.9645**	**1.0000**	0.9595
EBIC(DPD)	0.9653	**1.0000**	0.9700
Cp(DPD)	0.9757	**1.0000**	0.8020

Table 4.29: Results for n = 500, SNR = 1, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.0063	-	-
Huber	1.0098	-	-
Tukey	1.0099	-	-
LAD	1.1623	-	-
LASSO	0.9689	**1.0000**	0.6665
LADlasso	1.0148	0.9760	**0.9985**
AIC(DPD)	0.9777	**1.0000**	0.8165
BIC(DPD)	**0.9635**	**1.0000**	0.9870
EBIC(DPD)	0.9643	**1.0000**	0.9920
Cp(DPD)	0.9776	**1.0000**	0.8175

Table 4.30: Results for n = 500, SNR = 5, $\epsilon = 0$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.0345	-	-
Huber	1.0382	-	-
Tukey	1.0382	-	-
LAD	1.8439	-	-
LASSO	0.9956	**1.0000**	0.6645
LADlasso	1.2049	**1.0000**	**1.0000**
AIC(DPD)	1.0056	**1.0000**	0.8175
BIC(DPD)	0.9900	**1.0000**	0.9960
EBIC(DPD)	**0.9897**	**1.0000**	0.9995
Cp(DPD)	1.0056	**1.0000**	0.8175

Table 4.31: Results for n = 500, SNR = 0.5, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	1.1874	-	-
Huber	1.0162	-	-
Tukey	1.0080	-	-
LAD	1.0957	-	-
LASSO	1.1002	0.9860	0.6765
LADlasso	1.0419	0.8120	**0.9880**
AIC(DPD)	1.0188	**0.9960**	0.8195
BIC(DPD)	**0.9983**	**0.9960**	0.9520
EBIC(DPD)	0.9986	0.9940	0.9690
Cp(DPD)	1.0218	**0.9960**	0.8015

Table 4.32: Results for n = 500, SNR = 1, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.3775	-	-
Huber	1.0148	-	-
Tukey	1.0113	-	-
LAD	1.1791	-	-
LASSO	1.2382	0.9940	0.6790
LADlasso	1.0891	0.9300	0.9930
AIC(DPD)	1.0189	**1.0000**	0.7540
BIC(DPD)	0.9865	**1.0000**	0.9870
EBIC(DPD)	**0.9855**	**1.0000**	**0.9935**
Cp(DPD)	1.0199	**1.0000**	0.7475

Table 4.33: Results for n = 500, SNR = 5, $\epsilon = 0.05$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	2.8938	-	-
Huber	1.0230	-	-
Tukey	1.0393	-	-
LAD	1.9024	-	-
LASSO	2.3577	**1.0000**	0.6855
LADlasso	1.5305	**1.0000**	0.9915
AIC(DPD)	1.0203	**1.0000**	0.7765
BIC(DPD)	0.9968	**1.0000**	0.9935
EBIC(DPD)	**0.9961**	**1.0000**	**0.9995**
Cp(DPD)	1.0202	**1.0000**	0.7790

Table 4.34: Results for n = 500, SNR = 0.5, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	1.5680	-	-
Huber	1.0546	-	-
Tukey	**1.0118**	-	-
LAD	1.1301	-	-
LASSO	1.4238	**0.9560**	0.6760
LADlasso	1.0970	0.746	**0.9840**
AIC(DPD)	1.1730	0.9380	0.8830
BIC(DPD)	1.1747	0.8780	0.9350
EBIC(DPD)	1.1841	0.8380	0.9545
Cp(DPD)	1.1733	0.9480	0.8380

Table 4.35: Results for n = 500, SNR = 1, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	**RMPE**	**Sensitivity**	**Specificity**
OLS	2.1449	-	-
Huber	1.0480	-	-
Tukey	**1.0153**	-	-
LAD	1.2181	-	-
LASSO	1.8983	0.9640	0.6790
LADlasso	1.1996	0.8380	0.9825
AIC(DPD)	1.0674	**0.9980**	0.7435
BIC(DPD)	1.0331	**0.9980**	0.9805
EBIC(DPD)	1.0373	0.9900	**0.9915**
Cp(DPD)	1.0679	**0.9980**	0.7365

Table 4.36: Results for n = 500, SNR = 5, $\epsilon = 0.1$, p = 25, $p_0 = 20$

Estimators	RMPE	Sensitivity	Specificity
OLS	6.7639	-	-
Huber	1.0339	-	-
Tukey	1.0446	-	-
LAD	1.9843	-	-
LASSO	5.6969	0.9780	0.6755
LADlasso	1.9802	0.9840	0.9670
AIC(DPD)	1.0383	**1.0000**	0.7840
BIC(DPD)	1.0145	**1.0000**	0.9980
EBIC(DPD)	**1.0141**	**1.0000**	**1.0000**
Cp(DPD)	1.0376	**1.0000**	0.7875

Tables 4.28 to 4.33 show that with $n = 500$, both **BIC(DPD)** and **EBIC(DPD)** outperformed other estimators, predicting lower values of RMPE. Also, all DPD methods demonstrated higher sensitivity rates in identifying positive cases. Table 4.36 highlights that **EBIC(DPD)** is the only estimator with a lower RMPE of **1.0141** and achieves **100%** sensitivity and specificity rates.

4.2 Real Data Analysis

Data Source and Description

A dataset on Bias correction of numerical prediction model temperature forecast is obtained from UCI which has 7752 observations across 25 variables, with two different response variables. The study involves two distinct output variables: the next-day maximum air temperature and the next-day minimum air temperature. Among the predictor variables are parameters such as the maximum air temperature between 0 and 21 hours on the present day, the minimum air temperature between 0 and 21 hours on the present day, date, weather station number, latitude, longitude, slope, elevation, daily incoming solar radiation, and more. All variables, with the exception of the date, are continuous in nature.

4.2.1 Application of Real Data on Models

We employed a robust penalized DPD regression model with an SCAD penalty on the bias correction dataset. Our analysis focused on utilizing the next-day maximum air temperature as the primary response variable. This modeling process was conducted on a subset of the data, specifically 513 randomly selected observations from a total of 7752. In our analysis, we excluded variables such as weather station number, date, and the next-day minimum air temperature.

The methodology encompasses two key steps: initially estimating regression coefficients through the DPD method and subsequently incorporating them into the SCAD penalty structure. This integration results in the derivation of a robust SCAD penalty. The resultant model proficiently shrinks coefficients associated with irrelevant parameters to zero, effectively achieving dimension reduction.

The construction of the model transpires on a training dataset, with subsequent validation carried out on a separate testing dataset, benchmarked by the Root Mean Prediction Error (RMPE). To evaluate the model's performance, a comparative analysis is conducted against other robust estimators such as Tukey, Huber, LAD, and LAD-LASSO.

Furthermore, on the validation of the real-world data, we compute the mean of the dimension reduction across our DPD methods and alternative techniques (LASSO and LAD-LASSO). This computation serves as a yardstick to determine the model's efficacy in driving irrelevant coefficients to zero, thereby ascertaining its superiority.

4.2.2 RMPE and Mean Dimension Reduction of the Models

The Root Mean Prediction Error (RMPE) serves as a crucial tool in research for evaluating the accuracy and efficacy of various predictive models. To gauge the effectiveness of our approach, we compare its RMPE against that of robust models like Tukey, Huber, LAD, and LAD-LASSO. Additionally, as our focus lies in dimension reduction, we ascertain the merit of our method by juxtaposing it with LASSO and LAD-LASSO. Upon analysis,

our robust DPD methods (**AIC(DPD)**, **BIC(DPD)**, and **EBIC(DPD)**) consistently yield small RMPE values. While the other robust estimators, Huber and Tukey exhibit slightly smaller values than our method, the distinctions between these models are relatively subtle. In the context of dimension reduction, **Cp(DPD)** estimator showcases superior performance compared to its counterparts, showcasing adept variable selection capabilities. To encapsulate the outcomes, the summarized table outlines the results of RMPE and the mean dimensions after reduction.

Table 4.37: RMPE and Mean Dimension Reduction of Models

Estimators	**RMPE**	**Mean Dimension Reduction**
OLS	1.3755	-
Huber	**1.3426**	-
Tukey	**1.3311**	-
LAD	**2.9956**	-
LASSO	1.3847	14.2857
LADlasso	**1.6406**	71.4286
AIC(DPD)	**1.4122**	19.0476
BIC(DPD)	**1.4122**	19.0476
EBIC(DPD)	**1.4122**	19.0476
Cp(DPD)	5.1922	**97.6191**

Chapter 5

Conclusion

In this research endeavor, we have introduced the Robust Density Power Divergence Regression with SCAD penalty as a proficient approach for both dimension reduction and resilience against outliers. Our investigation encompassed a series of simulation studies designed to juxtapose our method against existing alternatives, utilizing three distinct metrics: Root Mean Prediction Error (RMPE), sensitivity, and specificity.

Throughout the diverse range of simulations carried out, our *BIC(DPD)* and *EBIC(DPD)* techniques consistently demonstrated the lowest RMPE values across scenarios encompassing outlier proportions of 0%, 5%, and 10%. These findings held steady across varying signal-to-noise ratio values of 0.5, 1, and 5, particularly as the sample size expanded to 500. Regarding sensitivity metrics, the *Cp(DPD)* approach consistently showcased strong performance, consistently achieving higher rates of accurate identification.

Subsequently, we utilized data from the UCI dataset to assess the performance of our methodology. During this evaluation, we conducted a comparative analysis by measuring the Root Mean Prediction Error (RMPE) values of our robust model estimators against other robust techniques, including Huber, Tukey, LAD, and LAD-LASSO. This comparison showcased the robustness of our DPD estimator with SCAD penalty, demonstrating its efficacy in handling outlier impact.

Acknowledgment: I would like to express my sincere gratitude to the collaborators of Dr. Abhijit Mandal for providing me with their invaluable research proposal, which served as the foundation for my thesis. Their contribution to developing the statistical algorithms (based on the DPD measure) and R codes was critical to the success of this project. I do

not claim any copywriting for these methods and the corresponding codes, as it was solely created by Dr. Mandal and his team.

Bibliography

Akaike, H. (1973), 'Information theory and an extension of the maximum likelihood principle pp. 267-281 in 2nd international symposium on information theory, edited by bn petriv and f', *Csaki, Akademia Kiado, Budapest* .

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Almetwally, E. M. & Almongy, H. (2018), 'Comparison between m estimation, s estimation, and mm estimation methods of robust estimation with application and simulation', *International Journal of Mathematical Archive* **9**(11), 1–9.

Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. (1998), 'Robust and efficient estimation by minimising a density power divergence', *Biometrika* **85**(3), 549–559.

Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W. & Wells, M. T. (2013), 'Aic, cp and estimators of loss for elliptically symmetric distributions', *arXiv preprint arXiv:1308.2766* .

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chen, J. & Chen, Z. (2008), 'Extended bayesian information criteria for model selection with large model spaces', *Biometrika* **95**(3), 759–771.

Chen, K., Ying, Z., Zhang, H. & Zhao, L. (2008), 'Analysis of least absolute deviation', *Biometrika* **95**(1), 107–122.

Dasgupta, M. & Mishra, S. K. (2004), 'Least absolute deviation estimation of linear econometric models: A literature review', *Available at SSRN 552502* .

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *biometrika* **81**(3), 425–455.

Durio, A. & Isaia, E. D. (2011), 'The minimum density power divergence approach in building robust regression models', *Informatica* **22**(1), 43–56.

Elsaied, H. & Fried, R. (2016), 'Tukey's m-estimator of the poisson parameter with a special focus on small means', *Statistical Methods & Applications* **25**, 191–209.

Emmert-Streib, F. & Dehmer, M. (2019), 'High-dimensional lasso-based computational regression models: regularization, shrinkage, and selection', *Machine Learning and Knowledge Extraction* **1**(1), 359–383.

Fan, J. (1997), 'Comments on «wavelets in statistics: A review» by a. antoniadis', *Journal of the Italian Statistical Society* **6**(2), 131.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

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

Ghosh, A. & Basu, A. (2016), 'Robust estimation in generalized linear models: the density power divergence approach', *Test* **25**, 269–290.

Gilmour, S. G. (1996), 'The interpretation of mallows's cp-statistic', *Journal of the Royal Statistical Society Series D: The Statistician* **45**(1), 49–56.

Gokcesu, K. & Gokcesu, H. (2021), 'Generalized huber loss for robust learning and its efficient minimization for a robust statistics', *arXiv preprint arXiv:2108.12627* .

Hocking, R. R. (1976), 'A biometrics invited paper. the analysis and selection of variables in linear regression', *Biometrics* pp. 1–49.

Huber, P. J. (1964), 'Robust Estimation of a Location Parameter', *The Annals of Mathematical Statistics* **35**(1), 73 – 101.
URL: *https://doi.org/10.1214/aoms/1177703732*

Johnstone, I. M. & Titterington, D. M. (2009), 'Statistical challenges of high-dimensional data'.

Li, Y. & Arce, G. R. (2004), 'A maximum likelihood approach to least absolute deviation regression', *EURASIP Journal on Advances in Signal Processing* **2004**, 1–8.

Luo, B. (2020), *Robust Penalized Regression for Complex High-Dimensional Data*, The University of North Carolina at Greensboro.

Mallows, C. L. (1973), 'Some comments on cp', *Technometrics* **15**(4), 661–675.
URL: *http://www.jstor.org/stable/1267380*

Nahar, J. & Purwani, S. (2017), 'Application of robust m-estimator regression in handling data outliers', *4th ICRIEMS Proceedings* pp. 53–60.

Neath, A. A. & Cavanaugh, J. E. (2012), 'The bayesian information criterion: background, derivation, and applications', *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(2), 199–203.

Rahardiantoro, S. & Kurnia, A. (2015), Lad-lasso: Simulation study of robust regression in high dimensional data, *in* 'Forum Statistika dan Komputasi', Vol. 20.

Ranstam, J. & Cook, J. (2018), 'Lasso regression', *Journal of British Surgery* **105**(10), 1348–1348.

Riani, M., Atkinson, A. C., Corbellini, A. & Perrotta, D. (2020), 'Robust regression with density power divergence: Theory, comparisons, and data analysis', *Entropy* **22**(4), 399.

Signorino, C. S. & Kirchner, A. (2018), 'Using lasso to model interactions and nonlinearities in survey data', *Survey Practice* **11**(1).

Thanoon, F. H. (2015), 'Robust regression by least absolute deviations method', *International Journal of Statistics and Applications* **5**(3), 109–112.

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

Venkat, N. (2018), 'The curse of dimensionality: Inside out', *Pilani (IN): Birla Institute of Technology and Science, Pilani, Department of Computer Science and Information Systems* **10**.

Wang, H., Li, G. & Jiang, G. (2007), 'Robust regression shrinkage and consistent variable selection through the lad-lasso', *Journal of Business & Economic Statistics* **25**(3), 347–355.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429.

Curriculum Vitae

Maxwell Kwesi Mac-Ocloo, born on July 23, 1995, is the eldest child of Maxwell Kwesi Mac-Ocloo (Snr) and Victoria Amerley Akorley. After completing his secondary education at St. Thomas Aquinas High School in 2014, he pursued higher education at the University of Energy and Natural Resources (UENR) in Ghana. There, he enrolled in a four-year bachelor's degree program in Statistics. During his time at UENR, Maxwell actively participated in the Mathematical Science Student Association activities and achieved notable recognition. Upon completion of his studies, Maxwell received three prestigious awards, including the distinction of being the overall best graduating student of the 2019 class. Following his graduation, he served as a teaching assistant in the Mathematics and Statistics Department at UENR as part of his national service commitment. In the fall of 2021, Maxwell commenced his Master's program in Statistics and Data Science at The University of Texas at El Paso. Throughout his studies, he developed a strong foundation in statistics and machine learning, which further fueled his interest in the field of Biostatistics. As a teaching and research assistant, Maxwell actively contributed to the academic community and had the opportunity to collaborate on a research publication with Dr. Abhijit Mandal. With a clear career goal of becoming a senior biostatistician and data scientist in the health sector, Maxwell plans to pursue a doctorate program in biostatistics at the College of Human Medicine, Michigan State University. His dedication to advancing his knowledge and skills in the field of biostatistics reflects his commitment to making a meaningful impact on healthcare research and data analysis.

E-mail address: *mkmacocloo@miners.utep.edu*