Original Article

Mammalian eye gene expression using support vector regression to evaluate a strategy for detecting human eye disease

Mahdi Roozbeh[1*] Monireh Maanavi[2]

1. Associate Professor, Faculty of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran
2. MSc. of statistics, Social Determinants of Health Research Center, Semnan University of Medical Sciences, Semnan,Iran

***Correspondence to**: Mahdi Roozbeh
mahdi.roozbeh@semnan.ac.ir

(Received: 4 Jan. 2022; Revised: 12 Apr. 2022; Accepted: 7 May. 2022)

Abstract

***Background and purpose*:** Machine learning is a class of modern and strong tools that can solve many important problems that nowadays Humans may be faced with. Support Vector Regression (SVR) is a way to build a regression model which is an incredible member of the machine learning family. SVR has been proven to be an effective tool in real-value function estimation. As a supervised-learning approach, SVR trains using a symmetrical loss function, which equally penalizes high and low misestimates. Recently, high-dimensional datasets are the most challenging problem that may be faced. The main problems in high-dimensional data are the estimation of the coefficients and interpretation. In the high-dimension problems, classical methods are not applicable because of a large number of predictor variables. SVR is an excellent alternative method to analyze such datasets. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy.

Methods: SVR is one of the best methods to analyze high-dimensional datasets. It is a really reliable and robust approach to have a good fit with high accuracy. SVR uses the same principles as the support vector machine for classification, with only a few minor differences.

Results: The techniques for analyzing the high-dimension datasets are really important methods because we frequently face such datasets in medical science and gene expression. It is not easy to analyze the high-dimension datasets because the classic methods cannot be used to estimate and interpret them. Therefore, we have to use alternative methods to analyze them. SVR is one of the best methods that can be applied. In this research, SVR is used in a real high-dimension dataset about the gene expression in eye disease, and then it is compared with well-known methods: LASSO and Sparse least trimmed squared (sparse LTS) methods. Based on the numerical result, SVR and Sparse LTS were better than LASSO, since the real dataset contained outliers (bad observation with big residuals).

Conclusions: SVR method was the best method to model and predict the high-dimensional mammalian eye dataset, because it was not affected by the outliers' corruptive impact, and it has minimum MSE (mean squares error), MAE (mean absolute error) and RMSE (root mean squared error) fitting criteria in comparison with the classical methods such as LASSO and sparse LTS estimations. Thus, sparse LTS was found to act better than the LASSO method. Moreover, stabilization of the data and freedom from obtaining the regularization parameter by running a complicated algorithmic program, which decreased the computational costs dramatically, were the invaluable advantages of this technique in comparison with the classical methods.

***Keywords*:** High-dimensional data set; Ordinary least square method; Outliers; Robust regression

Citation: **Roozbeh M*,** Maanavi M. Mammalian eye gene expression using support vector regression to evaluate a strategy for detecting human eye disease. Iran J Health Sci. 2022; 10(2): 14-28.

[DOI: 10.18502/jhs.v10i2.9763]

1. *Introduction*

Machine learning theory is the broad framework for studying the concept of statistical inference. Inference covers the entire spectrum of machine learning, from gaining knowledge, making predictions or decisions and constructing models from a set of labeled or unlabeled data. The entire process is stated in a statistical framework, with every assumption stated mathematically as a null or alternative hypothesis. In other words, Machine Learning is a set of tools for understanding data. These tools broadly come under two classes: supervised learning and unsupervised learning. Generally, supervised learning refers to predicting or estimating an output based on one or more inputs. Unsupervised learning, on the other hand, provides a relationship or finds a pattern within the given data without a supervised output. Machine learning plays a key role in many areas of science, finance and industry. Here are some examples of learning problems (1):

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

The science of learning plays a key role in the fields of statistics, data mining, and artificial intelligence, intersecting with areas of engineering and other disciplines (1).

Machine learning involves predicting and classifying data, and in order to do so, we employ various machine learning algorithms according to the dataset. A support vector machine is a very important and versatile machine learning algorithm which is capable of doing linear and nonlinear classification, regression and outlier detection. Support vector machine also known as SVM is another algorithm widely used by machine learning people for both classification and regression problems, while it is widely used for classification tasks. It is preferred over other classification algorithms because it uses less computation and gives notable accuracy. It is good because it yields reliable results even if there is less data.

Support Vector Machine is a promising pattern classification technique proposed by Boser et al. (1992), Cortes and Vapnik (1995), and Vapnik (2013). Support vector machines have strong theoretical foundations and excellent empirical successes. SVM is a computer algorithm that learns by example to assign labels to objects. For instance, SVM can learn to recognize fraudulent credit card activity by examining hundreds or thousands of fraudulent and nonfraudulent credit card activity reports. Alternatively, SVM can learn to recognize handwritten digits by examining a large collection of scanned images of handwritten zeroes, ones and so forth. SVMs have also been successfully applied to an increasingly wide variety of biological applications.

[DOI: 10.18502/jhs.v10i2.9763]

A common biomedical application of SVMs is the automatic classification of microarray gene expression profiles. Theoretically, SVM can examine the gene expression profile derived from a tumor sample or from peripheral fluid and arrive at a diagnosis or prognosis. Other biological applications of SVMs involve classifying objects as diverse as protein and DNA sequences, microarray expression profiles, and mass spectra (5).

Training a support vector machine requires solving a quadratic programming (QP) problem in a number of coefficients equal to the number of training examples. In other words, a support vector machine is a mathematical entity, an algorithm (or recipe) for maximizing a particular mathematical function with respect to a given collection of data.

In general, we can divide SVMs into two parts:

- **Linear SVM:** Linear SVM is used for the data that are linearly separable i.e. for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data, and the classifier used is described as a Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for the data that are non-linearly separable, i.e. a straight line cannot be used to classify the dataset. For this, we use something known as a kernel trick that sets data points in a higher dimension where they can be separated using planes or other mathematical functions. Such data points are termed as non-linear data, and the classifier used is termed as a non-linear SVM classifier.

In both upper parts we have two kinds of datasets: dataset without overlap and dataset with overlap. You can see them in Figure 1 where the top plots are linear SVM, the left sides are the dataset without overlap, and the right sides are the dataset with overlap. For the datasets without overlap, we used hard margin SVM, and for datasets with overlap we used soft margin SVM to analyze them. In Figure 1, the bottom plots are non-linear SVM.

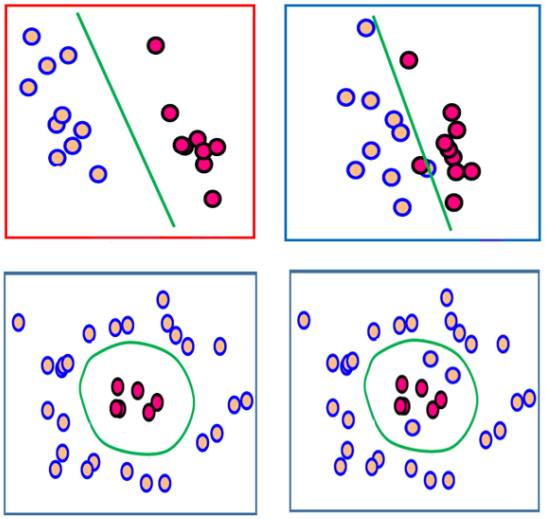

Figure 1. Types of data in SVM problems.

[DOI: 10.18502/jhs.v10i2.9763]

The most important types of datasets that we can face with are high-dimensional datasets. Nowadays, we can gather and record many features of a special subject easily, so we can build many high-dimensional datasets (6). A high-dimensional dataset is characterized by multiple dimensions. There can be thousands, if not millions, of dimensions. In other words, High-dimensional statistics focuses on datasets in which the number of features is of comparable size, or larger than the number of observations (6). Data sets of this type present a variety of new challenges, since classical theory and methodology can break down in surprising and unexpected ways. So, we have to find the alternative techniques to analyze them. It is so important how we behave them. Machine learning methods have powerful tools to analyze them, and SVM and support vector regression (SVR) methods are some of these amazing tools that can analyze the high-dimensional datasets very fast with the highest accuracy.

Notice that there is really important difference between SVM and SVR. Regression and Classification algorithms are supervised learning algorithms, both of which are used for prediction in machine learning and work with the labeled datasets. But the difference between them is about the ways they are used for different machine learning problems. The main difference between regression and classification algorithms is that the regression algorithms are used to predict the continuous variables, such as price, salary, age, etc. and classification algorithms are used to predict/classify the discrete variables, such as male or female, true or false, etc.

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. In fact, logistic regression models the probabilities for classification problems with two possible outcomes. It is an alternative of the linear regression model for classification problems.

Some people imagine that logistic regression is simpler and more useful than SVM, but actually it is not always true. While linear logistic regression has been the mainstay in biostatistics and epidemiology, it has had a mixed reception in the machine learning community. Logistic regression builds a classifier in two steps: fit a conditional probability model for $p(Y = 1|X = x)$, and then classify as one if $\hat{p}(Y = 1|X = x) \geq 0.05$. Instead, SVMs bypass the first step, and build a classifier directly. Another rather awkward issue with logistic regression is that it fails if the training data are linearly separable. What this means is that, in the feature space, one can separate two classes by a linear boundary. In cases, such as this, maximum likelihood fails and some parameters march off to infinity, while this might have seemed an unlikely scenario to the early users of logistic regression and it becomes almost a certainty with modern wide genomics data (7).

The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Hence, we have a dataset with two explanatory variables, X_1 and X_2 , and qualitative response variable that has two classes (class1 and calss2). Numerically, we would score these classes as $y = +1$ for say class1, and $y = -1$ for class2. We want to separate these two classes from each other and then make an exact prediction for new data. Thus, we need to fit a good model. In other words, we are searching a straight line. There are many

[DOI: 10.18502/jhs.v10i2.9763]

lines that can separate these two classes, but we want to find the best one. The main question is "Which line is the best?" or "What features does the best line have?"

First, let us call the best line (the optimal line) as a decision boundary (with two explanatory variables), but when we have more than two variables, it can be called the optimal hyperplane.

Basically, select the hyperplane which separates the two classes better. We do this by maximizing the distance between the closest data points and the hyperplane. The greater the distance, the better is the hyperplane and better classification results ensue. In other words, we need to find the optimal hyperplane. The optimal separating hyperplane is the linear classifier that creates the largest margin between the two classes, and is shown in the right panel (it is also known as an optimal-margin classifier). The underlying hope is that, by making a big margin on the training data, it will also classify future observations well. Let us call this margin M. Figure 2 shows this concept.

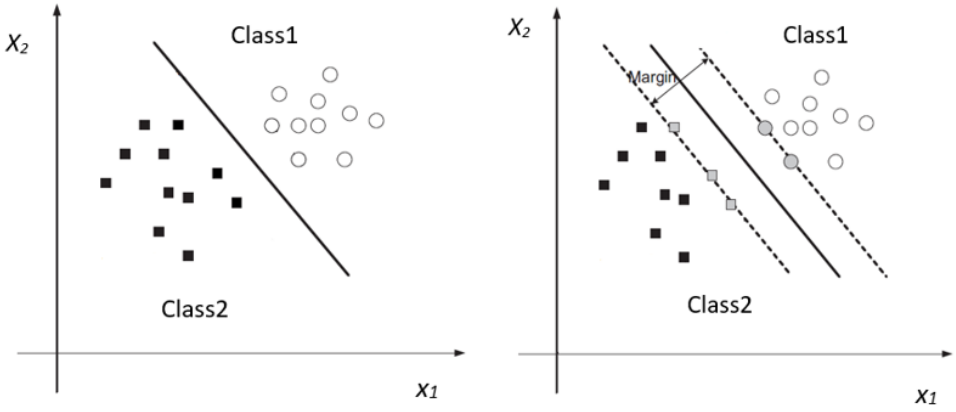

Figure 2. Decision boundary

Let $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is decision boundary so that $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and b is an intercept for decision boundary, which can be aero. Actually, it depends on dataset. We have:

$$\text{if } x_i \in \text{Calss1} \rightarrow \boldsymbol{w}^T\boldsymbol{x} + b \geq 0 \rightarrow y_i = +1, \ i = 1, \cdots, n,$$

$$\text{if } x_i \in \text{Calss2} \rightarrow \boldsymbol{w}^T\boldsymbol{x} + b \leq 0 \rightarrow y_i = -1, \ i = 1, \cdots, n. \quad (1)$$

Note that the points that are on the margin lines are called support vectors and they help us to find the optimal problem. We have 5 support vectors in Figure 3 shown in grey color, and we have considered $M = 2$.

So, we can rewrite (1):

$$\text{if } x_i \in \text{Calss1} \rightarrow \boldsymbol{w}^T\boldsymbol{x} + b \geq \frac{M}{2} = r \rightarrow y_i = +1, \ i = 1, \cdots, n,$$

$$\text{if } x_i \in \text{Calss2} \rightarrow \boldsymbol{w}^T\boldsymbol{x} + b \leq -\frac{M}{2} = r \rightarrow y_i = -1, \ i = 1, \cdots, n.$$

We can combine these two equations in the form:

$$\forall \boldsymbol{x}, \quad y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq +\frac{M}{2}, \ y_i = \pm 1, \ i = 1, \cdots, n.$$

Now it is needed to refer to the mathematical concepts of norm:

Norm: A norm on a real vector space V is a function $||.|| : V \rightarrow IR$ satisfying the following properties:

[DOI: 10.18502/jhs.v10i2.9763]

1) $\forall x \in X, \quad \|x\| \geq 0,$
$\|x\| = 0 \quad \Leftrightarrow \quad x = 0, \quad$ (positivity)

2) $\forall x \in X, \forall \alpha \ni \mathbb{R}, \quad \|\alpha x\| = |\alpha| \|x\|, \quad$ (homogeneous)

3) $\forall x, y \in X, \quad ||x + y|| < ||x|| + ||y||, \quad$ (triangle inequality).

The homogeneous condition ensures that the norm of the zero vector in V is 0; this condition is often included in the definition of a norm. A common example of norms on $\mathbb{R}^n$ are the ℓ_p norm, for $1 \leq p \leq \infty$, which is defined by:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

The distance between a point $A = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$ from a line $ay + bx + c = 0$ is also defined by:

$$Dis = \frac{|ay_1 + bx_1 + c|}{\|w\|_2},$$

where $w = \begin{bmatrix} a \\ b \end{bmatrix}$, $\|w\|_2$ is the ℓ_2 norm of vector w and $|.|$ denotes the absolute value.

By doing some simple calculations, using the concept of distance in mathematics, M can be obtained by:

$$r = \frac{|\boldsymbol{w}^T \boldsymbol{x}^+ + b|}{\|\boldsymbol{w}\|_2} = \frac{|\boldsymbol{w}^T \boldsymbol{x}^- + b|}{\|\boldsymbol{w}\|_2} = \frac{1}{\|\boldsymbol{w}\|_2}, \quad M = \mathbf{2}r = \frac{2}{\|\boldsymbol{w}\|_2},$$

Where $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ are the observations related to the first and second classes, respectively.

Now it is enough to minimize the margin, so, we have:

$$\min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|_{\mathbf{2}}^{\mathbf{2}},$$

$$s.t. \quad y_i(\boldsymbol{w}^T \boldsymbol{x} + b) \geq \frac{M}{2}, \qquad i = 1, \cdots, n.$$

By solving this optimization problem, we can find the decision boundary solution.

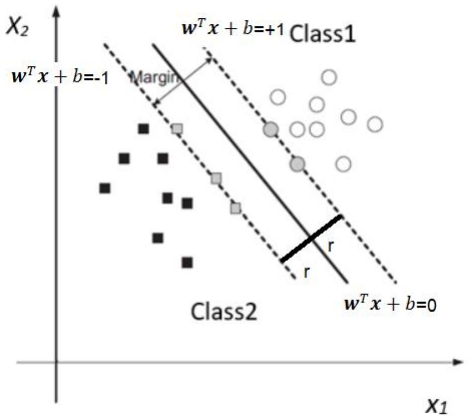

Figure 3. Supportvectors

Regression analysis is a statistical method used for the prediction of relationships between a dependent variable and one or more predictor variables. It can be utilized to assess the strength of the relationship between variables and modeling the future relationship between them. Regression is a statistical approach used in finance, investment, and other fields to identify the strength and type of a connection between one dependent variable (typically represented by y) and a sequence of other variables (known as independent variables). Regression analysis is a

[DOI: 10.18502/jhs.v10i2.9763]

statistical method for determining which predictor variables have a significant effect on dependent variable.

Before continuing this section, we need to explain the difference between regression and classification. A variable, same as y, can be broadly characterized as a quantitative or qualitative (also known as categorical) variable. Quantitative variables take on numerical values, e.g., age, height, income, price, and much more. Predicting the qualitative responses is often termed as a regression problem. Qualitative variables take on categorical values, e.g., gender, brand, parts of speech, and much more. Predicting the qualitative responses is often termed as a classification problem.

The linear regression model can be shown as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad (2)$$

Where $\boldsymbol{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of observations on the dependent or response variable, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of observations on the explanatory variables such that $\boldsymbol{x}_i = (x_{1i}, \dots, x_{ni})^{\mathrm{T}}$, $i = 1, \dots, p$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^{\mathrm{T}}$ is an $n \times 1$ vector of the error terms with $E(\boldsymbol{\epsilon}) = 0$ and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}) = \sigma^2 \boldsymbol{I}_p$.

Our goal in regression analysis is to estimate the regression coefficients. There are many ways to end this, but the most famous method is ordinary least square (OLS). It is a really simple way to estimate regression coefficients, and it provides the best linear unbiased estimator with minimum variance (BLUE) under Gauss-Markov theorem (1,7). Hence, everyone can simply apply this method. But it can sometimes be deceptive. In fact, there are some situations that OLS method acts really bad, same as in high-dimensional dataset, existence of outliers or multicollinearity, and so on (1,7). Thus, we need to find the alternative methods. Support vector machine is one of those methods that can act well to analyze high-dimensional datasets with outliers (1,6). In fact, the support vector regression uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number, it becomes very difficult to predict the response variable, which has infinite possibilities. In the case of regression, a margin of tolerance or tube (epsilon) is set to approximate the SVM which would have already requested in the problem. In addition to this fact, there is also more complication, and the algorithm is more complicated, and is therefore taken into consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated (1).

SVR has been proven to be an effective tool in real-value function prediction. As a supervised-learning approach, SVR trains using a symmetrical loss function, which equally penalizes high and low misestimates (8-10).

One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy. Heuristic methods for modelling high-dimensional regression models can be seen in (11).

2. *Methods and Result*

In this section, first of all, we would like to explain SVR method and then, compare

[DOI: 10.18502/jhs.v10i2.9763]

the performances of SVR with LASSO and Sparse LTS methods. So, we use eye tissue samples to evaluate the proposed method by comparing with the mentioned classical approaches. As it can be shown, this high-dimensional dataset contains outliers and can be a good choice for challenging the proposed method. In this dataset, based on the selected rats, the expression level of TRIM32 gene is considered as a response or dependent variable and 200 explanatory variables measuring the gene probes are considered as the independent variables (12). We use R Software to analyze this dataset.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample. Hence, now we can dive deep and understand how SVR actually works. Again, we are searching for a line that fits the dataset, but we want to find optimal line that minimizes the error value. In other words, the objective is to basically consider the points that are within the decision boundary line. The best fit line is the hyperplane that has a maximum number of points. Vapnik (2013) proposed the following optimization problem**:**

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2, \qquad (3)$$
$$\text{s.t. } |\boldsymbol{y}-\hat{\boldsymbol{y}}| \le \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\hat{\boldsymbol{y}}=\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$ are the real values and fitted values of response variable, respectively, and $\boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is the vector of tubes that reformulates the optimization problem to find the optimal tube to reach the best approximates and continuous-valued function while balancing model complexity and prediction error. More specifically, SVR is formulated as an optimization problem by first defining a convex epsilon-insensitive loss function to be minimized, and finding the flattest tube that contains most of the training instances. We can write $\hat{\boldsymbol{y}}=w\boldsymbol{x}+b$. Hence, the optimization problem was rewritten (3) in the following form:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2, \qquad (4)$$
$$s.t.\ (wx_i+b)-y_i \le \varepsilon_i, \qquad i=1,\cdots,n,$$
$$s.t.\ y_i-(wx_i+b) \le \varepsilon_i, \qquad i=1,\cdots,n.$$

By solving the above optimization problem, we can find the optimal solution. Support vector regression method is shown in Figure 4.

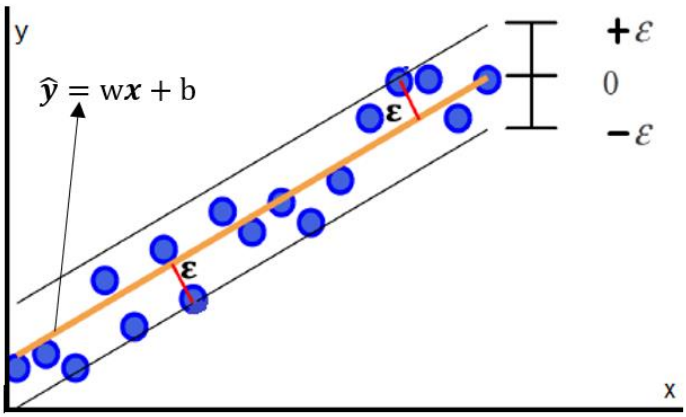

Figure 4. Support vector regression

[DOI: 10.18502/jhs.v10i2.9763]

Sometimes, we have special datasets like the dataset in Figure 5. In such a situation, we have a dataset with noisy observation(s) and so, we need to find an optimal solution. In Figure 5, we have too noisy observations. At first step, we consider them as slack variables, with ξ_i and ξ_i^*, corresponding to the size of the excess deviation for positive and negative deviations, respectively, as shown in Figure 5. Now the optimization problem (3) can be rewritten in the following form:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*), \quad (5)$$
$$s.t.\ \ (wx_i + b) - y_i \le \varepsilon_i + \xi_i,\ \ i = 1,\cdots,n,$$
$$s.t.\ \ y_i - (wx_i + b) \le \varepsilon_i + \xi_i^*,\ \ i = 1,\cdots,n,$$
$$s.t.\ \ \xi_i \ge 0,\ \ \xi_i^* \ge 0,\ \ i = 1,\cdots,n,$$

where $c\ (\ge 0)$ controls the size of the margin. By solving the above optimization problem, we can find the optimal coefficients.

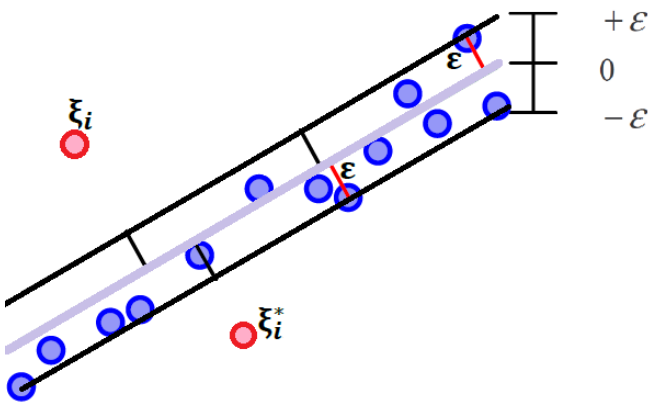

Figure 5. Support vector regression with noisy observations in dataset.

Since we cannot use classic methods for high-dimensional datasets, we had to use other methods like LASSO in such situations. On the other hand, LASSO method is not a robust method in the presence of outliers. Therefore, sparse least trimmed squares was used.

2.2. Sparse least trimmed squares

Sparse least trimmed squares (Sparse LTS) estimator is a combination of robust and sparse estimators. This method is introduced by adding an ℓ_1 penalty on the coefficient estimates to the well-known least trimmed squares (LTS) estimator. Sparse LTS can be calculated by minimizing the following optimization problem with respect to $\boldsymbol{\beta}$:

$$(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{Z}(\boldsymbol{y} - \boldsymbol{X\beta}) + h\lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\boldsymbol{Z} = \mathrm{diag}(z_1,\cdots,z_n)_{n\times n}$ is a diagonal matrix such that $z_i \in \{0,1\}$, h is a trimmed parameter and λ is a regularization parameter.

Note that:

$$z_i = \begin{cases} 0 & is\ i^{th}\ observation\ is\ an\ outlier \\ 1 & o.w. \end{cases}.$$

Increasing λ will increase bias and decrease variance. Likewise, decreasing λ will decrease bias and increase variance. A big part of the building, the best models in LASSO deals with the bias-variance tradeoff. Bias refers to how correct (or incorrect) the model is. A very simple model that makes a lot of mistakes is said

[DOI: 10.18502/jhs.v10i2.9763]

to have a high bias. A very complicated model that does well on its training data is said to have a low bias. There are several ways to choose the optimal λ, such as AIC, BIC, C_p and so on. For this purpose, one of the most popular methods is the cross-validation (CV) method.

We apply the cross-validation method to select the optimal λ which minimizes the following CV function:

$$\mathrm{CV} = \frac{1}{n}\left\|\boldsymbol{y} - \hat{\boldsymbol{y}}_{(-i)}\right\|_2^2,$$

where $\hat{\boldsymbol{y}}_{(-i)} = \boldsymbol{X}_{(-i)}\hat{\boldsymbol{\beta}}_{(-i)}$ with $\hat{\boldsymbol{\beta}}_{(-i)} = (\boldsymbol{X}_{(-i)}^{\mathrm{T}}\boldsymbol{X}_{(-i)})^{-1}\boldsymbol{X}_{(-i)}^{\mathrm{T}}\boldsymbol{y}_{(-i)}$. Note that $\boldsymbol{X}_{(-i)}$ is matrix of observations on the explanatory variables without i^{th} row of $\boldsymbol{X}$ and $\boldsymbol{y}_{(-i)}$ is vector of observations on the dependent variable without i^{th} element.

Scheetz et al. (12) used explanation of measurable attribute locus mapping in the laboratory rat (Rattus norvegicus) to obtain a broad aspect of gene arrangement in the mammalian eye and to determine genetic change relevant to human eye illness. Of >31,000 gene researchers displayed on an Affymetrix expression microarray, 18, 976 exhibited acceptable noteworthy for reliable analysis, and at least 2-fold variation in expression among 120 F_2 rats were developed from an $SR/JrHsd \times SHRSP$ intersect. Genome-wide linkage analysis with 399 genetic markers revealed significant linkage with at least one marker for 1,300 probes ($\alpha = 0.001$; estimated empirical false discovery rate= 2%). Both contiguous and noncontiguous loci were found to be important in regulating mammalian eye gene expression. One locus of each type was discovered in greater detail and identified putative transcription-altering variations in both cases. Also, an inserted cREL binding sequence was obtained in the 5 flanking sequence of the Abca4 gene associated with an increased expression level of that gene. Moreover, a pairwise analysis of gene expression to identify genes was applied to adjust genes in a coordinated appearance and apply this method to verify two formerly unknown genes involved in the human disease Bardet-Biedl syndrome. These data and analytical methods can then be used to accelerate the detection of further genes involved in human eye disease.

First, we apply LASSO method in the eye dataset. Figure 6 shows cross-validation method to choose the regularization parameter. 27 explanatory variables were chosen by LASSO method. Now, we would like to check and find outliers. We draw some plots to identify them. You can see cell map plot in Figure 7. Red and orange cells are outliers. As seen in Figure 8 about stalactite plot, last stars were outliers. These plots were multivariate graphics that were designed for the detection and identification of multivariate outliers. Figure 9 shows that boxplot and red points are considered as the outliers. It is clear that there were some outliers in dataset, and hence, we had to use robust methods. Diagnostic plots of Sparse LTS Method which clearly revealed that the data contain some outliers are shown in Figure 10. Also, Figure 10 gives the residual plots to diagnose outliers for the mammalian eye gene expression data set based on the model with effective genes. In normal QQ plot, if the resulting plot produces points 'close' to a straight line, then the data are said to be consistent with that from a normal distribution without outliers. On the other hand, departures from linearity provide evidence of skewed distributions which can lead to the observation of outlier. According to Figure

[DOI: 10.18502/jhs.v10i2.9763]

10, it can be seen that there were some standardized residuals that lied quite far from the rest (e.g. observations 43, 33, 94, 69, 58, 75, …, 72) in normal Q–Q plot. These points were potentially considered as outliers, and tended to give the normal probability plot the appearance of one for skewed data. The main tools we used to validate the sparse regression model were plots of standardized residuals. The plots enabled us to assess visually whether the assumptions were being violated and pointed to what should be done to overcome these violations. We followed the common practice of labelling points as outliers in case the standardized residual for the point fell outside the interval from 2.5 to 2.5. Identification and examination of any outliers is a key part of regression analysis. The plots suggest there are some outliers in the data set. Hence, developing the efficient robust estimation strategies are required for data predicting.

Table 1 shows the summarized results. In the table, MSE (mean squares error), $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), i = 1 \cdots, n$, MAE (mean absolute error), $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, i = 1 \cdots, n,$ and RMSE (root mean squared error) $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}, i = 1 \cdots, n,$ as fitting criteria. Note that y_i and $\hat{y}_i$ are observed values and fitted values of response variable, respectively. According to Table 1, based on the introduced fitting criteria, SVR method performed better than LASSO and Sparse LTS methods.

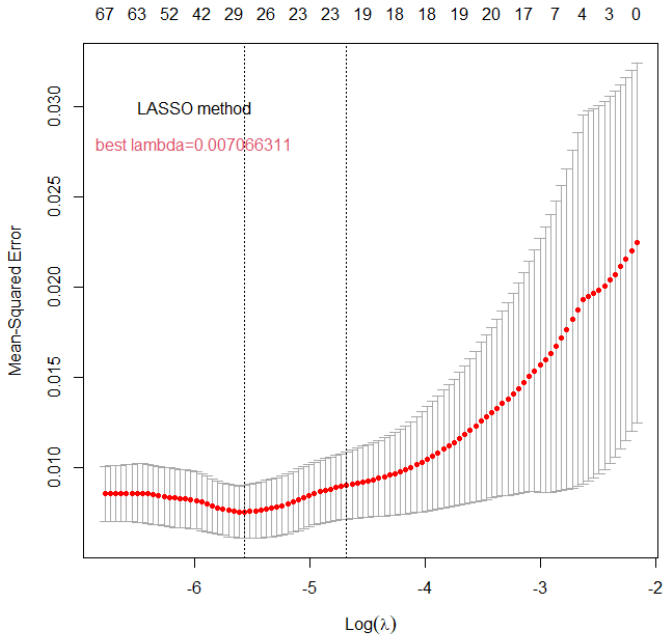

Figure 6. Cross-validation method to choose regularization parameter

[DOI: 10.18502/jhs.v10i2.9763]

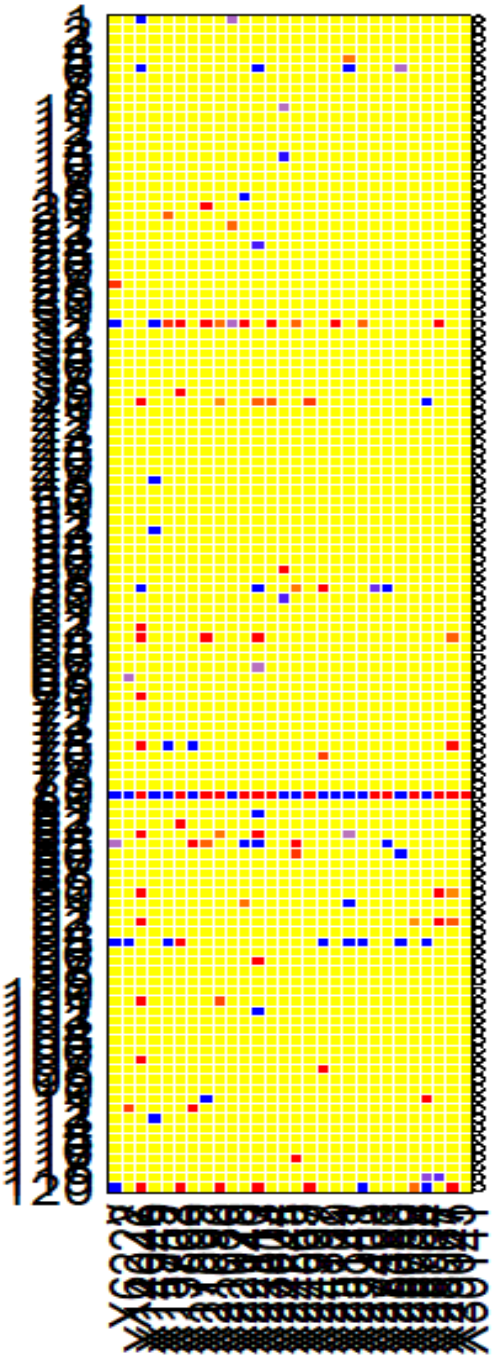

Figure 7. Cell map plot

[DOI: 10.18502/jhs.v10i2.9763]

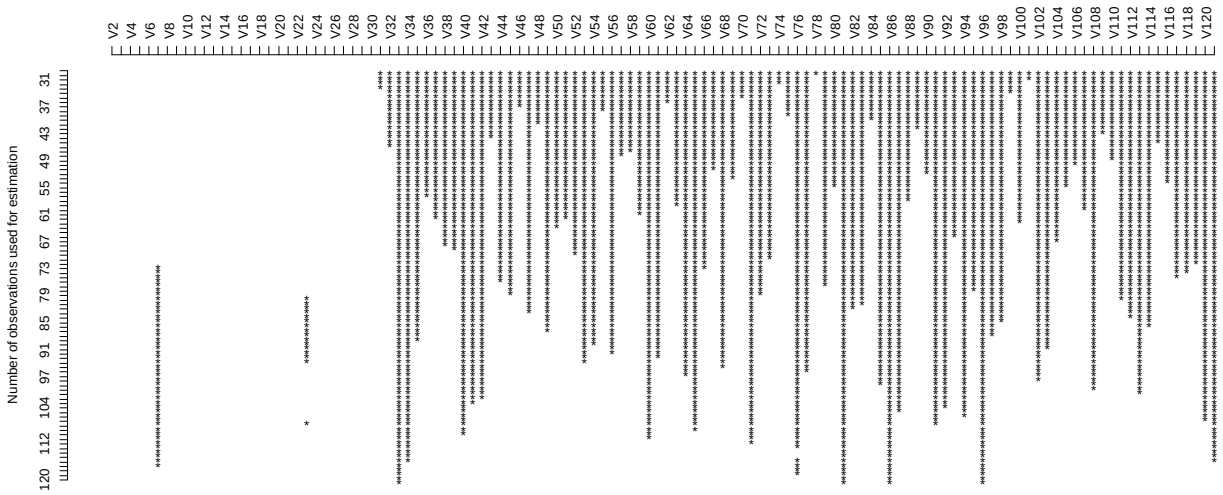

Figure 8. Stalactite plot

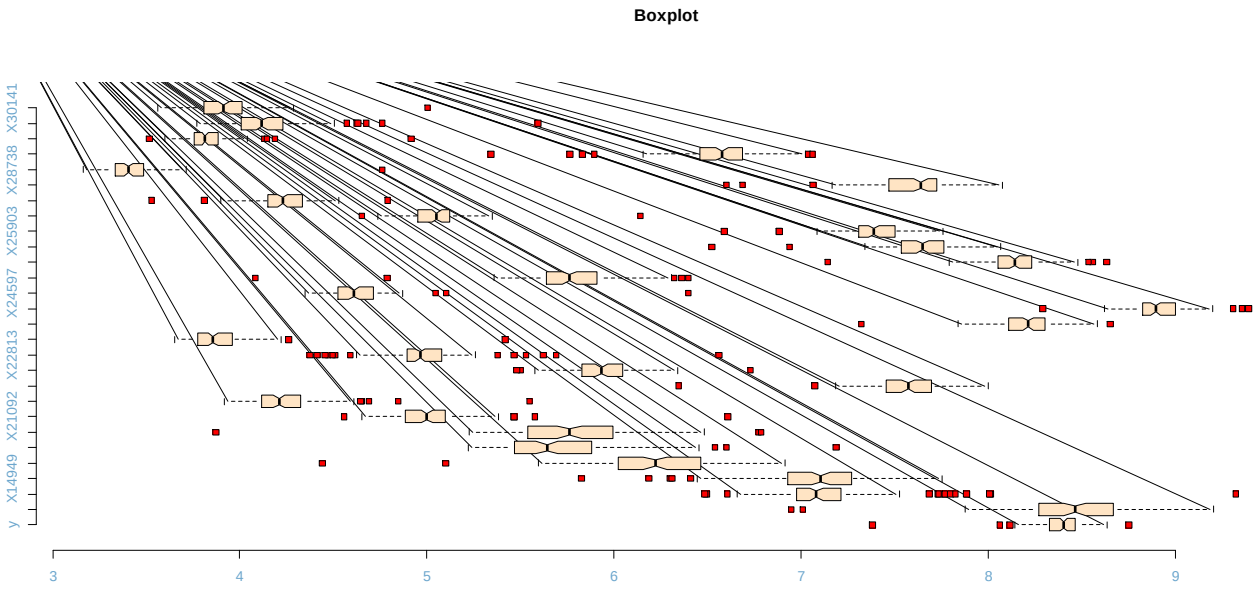

Figure 9. Boxplot

[DOI: 10.18502/jhs.v10i2.9763]

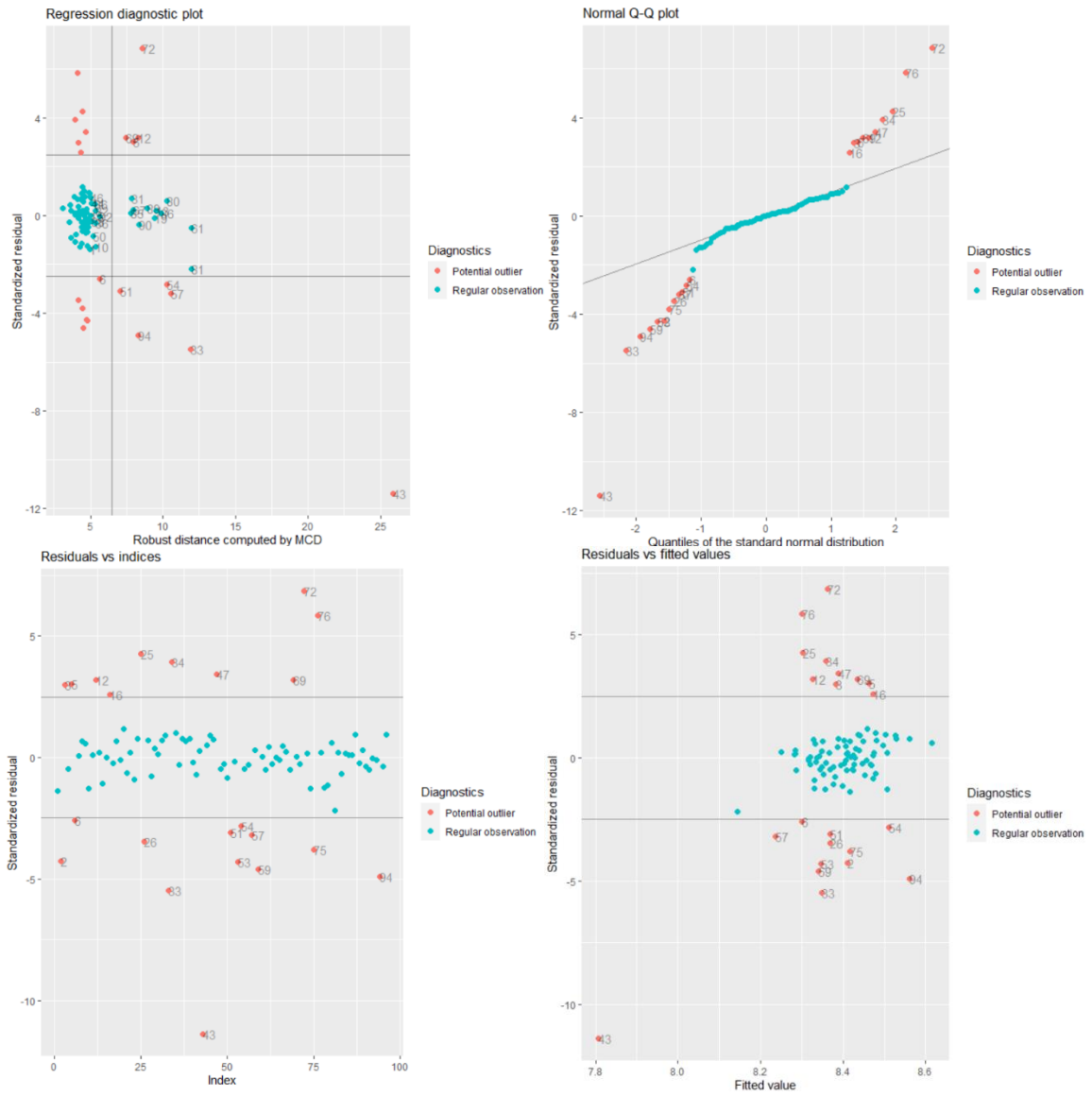

Figure 10. Diagnostic plots of Sparse LTS method

Table 1. The results of LASSO, sparse LTS and SVR methods

Method	LASSO	Sparse LTS	SVR
MSE	0.2008396	0.01774238	0.008679111
RMSE	0.4481513	0.1185427	0.09316175
MAE	0.07169988	0.1630347	0.06598934

3.Discussion

Nowadays, the analysis of high-dimensional datasets, in which the number of observations is smaller than the number of parameters, is very usual and conventional in gene expression modelling in human diseases. The analysis of such datasets is often extremely complicated and impossible through classical approaches. In this regard, we have proposed a statistical machine learning approach for the regression models to simultaneously combat high-dimension and outliers in the dataset which can make the estimators resistant against the outlying

[DOI: 10.18502/jhs.v10i2.9763]

observations and non-normal error distributions. Stabilization of the data and freedom from obtaining the regularization parameter by running a complicated algorithmic program are the invaluable advantages of this technique in comparison with the classical methods, such as LASSO and Sparse LTS estimations. Especially, based on a support vector regression technique, we have suggested a robust programming approach to model the high-dimensional mammalian eye dataset which can be efficiently applied to estimate and predict the high-dimensional regression models, not affected by the outliers' corruptive impact. The results of this study can potentially be extended to investigate the role of gene regulation in a number of human eye diseases.

Conflicts of interest

None declared

References

1. Hastie T, Tibshirani B, Friedman J. The elements of statistical learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer; 2017.
2. Boser BE, Guyon IM, Vapnik, VN. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on computational learning theory. 1992 July 27-29; Pittsburgh, United States. 1992. p. 144-152.
3. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3): 273-297.
4. Vapnik VN. The nature of statistical learning theory. 2nd ed. New York: Springer; 2000.
5. Noble WS. What is a support vector machine?. Nature biotechnology. 2006;24(12): 1565-1567.
6. Roozbeh M, Maanavi M, Babaie-Kafaki S. Robust high-dimensional semiparametric regression using optimized differencing method applied to the vitamin B2 production data. Iranian Journal of Health Sciences. 2020;8(2):9-22.
7. Efron B, Hastie T. Computer age statistical inference: Data mining, inference and prediction. Cambridge: Cambridge University Press; 2016.
8. Awad M, Khanna R. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley: Apress, CA; 2015.
9. Roozbeh M, Babaie-Kafaki S and Aminifard Z. Improved high-dimensional regression models with matrix approximations applied to the comparative case studies with support vector machines. Optimization Methods and Software, 2022; 1-18.
10. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with applications in R. 2nd ed. Springer, New York, 2021.
11. Roozbeh M, Babaie-Kafaki S and Manavi M. A heuristic algorithm to combat outliers and multicollinearity in regression model analysis. Iranian Journal of Numerical Analysis and Optimization. 2022;12 (1):173-186.
12. Scheetz TE, Kim KYA, Swiderski RE, Philp AR, Braun TA, Knudtson KL et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(39):14429-14434

[DOI: 10.18502/jhs.v10i2.9763]