On the Robustness of Support Vector Regression

Sanghun Jeong : Ph.D. Candidate, Department of Statistics, Pusan National University
Choongrak Kim : Professor, Department of Statistics, Pusan National University

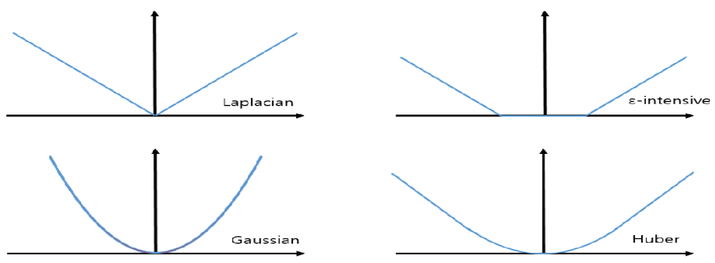
1. INTRODUCTION

- The Support Vector Machine (SVM) is a universal approach for solving the problems of multidimensional function estimation, and has also been applied to various fields successfully such as classification and regression.
- There are several types in SVM for regression, i.e. Support Vector Regression (SVR), and we introduce SVR using ε-intensive loss function (ε-SVR).
- ε-intensive loss function does not care about errors as long as they are less than ε, but does not accept any deviation larger than this, and its advantage is less sensitive to outliers in data than the squared error loss like Huber loss function.
- In this study, we give an overview of the basic ideas underlying ε-SVR and include a summary of algorithms for solving the optimization problem through dual problem, and confirm the performance of ε-SVR as robust regression.
- In numerical study, we compare the ε-SVR to the Simple Linear Regression as a linear case and the Local Linear Regression as a non-linear case and calculate the variance of slope and residual as robustness metric.

2. METHODS

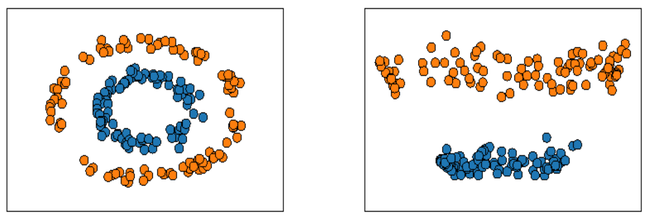
- Support Vector Regression (SVR) is a regression method to calculate the quantitative response using the loss function, kernel trick, based on Support Vector Machine which is used to classification problems.

- Commonly used loss functions in the SVR : the ε-intensive and the Huber loss functions require more parameters than the Laplacian and the Gaussian loss function but more robust.



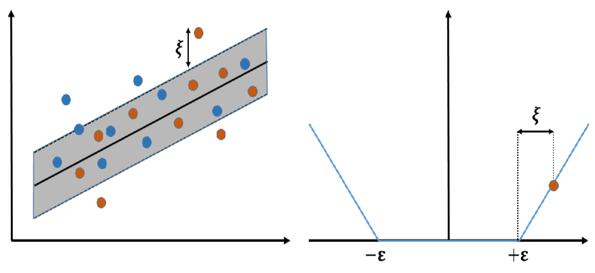
- There are two classes of observations, shown in blue and in orange. Even if allow some misclassified data, there is no suitable hyperplane. However, we can easily find the hyperplanes by using a mapping function Φ

$$\Phi : (x, y) \rightarrow (x, x^2 + y^2).$$



- Let us consider a similarity measure of the form $k : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$, that is, a function that, given two vectors x and x', returns a scalar as their similarity, i.e., $(x, x') \rightarrow k(x, x')$. The function k is called a kernel, and there is a kind of kernel like linear, polynomial, and radial basis function (RBF). but RBF is mainly used.

- In SVR using ε-intensive loss function (ε-SVR), our goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained response y_i for all the training data $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{H} \times \mathbb{R}$,where $\mathcal{H}$ denotes the space of the input data (e.g. $\mathcal{H} = \mathbb{R}^p$), and at the same time is as flat as possible.



- Given training vectors x_i, and a response y_i, $i = 1, \ldots, n$, our goal is to find $\beta \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$ such that minimize to measure of error for prediction $x_i^T\beta + \alpha$. Then, the ε-SVR solves the following optimization problem :

$$\underset{\alpha,\beta,\xi^{(*)}}{minimize} \left\{\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right\}$$

$$subject\ to \begin{cases} y_i - x_i^T\beta - \alpha \le \varepsilon + \xi_i \\ x_i^T\beta + \alpha - y_i \le \varepsilon + \xi_i^* \\ \xi_i^{(*)} \ge 0 \end{cases}$$

where $C > 0$ determines the trade-off between the flatness of f, and $\xi^{(*)} \in \mathbb{R}^n$ is slack variable to cope with the infeasible optimization problem.

- The ε-intensive loss function $|\xi|_\varepsilon$ is defined as

$$|\xi|_\varepsilon = \begin{cases} 0, & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & otherwise \end{cases}$$

- We can easily solve the optimization problem by using the transformation to Lagrangian dual problem, and then the Lagrangian primal function :

$$\mathcal{L} := \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\lambda_i\xi_i + \lambda_i^*\xi_i^*) - \sum_{i=1}^{n}\nu_i(\varepsilon + \xi_i - y_i + x_i^T\beta + \alpha) - \sum_{i=1}^{n}\nu_i^*(\varepsilon + \xi_i^* - y_i + x_i^T\beta + \alpha)$$

where $\lambda_i, \lambda_i^*, \nu_i, \nu_i^* \ge 0$ are Lagrange multipliers.

- The partial derivatives of $\mathcal{L}$ with respect to the primal variables $(\alpha, \beta, \xi^{(*)})$ have to vanish for optimality.

$$\frac{\partial\mathcal{L}}{\partial\alpha} = \sum_{i=1}^{n}(\nu_i^* - \nu_i) = 0$$
$$\frac{\partial\mathcal{L}}{\partial\beta} = \beta - \sum_{i=1}^{n}(\nu_i - \nu_i^*)x_i = 0$$
$$\frac{\partial\mathcal{L}}{\partial\xi_i^{(*)}} = C - \nu_i^{(*)} - \lambda_i^{(*)} = 0$$

- The optimization problem can be rewritten to dual form as follows :

$$\underset{\nu^{(*)}\in\mathbb{R}^n}{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\nu_i - \nu_i^*)(\nu_j - \nu_j^*)x_i^Tx_j \\ -\varepsilon\sum_{i=1}^{n}(\nu_i + \nu_i^*) + \sum_{i=1}^{n}y_i(\nu_i - \nu_i^*) \end{cases}$$
$$subject\ to \sum_{i=1}^{n}(\nu_i - \nu_i^*) = 0\ and\ \nu_i^{(*)} \in [0, C]$$

- To compute α, we use some constraints via Karush-Kuhn-Tucker conditions

$$\nu_i(\varepsilon + \xi_i - y_i + x_i^T\beta + \alpha) = 0$$
$$\nu_i^*(\varepsilon + \xi_i^* + y_i - x_i^T\beta - \alpha) = 0$$
$$\left(C - \nu_i^{(*)}\right)\xi_i^{(*)} = 0$$

- Now, we deduce that

$$\nu_i^{(*)} = 0\ if\ \xi_i^{(*)} > 0$$
$$\nu_i\nu_i^* = 0$$

- Finally, we have

$$\beta = \sum_{i=1}^{n}(\nu_i - \nu_i^*)x_i$$
$$\alpha = \frac{1}{k}\sum_{j=1}^{k}\alpha_j$$

where $\alpha_j \in \{-\varepsilon + y_i - x_i^T\beta \mid 0 < \nu_i < C\ or\ 0 < \nu_i^* < C, i = 1, \ldots n\}$, i.e. $\alpha_j \in \{-\varepsilon + y_i - x_i^T\beta \mid x_i\ is\ support\ vector\}$.

- In summary, formulating the optimization problem for the ε-SVR, using the RBF as kernel, we arrive at the as follows:

$$\underset{\nu^{(*)}\in\mathbb{R}^n}{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\nu_i - \nu_i^*)(\nu_j - \nu_j^*)k(x_i, x_j) \\ -\varepsilon\sum_{i=1}^{n}(\nu_i + \nu_i^*) + \sum_{i=1}^{n}y_i(\nu_i - \nu_i^*) \end{cases}$$
$$subject\ to \sum_{i=1}^{n}(\nu_i - \nu_i^*) = 0\ and\ \nu_i^{(*)} \in [0, C]$$

where $k(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$.

- The decision function is

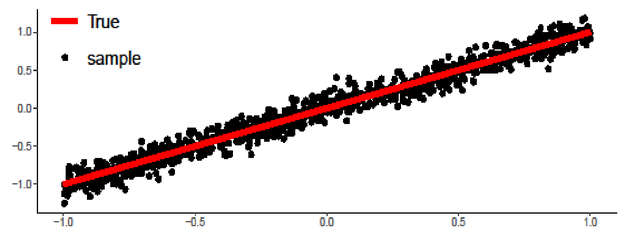
$$f(x) = \sum_{i=1}^{n}(\nu_i - \nu_i^*)k(x_i, x) + \alpha$$

3. NUMERICAL STUDY

- As linear case, the artificial data was generated as

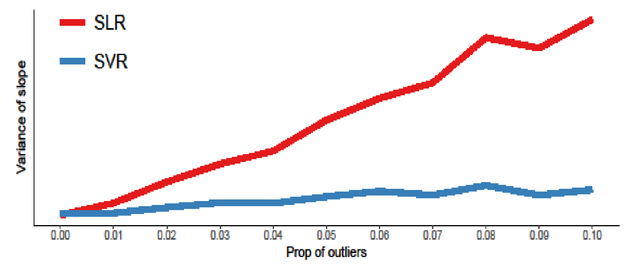
$$y = x + error$$

where $x \sim U(-1, 1),\ error \sim N(0, (0.1)^2)$ with sample size $n = 1000$.



- We diagnosis the sample that the absolute value of residual is greater than 3σ (i.e. 0.3) as an outlier, and calculate the variance of estimated slope in replication $k = 100$.

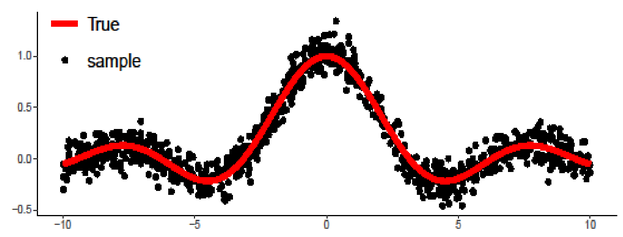
- In proportion of outliers from 0 to 0.1, the variance of slope in Simple Linear Regression (SLR) increases faster than the variance of slope in ε-SVR using the linear kernel, indicating that ε-SVR is more robust than SLR. Because, even if there is an outlier, there is less perturbation of the slope, this means robust.



- As non-linear case, the artificial data was generated as

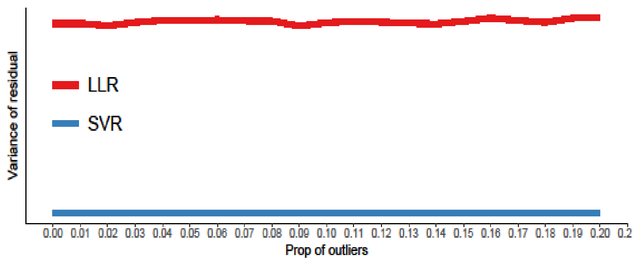
$$y = \frac{\sin(x)}{x} + error$$

where $x \sim U(-10, 10),\ error \sim N(0, (0.1)^2)$ with sample size $n = 1000$.



- We diagnosis outlier same as linear case, and calculate the variance of residual between true function $\frac{\sin(x)}{x}$ and estimated function $\hat{f}(x)$ in replication $k = 100$.

- In proportion of outliers from 0 to 0.2, the variance of residual in Local Linear Regression (LLR) using bandwidth $h = 0.5$ increases a little faster than the variance of residual in ε-SVR using the RBF kernel, indicating that ε-SVR is more robust than LLR. Because, even if there is an outlier, there is less difference to estimated function $\hat{f}(x)$ when there is no outlier, this means robust.



- However, in non-linear case, there is a problem. Hyper-parameters must be given to estimate LLR and ε-SVR such as bandwidth h in LLR, cost C, ε in ε-SVR. The performance of the estimated function $\hat{f}(x)$ depends on hyper-parameters, so the choice of hyper-parameters is an important problem.

- Furthermore, the measurement error model can be applied to the ε-SVR to make a much more robust estimate. Similar to the basic process, but the difference is to put $\bar{x}_i + \theta_i$ instead of x_i, and solve the optimization problem by using Second -Order Cone Programming.

4. CONCLUSION

- In this study, we briefed the concept of SVM and the process that solving the optimization problem of SVR and confirmed the robustness of ε-SVR by comparing it to SLR and LLR.
- In simulation study, the performance of ε-SVR as robust regression outperformed SLR and LLR, in particular, when more outliers exist.
- The results of the simulation indicated that the ε-SVR is robust or not by controlling the hyper-parameter.
- Additionally, ε-SVR can be more robust by applying the measurement error model, but the number of hyper-parameter increases.