Communications of the Korean Statistical Society
2011, Vol. 18, No. 2, 165–170

DOI: 10.5351/CKSS.2011.18.2.165

Support Vector Quantile Regression Using Asymmetric e-Insensitive Loss Function

Jooyong Shim[a], Kyung Ha Seok[1,b], Changha Hwang[c], Daehyeon Cho[b]

[a]Department of Data Science, Inje University
[b]Department of Data Science and Institute of Statistical Information, Inje University
[c]Department of Statistics, Dankook University

Abstract

Support vector quantile regression(SVQR) is capable of providing a good description of the linear and nonlinear relationships among random variables. In this paper we propose a sparse SVQR to overcome a limitation of SVQR, nonsparsity. The asymmetric e-insensitive loss function is used to efficiently provide sparsity. The experimental results are presented to illustrate the performance of the proposed method by comparing it with nonsparse SVQR.

Keywords: Asymmetric e-insensitive loss function, quantile regression, support vector machine, support vector quantile regression.

1. Introduction

Quantile regression has been a popular method for estimating the quantiles of a conditional distribution on the values of covariates since Koenker and Bassett (1978) introduced linear quantile regression. Just as classical linear regression methods based on the minimizing sum of squared residuals enable us to estimate a wide variety of models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the full range of conditional quantile functions, including the conditional median function. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a better statistical analysis of the stochastic relationships among random variables. An introduction and examination of current research areas of quantile regression can be found in Yu *et al*. (2003) and Koenker (2005). Support vector machine(SVM) is used as a new technique for regression and classification problems. The SVM is based on the structural risk minimization(SRM) principle that is shown to be superior to the traditional empirical risk minimization(ERM) principle. SRM minimizes an upper bound on the expected risk, unlike ERM, which minimizes the error on the training data. By minimizing this bound, high generalization performance can be achieved. In particular, for the SVM regression case, SRM results in regularized ERM with e-insensitive loss function. Introductions to and overviews of recent developments of SVM can be found in Vapnik (1995, 1998), Smola and Schölkopf (1998) and Wang (2005)

Sparsity is known as an important feature of kernel regression models. It provides efficiency in predicting the regression function, which implies that the predicted regression function of the test data

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0072369).

[1] Corresponding author: Professor, Department of Data Science and Institute of Statistical Information, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: statskh@paran.com

can be obtained with the small number of data in the training data set. SVM provides sparsity in which the number of support vectors depends on the number of training data and the size of insensitivity. A small number of support vectors implies sparsity of the model. Tipping (2001) proposed a Bayesian approach referred to as the relevance vector machine(RVM), providing more sparsity. However, RVM has computational problems since there are no closed-form solutions for maximizing the marginal likelihood. SVQR can be obtained by applying SVR with a check function instead of an e-insensitive loss function into the quantile regression (Takeuchi *et al*., 2006). However, SVQR does not provide sparsity due to zero insensitiveness of the check function. Here we define the support vectors as the index numbers corresponding to nonzero Lagrange multiplier differences. By using an asymmetric e-insensitive loss function we can take the support vectors efficiently depending on the value of quantiles. In this paper we use an asymmetric e-insensitive loss function in SVQR to provide the sparsity. The proposed loss function is designed to provide more sparsity by adjusting insensitiveness according the sign of residuals. In Section 2 we propose a sparse SVQR using an asymmetric e-insensitive loss function and perform numerical studies through examples. In Section 3 we give the conclusions.

2. Support Vector Quantile Regression

2.1. SVQR

Let the training data set D be denoted by $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ with each input vector $\boldsymbol{x} \in R^d$ and the output $y_i \in R$, which is linearly or nonlinearly related to the input vector $\boldsymbol{x}_i$. Here the feature mapping function $\phi(\cdot) : R^d \to R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. An inner product in feature space has an equivalent kernel in input space, $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (Mercer, 1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible. We consider the nonlinear case, in which the θ^{th} quantile function, given $\boldsymbol{x}$, $q_\theta(\boldsymbol{x})$ for $\theta \in (0, 1)$, can be regarded as a nonlinear function of input vector $\boldsymbol{x}$.

With a check function $h_\theta(\cdot)$, the θ^{th} quantile function can be defined as a function of any solution to the optimization problem,

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} h_\theta(y_i - q_\theta(\boldsymbol{x}_i)), \tag{2.1}$$

where $h_\theta(r) = \theta r I(r > 0) + (\theta - 1) r I(x \le 0)$ for $\theta \in (0, 1)$, where $I(\cdot)$ is the indicated function. We can express the quantile regression problem by formulation for SVM as follows.

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) \tag{2.2}$$

subject to

$$y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b \le \xi_i, \quad \boldsymbol{w}'\phi(\boldsymbol{x}_i) + b - y_i \le \xi_i^*, \quad \xi_i \ge 0,\ \xi_i^* \ge 0,$$

where C is a regularization parameter penalizing the training errors. We construct a Lagrange function as follows:

$$\begin{aligned} L = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\xi_i - (y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b)) \\ - \sum_{i=1}^{n}\alpha_i^*(\xi_i^* - (\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b - y_i)) - \sum_{i=1}^{n}\eta_i\xi_i - \sum_{i=1}^{n}\eta_i^*\xi_i^*. \end{aligned} \tag{2.3}$$

We notice that the non-negative constraints $\alpha_i, \eta_i, \alpha_i^*, \eta_i^* \geq 0$ should be satisfied. After taking partial derivatives of Equation (2.3) with regard to the primal variables $(\boldsymbol{w}, \xi_i, b)$ and plugging them into Equation (2.3), we have the optimization problem below.

$$\max\ L = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \tag{2.4}$$

subject to $\sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i^* = 0$, $0 \leq \alpha_i \leq \theta C$ and $0 \leq \alpha_i^* \leq (1-\theta)C$, $i = 1, \ldots, n$.

Solving the above problem with the constraints determines the optimal Lagrange multipliers $\hat{\alpha}_i$ and $\hat{\alpha}_i^*$. Thus, the estimated θ^{th} quantile function given the input vector $\boldsymbol{x}_0$ is obtained as

$$\hat{q}_\theta(\boldsymbol{x}_0) = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}_0) + \hat{b}, \tag{2.5}$$

where $\hat{b}$ is obtained via Kuhn-Tucker conditions (Kuhn and Tucker, 1951) such as,

$$\hat{b} = \frac{1}{n_s}\sum_{i \in I_s}(y_i - K_i(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^*)),$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)'$, $\hat{\boldsymbol{\alpha}}^* = (\hat{\alpha}_1^*, \ldots, \hat{\alpha}_n^*)'$ and n_s is the size of the set $I_s = \{i = 1, \ldots, n | C(\theta - 1) < \hat{\alpha}_i - \hat{\alpha}_i^* < C\theta\}$ and K_i is the i^{th} row of the kernel matrix $K = \{K(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{n \times n}$. In the nonlinear case, $\boldsymbol{w}$ is no longer explicitly given. However, it is uniquely defined in the weak sense by the dot products. Here the linear regression model can be regarded as a special case of the nonlinear regression model by using the identity feature mapping function, that is, $\phi(\boldsymbol{x}) = \boldsymbol{x}$ which implies the linear kernel such that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{x}_2$.

2.2. Sparse SVQR

With an asymmetric e-insensitive loss function $g_{\theta,e}(\cdot)$, shown in Figure 1, the θ^{th} quantile function can be defined as a function of any solution to the optimization problem,

$$\min\ \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} g_{\theta,e}(y_i - q_\theta(\boldsymbol{x}_i)), \tag{2.6}$$

where $g_{\theta,e}(r) = 0$ if $\theta/(\theta-1)\,e \leq r \leq (1-\theta)/\theta\,e$, $g_{\theta,e}(r) = \theta r - (1-\theta)e$ if $r > (1-\theta)/\theta\,e$ and $g_{\theta,e}(r) = (\theta-1)r - \theta e$ if $r < \theta/(\theta-1)e$ for $\theta \in (0,1)$. The check function used in nonsparse SVQR and the asymmetric e-insensitive loss function are illustrated in Figure 1 with $\theta = 0.35$, 0.75 and $e = 0.2$.

We can express the quantile regression problem by the formulation for SVM as follows.

$$\min\ \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) \tag{2.7}$$

subject to

$$\begin{aligned} &y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b \leq \xi_i + \frac{1-\theta}{\theta}e, \\ &\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b - y_i \leq \xi_i^* + \frac{\theta}{1-\theta}e, \\ &\xi_i \geq 0, \quad \xi_i^* \geq 0, \end{aligned}$$

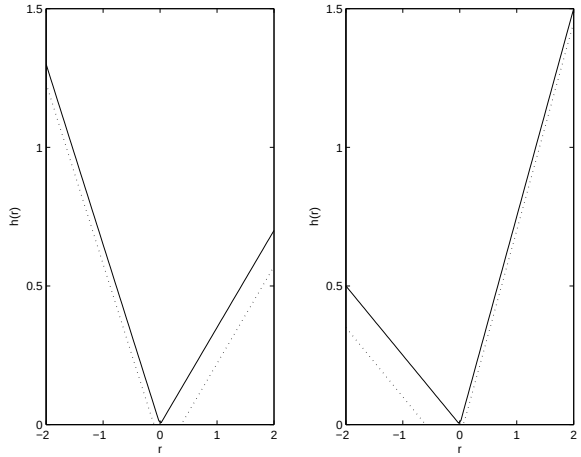

Figure 1: *The check function(solid line) and the asymmetric 0.2-insensitive loss function(dotted line) with* $\theta = 0.35$*(Left) and* $\theta = 0.75$*(Right).*

where $e > 0$ and C is a regularization parameter penalizing the training errors. We construct a Lagrange function as follows:

$$L = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n}(\theta\xi_i + (1-\theta)\xi_i^*) - \sum_{i=1}^{n}\alpha_i\left(\xi_i + \frac{1-\theta}{\theta}e - (y_i - \boldsymbol{w}'\phi(\boldsymbol{x}_i) - b)\right) \tag{2.8}$$
$$- \sum_{i=1}^{n}\alpha_i^*\left(\xi_i^* + \frac{\theta}{1-\theta}e - (\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b - y_i)\right) - \sum_{i=1}^{n}\eta_i\xi_i - \sum_{i=1}^{n}\eta_i^*\xi_i^*.$$

We notice that the non-negative constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$ should be satisfied. After taking partial derivatives of Equation (2.8) with regard to the primal variables $(\boldsymbol{w}, \xi_i, b)$ and plugging them into Equation (2.8), we have the optimization problem below.

$$\min L = \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i\left(\frac{1-\theta}{\theta}e - y_i\right) + \sum_{i=1}^{n}\alpha_i^*\left(\frac{\theta}{1-\theta}e + y_i\right) \tag{2.9}$$

subject to $\sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i^* = 0$, $0 \leq \alpha_i \leq \theta C$ and $0 \leq \alpha_i^* \leq (1-\theta)C$, $i = 1, \ldots, n$.

Solving the above problem with the constraints determines the optimal Lagrange multipliers $\hat{\alpha}_i$ and $\hat{\alpha}_i^*$. Here the input vector $\boldsymbol{x}_i$, corresponding to positive $\hat{\alpha}_i$ or $\hat{\alpha}_i^*$, is called the support vector. Thus, the estimated θ^{th} quantile function, given the input vector $\boldsymbol{x}_0$, is obtained as

$$\hat{q}_\theta(\boldsymbol{x}_0) = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}_0) + \hat{b}, \tag{2.10}$$

where $\hat{b}$ is obtained via Kuhn-Tucker conditions (Kuhn and Tucker, 1951) such as,

$$\hat{b} = \frac{1}{n_1 + n_2}\left\{\sum_{i\in I_1}\left(y_i - K_i(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^*) - \frac{1-\theta}{\theta}e\right) + \sum_{i\in I_2}\left(y_i - K_i(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^*) + \frac{\theta}{1-\theta}e\right)\right\},$$

where n_1 is the size of the set $I_1 = \{i = 1, \ldots, n \,|\, 0 < \hat{\alpha}_i < C\theta\}$ and n_2 is the size of the set $I_2 = \{i = 1, \ldots, n \,|\, 0 < \hat{\alpha}_i^* < C(1-\theta)\}$, $i = 1, \ldots, n$.

Table 1: The average of 100 MSEs of $\hat{q}_\theta(x)$ for $\theta = 0.1, 0.5, 0.9$ (standard error of MSEs in parenthesis) and the number of support vectors in spars SVQR

θ	Sparse SVQR	No. of SVs	Nonsparse SVQR
0.1	0.0248(0.002644)	53	0.0422(0.003429)
0.5	0.0087(0.000874)	82	0.0144(0.001454)
0.9	0.0019(0.0003387)	50	0.0071(0.0009113)

We can see that $\{i = 1, \ldots, n \,|\, 0 < \hat{\alpha}_i^* \le (1-\theta)C\} = \{i = 1, \ldots, n \,|\, y_i \ge \hat{q}_\theta(\boldsymbol{x}_i) + (1-\theta)/\theta\, e\}$ and $\{i = 1, \ldots, n \,|\, 0 < \hat{\alpha}_i \le \theta C\} = \{i = 1, \ldots, n \,|\, y_i \le \hat{q}_\theta(\boldsymbol{x}_i) - \theta/(1-\theta)e\}$, which are indices of data points with support vectors that are not in the asymmetric e-tube.

2.3. Numerical studies

We illustrate the performance of the sparse quantile regression estimation through the simulated data for nonlinear regression cases. A total of 101 data sets (1 training data set and 100 test data sets) are generated to present the prediction performance of the proposed method. Each data set consists of 100 x's and 100 y's. Here x's are generated from a uniform distribution $U(0, \pi)$; y's are generated from a normal distribution $N(1 + \sin(x), 0.1)$. The true θ^{th} quantile function is given as

$$q_\theta(x) = 1 + \sin(x) + 0.1\Phi^{-1}(\theta) \quad \text{for } \theta \in (0, 1),$$

where $\Phi(\cdot)$ is the cdf of $N(0, 1)$ distribution. The radial basis kernel function is utilized in this example, which is

$$K(x_1, x_2) = e^{-\frac{(x_1 - x_2)^2}{\sigma^2}}.$$

For the training data set the hyperparameters (e, C, σ^2) were chosen as $(0.05, 100, 1)$ by 5-fold cross-validation. Figure 2 shows the estimated θ^{th} quantile regression functions imposed on the scatter plots of a test data set for $\theta = 0.1$ (Left), $\theta = 0.9$ (Middle) and $\theta = 0.5$ (Right). From Figure 2 we can see that the proposed method provides the sparsity. In Figure 2 the e-tube is obtained as $(\hat{q}_{0.1}(x) - \theta/(1-\theta)e, \hat{q}_{0.1}(x) + (1-\theta)/\theta\, e)$. For $\theta = 0.1$ the e-tube is $(\hat{q}_{0.1}(x) - 0.0055,\ \hat{q}_{0.1}(x) + 0.45)$, this makes the upper bound distinguishable in Figure 2 (Left) but the lower bound indistinguishable in Figure 2 (Left). For $\theta = 0.9$ the e-tube is $(\hat{q}_{0.9}(x) - 0.45, \hat{q}_{0.9}(x) + 0.0055)$, this makes the lower bound distinguishable in Figure 2 (Middle) but the upper bound indistinguishable in Figure 2 (Middle). From 100 test data sets we obtain a mean squared error of $\hat{q}_\theta(x)$ to compare the performance of sparse SVQR to SVQR, with the results shown in Table 1. From the table we can see that the proposed sparse SVQR provides smaller MSE and error of MSE.

3. Conclusion

In this paper, we dealt with estimating the quantile regression function by SVQR using an asymmetric e-insensitive loss function. Through example we showed that the proposed method provides sparsity and better performance than that of SVQR. The model selection of SVQR using an asymmetric e-insensitive loss function takes n_e times of the model selection of SVQR in computing time, where n_e is the number of candidates of e. We must consider this fact in application of both methods to a large data set. The model selection method, such as the generalized approximate cross-validation function, will be studied in further research using the effective dimensionality of the fitted model.

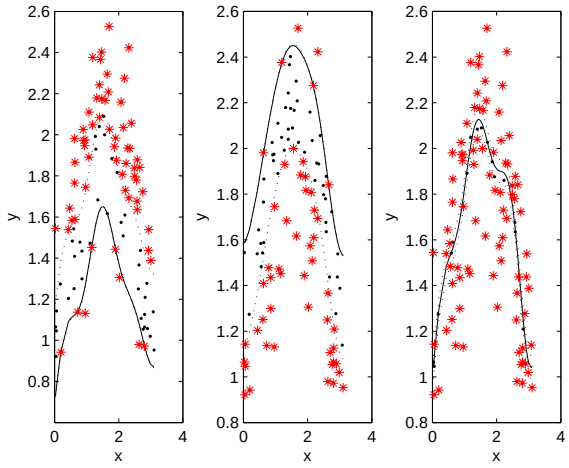

Figure 2: *Estimated θ^{th} quantile regression functions ($\hat{q}_\theta(x)$) imposed on the scatter plots of 100 data points of a data set(dots = data points with nonsupport vector, stars = data points with support vectors, solid line = $\hat{q}_\theta(x)$, dotted line(upper) = $\hat{q}_\theta(x) + (1-\theta)/\theta\, e$ and dotted line(lower) = $\hat{q}_\theta(x) - \theta/(1-\theta)e$)*

References

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression quantile, *Econometrica*, **46**, 33–50.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming, In *Proceedings of 2nd Berkeley Symposium*, Berkeley: University of California Press, 481–492.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. Philosophical Transactions of Royal Society, A:415-446.

Smola, A. and Schölkopf, B. (1998). On a Kernel-based method for pattern recognition, regression, approximation and operator inversion, *Algorithmica*, **22**, 211–231.

Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric quantile estimation, *Journal of Machine Learning Research*, **7**, 1231–1264.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine, *Journal of Machine Learning Research*, **1**, 211–244.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*, John Wiley, New York.

Wang, L.(Ed.) (2005). *Support Vector Machines: Theory and Application*, Springer, New York.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research area, *Journal of the Royal Statistical Society, Series D(The Statistician)*, **52**, 331–350.

Received November 2010; Accepted January 2011