Journal of the Korean Data & Information Science Society
2011, **22**(5), 999–1005

한국데이터정보과학회지

Asymmetric least squares regression estimation using weighted least squares support vector machine[†]

Changha Hwang[1]

[1]Department of Statistics, Dankook University

Received 31 August 2011, revised 17 September 2011, accepted 21 September 2011

Abstract

This paper proposes a weighted least squares support vector machine for asymmetric least squares regression. This method achieves nonlinear prediction power, while making no assumption on the underlying probability distributions. The cross validation function is introduced to choose optimal hyperparameters in the procedure. Experimental results are then presented which indicate the performance of the proposed model.

Keywords: Asymmetric least squares regression, cross validation, expectile, least squares support vector machine, percentile.

1. Introduction

Asymmetric least squares (ALS) regression is the least squares analogue of quantile regression. The solution of an ALS regression is known as an expectile. This name was originally proposed by Newey and Powell (1987) who note that the ALS solution is determined by the properties of the expectation of exceedances beyond the solution. It has been shown that there exists a one-to-one mapping from expectiles to quantiles. Thus this is used as the basis for estimating value-at-risk (VaR) and expected shortfall (ES). See for details Efron (1991), Yao and Tong, (1996) and Taylor (2008).

Support vector machine (SVM), introduced by Vapnik (1995), is a useful tool for data mining, especially in the fields of pattern recognition and regression. During the past few years, SVM has gained a lot of popularity due to its solid theoretical foundation and good behaviors. Kernel trick of SVM also has been demonstrated to be an effective method for solving nonlinear statistical problems. See for details Hwang (2010a, 2010b), Seok (2010) and Shim and Lee (2009). Suykens and Vandewalle (1999), from another perspective, proposed least-squares SVM (LS-SVM), which instead uses a nonsparse loss function: sum square error (SSE). This trick converts the inequality constraints in classical SVM to equality ones.

In this paper we propose to use a weighted LS-SVM to define a generalization of the ALS regression method. The rest of this paper is organized as follows. In Section 2 we give a brief review of ALS regression. In Section 3 we show how this can be extended to a nonparametric ALS regression model, whose estimation is based on an iterative weighted LS-SVM. In Section 4 we perform the numerical studies through two examples.

[†] The present research was conducted by the research fund of Dankook university in 2010.

[1] Professor, Department of Statistics, Dankook University, Gyeonggido 448-160, Korea.
E-mail: chwang@dankook.ac.kr

2. Asymmetric least squares regression

We begin with the data set $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$, thought of as a point cloud in $(d+1)$ -dimensional Euclidean space R^{d+1}, the $\boldsymbol{x}_i$ being $d \times 1$ covariate vectors and the y_i being scalar responses. The linear ALS regression method directly builds the functional relations between the predictors $\boldsymbol{X}$ and the τth expectile $\mu_\tau(\boldsymbol{x})$ by the following minimization problem

$$\min_{\mu_\gamma \in \Omega} \sum_{i=1}^{n} \rho_\tau(y_i - \mu_\tau(\boldsymbol{x_i})), \tag{2.1}$$

where Ω is the function space from $R^d \rightarrow R$ and $\rho_\tau(\cdot)$ is the asymmetric least squared error loss function

$$\rho_\tau(r) = \begin{cases} \tau r^2, & r > 0 \\ (1-\tau) r^2, & \text{otherwise} \end{cases}, \tag{2.2}$$

The minimization problem (2.1) can be expressed in the way of least asymmetrically weighted squares format as follows:

$$\min_{\mu_\gamma \in \Omega} \sum_{i=1}^{n} v_i(\tau)(y_i - \mu_\tau(\boldsymbol{x_i}))^2 \tag{2.3}$$

with

$$v_i(\tau) = \begin{cases} \tau, & y_i > \mu_\tau(\boldsymbol{x_i}) \\ 1-\tau, & y_i \leq \mu_\tau(\boldsymbol{x_i}) \end{cases}. \tag{2.4}$$

To make the above minimization problem tractable, Newey and Powell (1987) suggested a linear function form of $\mu_\tau(\boldsymbol{x})$

$$\mu_\tau(\boldsymbol{x}) = b + \boldsymbol{w}^t \boldsymbol{x}, \tag{2.5}$$

where $b \in R$ and $\boldsymbol{w} \in R^d$ are the parameters to be determined according to the following optimization problem

$$\min_{b, \boldsymbol{w}} \sum_{i=1}^{n} \rho_\tau(y_i - b - \boldsymbol{w}^t \boldsymbol{x}_i). \tag{2.6}$$

See for details Efron (1991), Yao and Tong (1996) and Taylor (2008).

3. ALS regression using weighted LS-SVM

3.1. Weighted LS-SVM

Given a training data set $\{\boldsymbol{x_i}, y_i\}_{i=1}^{n}$ with each input $\boldsymbol{x_i} \in R^d$ and corresponding response $y_i \in R$, we consider the following optimization problem in primal weight space given weight on each observation:

$$L(\boldsymbol{w}, b, \boldsymbol{e}) = \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \frac{\gamma}{2}\sum_{i=1}^{n} v_{ii}e_i^2 \tag{3.1}$$

subject to equality constraints $y_i - \boldsymbol{w}^t\boldsymbol{\phi}(\boldsymbol{x_i}) - b = e_i, i = 1, \cdots, n$, where v_{ii} are weights determined in an appropriate way according to the given problem. Here $\boldsymbol{\phi} : R^d \to R^{d_f}$ is a nonlinear feature mapping function which maps the input space into a higher dimensional(possibly infinite dimensional) feature space, weight vector $\boldsymbol{w}$ in R^{d_f} in primal weight space, error variables $e_i \in R$ and bias term b. It is well known that $\boldsymbol{\phi}(\boldsymbol{x_i})^t\boldsymbol{\phi}(\boldsymbol{x_j}) = K(\boldsymbol{x_i}, \boldsymbol{x_j})$, which are obtained from the application of Mercer (1909)'s conditions. The cost function with squared error and regularization corresponds to a form of ridge regression. To find minimizers of the objective function, we can construct the Lagrangian function as follows:

$$L(\boldsymbol{w}, b, \boldsymbol{e}; \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \frac{\gamma}{2}\sum_{i=1}^{n} v_{ii}e_i^2 - \sum_{i=1}^{n} \alpha_i(\boldsymbol{w}^t\boldsymbol{\phi}(\boldsymbol{x}_i) + b + e_i - y_i), \tag{3.2}$$

where α_i's are the Lagrange multipliers. Then, the conditions for optimality are given by

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{0} \to \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\boldsymbol{\phi}(\boldsymbol{x}_i) \\
&\frac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{n} \alpha_i = 0 \\
&\frac{\partial L}{\partial e_i} = 0 \to e_i = \frac{1}{\gamma v_{ii}}\alpha_i, \quad i = 1, \cdots, n \\
&\frac{\partial L}{\partial \alpha_i} = 0 \to e_i - y_i + \boldsymbol{w}^t\boldsymbol{\phi}(\boldsymbol{x}_i) + b = 0, \quad i = 1, \cdots, n
\end{aligned} \tag{3.3}$$

After eliminating e_i and $\boldsymbol{w}$, we have the solution by the following linear equations,

$$\begin{bmatrix} \boldsymbol{K} + \frac{1}{\gamma}\boldsymbol{V}^{-1} & \boldsymbol{1} \\ \boldsymbol{1}' & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{3.4}$$

where $K_{ij} = \boldsymbol{\phi}(\boldsymbol{x_i})^t\boldsymbol{\phi}(\boldsymbol{x_j}) = K(\boldsymbol{x_i}, \boldsymbol{x_j})$ and $\boldsymbol{V}$ is a diagonal matrix of v_{ii}'s.

Finally, for a given $\boldsymbol{x}$ in dual space the nonlinear LS-SVM becomes

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{3.5}$$

Then, for the given training data set, $\boldsymbol{f_\lambda} = (f(\boldsymbol{x_1}), \cdots, f(\boldsymbol{x_n}))^t$ can be expressed as the linear combination of y_i's as follows:

$$\boldsymbol{f_\lambda} = \boldsymbol{H}\boldsymbol{y}, \tag{3.6}$$

where $\boldsymbol{\lambda}$ is the set of the regularization parameter and the kernel parameter, $\boldsymbol{H} = (\boldsymbol{K}, \boldsymbol{1})\boldsymbol{S_1}$ and $\boldsymbol{S}_1$ is a $(n+1) \times n$ submatrix of the inverse of the leftmost matrix $\boldsymbol{S}$ of (3.4) such that $\boldsymbol{S^{-1}} = \big(\boldsymbol{S_1}, \boldsymbol{S_2}\big)$.

Note that the ordinary LS-SVM can be seen a weighted LS-SVM with the identity weight matrix such that $\boldsymbol{V} = \boldsymbol{I}$.

3.2. Iterative weighted LS-SVM for nonlinear ALS regression

A distribution of random variable is characterized by its expectiles similar to its characterization by quantiles. Quantiles have a strong intuitive appealing, but expectiles are known to be easier to compute and more efficient.

Given a training data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with each input $\boldsymbol{x}_i \in R^d$ and corresponding response $y_i \in R$, if the distribution of y_i's are known, the θth quantile regression function given $\boldsymbol{x}$, $q_\theta(\boldsymbol{x})$ can be obtained by minimizing the following function:

$$(1-\theta)\int_{-\infty}^{q_\theta(\boldsymbol{x})} |y - q_\theta(\boldsymbol{x})| f(y|\boldsymbol{x}) dy + \theta \int_{q_\theta(\boldsymbol{x})}^{\infty} |y - q_\theta(\boldsymbol{x})| f(y|\boldsymbol{x}) dy. \tag{3.7}$$

Similarly, the τth expectile regression function given $\boldsymbol{x}$, $\mu_\tau(\boldsymbol{x})$ can be obtained by minimizing the following function:

$$(1-\tau)\int_{-\infty}^{\mu_\tau(\boldsymbol{x})} (y - \mu_\tau(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) dy + \tau \int_{\mu_\tau(\boldsymbol{x})}^{\infty} (y - \mu_\tau(\boldsymbol{x}))^2 f(y|\boldsymbol{x}) dy. \tag{3.8}$$

If we apply kernel trick to define a nonlinear generalization of linear ALS regression, the τth expectile regression function given $\boldsymbol{x}$, $\mu_\tau(\boldsymbol{x})$ can be obtained as $\mu_\tau(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{\phi}(\boldsymbol{x}) + b$, where $\boldsymbol{w}$ and b are solutions to the objective function (3.1) of the weighted LS-SVM with $v_{ii} = \tau I(y_i - \widehat{\mu}_\tau(\boldsymbol{x_i}) > 0) + (1-\tau) I(y_i - \widehat{\mu}_\tau(\boldsymbol{x_i}) \leq 0)$. From the application of Mercer (1909)'s conditions, the τth expectile regression function can be expressed as $\mu_\tau(\boldsymbol{x}) = K\boldsymbol{\alpha} + b$ where $\boldsymbol{\alpha}$ and b are the solutions to (3.4) with $\boldsymbol{V} = diag\{v_{ii}\}$. But since v_{ii} contains $\boldsymbol{\alpha}$ and b, $\boldsymbol{\alpha}$ and b cannot be obtained in a step using a weighted LS-SVM but from the iterative method including a weighted LS-SVM in each step as follows:

(0) Find $\mu_\tau^{(0)}(\boldsymbol{x}) = K\boldsymbol{\alpha} + b$ using the weighted LS-SVM with $\boldsymbol{V} = \boldsymbol{I}$.
(1) Find $v_{ii}^{(k)} = \tau I(y_i - \widehat{\mu}_\tau^{(k-1)}(\boldsymbol{x_i}) > 0) + (1-\tau) I(y_i - \widehat{\mu}_\tau^{(k-1)}(\boldsymbol{x_i}) \leq 0)$.
(2) Find $\mu_\tau^{(k)}(\boldsymbol{x}) = K\boldsymbol{\alpha} + b$ using the weighted LS-SVM with $\boldsymbol{V} = diag\{v_{ii}^{(k)}\}$.
(3) Iterate until convergence.

The functional structures of the proposed ALS regression is characterized by the regularization parameter and the kernel parameter. To select these parameters of LS-SVM we define the cross validation (CV) function as follows:

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^n v_{ii}(y_i - \widehat{\mu}_\tau^{(-i)}(\boldsymbol{x}_i))^2 \tag{3.9}$$

where v_{ii} is the final estimate of $v_{ii} = \tau I(y_i - \widehat{\mu}_\tau(\boldsymbol{x_i}) > 0) + (1-\tau) I(y_i - \widehat{\mu}_\tau(\boldsymbol{x_i}) \leq 0)$ and $\widehat{\mu}_\tau^{(-i)}(\boldsymbol{x_i})$ is the expectile regression function estimated without ith observation. Since for each candidates of parameters, $\widehat{\mu}_\tau^{(-i)}(\boldsymbol{x_i})$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally formidable. By leaving-out-one lemma (Craven and Wahba, 1979)

$$y_i - \widehat{\mu}_\tau^{(-i)}(\boldsymbol{x}_i) \approx \frac{y_i - \widehat{\mu}_\tau(\boldsymbol{x}_i)}{1 - \dfrac{\partial \widehat{\mu}_\tau(\boldsymbol{x}_i)}{\partial y_i}} = \frac{y_i - \widehat{\mu}_\tau(\boldsymbol{x}_i)}{1 - h_{ii}}, \tag{3.10}$$

where h_{ii} is the ith diagonal element of $\boldsymbol{H}$ which is a hat matrix such that $\widehat{\boldsymbol{\mu}}_\tau = (\widehat{\mu}_\tau(\boldsymbol{x_i}), \cdots, \widehat{\mu}_\tau(\boldsymbol{x_n}))^t = \boldsymbol{Hy}$. Then the ordinary cross validation (OCV) function is obtained as

$$OCV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} v_{ii}\left(\frac{y_i - \widehat{\mu}_\tau(\boldsymbol{x}_i)}{1-h_{ii}}\right)^2. \tag{3.11}$$

Replacing h_{ii} by their average $tr(\boldsymbol{H})/n$, the generalized cross validation (GCV) function can be obtained as

$$GCV(\lambda) = \frac{n}{(n - tr(\boldsymbol{H}))^2}\sum_{i=1}^{n} v_{ii}(y_i - \widehat{\mu}_\tau(\boldsymbol{x}_i))^2. \tag{3.12}$$

4. Numerical studies

In this section, we illustrate the performance of the iterative weighted LS-SVM for ALS regression through the simulated data sets and a well-known motorcycle data set in Table 1 on page 302 of Haerdle (1989).

Example 4.1 For this example we generate 100 data sets of size 150 in a similar manner to Shim *et al.* (2009). The univariate input observations x's are drawn from a uniform distribution on the interval $(0, 1)$, the corresponding responses y's are drawn from a univariate normal distribution with mean and variance that vary smoothly with x as follows:

$$y \sim N(f(x), \sigma^2),$$

where $f(x) = \sin(1.5x)\sin(2.5x)$ and $\sigma^2 = 0.01 + 0.25(1 - \sin^2(2.5x))$. The τth expectile regression function given x, $\mu_\tau(x)$, is obtained by solving the following equation (Schnabel and Eilers, 2009),

$$\tau = \frac{G(\mu_\tau(x)) - \mu_\tau(x)F(\mu_\tau(x))}{2(G(\mu_\tau(x)) - \mu_\tau(x)F(\mu_\tau(x))) + (\mu_\tau(x) - \mu)}$$

where $F(u)$ and $G(u)$ are the distribution function and the partial moment function of $N(f(u), \sigma^2)$, respectively. Here μ is the mean of the underlying distribution F and satisfies $G(\infty) = \mu$. The RBF kernel, $K(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/s^2)$, is utilized and the optimal values of (γ, s^2) are chosen by GCV function in (3.12). We obtained the mean squared error of $(\widehat{\mu}_\tau(x) - \mu_\tau(x))$ and its standard deviation for $\tau = 0.05, 0.5, 0.95$ in each data set. As results we obtained the averages of 100 mean squared errors and their standard deviations for the proposed method as (0.0064, 0.0042) for $\tau = 0.05$, (0.0044, 0.0029) for $\tau = 0.5$ and (0.0082, 0.0052) for $\tau = 0.95$, respectively, this implies that the proposed method has good estimation performance in this example. Figure 4.1 shows that the estimated expectile regression functions (dotted lines) and true expectile regression functions (solid lines) are superimposed on the scatter plots. As seen from Figure 4.1, the estimated expectile regression functions reflect well the heteroscedastic structure of the error terms. They have their maxima at different x values. For example, the 0.05th, 0.5th and 0.95th expectile regression functions have maxima at $x = 0.67$, 0.74 and 0.88, respectively.

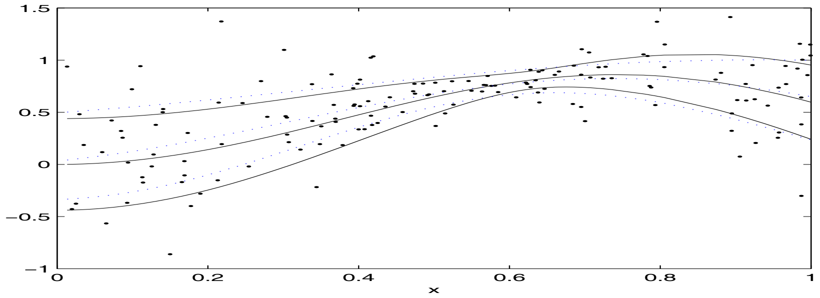

Figure 4.1 An illustration of the proposed method for expectile regression analysis for a simulated data set of size 150 generated from the process in Example 4.1

Example 4.2 In this example we consider the motorcycle data, which have been widely used to demonstrate the performance of nonparametric regression methods. The data were collected performing crash tests with dummies sitting on motorcycles. The head acceleration (y) of the dummies (in g) was recorded a certain time measured in milliseconds (x) after they had hit a wall. The RBF kernel is utilized and from GCV function in (3.12) the value of (γ, s^2) is chosen as (1100, 0.5) for $\tau = 0.05$, (100, 0.75) for $\tau = 0.5$ and (100, 0.5) for $\tau = 0.95$, respectively. In Figure 4.2 the estimated expectile regression functions for $\tau = 0.05, 0.5, 0.95$ are superimposed on the scatter plots. As seen from Figure 4.2, as x increases the variance of y increases when $x < 33$ and decreases when $x > 33$. The estimated expectile regression functions do reasonably well even in the region beyond 50 milliseconds where the data points are so sparse that all the expectile functions want to coalesce. As a whole, the proposed method seems to give a good estimation of true expectile regression functions.

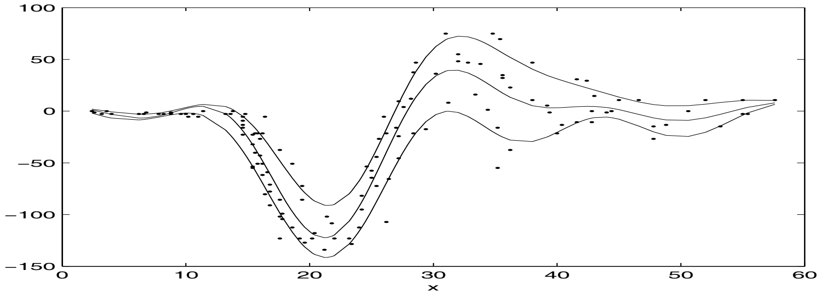

Figure 4.2 The estimated expectile regression functions for $\tau = 0.05$, 0 5, 0.95 are superimposed on the scatter plots of the motorcycle data

5. Conclusion

ALS regression is an increasingly popular method for estimating the expectiles of a distribution conditional on the values of covariates. In this paper, we dealt with estimating expectile regression function using a weighted LS-SVM, which has lots of potential applications.

Through the examples we recognized that the proposed procedure can be useful in characterizing the relationship between a response variable and covariates when the behaviour of nonaverage individuals is of interest. We also recognized that the proposed precedure derives the satisfying solutions.

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics*, **31**, 377-403.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, **1**, 93-125.

Haerdle, W. (1989). *Applied nonparametric regression*, Cambridge University Press, New York.

Hwang, C. (2010a). *M*-quantile regression using kernel machine technique. *Journal of the Korean Data & Information Science Society*, **21**, 973-981.

Hwang, C. (2010b). Support vector quantile regression for longitudinal data. *Journal of the Korean Data & Information Science Society*, **21**, 309-316.

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.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.

Schnabel, S. K. and Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics and Data Analysis*, **53**, 4168-4177.

Seok, K. H. (2010). Semi-supervised classification with LS-SVM formulation. *Journal of the Korean Data & Information Science Society*, **21**, 461-470.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of the Korean Data & Information Science Society*, **20**, 949-964.

Shim, J., Seok, K. H. and Hwang, C. (2009). Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics*, **24**, 83-94.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293-300.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, **6**, 231-252.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer, New York.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, **6**, 273-292.