Asymmetric ν-tube Support Vector Regression

Xiaolin Huang[a,*], Lei Shi[a,c], Kristiaan Pelckmans[b], Johan A.K. Suykens[a]

[a] *KU Leuven, Department of Electrical Engineering, ESAT-STADIUS, B-3001, Leuven, Belgium*
[b] *Department of Information Technology, Uppsala University, SE-751 05, Uppsala, Sweden*
[c] *School of Mathematical Sciences, Fudan University, 200433, Shanghai, P.R. China*

Abstract

Finding a tube of small width that covers a certain percentage of the training data samples is a robust way to estimate a location: the values of the data samples falling outside the tube have no direct influence on the estimate. The well-known ν-tube Support Vector Regression (ν-SVR) is an effective method for implementing this idea in the context of covariates. However, the ν-SVR considers only one possible location of this tube: it imposes that the amount of data samples above and below the tube are equal. The method is generalized such that those outliers can be divided asymmetrically over both regions. This extension gives an effective way to deal with skewed noise in regression problems. Numerical experiments illustrate the computational efficacy of this extension to the ν-SVR.

Keywords: robust regression, ν-tube support vector regression, asymmetric loss, quantile regression

1. Introduction

Since its introduction by Schölkopf et al. (2000), the ν-tube Support Vector Regression (ν-SVR) has become a standard tool in nonparametric regression tasks. The ν-SVRs extend standard Support Vector Regression techniques given by Vapnik (1995) via (i) enforcing a fraction of the data samples to lie inside a tube, as well as (ii) minimizing the width of this tube. Mathematically, training such a tube $[f(x)-\varepsilon, f(x)+\varepsilon]$ can be formulated as the following optimization problem,

$$\begin{aligned} \min_{f,\varepsilon} \quad & \varepsilon \\ \text{s.t.} \quad & \sum\nolimits_{i=1}^{n} \mathcal{I}\left(y_i \in [f(x_i)-\varepsilon, f(x_i)+\varepsilon]\right) \geq \rho n, \end{aligned} \tag{1}$$

where $\{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ are the data samples, $0 \leq \rho \leq 1$ is a user defined constant, and $\mathcal{I}(a)$ stands for an indicator function, which equals one when a is true and equals zero otherwise. Unlike traditional point-regression methods, (1) focuses on estimation of the confidence region directly, which is called *support tube* by Pelckmans et al. (2009). One can find the corresponding statistical discussion therein.

Apparently, the values of the data samples falling outside the tube have no direct influence on the result of (1), which is quite robust to *outliers*, since the outliers probably fall outside. In fact, this idea has appeared in robust regression and is known as the least median squares regression, which is proposed by Rousseeuw (1984), Rousseeuw and Leroy (1987). Denote the k-th maximum of $\{u_i\}_{i=1}^n$ by $\max_{1\leq i\leq n}^k\{u_i\}$:

$$\max\nolimits_{1\leq i\leq n}^k\{u_i\} = u_{\Gamma(k)} \quad \text{with} \quad u_{\Gamma(1)} \geq u_{\Gamma(2)} \geq \cdots \geq u_{\Gamma(n)}.$$

*Corresponding author. ESAT-STADIUS, Kasteelpark Arenberg 10, bus 2446, 3001 Heverlee, Belgium; Tel: +32-16328653, Fax: +32-16321970.

Email addresses: huangxl06@mails.tsinghua.edu.cn (Xiaolin Huang), leishi@fudan.edu.cn (Lei Shi), kp@it.uu.se (Kristiaan Pelckmans), johan.suykens@esat.kuleuven.be (Johan A.K. Suykens)

Then the least median squares estimator can be written as

$$\min_{f} \quad \max_{1\leq i\leq n}^{k} \left\{ (y_i - f(x_i))^2 \right\}. \tag{2}$$

One can observe the equivalence between (2) and (1) when $\rho = k/n$. If the median squared error is minimized, (2) is regarded as the most robust estimator in view of the breakdown point defined by Donoho and Huber (1982). The idea of minimizing the median error also has been discussed for classification tasks by Ma et al. (2011) and Tsyurmasto et al. (2013). For the least median squares regression, there have been some approximation algorithms proposed by Tichavsky (1991), Boček and Lachout (1995), Olson (1997), Verardi and Croux (2009), and Winker et al. (2011). The most popular method for computing the least median squares estimator is PROGRESS suggested by Rousseeuw and Leroy (1987) and modified by Rousseeuw and Hubert (1997). Besides, several algorithms have been developed for finding the global optimum, see the works of Steele and Steiger (1986) and Stromberg (1993).

Regression method (1) enjoys robustness to outliers, but it is non-convex. It can be modeled as a mixed integer linear programming (MILP), which has been proved to be NP-hard, see the discussion given by Huang et al. (2012). The solving time of an NP-hard problem is not acceptable for large-scale problems. Thus, we need a convex proxy for (1) for computational efficacy. Actually, the ν-SVR can be regarded as such a convex approximation, i.e., one can get a narrow tube, though not the optimal one, to cover ρ percentage samplings via solving a convex problem. Let us consider the case that f is chosen from the set of affine functions. Then the ν-SVR in primal space refers to the following optimization problem,

$$\begin{aligned} \min_{w,b,\varepsilon} \quad & \frac{1}{2\gamma} w^T w + \nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon\left(y_i - (w^T x_i + b)\right) \\ \text{s.t.} \quad & \varepsilon \geq 0, \end{aligned} \tag{3}$$

where $\nu \geq 0$ is a user defined parameter and $L_\varepsilon(u)$ is the ε insensitive zone loss as defined below:

$$L_\varepsilon(u) = \begin{cases} u - \varepsilon, & u \geq \varepsilon, \\ 0, & -\varepsilon < u < \varepsilon, \\ -u - \varepsilon, & u \leq -\varepsilon. \end{cases}$$

It has been proved by Schölkopf et al. (2000) that the minimizer of (3) satisfies:

$$\sum_{i=1}^{n} \mathcal{I}\left(y_i \in [w^T x_i + b - \varepsilon, w^T x_i + b + \varepsilon]\right) \geq (1-\nu)n,$$

which means that by setting $\nu = 1 - \rho$, the ν-SVR provides a feasible solution to (1). The difference between (1) and (3) can be observed by a linear regression task shown in Fig.1. In this example, we pursue a tube covering half of the data samples, which are displayed by blue crosses. With a suitable ν, the ν-SVR (3) results in a good solution to (1). This tube is shown by red solid lines and covers 50% of the data. This solution is not yet the optimal to (1). Since this problem scale is small, we can get the optimal tube via solving the MILP formulation of (1) by iLog CPLEX. The optimal tube is shown by black dotted lines, which has the smallest width among all the tubes covering 50% of the data. One noticeable point is that there are 8 points above the optimal tube and 2 points below it. However, the amount of outliers above and below the red solid tube are imposed to be equal by (3), which reserves from the symmetry of L_ε.

Motivated by these observations, this paper extends L_ε to an asymmetric loss. Then an asymmetric ν-tube support vector regression (asymmetric ν-SVR) is established. By the proposed method, we can find an asymmetric tube, above and below which the outliers are distributed asymmetrically. The asymmetric tube is more flexible and can give a better solution to (1). This is especially suitable for dealing with asymmetric noise. Many applications have asymmetric noise. For example, when the measurement is close to saturation, then the noise may follow a distribution of which one tail is long but the other tail is truncated. Some existing discussions about asymmetric noise and the related methods can be found in literature, e.g.,

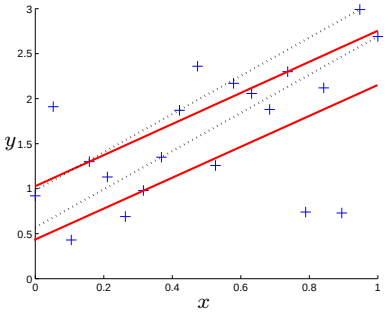


Figure 1: We require 50% of the data (blue crosses) are covered by a tube. The result of the ν-SVR (3) is illustrated by red solid lines. Black dotted lines display the optimal solution to (1). The two tubes cover the same amount of the data samples, but the distributions of the outliers are different.

Kassam et al. (1982), Hubert et al. (2009), Le Masne et al. (2009), and Solli et al. (2010). For these skewed noises, it is reasonable to require an asymmetric tube.

The remainder of the paper is organized as follows: the formulation of the asymmetric ν-SVR is given and the properties are discussed in Section 2. Section 3 gives its dual problem and an algorithm. Then the proposed method is evaluated by numerical experiments in Section 4. Section 5 ends the paper with concluding remarks.

2. Asymmetric ν-tube Support Vector Regression

As discussed previously, the ν-SVR (3) results in a symmetric tube. To pursue an asymmetric tube, we extend L_ε into an asymmetric loss:

$$L_\varepsilon^p(u) = \begin{cases} \frac{1}{2p}(u-\varepsilon), & u \geq \varepsilon, \\ 0, & -\varepsilon < u < \varepsilon, \\ \frac{1}{2(1-p)}(-u-\varepsilon), & u \leq -\varepsilon, \end{cases}$$

where p is the parameter related to asymmetry. When $p = 0.5$, L_ε^p reduces to L_ε. The plots of L_ε^p for some p values are shown in Fig.2. L_ε^p also can be constructed by introducing an ε insensitive zone to the pinball loss, which has been widely applied and studied in the field of quantile regression, see Koenker (2005) for parametric methods. For nonparametric methods, one can refer to the literature of Steinwart and Christmann (2008), Steinwart and Christmann (2011). Introducing insensitive zone into quantile regression brings sparseness and its approximation behavior has been discussed by Xiang et al. (2012).

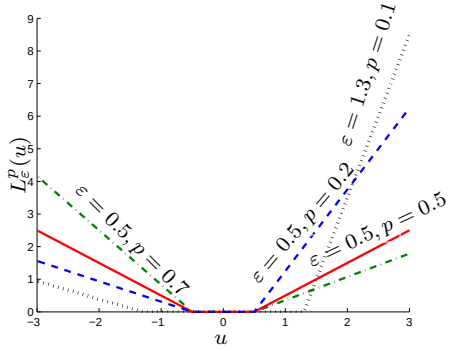


Figure 2: Loss value of $L_\varepsilon^p(u)$ for different ε and p. When $p = 0.5$, $L_\varepsilon^p(u)$ reduces to the ε insensitive zone loss $L_\varepsilon(u)$.

Replacing L_ε in (3) by L^p_ε, we obtain the following asymmetric ν-tube support vector regression (asymmetric ν-SVR),

$$\begin{aligned} \min_{w,b,\varepsilon} \quad & \frac{1}{2\gamma} w^T w + \nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n} L^p_\varepsilon\left(y_i - (w^T x_i + b)\right) \\ \text{s.t.} \quad & \varepsilon \geq 0. \end{aligned} \tag{4}$$

We can introduce a nonlinear feature map $\phi(x)$ and then solve the asymmetric ν-SVR to find a good tube in the feature space. In this case, the proposed asymmetric ν-SVR (4) can be equivalently transformed into a quadratic programming (QP) problem as below,

$$\begin{aligned} \min_{w,b,\varepsilon,e^+,e^-} \quad & \frac{1}{2\gamma} w^T w + \nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(e_i^+ + e_i^-) \\ \text{s.t.} \quad & y_i - (w^T\phi(x_i) + b + \varepsilon) \leq 2pe_i^+, \forall i, \\ & (w^T\phi(x_i) + b - \varepsilon) - y_i \leq 2(1-p)e_i^-, \forall i, \\ & \varepsilon \geq 0, e_i^+ \geq 0, e_i^- \geq 0, \forall i. \end{aligned} \tag{5}$$

In the above formulation, e_i^+ characterizes the sample above the upper boundary. Similarly, e_i^- is related to the distance to the lower boundary. The ν-SVR gives equal emphasis to e_i^+ and e_i^-, which results in an equal amount of points above and below the tube. In (5), we give different weights to e_i^+ and e_i^-. Heuristically, when p is larger than 0.5, the penalties to the samples above the upper boundary are smaller, which implies that we tolerate more outliers above the tube than below it. In the asymmetric ν-SVR, p and ν control the fraction of points above and below the tube, guaranteed by the following proposition.

Proposition 1. *The optimal solution to (5) satisfies:*

$$\begin{aligned} \sum_{i=1}^{n} \mathcal{I}\left(y_i > w^T\phi(x_i) + b + \varepsilon\right) &\leq p\nu n, \\ \sum_{i=1}^{n} \mathcal{I}\left(y_i < w^T\phi(x_i) + b - \varepsilon\right) &\leq (1-p)\nu n. \end{aligned}$$

Proof. Introducing the Lagrange multipliers $\alpha_i^+, \beta_i^+, \alpha_i^-, \beta_i^-, \zeta$, which correspond to the constraints of (5) and are nonnegative, we get the Lagrangian

$$\begin{aligned} & \mathcal{L}(w, b, \varepsilon, e^+, e^-; \alpha^+, \beta^+, \alpha^-, \beta^-, \zeta) \\ = \; & \frac{1}{2\gamma} w^T w + \nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n} e_i^+ + \frac{1}{n}\sum_{i=1}^{n} e_i^- - \sum_{i=1}^{n} \alpha_i^+\left(w^T\phi(x_i) + b + \varepsilon - y_i + 2pe_i^+\right) \\ & + \sum_{i=1}^{n} \alpha_i^-\left(w^T\phi(x_i) + b - \varepsilon - y_i - 2(1-p)e_i^-\right) - \zeta\varepsilon - \sum_{i=1}^{n}\beta_i^+ e_i^+ - \sum_{i=1}^{n}\beta_i^- e_i^-. \end{aligned} \tag{6}$$

From the saddle point condition, we have that

$$\frac{\partial\mathcal{L}}{\partial\varepsilon} = \nu - \zeta - \sum_{i=1}^{n}\left(\alpha_i^+ + \alpha_i^-\right) = 0, \tag{7}$$

$$\frac{\partial\mathcal{L}}{\partial b} = \sum_{i=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right) = 0, \tag{8}$$

$$\frac{\partial\mathcal{L}}{\partial e_i^+} = \frac{1}{n} - 2p\alpha_i^+ - \beta_i^+ = 0, i = 1, 2, \ldots, n, \tag{9}$$

$$\frac{\partial\mathcal{L}}{\partial e_i^-} = \frac{1}{n} - 2(1-p)\alpha_i^- - \beta_i^- = 0, i = 1, 2, \ldots, n. \tag{10}$$

According to (7) and (8) and the fact $\zeta \geq 0$, we know that

$$\sum_{i=1}^{n} \alpha_i^+ = \sum_{i=1}^{n} \alpha_i^- \leq \frac{\nu}{2}. \tag{11}$$

For any point above the tube, there is

$$y_i - (w^T \phi(x_i) + b + \varepsilon) = 2pe_i^+ \ \text{ and } \ e_i^+ > 0.$$

According to the complementary slackness condition, we have $\beta_i^+ = 0$. Then (9) tells us $\alpha_i^+ = \frac{1}{2np}$. Therefore, the amount of the data above the tube cannot exceed $p\nu n$, otherwise

$$\sum_{i=1}^{n} \alpha_i^+ > p\nu n \frac{1}{2np} = \frac{\nu}{2},$$

which conflicts with (11). Thus, p and ν control the fraction of points located above the tube, i.e.,

$$\sum_{i=1}^{n} \mathcal{I}\left(y_i > w^T \phi(x_i) + b + \varepsilon\right) \leq p\nu n.$$

Similarly, the complementary slackness condition and (10) together lead to

$$\sum_{i=1}^{n} \mathcal{I}\left(y_i < w^T \phi(x_i) + b - \varepsilon\right) \leq (1-p)\nu n.$$

□

From Proposition 1, we further conclude that the solution of the asymmetric ν-SVR satisfies

$$\sum_{i=1}^{n} \mathcal{I}\left(y_i \in [w^T \phi(x_i) + b - \varepsilon, w^T \phi(x_i) + b + \varepsilon]\right) \geq (1-\nu)n,$$

which means that ν controls the fraction of the data falling outside the tube. The behavior of ν in (4) is the same as in the ν-SVR. One noticeable point is that the computational complexity of the asymmetric ν-SVR (4) is similar to that of the ν-SVR (3), since the corresponding QPs have the same numbers of optimization variables and inequality constraints.

3. Nonparametric Formulation

In the above, we analyze the performance of the asymmetric ν-SVR in the primal space. Now let us focus on the Lagrangian (6). Its saddle point conditions include (7) – (10) and

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{\gamma} w - \sum_{i=1}^{n} \left(\alpha_i^+ - \alpha_i^-\right) \phi(x_i) = 0.$$

Then we let $\lambda_i^+ = \gamma\alpha_i^+, \lambda_i^- = \gamma\alpha_i^+$, get the dual problem to (4), and establish the following nonparametric asymmetric ν-SVR,

$$\begin{aligned}
\min_{\lambda^+,\lambda^-} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\lambda_i^+ - \lambda_i^-)^T\mathcal{K}(x_i,x_j)(\lambda_j^+ - \lambda_j^-) - \sum_{i=1}^{n} y_i(\lambda_i^+ - \lambda_i^-) \\
\text{s.t.} \quad & \sum_{i=1}^{n}(\lambda_i^+ - \lambda_i^-) = 0, \\
& \sum_{i=1}^{n}(\lambda_i^+ + \lambda_i^-) \le \nu\gamma, \\
& 0 \le \lambda_i^+ \le \frac{\gamma}{2p}, \forall i = 1,2,\ldots,n, \\
& 0 \le \lambda_i^- \le \frac{\gamma}{2(1-p)}, \forall i = 1,2,\ldots,n,
\end{aligned} \tag{12}$$

where $\mathcal{K}(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ corresponds to a positive definite kernel. Any positive-definite kernel, such as radial basis function (RBF) kernel or polynomial kernel, is applicable to (12). After solving (12), we get the optimal dual variables λ^+, λ^-, and calculate $w^T\phi(x)$ by

$$w^T\phi(x) = \sum_{i=1}^{n}\left(\lambda_i^+ - \lambda_i^-\right)\mathcal{K}(x,x_i).$$

To compute the bias term b and the width ε, we consider the sample data (x_i,y_i) with $0 < \lambda_i^+ < \frac{\gamma}{2p}$, denoted by $\mathcal{S}_0^+ = \left\{i : 0 < \lambda_i^+ < \frac{\gamma}{2p}\right\}$. These points are located on the upper boundary of the tube, i.e.,

$$\sum_{j=1}^{n}\left(\lambda_j^+ - \lambda_j^-\right)\mathcal{K}(x_j,x_i) + b + \varepsilon = y_i, \forall i \in \mathcal{S}_0^+.$$

Similarly, for samples in $\mathcal{S}_0^- = \left\{i : 0 < \lambda_i^- < \frac{\gamma}{2(1-p)}\right\}$, there is

$$\sum_{j=1}^{n}\left(\lambda_j^+ - \lambda_j^-\right)\mathcal{K}(x_j,x_i) + b - \varepsilon = y_i, \forall i \in \mathcal{S}_0^-.$$

Using one element in $\mathcal{S}_0^+$ and one element in $\mathcal{S}_0^-$, the optimal b and ε can be calculated. As a result, we obtain a nonparametric formulation of a tube covering a certain percentage of the data. The center of the tube is expressed using the dual variables as

$$f(x) = \sum_{i=1}^{n}\left(\lambda_i^+ - \lambda_i^-\right)\mathcal{K}(x,x_i) + b. \tag{13}$$

Consequently, the upper and the lower boundary of the tube is obtained as $f(x) + \varepsilon$ and $f(x) - \varepsilon$.

Similarly to Proposition 1, the fraction of the data samples above and below the tube obtained from (12) is described by the following proposition. This proposition comes from the constraints in (12) and the complementary slackness condition. The proof is similar to that of Proposition 1 and is omitted here.

Proposition 2. *The optimal solution to (12) satisfies:*

$$\begin{aligned}
\sum_{i=1}^{n}\mathcal{I}\left(y_i > \sum_{j=1}^{n}(\lambda_j^+ - \lambda_j^-)\mathcal{K}(x_i,x_j) + b + \varepsilon\right) &\le p\nu n, \\
\sum_{i=1}^{n}\mathcal{I}\left(y_i < \sum_{j=1}^{n}(\lambda_j^+ - \lambda_j^-)\mathcal{K}(x_i,x_j) + b - \varepsilon\right) &\le (1-p)\nu n,
\end{aligned}$$

and

$$\sum_{i=1}^{n} \mathcal{I}\left(y_i \in \left[\sum_{j=1}^{n}(\lambda_j^+ - \lambda_j^-)\mathcal{K}(x_i, x_j) + b - \varepsilon, \sum_{j=1}^{n}(\lambda_j^+ - \lambda_j^-)\mathcal{K}(x_i, x_j) + b + \varepsilon\right]\right) \geq (1-\nu)n.$$

The sparseness of the asymmetric ν-SVR is similar to that of the ν-SVR: for a training point (x_i, y_i) inside the tube, i.e., $f(x_i) - \varepsilon < y_i < f(x_i) + \varepsilon$, we have that $\lambda_i^+ = \lambda_i^- = 0$. Otherwise, $\lambda_i^+ - \lambda_i^- \neq 0$, which corresponds to a support vector. Similarly to the ν-SVR, the parameter ν in (12) bounds the fraction of support vectors. Additionally, p controls the location of support vectors.

The value of p should coincide with the skewness of the noise. Consider additive noise δ and assume that the mean of δ is zero. $\text{Prob}(\delta > 0)/\text{Prob}(\delta < 0)$ generally reflects the skewness of δ. Thus, we can first estimate the mean function by a least squares method and count the positive and negative residuals. Then p is heuristically set to be the fraction of positive residuals. In this paper, we tune p with the help of the least squares support vector machine (LS-SVM, Suykens and Vandewalle, 1999; Suykens et al., 2002). After estimating the mean function, one can measure the skewness by some robust methods proposed by Brys et al. (2004), Kim and White (2004). Another potential choice is the methods based on a Huber loss, which enhances the robustness. These robust methods are suitable to deal with outlier corrupted data but require more computational time. In this paper, we simply set p based on the LS-SVM for computational efficacy. If the outliers are heavy, one can consider robust methods or evaluate different p values by cross validation.

Now we summarize the above discussions and give the following asymmetric ν-SVR algorithm.

Algorithm 1: Asymmetric ν-SVR Algorithm

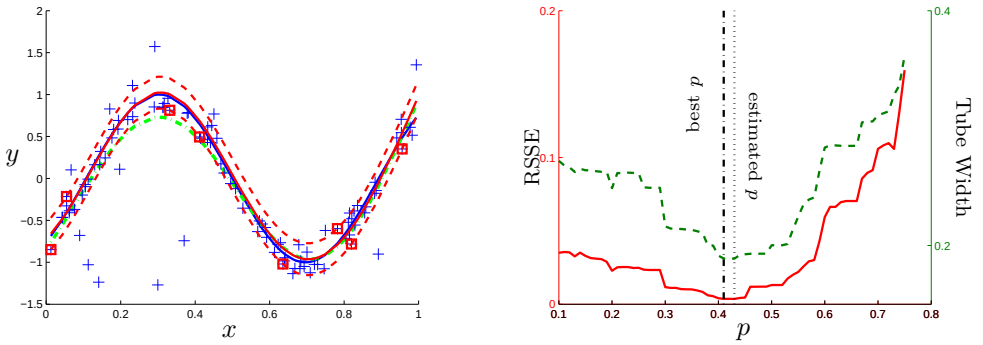
- Input $\{(x_i, y_i)\}_{i=1}^n$ (the training data), ρ (the required percentage of the covered training data);
- Choose the regularization constant γ, the positive definite kernel, and the kernel parameters;
- Use the LS-SVM to do estimation and denote the result by $\hat{y}_i$;
- Set $\nu := 1 - \rho$ and $p := \sum_i \mathcal{I}(y_i > \hat{y}_i)/n$;
- Solve the asymmetric ν-SVR (12);
- Return the tube $\left[f(x) - \varepsilon, f(x) + \varepsilon\right]$, where $f(x)$ is calculated by (13).

Next we illustrate the performance of the nonparametric asymmetric ν-SVR as a nonlinear example shown in Fig.3(a). The underlying function is displayed by a blue solid line. The observed data (blue crosses) are corrupted by noise and contain outliers. The existence of outliers makes the result of a least squares method using the RBF kernel

$$\mathcal{K}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

with $\sigma = 0.5$ (illustrated by a green dash-dotted line) deviate significantly from the underlying function. Then the nonparametric asymmetric ν-SVR (12) is applied to address this problem. We want the tube to cover 80% of the data samples, which also means that almost 80% of the dual variables equal zero. Different p values are related to different locations of the tube. The least squares method is first applied to estimate the mean value. We then count the positive and negative residuals and set p as the fraction of positive residuals: there are 43 data points above the least square regressor (green dash-dotted line) and 57 data below it, we thus set $p = \frac{43}{100}$. The result of the asymmetric ν-SVR (12) with $p = 0.43, \gamma = 100$ and RBF kernel ($\sigma = 0.5$) is displayed by its middle (red solid line) and boundaries (red dotted lines). There are 8 points above the obtained nonparametric tube and 11 below it. The obtained tube is narrow and the difference between the middle and the underlying function is small, illustrating the good performance of the asymmetric ν-SVR for asymmetric noise distribution.

We also test the performance of the asymmetric ν-SVR with different p values. For each p value, we solve (12) and plot the width of the obtained tube in Fig.3(b) by the dashed line. In this example, the



(a) (b)

Figure 3: An example of the nonparametric asymmetric ν-SVR. (a) The data are shown by blue crosses. The existence of outliers makes the result of the least squares method (green dash-dotted line) deviate from the underlying function (blue solid line). The tube obtained by (12) is illustrated by its middle (red solid line) and boundaries (red dotted lines). The points marked by red squares are located on top of the boundaries. In between the boundaries, the points correspond to zero dual variables. (b) The RSSEs and the tube widths corresponding to different p values are displayed by red solid and green dashed lines, respectively. The best choice is $p = 0.41$ and the p value selected based on the LS-SVM is $p = 0.43$.

underlying function $f(x)$ is known. Then for the estimation $\hat{f}(x)$, we can calculate the relative sum of the squared errors, defined by

$$\mathrm{RSSE} = \frac{\sum_{x\in V}\big(f(x) - \hat{f}(x)\big)^2}{\sum_{x\in V}\big(f(x) - \mathrm{E}(f)\big)^2},$$

where V is the set of concerned data and $\mathrm{E}(f)$ is the average value of $f(x), x \in V$. In this example, we consider the training set and report the RSSEs in Fig.3(b) by the solid line. Generally, a small tube width corresponds to a small RSSE. In the viewpoint of the tube width, the best choice is $p = 0.41$. But the performance of the asymmetric ν-SVR is not very sensitive to p. For example, based on the LS-SVM, we set $p = 0.43$ which results in good performance as well.

4. Numerical Experiments

In the numerical experiments, we evaluate the proposed asymmetric ν-SVR from the following three aspects:

- The asymmetric ν-SVR can be regarded as a convex approach to the non-convex problem (1). We compare the asymmetric ν-SVR with existing heuristics;
- We discuss the robustness of the result of the asymmetric ν-SVR on a linear problem;
- Finally, we compare the asymmetric ν-SVR with the ν-SVR and the LS-SVM on some standard data sets with outliers and asymmetric noise distributions.

In this paper, all the experiments are done using Matlab R2011a with Core 2-2.83 GHz, 2.96G RAM. The ν-SVR and the asymmetric ν-SVR are solved by the QP solver embedded in Matlab. The LS-SVM is implemented by LS-SVMLab v1.8, designed by De Brabanter et al. (2011).

4.1. Optimization performance

The proposed asymmetric ν-SVR can be regarded as a convex approach to (1), which tries to find a tube with small width so as to cover a required amount of samples. There have been some algorithms for this non-convex problem. Among these methods, PROGRESS established by Rousseeuw and Leroy (1987) Rousseeuw and Hubert (1997) is the most popular one. Support Vector Tube (SVT, Pelckmans et al.,

2009) is applicable as well. In the following, we apply these algorithms to real data sets downloaded from http://www.agoras.ua.ac.be/ and search an affine tube which contains at least k samples of each data set.

This experiment focuses on optimization performance not the generalization error. Thus, a large enough γ is used for the asymmetric ν-SVR. We test $p = 0.2, 0.4, 0.5, 0.6, 0.8$ and choose the best one as the optimization result. For SVT, there is one parameter for bounding the amount of points outside. We set it to be k. PROGRESS is based on random sampling. To make a fair comparison, we run PROGRESS 2000 times and let PROGRESS has the similar computation time as the asymmetric ν-SVR. The widths of the tubes containing the required amount of the data are compared in Table 1, where the best one is underlined. For each data set, we report the data name, the dimension d, and the number of the data samples n. The results of the descent algorithm using linear programming (DALP), which is a global algorithm proposed by Huang et al. (2012), are given as well for reference.

Table 1: Width of the Tube Covering k Data and the Computation Time (in millisecond)

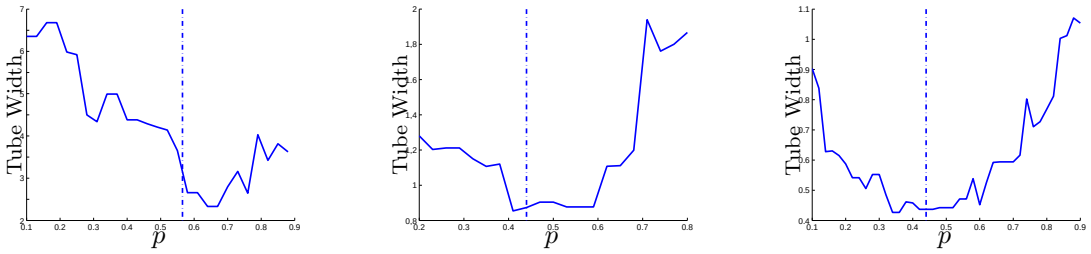
data set	d	n	k	PROG.		SVT		asymmetric ν-SVR		DALP	
AIRCRAFT	4	23	12	3.164	(106)	4.156	(17)	2.587	(93)	1.043	(279)
			16	5.450	(109)	5.544	(18)	5.179	(98)	2.904	(248)
			20	7.400	(119)	11.03	(21)	8.854	(99)	5.289	(245)
AIRMAY	3	31	16	11.45	(134)	14.07	(31)	11.13	(150)	7.231	(224)
			20	14.91	(141)	111.7	(28)	11.92	(137)	11.92	(251)
			24	21.99	(131)	306.8	(32)	20.96	(165)	14.66	(336)
COLEMAN	6	20	11	0.283	(102)	0.699	(17)	0.561	(84)	0.120	(245)
			15	1.029	(104)	0.927	(18)	0.726	(86)	0.509	(257)
			18	2.003	(101)	3.130	(17)	1.928	(82)	0.906	(385)
DELIVERY	2	25	13	1.005	(117)	0.920	(19)	0.904	(97)	0.758	(269)
			17	1.911	(111)	3.165	(18)	2.057	(98)	1.350	(193)
			21	2.984	(116)	3.825	(17)	4.016	(96)	2.984	(192)
EDUCAT	3	50	26	20.70	(146)	27.04	(37)	18.56	(122)	16.11	(234)
			36	37.28	(143)	34.85	(27)	35.68	(152)	32.28	(289)
			43	60.39	(146)	58.65	(38)	61.27	(163)	44.06	(265)
ENKNOCK	4	16	9	0.184	(91)	0.669	(14)	0.110	(71)	0.091	(219)
			11	0.932	(93)	0.981	(14)	0.663	(68)	0.228	(240)
			14	2.514	(89)	1.881	(13)	1.297	(62)	1.297	(212)
EXACT	3	25	13	0.500	(193)	0.700	(20)	0.500	(96)	0.250	(396)
			18	0.889	(201)	1.039	(21)	0.500	(97)	0.500	(268)
			23	2.000	(232)	1.375	(17)	1.375	(93)	1.300	(252)
HAWKINS	3	75	38	0.597	(379)	0.577	(74)	0.577	(280)	0.419	(573)
			51	0.736	(363)	0.851	(57)	0.799	(263)	0.716	(622)
			66	1.337	(368)	2.398	(64)	1.967	(327)	1.047	(494)
HEART	2	12	7	0.766	(78)	1.646	(10)	1.237	(52)	0.542	(261)
			9	2.565	(77)	2.680	(11)	2.149	(51)	1.974	(201)
			11	5.195	(85)	5.738	(11)	4.715	(53)	3.591	(227)
PHOSRHR	2	18	10	6.573	(90)	4.752	(13)	4.084	(70)	4.084	(260)
			13	11.27	(88)	13.38	(15)	11.21	(80)	9.678	(203)
			16	18.44	(95)	27.37	(15)	16.09	(78)	16.09	(191)
SALINITY	4	28	15	0.509	(124)	0.668	(22)	0.468	(190)	0.258	(231)
			20	1.028	(131)	1.030	(22)	0.824	(116)	0.808	(237)
			25	1.813	(131)	1.796	(35)	1.711	(163)	1.354	(308)
STACKLOSS	4	21	11	1.000	(119)	1.840	(15)	1.079	(81)	0.500	(288)
			15	2.244	(119)	2.555	(19)	2.208	(87)	1.816	(266)
			19	4.500	(128)	4.362	(28)	4.333	(168)	2.844	(363)
WOOD	6	20	11	0.004	(97)	0.014	(16)	0.009	(76)	0.002	(292)
			15	0.012	(98)	0.018	(19)	0.016	(82)	0.006	(221)
			18	0.035	(97)	0.030	(20)	0.027	(81)	0.021	(397)

From Table 1, one can see that the asymmetric ν-SVR generally gives good results for (1). Compared to SVT, the advantage of the asymmetric ν-SVR is that various locations are considered. For some data sets, PROGRESS provides good results as well. However, the nonparametric formulation based on PROGRESS is hard to compute. Similarly, DALP is not applicable in the dual space.

DALP is a global optimization method and requires more computation time, especially when the data size increases. As convex approaches for the least median squares (2), both SVT and the asymmetric ν-SVR can be solved effectively. Notice that in this experiment, we consider 5 different p values. If we only use one p value, its computing time will be similar to that of SVT. A simple and effective way of selecting p is based on the LS-SVM, as described in Algorithm 1. In Fig.4, we illustrate the obtained tube widths for different p values and show the selected p for some problems. Though there are outliers in these data sets, the selected p value is generally suitable.

4.2. Robustness

As analyzed by Rousseeuw and Leroy (1987), (1) is a very robust regression method. The asymmetric ν-SVR is a convex approximation and its robustness can be expected. To evaluate the robustness, a linear



(a) (b) (c)

Figure 4: The tube widths for different p values are shown by solid lines. The p values selected by Algorithm 1 are displayed by dot-dashed lines. (a) AIRCARFT $k = 12$; (b) DELIVERY $k = 13$; (c) SALINITY $k = 11$.

function $f(x) = w^T x + b$ with $w = [1, 0.5, -0.5, -1, 2]^T$ and $b = -3$ is considered. We generate $n = 100$ data following a uniform distribution in $[0, 1]^5$. In this experiment, a typical asymmetric noise following a chi-square distribution is considered: $y_i = f(x_i) + (\delta_{\chi^2} - 4)$, where δ_{χ^2} follows a chi-square distribution with 4 degrees of freedom. The mean of δ_{χ^2} is 4 and hence $(\delta_{\chi^2} - 4)$ provides a mean-zero noise, but $\text{Prob}\{\delta_{\chi^2} - 4 < 0\} = 59.4\%$, $\text{Prob}\{\delta_{\chi^2} - 4 < 0\} = 40.6\%$. Fig.5 displays the probability density function with variance 8.00. We let the ratio of the variance of the noise $(\delta_{\chi^2} - 4)$ to the variance of $f(x)$ denoted as r_{noise} equal to 0.05 and 0.1. Besides the noise, outliers are added into the training set. We randomly select r_o percent of the samples and replace their observed values by random values following a uniform distribution in the range of $f(x), x \in [0, 1]^5$.

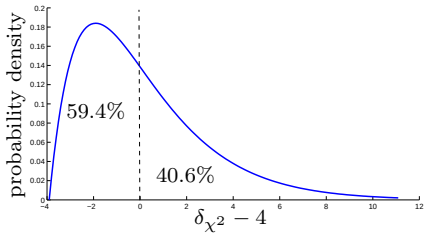


Figure 5: The probability density function of $\delta_{\chi^2} - 4$. The mean of $\delta_{\chi^2} - 4$ is zero and the variance of δ_{χ^2} is 8.00. This distribution is asymmetric and the probability of $\delta_{\chi^2} - 4$ being positive is 40.6%.

We pursue a linear tube covering 80% of the above data samples and use Algorithm 1 with a linear kernel to find the tube. We also consider the ν-SVR (3), robust regression using the Huber loss, and the least trimmed squares estimation. The robust regression using the Huber loss is implemented by the statistics toolbox of Matlab. The least trimmed squares is closely related to the least median squares. Actually, they came from the same paper of Rousseeuw (1984). The least trimmed squares is to minimize the sum of the smallest h squared residuals, which is apparently a robust method. However, it is also a non-convex problem and the nonparametric model has not been established. In this experiment, we set $h = 0.8n$ and use the algorithm established by Rousseeuw and Van Driessen (2006). This algorithm is included in LIBRA, a robust analysis toolbox developed by Verboven and Hubert (2005). For each group of r_{noise} and r_o, we repeat the above process 10 times, report the average and the standard deviation of the results in Table 2.

The least squares method is suitable to deal with Gaussian noise. When the noise follows another distribution or there exist outliers, the other robust methods perform better, as shown in Table 2. The standard deviation of the results of the ν-SVR is small, illustrating its robustness. But the average value is biased due to the asymmetry of the noise, which motivates us establish the asymmetric ν-SVR. Robust regression methods, including the least trimmed squares, PROGRESS, and the Huber loss, are also good for handling outliers and have similar regression performance as the ν-SVRs. Compared with these methods, the ν-SVR and the asymmetric ν-SVR give not only the predictive value, but also a tube covering a required

Table 2: The Average and Standard Deviation of the Regression Results

	Real value	1	0.5	−0.5	−1	2	−3
$r_{\text{noise}} = 0.05$	Least Squares	0.947 ± 0.125	0.487 ± 0.060	-0.434 ± 0.107	-0.928 ± 0.136	1.835 ± 0.075	-2.942 ± 0.066
$r_o = 5\%$	Least Trimmed Squares	1.002 ± 0.042	0.494 ± 0.354	-0.503 ± 0.058	-0.993 ± 0.059	2.031 ± 0.051	-3.058 ± 0.065
	PROGRESS	0.955 ± 0.117	0.529 ± 0.122	-0.491 ± 0.069	-1.040 ± 0.140	2.063 ± 0.144	-3.088 ± 0.094
	Huber Loss	0.977 ± 0.085	0.519 ± 0.048	-0.500 ± 0.051	-0.952 ± 0.060	1.943 ± 0.050	-3.009 ± 0.062
	ν-SVR	0.986 ± 0.038	0.499 ± 0.053	-0.475 ± 0.036	-0.989 ± 0.043	2.010 ± 0.055	-2.903 ± 0.116
	asymmetric ν-SVR	0.992 ± 0.036	0.507 ± 0.022	-0.502 ± 0.029	-1.001 ± 0.055	1.994 ± 0.042	-2.997 ± 0.055
$r_{\text{noise}} = 0.05$	Least Squares	0.877 ± 0.103	0.477 ± 0.146	-0.514 ± 0.147	-0.904 ± 0.143	1.868 ± 0.120	-2.880 ± 0.163
$r_o = 10\%$	Least Trimmed Squares	0.981 ± 0.054	0.487 ± 0.040	-0.512 ± 0.033	-1.032 ± 0.041	2.010 ± 0.025	-3.008 ± 0.056
	PROGRESS	0.972 ± 0.039	0.508 ± 0.064	-0.506 ± 0.067	-0.974 ± 0.060	1.987 ± 0.087	-3.024 ± 0.073
	Huber Loss	1.014 ± 0.096	0.460 ± 0.082	-0.475 ± 0.043	-1.021 ± 0.090	1.970 ± 0.095	-3.059 ± 0.089
	ν-SVR	0.996 ± 0.047	0.487 ± 0.038	-0.507 ± 0.076	-1.015 ± 0.058	1.990 ± 0.069	-2.909 ± 0.086
	asymmetric ν-SVR	1.000 ± 0.035	0.500 ± 0.031	-0.518 ± 0.028	-0.998 ± 0.035	2.013 ± 0.028	-2.987 ± 0.040
$r_{\text{noise}} = 0.10$	Least Squares	1.027 ± 0.111	0.514 ± 0.121	-0.489 ± 0.116	-1.014 ± 0.111	1.926 ± 0.076	-2.983 ± 0.077
$r_o = 5\%$	Least Trimmed Squares	1.012 ± 0.101	0.492 ± 0.084	-0.545 ± 0.066	-0.988 ± 0.065	2.002 ± 0.059	-3.053 ± 0.106
	PROGRESS	0.890 ± 0.168	0.495 ± 0.166	-0.509 ± 0.138	-1.130 ± 0.368	1.895 ± 0.275	-2.909 ± 0.255
	Huber Loss	0.989 ± 0.075	0.525 ± 0.047	-0.515 ± 0.057	-0.971 ± 0.068	2.002 ± 0.041	-3.026 ± 0.049
	ν-SVR	1.013 ± 0.058	0.496 ± 0.105	-0.532 ± 0.092	-0.990 ± 0.078	1.989 ± 0.086	-2.828 ± 0.162
	asymmetric ν-SVR	1.007 ± 0.054	0.497 ± 0.103	-0.488 ± 0.083	-0.992 ± 0.085	1.997 ± 0.056	-3.024 ± 0.081
$r_{\text{noise}} = 0.10$	Least Squares	0.874 ± 0.196	0.491 ± 0.158	-0.426 ± 0.136	-0.960 ± 0.094	1.798 ± 0.103	-2.867 ± 0.118
$r_o = 10\%$	Least Trimmed Squares	1.015 ± 0.085	0.516 ± 0.089	-0.497 ± 0.079	-1.022 ± 0.082	1.955 ± 0.070	-3.049 ± 0.035
	PROGRESS	0.936 ± 0.130	0.507 ± 0.087	-0.450 ± 0.035	-1.041 ± 0.070	1.952 ± 0.066	-2.908 ± 0.081
	Huber Loss	0.928 ± 0.123	0.564 ± 0.147	-0.532 ± 0.125	-1.011 ± 0.207	2.125 ± 0.180	-3.208 ± 0.231
	ν-SVR	1.025 ± 0.068	0.463 ± 0.061	-0.503 ± 0.078	-0.971 ± 0.076	2.014 ± 0.097	-2.921 ± 0.105
	asymmetric ν-SVR	0.992 ± 0.052	0.499 ± 0.070	-0.492 ± 0.061	-0.980 ± 0.060	1.997 ± 0.082	-2.996 ± 0.101

percentage of the data. The potential applications can be found in the field involving l_∞ regression or minimax approximation, which is an extreme case of (1) with $\rho = 1$. Generally, l_∞ regression pursues a narrow tube to cover all the training data. It originates from the worst-case analysis and has been applied widely in e.g., circuit design (Antreich et al., 1994; Papamarkos and Chamzas, 1996), signal processing (Kollar et al., 1990; Dvorkind et al., 2007), and portfolio optimization (Young, 1998; Cai et al., 2000). Obviously, l_∞ regression is sensitive to outliers. Hence, the asymmetric ν-SVR provides a potential tool to handle outliers in these fields.

In regression, both y_i and x_i could contain outliers. If x_i is a outlier, it is usually called a leverage point. If the observation value of a leverage point is far away from the regression line, it will significantly reduce the precision of many regression methods. The performance of the least median squares regression (2) for leverage points has been discussed by Rousseeuw and Leroy (1994), Coakley et al. (1996). The asymmetric ν-SVR is a good convex approximation for the least median squares. In this paper, we only focus on the vertical outliers but its robustness for leverage points can be expected as well.

4.3. Nonlinear regression with outliers

In the last subsection, we evaluate the LS-SVM, the ν-SVR, and the asymmetric ν-SVR on the test functions provided in Cherkassky et al. (1996). These functions are listed below and have been used in many papers to examine the performance of different regression methods, see, e.g., Hush and Horne (1998) Martínez-Estudillo et al. (2006), and Wang et al. (2010).

Fun. 1: $f_1(x) = \sin(x(1)x(2))$, $D = [-2, 2]^2$.

Fun. 2: $f_2(x) = \exp\big(x(1)\sin(\pi x(2))\big)$, $D = [-1, 1]^2$.

Fun. 3: $f_3(x) = \dfrac{40 f_3^1(x)}{f_3^2(x) + f_3^3(x)}$, $D = [0, 1]^2$, where

$$\begin{aligned} f_3^1(x) &= \exp\Big(8\big((x(1) - 0.5)^2 + (x(2) - 0.5)^2\big)\Big), \\ f_3^2(x) &= \exp\Big(8\big((x(1) - 0.2)^2 + (x(2) - 0.7)^2\big)\Big), \\ f_3^3(x) &= \exp\Big(8\big((x(1) - 0.7)^2 + (x(2) - 0.2)^2\big)\Big). \end{aligned}$$

Fun. 4: $f_4(x) = 42.659\big(0.1 + (x(1) - 0.5) f_4^1(x)\big)$, $D = [-0.5, 0.5]^2$, where

$$f_4^1(x) = 0.05 + (x(1) - 0.5)^4 - 10(x(1) - 0.5)^2(x(2) - 0.5)^2 + 5(x(2) - 0.5)^4.$$

Fun. 5: $f_5(x) = 1.3356\big(f_5^1(x(1)) + f_5^2(x(2))\big)$, $D = [0,1]^2$, where

$$\begin{aligned} f_5^1(x(1)) &= 1.5(1-x(1)) + \exp(2x(1)-1)\sin(3\pi(x(1)-0.6)^2), \\ f_5^2(x(2)) &= \exp(3(x(2)-0.5))\sin(4\pi(x(2)-0.9)^2). \end{aligned}$$

Fun. 6: $f_6(x) = 10\sin(\pi x(1)x(2)) + 20(x(3)-0.5)^2 + 5x(4) + 10x(5) + 0x(6)$, $D = [-1,1]^6$.

Fun. 7: $f_7(x) = \exp\big(2x(1)\sin(\pi x(4))\big) + \sin(x(2)x(3))$, $D = [-0.25, 0.25]^4$.

For the two-dimensional functions above, we generate 400 deterministic samples $x_i \in \mathbb{R}^2$, $1 \le i \le 400$, evenly spaced along its domain axes. Then we add noise and outliers following the same way as used in subsection 4.2. For the high-dimensional functions, the same procedure is conducted except that 400 data are randomly taken from a uniform probability distribution within the domain of interest.

For each group of data, we use the LS-SVM, the ν-SVR, and the asymmetric ν-SVR to obtain the approximation $\hat{f}$. For outliers corrupted data, the reweighted strategies for the LS-SVMs have been discussed by Suykens et al. (2002) and Valyon and Horváth (2005). In this experiment, we apply the robust cross validation and robust training method in the LS-SVMLab toolbox. The RBF kernel is used in the three methods. Since the computation time of the LS-SVM is significantly smaller than that of the other two methods, we apply 10-fold cross validation based on the LS-SVM to find the suitable kernel parameter and the coefficient of the regularization term, then we use them in the LS-SVM and the ν-SVRs. For the ν-SVR and the asymmetric ν-SVR, we set $\nu = 0.2$. In (12), p is chosen based on the result of the LS-SVM, as described in Algorithm 1. After obtaining the approximation function $\hat{f}$, 100 test data samples uniformly distributed in the domain are generated and the relative sum of the squared errors. For each case, we repeat the process above 10 times and report the average and the standard deviation of RSSEs in Table 3.

Table 3: Relative Sum of the Square Errors on Test Data

	r_{noise}	r_o	LS-SVM	Robust LS-SVM	ν-SVR	asymmetric ν-SVR
f_1	0.05	0.05	0.0156 ± 0.0051	0.0066 ± 0.0018	0.0143 ± 0.0049	0.0102 ± 0.0034
	0.05	0.10	0.0263 ± 0.0067	0.0087 ± 0.0012	0.0226 ± 0.0064	0.0210 ± 0.0091
	0.10	0.05	0.0170 ± 0.0074	0.0158 ± 0.0060	0.0202 ± 0.0048	0.0169 ± 0.0080
	0.10	0.10	0.0292 ± 0.0119	0.0193 ± 0.0069	0.0277 ± 0.0074	0.0259 ± 0.0128
f_2	0.05	0.05	0.0231 ± 0.0083	0.0094 ± 0.0037	0.0156 ± 0.0038	0.0121 ± 0.0046
	0.05	0.10	0.0484 ± 0.0100	0.0173 ± 0.0022	0.0302 ± 0.0089	0.0224 ± 0.0066
	0.10	0.05	0.0357 ± 0.0119	0.0184 ± 0.0044	0.0325 ± 0.0084	0.0214 ± 0.0052
	0.10	0.10	0.0585 ± 0.0246	0.0301 ± 0.0064	0.0410 ± 0.0187	0.0360 ± 0.0114
f_3	0.05	0.05	0.0238 ± 0.0127	0.0096 ± 0.0035	0.0152 ± 0.0084	0.0095 ± 0.0052
	0.05	0.10	0.0445 ± 0.0136	0.0171 ± 0.0044	0.0302 ± 0.0122	0.0223 ± 0.0109
	0.10	0.05	0.0248 ± 0.0075	0.0127 ± 0.0033	0.0177 ± 0.0095	0.0121 ± 0.0059
	0.10	0.10	0.0485 ± 0.0138	0.0235 ± 0.0081	0.0375 ± 0.0118	0.0313 ± 0.0165
f_4	0.05	0.05	0.0243 ± 0.0149	0.0055 ± 0.0014	0.0136 ± 0.0078	0.0085 ± 0.0081
	0.05	0.10	0.0588 ± 0.0172	0.0162 ± 0.0028	0.0312 ± 0.0121	0.0233 ± 0.0106
	0.10	0.05	0.0252 ± 0.0083	0.0092 ± 0.0032	0.0192 ± 0.0121	0.0169 ± 0.0068
	0.10	0.10	0.0650 ± 0.0186	0.0209 ± 0.0059	0.0330 ± 0.0095	0.0242 ± 0.0117
f_5	0.05	0.05	0.0266 ± 0.0107	0.0095 ± 0.0023	0.0193 ± 0.0043	0.0169 ± 0.0084
	0.05	0.10	0.0487 ± 0.0204	0.0161 ± 0.0069	0.0317 ± 0.0127	0.0220 ± 0.0151
	0.10	0.05	0.0336 ± 0.0098	0.0218 ± 0.0045	0.0248 ± 0.0076	0.0214 ± 0.0085
	0.10	0.10	0.0565 ± 0.0551	0.0302 ± 0.0068	0.0346 ± 0.0168	0.0326 ± 0.0182
f_6	0.05	0.05	0.0533 ± 0.0125	0.0447 ± 0.0093	0.0478 ± 0.0128	0.0443 ± 0.0131
	0.05	0.10	0.0742 ± 0.0154	0.0573 ± 0.0058	0.0798 ± 0.0161	0.0611 ± 0.0118
	0.10	0.05	0.0589 ± 0.0121	0.0530 ± 0.0059	0.0667 ± 0.0291	0.0518 ± 0.0097
	0.10	0.10	0.0789 ± 0.0102	0.0615 ± 0.0148	0.0873 ± 0.0141	0.0610 ± 0.0167
f_7	0.05	0.05	0.0241 ± 0.0186	0.0098 ± 0.0013	0.0164 ± 0.0074	0.0220 ± 0.0062
	0.05	0.10	0.0328 ± 0.0160	0.0250 ± 0.0073	0.0268 ± 0.0059	0.0272 ± 0.0102
	0.10	0.05	0.0312 ± 0.0048	0.0144 ± 0.0040	0.0248 ± 0.0131	0.0228 ± 0.0069
	0.10	0.10	0.0484 ± 0.0218	0.0274 ± 0.0083	0.0342 ± 0.0134	0.0294 ± 0.0140

The LS-SVM is the best for the Gaussian noise and enjoys low computational complexity. This experiment contains outliers and hence robust methods, i.e., the robust LS-SVM, the ν-SVR, and the asymmetric ν-SVR, perform well. The asymmetric ν-SVR gives a better result than the ν-SVR, since it is flexible with respective to the tube location and is suitable for handling asymmetric noise. The accuracy of the asymmetric ν-SVR is slightly worse than that of the robust LS-SVM. But it can also output a tube covering a required percentage of the data. Moreover, the ν-SVR and the asymmetric ν-SVR are sparse (in this experiment, only about 20% data points are support vectors). In Algorithm 1, the p value is selected based on the LS-SVM. If the outliers are heavy, we can use the robust LS-SVM or cross validation to tune p, then the accuracy of the asymmetric ν-SVR can be improved further, but more computation time is needed.

5. Conclusion and Further Study

As a robust regression method, the ν-tube Support Vector Regression can find a good tube covering a given percentage of the training data. However, equal amount of support vectors are located above and below the tube. To enhance the flexibility of the tube location, we extended the ν-SVR to the asymmetric ν-SVR, where one can use an additional parameter p to control the distribution of outliers. Enhancing the flexibility may result in a narrower tube which covers the required percentage of the data. Numerical experiments illustrated good performance of the asymmetric ν-SVR, especially when the samples were corrupted by asymmetric noise and outliers.

In the future, we would like to study efficient solving-algorithms for the asymmetric ν-SVR (4) and (12). Generally, the properties of the optimization problem associated with the asymmetric ν-SVR are similar to those of the ν-SVR. Thus, one can draw lessons from the optimization techniques developed for the ν-SVR, see, e.g., the works of Chang and Lin (2001), Chapelle (2007), and Tseng and Yun (2010).

Acknowledgements

The authors are grateful to the anonymous reviewers for insightful comments.

This work was supported in part by the scholarship of the Flemish Government; Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), G.0377.12 (Structured models), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); iMinds; EU: ERNSI; ERC AdG A-DATADRIVE-B, FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. L. Shi is also supported by the National Natural Science Foundation of China (No. 11201079) and the Fundamental Research Funds for the Central Universities of China (No.20520133238, No.20520131169). Johan Suykens is a professor at KU Leuven, Belgium.

References

Antreich, K., Graeb, H., Wieser, C., 1994. Circuit analysis and optimization driven by worst-case distance. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 13(1), 57–71.

Boček, P., Lachout, P., 1995. Linear programming approach to LMS-estimation. Computational Statistics and Data Analysis 19(2), 129–134.

Brys, G., Hubert, M., Struyf, A., Bartlett, P. L., 2004. A robust measure of skewness. Journal of Computational and Graphical Statistics 13(4), 996–1017.

Cai, X., Teo, K., Yang, X., Zhou, X., 2000. Portfolio optimization under a minimax rule. Management Science 46(7), 957–972.

Chang, C.-C., Lin, C.-J., 2001. Training ν-support vector classifiers: theory and algorithms. Neural Computation 13(9), 2119–2147.

Chapelle, O., 2007. Training a support vector machine in the primal. Neural Computation 19(5), 1155–1178.
Cherkassky, V., Gehring, D., Mulier, F., 1996. Comparison of adaptive methods for function estimation from samples. IEEE Transactions on Neural Networks 7(4), 969–984.
Coakley, C., Mili, L., Cheniae, M., 1994. Effect of leverage on the finite sample efficiencies of high breakdown estimators. Statistics and Probability Letters 19(5), 399–408.
De Brabanter, K., Karsmakers, P., Ojeda F., Alzate, C., De Brabanter, J., Pelckmans, K., De Moor, B., Vandewalle, J., Suykens, J.A.K., 2011. LS-SVMlab Toolbox User's Guide version 1.8, Internal Report 10-146, ESAT-SISTA, KU Leuven.
Donoho, D., Huber, P., 1982. The notion of breakdown point. A Festschrift for Erich L. Lehmann in Honor of His Sixty-fifth Birthday, 157–184.
Dvorkind, T., Kirshner, H., Eldar, Y., Porat, M., 2007. Minimax approximation of representation coefficients from generalized samples. IEEE Transactions on Signal Processing 55(9), 4430–4443.
Huang, X., Xu, J., Wang, S., Xu, C., 2012. Minimization of the k-th maximum and its application on LMS regression and VaR optimization. Journal of the Operational Research Society 63(11), 1479–1491.
Hubert, M., Rousseeuw, P., Verdonck, T., 2009. Robust PCA for skewed data and its outlier map. Computational Statistics and Data Analysis 53(6), 2264–2274.
Hush, D., Horne, B., 1998. Efficient algorithms for function approximation with piecewise linear sigmoidal networks. IEEE Transactions on Neural Networks 9(6), 1129–1141.
Kassam, S., Moustakides, G., Shin, J., 1982. Robust detection of known signals in asymmetric noise. IEEE Transactions on Information Theory 28(1), 84–91.
Kim, T.H., White, H., 2004. On more robust estimation of skewness and kurtosis. Finance Research Letters 1(1), 56–73.
Koenker, R., 2005. Quantile Regression. Cambridge University Press.
Kollar, I., Pintelon, R., Schoukens, J., 1990. Optimal FIR and IIR hilbert transformer design via LS and minimax fitting. IEEE Transactions on Instrumentation and Measurement 39(6), 847–852.
Le Masne, Q., Pothier, H., Birge, N., 2009. Asymmetric noise probed with a Josephson junction. Physical Review Letters 102(6), 067002.
Ma, Y., Li, L., Huang, X., Wang, S., 2011. Robust support vector machine using least median loss penalty. the 16th IFAC World Congress, 11208-11213.
Martínez-Estudillo, A., Martínez-Estudillo, F., Hervás-Martínez, C., García-Pedrajas, N., 2006. Evolutionary product unit based neural networks for regression. Neural Networks 19(4), 477–486.
Olson, C., 1997. An approximation algorithm for least median of squares regression. Information Processing Letters 63(5), 237–241.
Papamarkos, N., Chamzas, C., 1996. A new approach for the design of digital integrators. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 43(9), 785–791.
Pelckmans, K., De Brabanter, J., Suykens, J.A.K., De Moor, B., 2009. Least conservative support and tolerance tubes. IEEE Transactions on Information Theory 55(8), 3799–3806.
Rousseeuw, P., 1984. Least median of squares regression. Journal of the American Statistical Association 79(388), 871–880.
Rousseeuw, P., Hubert, M., 1997. Recent developments in PROGRESS. Lecture Notes-Monograph Series 31, 201–214.
Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection. Wiley-IEEE.
Rousseeuw, P., Wagner, J., 1994. Robust regression with a distributed intercept using least median of squares. Computational Statistics and Data Analysis 17(1), 65–76.
Rousseeuw, P., Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Mining and Knowledge Discovery 12(1), 29–45.
Schölkopf, B., Smola, A. J., Williamson, R. C., Bartlett, P. L., 2000. New support vector algorithms. Neural Computation 12(5), 1207–1245.
Solli, D.R., Ropers, C., Jalali, B., 2010. Rare frustration of optical supercontinuum generation. Applied Physics Letters 96(15), 151108.
Steele, J., Steiger, W., 1986. Algorithms and complexity for least median of squares regression. Discrete Applied Mathematics 14(1), 93–100.
Steinwart, I., Christmann, A., 2008. How SVMs can estimate quantiles and the median. Advances in Neural Information Processing Systems 20, 305–312.
Steinwart, I., Christmann, A., 2011. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17(1), 211–225.
Stromberg, A., 1993. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. SIAM Journal on Scientific Computing 14(6), 1289–1299.
Suykens, J.A.K., De Brabanter, J., Lukas, L., Vandewalle, J., 2002. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1), 85–105.
Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, J., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.
Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Processing Letters 9(3), 293–300.
Tichavsky, P., 1991. Algorithms for and geometrical characterization of solutions in the LMS and the LTS linear regression. Computational Statistics Quarterly 6, 139–151.
Tseng, P., Yun, S., 2010. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Computational Optimization and Applications 47(2), 179–206.
Tsyurmasto, P., Zabarankin, M., Uryasev, S., 2013. Value-at-Risk support vector machine: stability to outliers. Research Report #2013-2, Department of Industrial and Systems Engineering, University of Florida.
Valyon, J., Horváth, G., 2005. A robust LS-SVM regression. International Journal of Computer, Information Science and

Engineering 7, 1868–1873.
Vapnik, V., 1995. The Nature of Statistical Learning. Springer.
Verardi, V., Croux, C., 2009. Robust regression in Stata. The Stata Journal 9(3), 439–453.
Verboven, S., Hubert, M., 2005. LIBRA: a MATLAB library for robust analysis. Chemometrics and Intelligent Laboratory Systems 75(2), 127–136. https://wis.kuleuven.be/stat/robust/LIBRA/
Wang, S., Huang, X., Yam, Y., 2010. A neural network of smooth hinge functions. IEEE Transactions on Neural Networks 21(9), 1381–1395.
Winker, P., Lyra, M., Sharpe, C., 2011. Least median of squares estimation by optimization heuristics with an application to the CAPM and a multi-factor model. Computational Management Science 8(1-2), 103–123.
Xiang, D., Hu, T., Zhou, D.-X., 2012. Approximation analysis of learning algorithms for support vector regression and quantile regression. Journal of Applied Mathematics, doi:10.1155/2012/902139.
Young, M., 1998. A minimax portfolio selection rule with linear programming solution. Management Science 44(5), 673–683.