

Wang Shitong · Zhu Jiagang · F. L. Chung
Lin Qing · Hu Dewen

Theoretically Optimal Parameter Choices for Support Vector Regression Machines with Noisy Input

Published online: 13 August 2004
© Springer-Verlag 2004

Abstract With the evidence framework, the regularized linear regression model can be explained as the corresponding MAP problem in this paper, and the general dependency relationships that the optimal parameters in this model with noisy input should follow is then derived. The support vector regression machines Huber-SVR and Norm-r r-SVR are two typical examples of this model and their optimal parameter choices are paid particular attention. It turns out that with the existence of the typical Gaussian noisy input, the parameter μ in Huber-SVR has the linear dependency with the input noise, and the parameter r in the r-SVR has the inversely proportional to the input noise. The theoretical results here will be helpful for us to apply kernel-based regression techniques effectively in practical applications.

Keywords Regularized linear regression · Support vectors · Huber loss functions · Norm-r loss functions

1 Introduction

Due to the fact that their generalization capabilities do not depend on the dimensionality of the problems,

support vector regression machines have been obtaining various applications. In general, we can generalize these regression machines into *the regularized linear regression model* [15] in this paper. In various versions of support vector regression machines, three types of the loss functions, as shown in Fig. 1, are often used. They are the ε -insensitivity loss function, the norm-r (e.g, square) loss function and the Huber loss function. In fact, the norm-1 loss function is the Laplace one, and the norm-2 loss function is the often-used measure MSE. When the training data contain noise, if its distribution is unknown, the Huber loss function and its corresponding support vector regression machine Huber-SVR are a good choice. Otherwise, the ε -insensitivity loss function (corresponding to ε -SVR) and the norm-r loss function (corresponding to r-SVR) are extensively suggested due to its simplicity and human being's habit (we often use MSE in various learning schemes). Nowadays, ε -SVR has been studied well [1–8]. J.B.Gao and S.R.Gunn gave its error bar estimate in [7], V. Cherkassky and Y. Ma discussed the practical selection of parameters in ε -SVR, and A.J. Sloma and J.Kwok et al derived the linear dependency between ε and the input noise in ε -SVR. Moreover, ε -SVR is special in that its loss function gives identical zero penalty to small noise values [14]. Because of this, training samples with small noise that fall in this flat zero region are not involved in the representation of regression functions. This simplification of computational burden is usually referred to as the sparseness property. Another two loss functions mentioned above do not enjoy this property since they contribute a positive penalty to all noise values other than zero. However, on the other hand, Huber loss functions and Norm-r loss functions are attractive because they are differentiable. This property leads to the extensive applicability of Huber-SVR and r-SVR. Because of this applicability of Huber-SVR and r-SVR, we concentrate study on them in this paper.

Noise often occurs in real input data. With the existence of noisy input, one interesting and challenging

S. Wang (✉)
School of Information Engineering,
Southern Yangtze University, Wuxi, China
E-mail: wxwangst@yahoo.com.cn

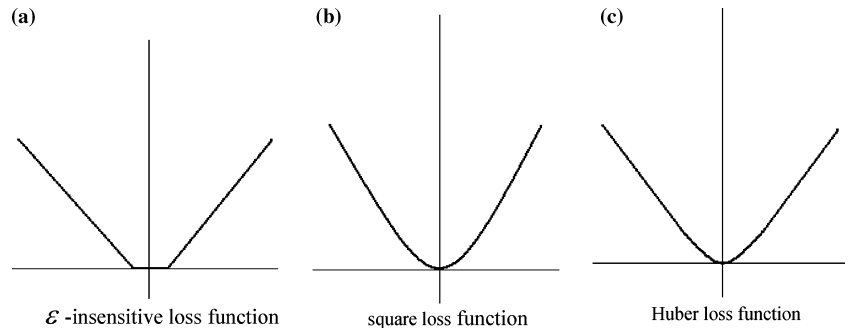
J. Zhu · Q. Lin · L. Qing
Dept. of Comp. Sci. and Engineering,
Nanjing Univ. of Sci. and Tech., Nanjing,
China

S. Wang · F. L. Chung
Dept. Computing, HongKong Polytechnic University,
HongKong, China

D. Hu
School of automation,
National Defense Univ. of Sci. and Tech.,
Changsha, China

S. Wang · J. Zhu
Lab. of Comp. Sci., Institute of Software,
Chinese Academy of Science, China

Fig. 1 3 loss functions



issue is how to determine the free parameters r and μ respectively in r -SVR and Huber-SVR. A bad choice for r and μ will heavily deteriorate the performances of r -SVR and Huber-SVR. Therefore, in this paper, we will pay attention on the theoretically optimal choices of parameters in Huber-SVR and r -SVR in particular. In sect. 2, we will introduce the regularized linear regression model and show the equivalent relationship between this model and MAP. The general dependency relationships that the optimal parameters in this model with noisy input should follow are also given in this section. Accordingly, in the case of the noisy input, the optimal choices of parameter μ in Huber-SVR and parameter r in r -SVR are respectively studied in sect. 3 and sect. 4. Sect. 5 concludes this paper.

2 Regularized linear regression model and MAP

2.1 Regularized linear regression model with convex risks

Assume we have a dataset with n -dimensional input vector \mathbf{x} and one-dimensional output variable y :

$$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}, \quad \mathbf{x}_i \in \mathbb{R}^n, \quad y_i \in \mathbb{R}, \\ i = 1, 2, \dots, l \quad (1)$$

we are interested in obtaining a weight vector w in the generalized linear regression model such that

$$y = \langle \phi(w^T \mathbf{x}) \rangle \quad \text{or} \quad y_i = \phi(\hat{w}^T \mathbf{x}_i) + \eta_i \quad (2)$$

for all the data in the dataset D , where ϕ is called a *link function* and all the data \mathbf{x}_i follow distribution $p(\cdot)$ and all η_i are i.i.d noise following some distribution $\eta(\cdot)$. Thus, the corresponding density function on y can be denoted as $p(y|\mathbf{x}) = \eta(y - \hat{w}^T \mathbf{x}_i)$. The degree of such an approximation can be measured by a loss function $L(\phi(w^T \mathbf{x}), y)$. A practical method to compute the weight vector w from the data is to find the minimum of the empirical expectation of the loss function:

$$\hat{w} = \arg \min_w \frac{1}{l} \sum_{i=1}^l L(\phi(w^T \mathbf{x}_i), y_i) \quad (3)$$

where $L(\phi(w^T \mathbf{x}_i), y_i)$ is a convex function. More generally, we can extend (3) into a regularized version of the generalized linear regression model with convex risks:

$$\hat{w} = \arg \min_w \frac{1}{l} \sum_{i=1}^l L(\phi(w^T \mathbf{x}_i), y_i) + \lambda g(w) \quad (4)$$

where g is a convex function of w and $\lambda > 0$ is a regularization parameter.

Let us discuss the regularized linear regression model in (4). If we take $g(w) = w^2/2$, and $L(\phi(w^T \mathbf{x}), y)$ as the ε -insensitive loss function in (4), then, we have the support vector regression machine ε -SVR. If we take $g(w) = w^2/2$, and $L(\phi(w^T \mathbf{x}), y)$ as the Huber loss function in (4), i.e.,

$$L_{\text{huber}}(\phi(w^T \mathbf{x}), y) = \begin{cases} \frac{1}{2}(\phi(w^T \mathbf{x}) - y)^2, & |\phi(w^T \mathbf{x}) - y| < \mu \\ \mu|\phi(w^T \mathbf{x}) - y| - \frac{1}{2}\mu^2, & \text{otherwise} \end{cases} \quad (5)$$

then we obtain the support vector regression machine Huber-SVR. If we take $g(w) = w^2/2$, and $L(\phi(w^T \mathbf{x}), y)$ as the loss function $L_r(\phi(w^T \mathbf{x}), y) = |\phi(w^T \mathbf{x}) - y|^r$ ($r > 0$) in (4), then we have the norm- r support vector regression machine r -SVR. In most cases, $r = 2$ in r -SVR.

In general, except for $g(w) = w^2/2$, other types of regularization $g(w)$ are also used in practice to satisfy different requirements. For example, in order to obtain a sparse weight vector \hat{w} , we may take the 1-norm regularization $g(w) = \|w\|_1$. More generally, we may take the general q -norm regularization $g(w) = \|w\|_q$. Another example is the maximum entropy framework for density estimation. In order to obtain a weight vector \hat{w} which is consistent with the data for density estimation, we may take $g(w)$ as the relative entropy [15].

2.2 The regression model and MAP

In this subsection, in terms of the evidence framework [7], we will demonstrate the regularized linear regression model is equivalent to *maximum a posteriori* MAP estimation, based on maximum likelihood estimation. Assume the loss function $L(\phi(w^T \mathbf{x}), y)$ leads to the following Gaussian probability density function on y :

$$p(y_i|\mathbf{x}_i, w, \beta, \theta) = \frac{1}{C(\beta, \theta)} \exp(-\beta L(\phi(w^T \mathbf{x}_i), y_i)) \quad (6)$$

where $C(\beta, \theta) = \iint_D \exp(-\beta L(\phi(w^T \mathbf{x}), y)) dx dy$ and θ denotes the free parameter in the loss function $L(\phi(w^T \mathbf{x}), y)$ (Note: when there are several free parameters in the loss function, θ denotes the corresponding vector on these parameters. Here we assume only one free parameter for simplicity). With the Gaussian prior on w

$$p(w|\alpha) = \frac{1}{M(\alpha)} \exp\left(-\frac{\alpha}{2} g(w)\right) \quad (7)$$

where $M(\alpha) = \int \exp(-\frac{\alpha}{2} g(w)) dw$, and on applying the Bayes rule:

$$p(w|D, \beta, \theta) \propto p(D|w, \beta, \theta) p(w|\alpha) \quad (8)$$

we have

$$\begin{aligned} \log p(w|D, \beta, \theta) = & -\frac{\alpha}{2} g(w) - \beta \sum_{i=1}^l L(\phi(w^T \mathbf{x}_i), y_i) \\ & + l \log C(\beta, \theta) + const \end{aligned} \quad (9)$$

On setting $\lambda = \beta/\alpha$, optimizing (4) can be interpreted as finding the MAP estimate of w at given values of β, θ . That is to say, the regularized linear regression model in (4) is equivalent to *maximum a posteriori* MAP estimation.

2.3 Estimating the optimal β and θ

Generally speaking, it is not easy for us to get the MAP estimate \hat{w} by directly solving (9) since it depends on the particular training set. For ease of the analysis, we replace $\frac{1}{l} \sum_{i=1}^l L(\phi(w^T \mathbf{x}_i), y_i)$ in (9) by its expectation:

$$E(L(\phi(w^T \mathbf{x}), y)) = \iint_D L(\phi(w^T \mathbf{x}), y) p(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dy \quad (10)$$

Thus, (9) becomes

$$\begin{aligned} G(w, \beta, \theta) = \log p(w|D, \beta, \theta) = & -\frac{\alpha}{2} g(w) - \beta l \\ & \times E(L(\phi(w^T \mathbf{x}), y)) + l \log C(\beta, \theta) + const \end{aligned} \quad (11)$$

In order to maximize (11), its derivatives with respect to w, β, θ must be zero. That is to say,

$$\begin{aligned} \frac{\partial G(w, \beta, \theta)}{\partial w} \Big|_{w=\hat{w}} = & -\frac{\alpha}{2} \frac{\partial g(\hat{w})}{\partial w} - \beta l \\ \frac{\partial}{\partial w} \left(\iint_D L(\phi(\hat{w}^T \mathbf{x}), y) p(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dy \right) / \partial w = & 0 \end{aligned} \quad (12)$$

$$\begin{aligned} \frac{\partial G(\hat{w}, \beta, \theta)}{\partial \beta} = & \left(\frac{\partial G(w, \beta, \theta)}{\partial w} \Big|_{w=\hat{w}} \right)^T \frac{\partial \hat{w}}{\partial \beta} - l \\ & \times E(L(\phi(\hat{w}^T \mathbf{x}), y)) + l \frac{\partial C(\beta, \theta) / \partial \beta}{C(\beta, \theta)} \\ = & -l \times E(L(\phi(\hat{w}^T \mathbf{x}), y)) + l \frac{\partial C(\beta, \theta) / \partial \beta}{C(\beta, \theta)} = 0 \end{aligned}$$

$$\text{i.e. } E(L(\phi(\hat{w}^T \mathbf{x}), y)) = \frac{\partial C(\beta, \theta) / \partial \beta}{C(\beta, \theta)} \quad (13)$$

$$\begin{aligned} \frac{\partial G(\hat{w}, \beta, \theta)}{\partial \theta} = & \left(\frac{\partial G(w, \beta, \theta)}{\partial w} \Big|_{w=\hat{w}} \right)^T \frac{\partial \hat{w}}{\partial \theta} - \beta l \\ & \times \frac{\partial E(L(\phi(\hat{w}^T \mathbf{x}), y))}{\partial \theta} + l \frac{\partial C(\beta, \theta) / \partial \theta}{C(\beta, \theta)} \\ = & -\beta l \frac{\partial E(L(\phi(\hat{w}^T \mathbf{x}), y))}{\partial \theta} + l \frac{\partial C(\beta, \theta) / \partial \theta}{C(\beta, \theta)} = 0 \end{aligned}$$

$$\text{i.e. } \frac{\partial E(L(\phi(\hat{w}^T \mathbf{x}), y))}{\partial \theta}$$

$$= \frac{1}{\beta} \frac{\partial C(\beta, \theta) / \partial \theta}{C(\beta, \theta)} \quad (14)$$

When $w = \hat{w}$, minimizing (11) actually becomes the following optimization problem:

$$\arg \min_{\beta, \theta} \beta E(L(\phi(\hat{w}^T \mathbf{x}), y)) - \log C(\beta, \theta) \quad (15)$$

Thus, (10), (13), (14) and (15) can be used to find out the optimal β and θ in the generalized support vector regression machines. In other words, the optimal β and θ should theoretically follow the dependency relationships contained in (10), (13), (14) and (15).

In [3], J. Kwok et al proved that the ε -insensitive loss function is taken, then there is a linear dependency between the optimal ε and the input noise in ε -SVR. However, due to the applicability of Huber-SVR and r-SVR, it will be beneficial for us if the dependency relationships between the optimal parameters and the input noise in Huber-SVR and r-SVR can be derived. So, in the following sections, we will study on the optimal choices of parameters in Huber-SVR and r-SVR with the input noise.

3 Estimating the optimal μ in Huber-SVR with the Gaussian noisy input

Without loss of generality, we take $g(w) = w^2/2$ and the Huber loss function, i.e

$$L_{\text{huber}}(\phi(w^T \mathbf{x}), y) = \begin{cases} \frac{1}{2} (\phi(w^T \mathbf{x}) - y)^2, & |\phi(w^T \mathbf{x}) - y| < \mu \\ \mu |\phi(w^T \mathbf{x}) - y| - \frac{1}{2} \mu^2, & \text{otherwise} \end{cases}$$

in Huber-SVR, thus, Huber-SVR attempts to minimizing the following problem

$$\min \Phi(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + c \sum_i (\xi_i^- + \xi_i^+),$$

s.t.

$$\begin{cases} y_i - f(\mathbf{x}_i) \leq \mu + \xi_i^-, \\ f(\mathbf{x}_i) - y_i \leq \mu + \xi_i^+, & i = 1, \dots, n \\ \xi_i^-, \xi_i^+ \geq 0 \end{cases}$$

where c is a predefined constant. According to its corresponding MAP, we have

$$C(\beta, \mu) = \left[\int_{-\infty}^{+\infty} -\beta L_{huber}(\phi(w^T \mathbf{x}), y) dt \right]^{-1}$$

$$= \left[2 \left(\int_0^\mu \exp\left(-\frac{1}{2}\beta \cdot t^2\right) dt + \int_\mu^\infty \exp\left(-\beta\mu t + \frac{1}{2}\beta\mu^2\right) dt \right) \right]^{-1}$$

$$\approx \left[2 \left(\mu \exp\left(-\frac{1}{8}\beta\mu^2\right) + \frac{1}{\beta\mu} \exp\left(-\frac{1}{2}\beta\mu^2\right) \right) \right]^{-1}$$

Applying its Taylor series expansion, we further have

$$C(\beta, \mu) \approx \frac{\beta\mu}{\beta\mu^2 + 2} \tag{16}$$

Moreover, we take the link function $\phi(\cdot) = \cdot$ for simplicity. However, the conclusions below still remain true for other link functions.

Let us consider

$$E(L_{huber}(y_i - \hat{w}^T \mathbf{x}_i)) = \int \int_D L_{huber}(y - w^T \mathbf{x}) p(y|\mathbf{x}) p(\mathbf{x}) dy d\mathbf{x}$$

$$= \int_D \left(\int_{-\infty}^{\hat{w}^T \mathbf{x} - \mu} \left[\mu(w^T \mathbf{x} - y) - \frac{1}{2}\mu^2 \right] p(y|\mathbf{x}) dy \right.$$

$$+ \int_{\hat{w}^T \mathbf{x} - \mu}^{\hat{w}^T \mathbf{x}} \frac{1}{2}(w^T \mathbf{x} - y)^2 p(y|\mathbf{x}) dy$$

$$+ \int_{\hat{w}^T \mathbf{x}}^{\hat{w}^T \mathbf{x} + \mu} \frac{1}{2}(y - w^T \mathbf{x})^2 p(y|\mathbf{x}) dy$$

$$\left. + \int_{\hat{w}^T \mathbf{x} + \mu}^\infty \left[\mu(y - w^T \mathbf{x}) - \frac{1}{2}\mu^2 \right] p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x}$$
(17)

Substituting (15) into (13) and (14), we obtain

$$E(L_{huber}(y - \hat{w}^T \mathbf{x})) = \frac{1}{\beta} - \frac{2\mu^2}{\beta\mu^2 + 2} \tag{18}$$

$$\int_D \left(\int_{-\infty}^{\hat{w}^T \mathbf{x} - \mu} (w^T \mathbf{x} - y - \mu) p(y|\mathbf{x}) dy \right.$$

$$+ \int_{\hat{w}^T \mathbf{x} - \mu}^{\hat{w}^T \mathbf{x}} (y - w^T \mathbf{x} - \mu) p(y|\mathbf{x}) dy \left. \right) p(\mathbf{x}) d\mathbf{x}$$

$$= \frac{1}{\mu} - \frac{2\beta\mu}{\beta\mu^2 + 2}$$
(19)

White noise often happens in real situations. Quite often, one utilizes the i.i.d Gaussian distribution with zero mean and variance σ in most robust analysis. We take this assumption here. Let

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \hat{w}^T \mathbf{x})^2}{2\sigma^2}\right) \tag{20}$$

after substituting (19) into (16), let $t = \frac{(y - \hat{w}^T \mathbf{x})^2}{2\sigma^2}$ and

$$y - w^T \mathbf{x} \approx \hat{w}^T \mathbf{x} - w^T \mathbf{x} = \delta(\mathbf{x}) \tag{21}$$

we can expand $E(L_{huber}(y - \hat{w}^T \mathbf{x}))$ using its Taylor series expression on $\delta(\mathbf{x})$ at $t = \frac{\mu}{\sqrt{2}\sigma}$.

Please note:

$$\int_D p(\mathbf{x}) d\mathbf{x} = 1$$

$$\int_D \delta(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = 0$$

$$\int_D \delta(\mathbf{x})^2 p(\mathbf{x}) d\mathbf{x} = \sigma^2$$

$$\exp\left(-\frac{\delta(\mathbf{x})^2}{2\sigma^2}\right) \approx 1 - \frac{\delta(\mathbf{x})^2}{2\sigma^2}$$

Thus, after a little tedious computation, we obtain

$$E(L_{huber}(y - \hat{w}^T \mathbf{x})) \approx \frac{2\sqrt{2}\mu\sigma}{\sqrt{\pi}} \int_{\frac{\mu}{\sqrt{2}\sigma}}^\infty t \exp(-t^2) dt$$

$$- \frac{\mu^2}{\sqrt{\pi}} \int_{\frac{\mu}{\sqrt{2}\sigma}}^\infty \exp(-t^2) dt$$

$$+ \frac{\mu\sigma}{2\sqrt{2}\pi} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

$$+ \frac{\mu^3}{2\sqrt{2}\pi\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \frac{\mu^3}{6\sqrt{2}\pi\sigma}$$

$$\approx \frac{3\mu\sigma}{\sqrt{2}\pi} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \frac{\mu^3}{2\sqrt{2}\pi\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

$$+ \frac{\mu^3}{6\sqrt{2}\pi\sigma} - \frac{\mu^2}{\sqrt{\pi}} \int_{\frac{\mu}{\sqrt{2}\sigma}}^\infty \exp(-t^2) dt$$
(22)

Now, let us observe $-\log C(\beta, \mu)$. In terms of (16), we have

$$-\log C(\beta, \mu) = \log\left(\frac{\beta\mu}{\beta\mu^2 + 2}\right)$$

$$= \log\left(\mu + \frac{2}{\beta\mu}\right) \tag{23}$$

Since β generally takes a comparatively large value, so $\left(\frac{2}{\beta\mu}\right)$ becomes a comparatively small value. Therefore, with its Taylor series expansion, $-\log C(\beta, \mu)$ can be approximated by

$$-\log C(\beta, \mu) = \log \mu + \frac{2}{\beta\mu^2} \tag{24}$$

By substituting (22) and (24) into (15), minimizing (15) becomes finding out β, μ such that

$$\beta E(L_{\text{huber}}(y - \hat{w}^T \mathbf{x})) + \frac{2}{\beta \mu^2} + \log \mu \quad (25)$$

achieves its minimum. Since

$$\begin{aligned} (25) &\geq 2\sqrt{\beta E(L_{\text{huber}}(y - \hat{w}^T \mathbf{x})) \frac{2}{\beta \mu^2} + \log \mu} \\ &= 2\sqrt{\left[\frac{3\mu\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \frac{\mu^3}{2\sqrt{2\pi}\sigma} \times \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \frac{\mu^3}{6\sqrt{2\pi}\sigma} - \frac{\mu^2}{\sqrt{\pi}} \int_{\frac{\mu}{\sqrt{2\sigma}}}^{\infty} \exp(-t^2) dt \right] \frac{2}{\mu^2}} \\ &\quad + \log \mu \\ &= 2\sqrt{\frac{3\sqrt{2}}{\sqrt{\pi}} \left(\frac{\mu}{\sigma}\right)^{-1} \exp\left(-\frac{1}{2} \left(\frac{\mu}{\sigma}\right)^2\right) + \frac{1}{\sqrt{2\pi}} \left(\frac{\mu}{\sigma}\right) \exp\left(-\frac{1}{2} \left(\frac{\mu}{\sigma}\right)^2\right) + \frac{1}{3\sqrt{2\pi}} \left(\frac{\mu}{\sigma}\right) - \frac{2}{\sqrt{\pi}} \int_{\frac{1}{\sqrt{2}} \left(\frac{\mu}{\sigma}\right)}^{\infty} \exp(-t^2) dt} \\ &\quad + \ln\left(\frac{\mu}{\sigma}\right) + \ln \sigma \end{aligned} \quad (26)$$

Obviously, when $\left(\frac{\mu}{\sigma}\right)$ takes some fixed value, (25) will achieve its minimum, which indicates that there is a linear dependency between the parameter μ and the variance σ of the Gaussian input noise. Moreover, with minimizing (25) using mathematical computation with MATLAB, we can determine the scope of σ within which such a linear dependency keeps as follows:

$$\mu \approx \begin{cases} 0.8\sigma, & \sigma < 0.5 \\ 1.7\sigma, & 0.5 \leq \sigma \leq 2 \\ \text{nonlinear dependency,} & \sigma > 2 \end{cases}$$

This computational result also demonstrates that when the input noise is comparatively small ($\sigma < 2.0$), the parameter μ in the Huber loss function of Huber-SVR has the distinctive linear dependency with the variance σ of the input noise. When the variance of the input noise is so large that the training data is heavily distorted, Huber-SVR can not actually give its meaningful result for regression problems, that is to say, such a linear dependency will not be kept.

4 Estimating the optimal r in r -SVR with the Gaussian noisy input

Now, let us turn into the similar estimate problem for r -SVR. We will see that the derivation for this estimate problem is not an easy work. For the same derivation convenience as in the above section, we take $g(w) = w^2/2$ and the norm- r loss function, i.e

$$L_r(\phi(w^T \mathbf{x}), y) = |\phi(w^T \mathbf{x}) - y|^r \quad (r > 0)$$

in r -SVR, thus, r -SVR is equivalent to the following minimization problem:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + c \sum_i (\xi_i), \quad r > 0$$

s.t.

$$\begin{aligned} |\phi(w^T \mathbf{x}_i) - y_i|^r &< \xi_i, \quad i = 1, 2, \dots, l \\ \xi_i &\geq 0 \end{aligned}$$

where c is a predefined constant. According to its corresponding MAP, we have

$$\begin{aligned} C(\beta, r) &= \left[\int_{-\infty}^{+\infty} -\exp(\beta L_r(\phi(w^T \mathbf{x}), y)) dt \right]^{-1} \\ &= \left[2 \int_0^{\infty} \exp(-\beta \cdot t^r) dt \right]^{-1} \\ &= \left[\frac{2}{r \cdot \beta^{\frac{1}{r}} \Gamma\left(\frac{1}{r}\right)} \right]^{-1} \\ &= \frac{r \cdot \beta^{\frac{1}{r}}}{2\Gamma\left(\frac{1}{r}\right)} \end{aligned}$$

Fig. 2 shows the curve of $\Gamma\left(\frac{1}{r}\right)$. When $r \in [0.5, 10]$, there is a roughly linear dependency between $\Gamma\left(\frac{1}{r}\right)$ and r .

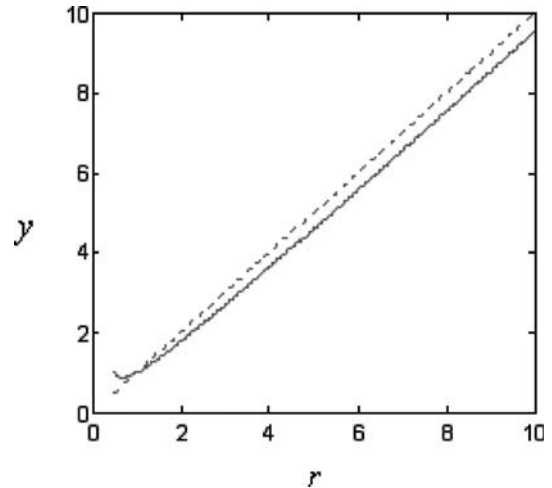


Fig. 2 Curve of $\Gamma\left(\frac{1}{r}\right)$, where the dotted line corresponds to $y = r$, and the other line corresponds to $y = \Gamma\left(\frac{1}{r}\right)$

Generally speaking, $r \in [0.5, 10]$ is enough in practical applications, so, we may take

$$\Gamma\left(\frac{1}{r}\right) \approx r$$

Thus, we have

$$C(\beta, r) \approx \frac{1}{2}\beta^{\frac{1}{r}} \quad (27)$$

Substituting (27) into (13) and with the same assumption on the link function as in Huber-SVR, we easily have

$$E(|y - \hat{w}^T \mathbf{x}|^r) = \int \left(\int_{-\infty}^{+\infty} |y - \hat{w}^T \mathbf{x}|^r p(y|\mathbf{x}) p(\mathbf{x}) dy \right) d\mathbf{x} \approx \frac{1}{\beta \cdot r} \quad (28)$$

$$\frac{\partial E(|y - \hat{w}^T \mathbf{x}|^r)}{\partial r} \approx -\frac{1}{r^2 \beta} \ln \beta \quad (29)$$

Assume the input noise and $p(y|\mathbf{x})$ has the same Gaussian distribution as in (20) and

$$y - w^T \mathbf{x} \approx \hat{w}^T \mathbf{x} - w^T \mathbf{x} = \delta(\mathbf{x})$$

let $t = \frac{(y - \hat{w}^T \mathbf{x})^2}{2\sigma^2}$, thus, we have

$$\begin{aligned} E(|y - \hat{w}^T \mathbf{x}|^r) &= \int_D \frac{1}{\sqrt{\pi}} \left[\int_{-\infty}^{\frac{\delta(\mathbf{x})}{\sqrt{2\sigma}}} (\delta(\mathbf{x}) - \sqrt{2}\sigma t)^r \exp(-t^2) dt \right. \\ &\quad \left. + \int_{\frac{\delta(\mathbf{x})}{\sqrt{2\sigma}}}^{\infty} (\sqrt{2}\sigma t - \delta(\mathbf{x}))^r \exp(-t^2) dt \right] p(\mathbf{x}) d\mathbf{x} \\ &\approx \int_D \frac{1}{\sqrt{\pi}} \left[\int_{-\infty}^0 (\delta(\mathbf{x}) - \sqrt{2}\sigma t)^r \exp(-t^2) dt \right. \\ &\quad \left. + \int_0^{\infty} (\sqrt{2}\sigma t - \delta(\mathbf{x}))^r \exp(-t^2) dt \right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Let $t = u + \frac{\delta(\mathbf{x})}{\sqrt{2\sigma}}$ and after simplifying the above formula we have

$$\begin{aligned} (|y - \hat{w}^T \mathbf{x}|^r) &\approx \int_D \frac{1}{\sqrt{\pi}} \left[\int_0^{\infty} (\sqrt{2}\sigma t)^r \right. \\ &\quad \left. \exp\left(-\left(t - \frac{\delta(\mathbf{x})}{\sqrt{2\sigma}}\right)^2\right) dt + \int_0^{\infty} (\sqrt{2}\sigma t)^r \right. \\ &\quad \left. \exp\left(-\left(t + \frac{\delta(\mathbf{x})}{\sqrt{2\sigma}}\right)^2\right) dt \right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Using its Taylor series expression on t , we further obtain

$$\begin{aligned} E\left(|y - \hat{w}^T \mathbf{x}|^r\right) &\approx \int_D \frac{2}{\sqrt{\pi}} p(\mathbf{x}) \left[\int_0^{\infty} (\sqrt{2}\sigma t)^r \right. \\ &\quad \times \left(\exp(-t^2) \left(1 - \frac{\delta^2(\mathbf{x})}{2\sigma^2} + \frac{\delta^4(\mathbf{x})}{8\sigma^2} - \frac{\delta^6(\mathbf{x})}{48\sigma^2}\right) \right. \\ &\quad \left. + t^2 \exp(-t^2) \left(\frac{\delta^2(\mathbf{x})}{\sigma^2} - \frac{\delta^4(\mathbf{x})}{2\sigma^2} + \frac{\delta^6(\mathbf{x})}{8\sigma^2}\right) \right. \\ &\quad \left. + t^4 \exp(-t^2) \left(\frac{\delta^4(\mathbf{x})}{6\sigma^2} - \frac{\delta^6(\mathbf{x})}{12\sigma^2}\right) \right. \\ &\quad \left. + \frac{\delta^6(\mathbf{x})}{90\sigma^2} t^6 \exp(-t^2) \right] dt d\mathbf{x} \end{aligned}$$

Since $\int_D p(\mathbf{x}) d\mathbf{x} = 1$ and $\int_0^{\infty} t^r \exp(-t^2) dt = \frac{1}{2} \Gamma\left(\frac{r+1}{2}\right)$, and we assume that δ approximately follows $N(0, \sigma^2)$, thus $E(\delta)^k = (k-1)!! \sigma^k$ [10], that is to say

$$\int_D \delta(\mathbf{x})^2 p(\mathbf{x}) d\mathbf{x} \approx \sigma^2$$

$$\int_D \delta(\mathbf{x})^4 p(\mathbf{x}) d\mathbf{x} \approx 3\sigma^4$$

$$\int_D \delta(\mathbf{x})^6 p(\mathbf{x}) d\mathbf{x} \approx 15\sigma^6$$

Therefore, we have

$$\begin{aligned} E(|y - \hat{w}^T \mathbf{x}|^r) &\approx \frac{1}{\sqrt{\pi}} (\sqrt{2}\sigma)^r \left[\frac{7}{16} \Gamma\left(\frac{r+1}{2}\right) \right. \\ &\quad \left. + \frac{7}{8} \Gamma\left(\frac{r+3}{2}\right) - \frac{3}{4} \Gamma\left(\frac{r+5}{2}\right) + \frac{1}{6} \Gamma\left(\frac{r+7}{2}\right) \right] \quad (30) \end{aligned}$$

$$\begin{aligned} \frac{\partial E(|y - \hat{w}^T \mathbf{x}|^r)}{\partial r} &\approx \frac{1}{\sqrt{\pi}} (\sqrt{2}\sigma)^r \ln(\sqrt{2}\sigma)^r \left[\frac{7}{16} \Gamma\left(\frac{r+1}{2}\right) \right. \\ &\quad \left. + \frac{7}{8} \Gamma\left(\frac{r+3}{2}\right) - \frac{3}{4} \Gamma\left(\frac{r+5}{2}\right) + \frac{1}{6} \Gamma\left(\frac{r+7}{2}\right) \right] \\ &\quad + \frac{1}{2\sqrt{\pi}} (\sqrt{2}\sigma)^r \log(\sqrt{2}\sigma)^r \left[\frac{7}{16} \Gamma\left(\frac{r+1}{2}\right) \psi\left(\frac{r+1}{2}\right) \right. \\ &\quad \left. + \frac{7}{8} \Gamma\left(\frac{r+3}{2}\right) \psi\left(\frac{r+3}{2}\right) - \frac{3}{4} \Gamma\left(\frac{r+5}{2}\right) \psi\left(\frac{r+5}{2}\right) \right. \\ &\quad \left. + \frac{1}{6} \Gamma\left(\frac{r+7}{2}\right) \psi\left(\frac{r+7}{2}\right) \right] \quad (31) \end{aligned}$$

where

$$\frac{\partial \Gamma(x)}{\partial r} = \Gamma(x) \Psi(x),$$

and

$$\Psi(x) = \sum_{m=0}^{\infty} \left(\frac{1}{m+1} - \frac{1}{m+x} \right) - \gamma$$

and γ is the Euler constant [10]. In this paper, we take

$$\Psi(x) \approx \sum_{m=0}^{1000} \left(\frac{1}{m+1} - \frac{1}{m+x} \right) - 0.57721567$$

In order to remove β , in terms of (28) and (29), we obtain

$$\frac{\partial E_{XY}(|y - \hat{w}^T \mathbf{x}|^r)}{\partial r} - \frac{1}{r} E_{XY}(|y - \hat{w}^T \mathbf{x}|^r) \quad (32)$$

$$\log(r E_{XY}(|y - \hat{w}^T \mathbf{x}|^r)) = 0$$

After substituting (30) and (31) into (32), let the left side of (32) be $f(r, \sigma)$, we have

$$f(r, \sigma) = 0 \quad (33)$$

Obviously, it is very difficult to obtain the direct dependency relationship between r and σ by solving the above (33). However, given a σ , we can get the corresponding r by plotting $f(r, \sigma)$ with MATLAB. With such a little tedious mathematical computation for the above (33), we can still observe the change trend between them. Fig. 3 depicts the curve of the computational result on solving (30). It should be noted that because we use approximation estimates several times in the above derivation, the change curve between r and σ in Fig. 3 is not exact. However, fortunately, it can well indicate a rough change trend between r and σ . That is to say, from Fig. 3, we can see that *the parameter r (i.e., the norm r) is basically inversely proportional to the variance σ of the input noise (see the dot line other than the curve in the figure). Especially, when noise is very small or (almost) no noise exists in the*

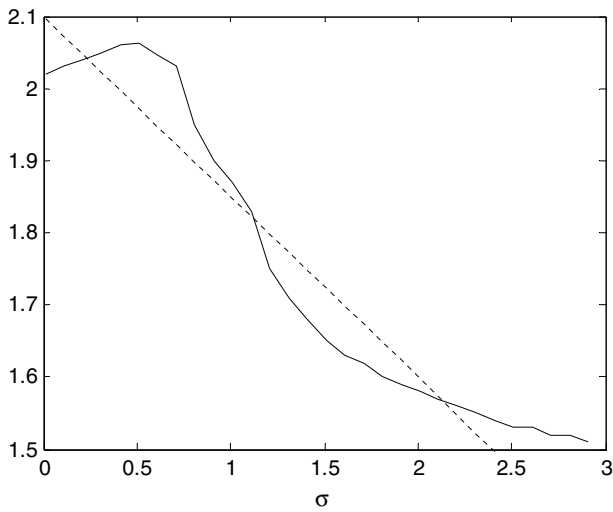


Fig. 3 The computational result for (32) using MATLAB

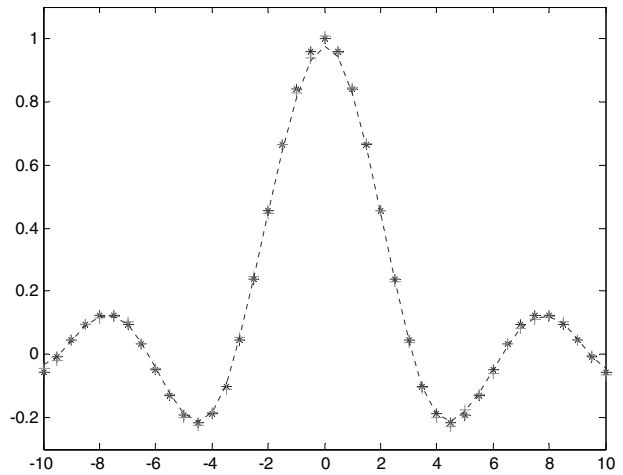


Fig. 5 The real output and the corresponding regression curve of $y = \frac{\sin(x)}{x}$ with the noisy input, where * and — are the same as above, + denotes the output of $y' = \frac{\sin(x)}{x} + k \cdot \eta$, $k = 0.01$

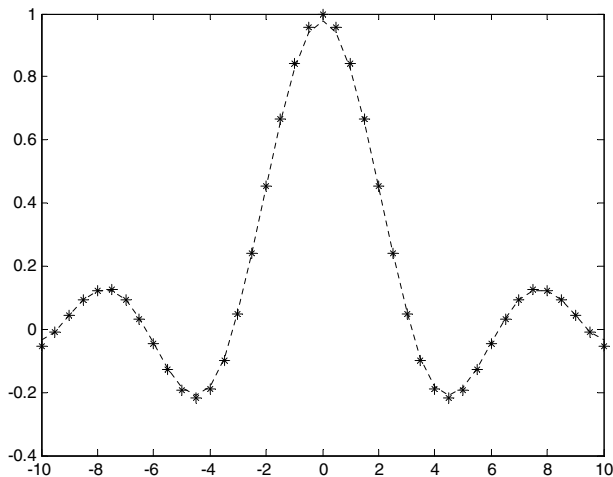


Fig. 4 The real output and the corresponding regression curve of $y = \frac{\sin(x)}{x}$, where * denotes the real output and — denotes the output curve of Huber-SVR

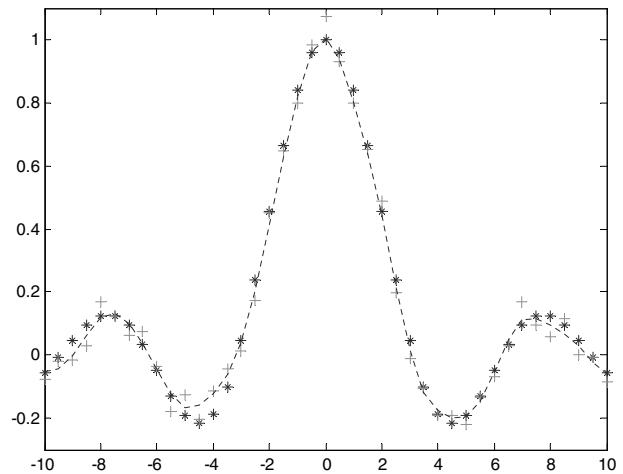


Fig. 6 The real output and the corresponding regression curve of $y = \frac{\sin(x)}{x}$ with the noisy input, where * and — are the same as above, + denotes the output of $y' = \frac{\sin(x)}{x} + k \cdot \eta$, $k = 0.05$

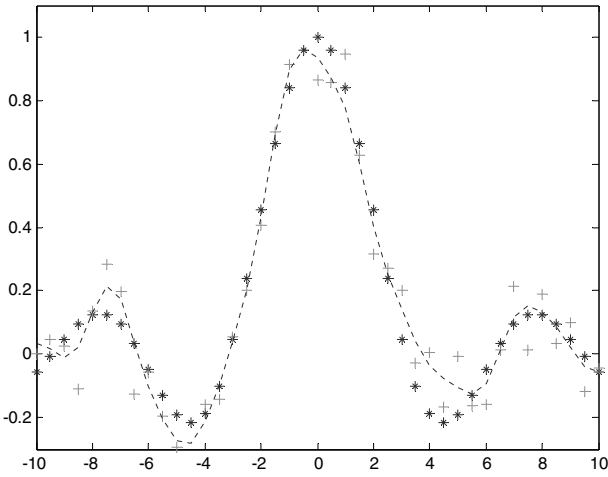


Fig. 7 The real output and the corresponding regression curve of $y = \frac{\sin(x)}{x}$ with the noisy input, where * and — are the same as above, + denotes the output of $y' = \frac{\sin(x)}{x} + k \cdot \eta$, $k = 0.10$

training data r should roughly take 2. As it is well known, SVM and/or SVR are equivalent to MLP to some extent. While training MLP based on the norm- r error function, the empirical result in [9] is that r should be 2 if no noise exists in the training data, otherwise, it should be less than 2 and go down with the increase of noise. Therefore, the rationale of the above analysis for r -SVR here can also be justified from the alternative angle where it is good in agreement with the empirical result in [9].

5 Experimental studies

In this section, for Huber-SVR, we will arrange an experiment on a typical example to validate the linear

dependency relationship between μ and σ . For norm- r r -SVR, when r takes a non-integer value, r -SVR may not be easily implemented using the classical QP optimization method [1]. We can implement norm- r r -SVR for the case of a non-integer r , using perceptron-like learning approach. In order to save the paper's space, we will introduce this implementation approach and report the related experimental results in another paper later.

Given a function $y = \frac{\sin(x)}{x}$, $x \in [-10, 10]$, let us generate its uniform dataset (x_i, y_i) , $i = 1, \dots, 41$, with x varying from -10 to 10 using the step length 0.5. First, for the dataset, we can easily construct its regression curve $f = f(x)$ using Huber-SVR. Next, in order to investigate the dependency relationship between μ and the noisy input, let $y' = \frac{\sin(x)}{x} + k \cdot \eta$, $x \in [-10, 10]$, where k is a noise-signal ratio and $\eta \sim N(0, \sigma)$ represents the Gaussian noise. Similarly, we can generate its corresponding sampling dataset (x_i, y'_i) , $i = 1, \dots, 41$, and obtain its Huber-SVR regression curve $f' = f(x)$. In order to make the experimental results fair, we take σ from $[0.1, 2.0]$ with the step length 0.1, and use the Gaussian noise distribution to generate 20 groups of the corresponding sampling datasets for each given σ . For each given σ , we take μ as the average result of all 20 μ values which can minimize $\sum_{i=1}^{41} |f'_i - f_i|$ respectively for each group of the sampling datasets.

Fig. 4 demonstrates the real output of $y = \frac{\sin(x)}{x}$ and its regression curve using Huber-SVR. Fig. 5–7 demonstrate the corresponding regression results using Huber-SVR with different noisy inputs. Fig. 8–10 depict the dependency relationships between μ and σ for all 20 σ values with different k (see + in the figures), where we use the curves to roughly indicate the change tendencies between μ and σ , respectively. We can easily see from these figures that when noise is small, i.e. k and σ is comparatively small, there is an obvious linear depen-

Fig. 8 The relationship between μ and σ when $k = 0.01$

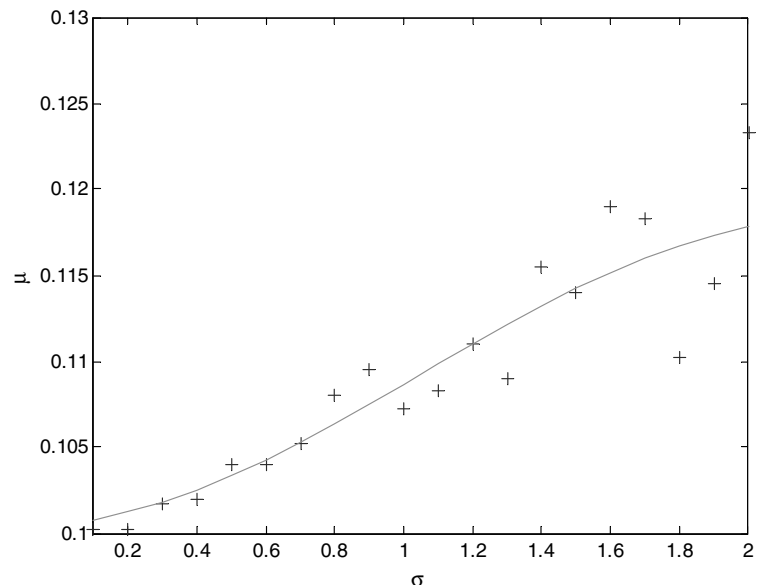


Fig. 9 The relationship between μ and σ when $k = 0.05$

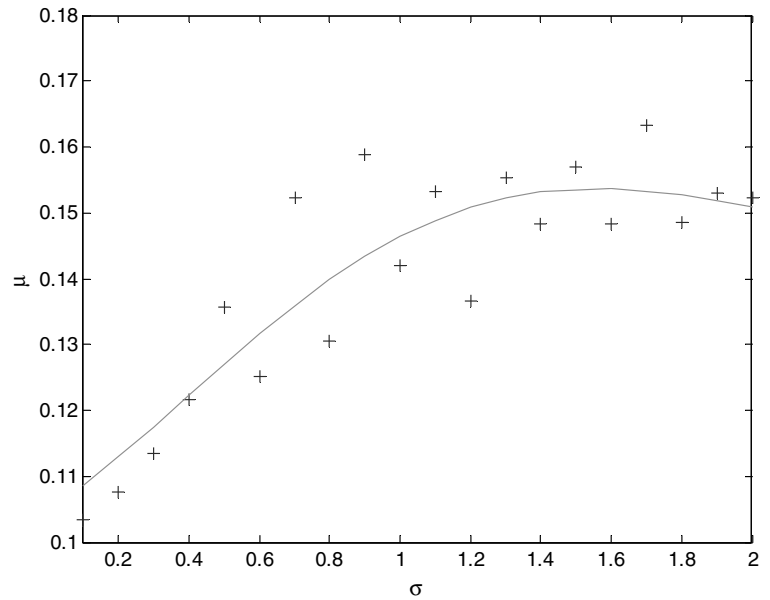
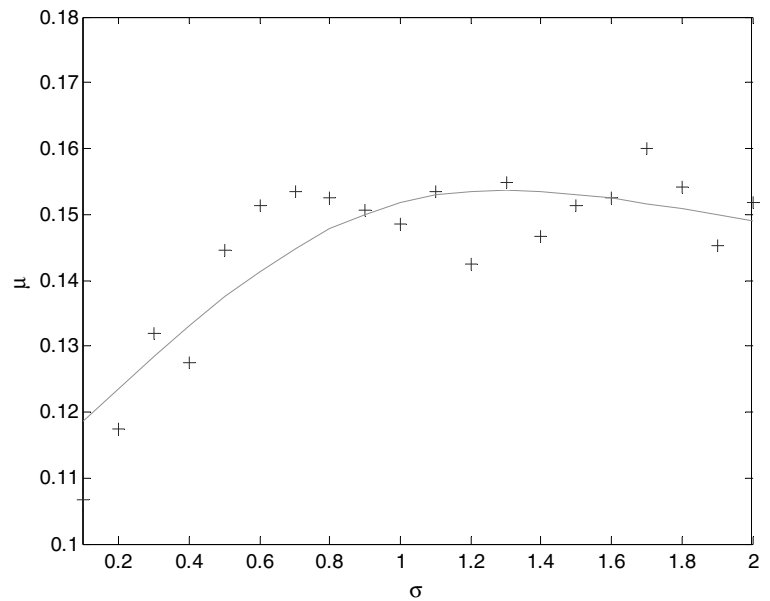


Fig. 10 The relationship between μ and σ when $k = 0.10$



dependency relationship between μ and σ . However, when k and/or σ are comparatively large, i.e., the datasets are seriously distorted, the linear relationship between μ and σ does not exist anymore (see Fig. 9 and Fig. 10). In other words, Huber-SVR may become ineffective for seriously distorted datasets. In summary, the experimental results here validate the above obtained conclusion on Huber-SVR.

6 Conclusions and future work

In this paper, based on the evidence framework, we study the general dependency relationship between the parameters in the regularized linear regression model

and the input noise, indicating that the optimal choices of these parameters are actually dependent on the variance of the input noise. Except for ϵ -SVM, Huber-SVR and norm-r r-SVR are another two typical examples of the regularized linear regression model. In this paper, we derive the linear dependency between the optimal μ in Huber-SVR and the Gaussian input noise, and the almost inversely linear dependency between the optimal norm r in r-SVR and the Gaussian input noise. The theoretical results here are very useful for practical applications of the corresponding support vector regression techniques.

Although Gaussian noise is typically adopted in most robust analysis, there remain other symmetric types of noise such as student-t-distribution noise and uniformly

noise in real datasets. In some particular situations, real datasets contain non-symmetric noise such as Dirichlet-distribution noise. The hard work on how to choose the optimal parameters in ε -SVR, Huber-SVR and r-SVR with such noisy input is worthy to be done in near future.

Acknowledgements This work is supported by the RGC Competitive Earmarked Research Grant (grant No. PolyU 5065/98 E), Natural Science Foundation of China (grant No. 60225015), Natural Science Foundation of JiangSu Province (grant No. BK2003017), National Key Lab. of Novel Software Technologies at NanJing University and Lab of Computer Science, Institute of Software, Chinese Academy of Science, China

References

1. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines. Cambridge University Press
2. Vapnik V (1998) Statistical Learning Theory. Wiley, New York
3. Kwok JT, Tsang IW (2003) Linear dependency between ε and the input noise in ε -support vector regression. *IEEE Trans. Neural Networks* 14(3): 544–553
4. Smola AJ, Murata N, Schölkopf B, Müller KR (1998) Asymptotically optimal choice of ε -loss for support vector machines. In: Proceedings of the International Conference on Artificial Neural Networks
5. Smola AJ, Schölkopf B (1998) A tutorial on support vector regression. *NeuroCOLT2 Technical Report NC2-TR-1998-030*, Royal Holloway College
6. Law MH, Kwok JT (2001) Bayesian support vector regression. In: Proceedings of the English International Workshop on Artificial Intelligence and Statistics, Florida pp 239–244
7. Gao JB, Gunn SR, Ham CJ (2002) A probabilistic framework for SVM regression and Error Bar Estimation. *Machine Learning*, 46: 71–89
8. Cherkassky V, Ma Y (2003) Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* (in press)
9. Yan Pinfan et al (2001) Artificial neural networks and evolutionary computation. Tsinghua University Press
10. Yonghuan S (2002) Handbook of practical mathematics, Chinese Science Press
11. Wang S, Chung FL et al Note on the relationship between probabilistic/fuzzy clustering. *Int J Soft Computing* (accepted)
12. Wang S et al Gene selection for cancer classification using the new SVM-based technique, *Chinese J Bioinformatics* (accepted)
13. Wang S (1998, 2000) Fuzzy systems, fuzzy neural networks and their programming, Press of Shanghai Sci Tech, 2nd edn
14. Wang S et al Robust maximum entropy clustering algorithm RMEC and its labeling for outliers, *Engineering Sci China* (accepted)
15. Zhang T (2002) On the dual formulation of regularized linear systems with convex risks, *Machine Learning* 46: 91–129