Square Penalty Support Vector Regression

Álvaro Barbero, Jorge López, José R. Dorronsoro *

Dpto. de Ingeniería Informática and Instituto de Ingeniería del Conocimiento
Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract. Support Vector Regression (SVR) is usually pursued using the ϵ–insensitive loss function while, alternatively, the initial regression problem can be reduced to a properly defined classification one. In either case, slack variables have to be introduced in practical interesting problems, the usual choice being the consideration of linear penalties for them. In this work we shall discuss the solution of an SVR problem recasting it first as a classification problem and working with square penalties. Besides a general theoretical discussion, we shall also derive some consequences for regression problems of the coefficient structure of the resulting SVMs and illustrate the procedure on some standard problems widely used as benchmarks and also over a wind energy forecasting problem.

1 Introduction

Standard ϵ–insensitive SVR [9, 12] seeks to minimize $\|W\|^2$ subject to the restrictions $W \cdot X^i + b - (y^i - \epsilon) \geq 0$, $W \cdot X^i + b - (y^i + \epsilon) \leq 0$. If it exists, the solution of this problem, that we shall refer to as ϵ–SVR, defines what is usually called a hard ϵ tube. However, in practical problems, hard tubes have to be replaced by soft ones, where besides ϵ insensitivity, extra slack terms have to be introduced. More precisely, the previous restrictions become now

$$W \cdot X^i + b - (y^i - \epsilon) + \xi_i \geq 0, \;\; W \cdot X^i + b - (y^i + \epsilon) - \mu_j \leq 0, \qquad (1)$$

and the function to be minimized is now $J_p(W, b, \xi, \mu) = \|W\|^2 + C \sum(\xi_i^p + \mu_j^p)$ for some $p \geq 1$ and where C is a properly chosen penalty factor. Obviously, minimizing $J_p(W, b, \xi, \mu)$ is equivalent to minimizing

$$\sum_i \left[y^i - f(X^i, W, b)\right]_\epsilon^p + \lambda \|W\|^2,$$

for $\lambda = 1/C$ and where $f(X, W, b) = W \cdot X + b$ and $[z]_\epsilon = \max(0, |z| - \epsilon)$. Thus, the problem that soft ϵ–insensitive SVR solves can be seen as a modelling problem where errors are measured in terms of the $[\cdot]_\epsilon^p$ function and a regularization term $\lambda \|W\|^2$ is added. As it is the case in standard soft margin SVMs, the usual choice in SVR is to take $p = 1$; however, in this work we shall take $p = 2$, which is

* All authors have been partially supported by Spain's TIN 2004–07676.

also a frequent election in SVM training. For either choice, rather than minimize the criterion $J_p(W, b, \xi, \mu)$ one defines a certain dual problem with a quadratic programming structure and that can be solved by standard packages or simply by gradient ascent on the dual function (see [13] for the linear penalty case). Here, however, we will take a different point of view. We note first that SVR can be transformed in a classification problem [2]. More precisely, if an ϵ hard tube exists, shifting the y_i values by $\pm\epsilon$ we obtain subsets $\mathcal{D}^+ = \{(X^i, y^i + \epsilon)\}$ and $\mathcal{D}^- = \{(X^i, y^i - \epsilon)\}$ that are linearly separable, and the ϵ–insensitive SVR problem can then be recast as that of minimizing the quantity $\|W\|^2 + \delta^2$ subject to the conditions

$$W \cdot X^i + b - \delta(y^i - \tilde{\epsilon}) \geq 1, \;\; W \cdot X^i + b - \delta(y^i + \tilde{\epsilon}) \leq -1, \tag{2}$$

where $\epsilon = \tilde{\epsilon} - 1/\delta$; we shall call this problem $\tilde{\epsilon}$–SVC. Its dual function is then

$$\Theta(\alpha, \beta) = -\frac{1}{2}\|X_\alpha - X_\beta\|^2 - \frac{1}{2}\left(\tilde{\epsilon}\sum(\alpha_i + \beta_i) - \sum(\alpha_i - \beta_i)y^i\right)^2 + \sum(\alpha_i + \beta_i) \tag{3}$$

subject to the restrictions $\alpha_i, \beta_j \geq 0$ and $\sum \alpha_i = \sum \beta_j$, and where $X_\alpha = \sum \alpha_i X^i$, $X_\beta = \sum \beta_j X^j$. We can get a more compact formulation of $\tilde{\epsilon}$–SVC writing points in $\mathcal{D}_\pm$ as $X^i_+ = (X^i, y^i + \tilde{\epsilon})$, $X^j_- = (X^j, y^j - \tilde{\epsilon})$ and the weight vector as $\widetilde{W} = (W, -\delta)$. Then $\tilde{\epsilon}$–SVC can be stated as minimizing $\|\widetilde{W}\|^2$ subject to $\widetilde{W} \cdot X^i_- + b \geq 1$, $\widetilde{W} \cdot X^j_+ + b \leq -1$. A possible way to solve it is to find [1] the closest points X^*_+ and X^*_- in the convex hulls $C(\mathcal{D}_+)$, $C(\mathcal{D}_-)$ of $\mathcal{D}_+$ and $\mathcal{D}_-$. We shall call this third convex hull formulation CH–SVM.

Our approach to solve square penalty ϵ–SVR will be based on the solution of CH–SVM. More precisely, we will show in section 2 the equivalence for the hard margin setting of $\tilde{\epsilon}$–SVC and CH–SVM, and how to reduce ϵ–SVR to $\tilde{\epsilon}$–SVC; in particular, we will see how their solutions are related. An advantage of using square penalties is that hard and soft SVMs can be treated formally in the same way. In section 3 we will recall how this is done and, turning our attention to ϵ–SVR, our main contribution will be Proposition 3, where we show that the coefficient structure of the optimal solution of soft CH–SVM can be seen as defining a certain regression tube, slightly larger than the original ϵ one. Patterns correctly classified by the CH–SVM solution will fall inside it, while not correctly classified patterns will fall outside. In section 4 we will illustrate the application of square penalty SVR to some publicly available regression datasets as well as in a wind energy prediction problem. The paper will end with a short discussion and some conclusions.

2 Hard Margin SV Regression and Classification

It is well known [9] that the optimal ϵ–SVR solution weight $\widehat{W}$ can be written as $\widehat{W} = X_{\widehat{\alpha}} - X_{\widehat{\beta}}$, with $\widehat{\alpha}, \widehat{\beta}$ the optimal dual solutions. Moreover, the Karush–

Kuhn–Tucker (KKT) conditions verified by the optimal $\widehat{W}, \widehat{b}, \widehat{\alpha}, \widehat{\beta}$ imply that if at some i we have, say, $\widehat{\alpha}_i > 0$, then $\widehat{W} \cdot X^i + \widehat{b} - (y^i - \epsilon) = 0$ and, therefore, $\widehat{b} = (y^i - \epsilon) - \widehat{W} \cdot X^i = 0$. Similarly, the optimal $\tilde{\epsilon}$–SVC solution weight W^o can be written as $W^o = X_{\alpha^o} - X_{\beta^o}$ [2], with α^o, β^o now the optimum solutions of the corresponding dual problem; moreover, the optimal δ^o value is given by

$$\begin{aligned} \delta^o &= \sum \left\{ \beta_i^o \left(y^i + \tilde{\epsilon} \right) - \alpha_i^o \left(y^i - \tilde{\epsilon} \right) \right\} \\ &= \tilde{\epsilon} \sum \left(\alpha_i^o + \beta_i^o \right) - \sum \left(\alpha_i^o - \beta_i^o \right) y^i. \end{aligned} \tag{4}$$

Finally, the KKT conditions are in this case

$$\begin{aligned} \alpha_i^o > 0 \Rightarrow W^o \cdot X^i - \delta^o (y^i - \tilde{\epsilon}) + b^o = 1, \\ \beta_j^o > 0 \Rightarrow W^o \cdot X^j - \delta^o (y^i + \tilde{\epsilon}) + b^o = -1, \end{aligned}$$

and just as before, the optimal b^o can be obtained as, say, $b^o = 1 + \delta^o (y^i - \tilde{\epsilon}) - W^o \cdot X^i$ if $\alpha_i^o > 0$. The following proposition relates the optimal solutions of ϵ–SVR and $\tilde{\epsilon}$–SVC. Although essentially known in the literature, we shall give here its proof for a lack of a proper reference.

Proposition 1. *Assume $\tilde{\epsilon}$ to be such that the shifted classes $\mathcal{D}_+$, $\mathcal{D}_-$ are linearly separable and let W^o, δ^o, b^o and $\widehat{W}, \widehat{b}$ be the optimal solution of $\tilde{\epsilon}$–SVC and ϵ–SVR respectively. We then have $W^o = \delta^o \widehat{W}$, $b^o = \delta^o \widehat{b}$ and $\epsilon = \tilde{\epsilon} - 1/\delta^o$.*

Proof. If W^o, δ^o, b^o is the optimal solution of $\tilde{\epsilon}$–SVC, it easily follows from (2) that $\tilde{\epsilon} \geq 1/\delta^o$. If $\tilde{\epsilon}\delta^o = 1$, the restrictions in (2) would become $W^o \cdot X^i - \delta^o y^i + b^o \geq 0$, $W^o \cdot X^i - \delta^o y^i + b^o \leq 0$ for all i. This would imply $\frac{W^o}{\delta^o} \cdot X^i + \frac{b^o}{\delta^o} = y^i$ for all i, i.e., we would have a perfect fit at all points, an unusual circumstance not likely to happen; hence, we will assume $\tilde{\epsilon} > 1/\delta^o$. Then $\tilde{W} = W^o/\delta^o$ and $\tilde{b} = b^o/\delta^o$ is a feasible solution of ϵ–SVR with $\epsilon = \tilde{\epsilon} - 1/\delta^o$. As a consequence, $\|\widehat{W}\| \leq \|\tilde{W}\| = \|W^o\|/\delta^o$.

On the other hand and as we have just mentioned (see also [9]), the optimal solution $\widehat{W}$ of ϵ–SVR can be written as $\widehat{W} = \sum_i \widehat{\alpha}_i X^i - \sum \widehat{\beta}_j X^j$, with $\sum_i \widehat{\alpha}_i = \sum \widehat{\beta}_j$. The KKT conditions imply that at an $\widehat{\alpha}_i > 0$ SV X^i we have $\widehat{W} \cdot X^i + \widehat{b} - (y^i - \epsilon) = 0$, while at a $\widehat{\beta}_j > 0$ SV X^j we have $\widehat{W} \cdot X^j + \widehat{b} - (y^j + \epsilon) = 0$. Writing now $\epsilon = \tilde{\epsilon} - 1/\delta^o$, it follows that

$$\widehat{W} \cdot X^j + \widehat{b} - (y^j - \tilde{\epsilon}) = \frac{1}{\delta^o}; \;\; \widehat{W} \cdot X^j + \widehat{b} - (y^j + \tilde{\epsilon}) = \frac{-1}{\delta^o},$$

and, therefore,

$$\delta^o \widehat{W} \cdot x^j + \delta^o \widehat{b} - \delta^o (y^j - \tilde{\epsilon}) = 1; \;\; \delta^o \widehat{W} \cdot x^j + \delta^o \widehat{b} - \delta^o (y^j + \tilde{\epsilon}) = -1.$$

Thus, $(W' = \delta^o \widehat{W}, \delta^o, b' = \delta^o \widehat{b})$ is a feasible solution of $\tilde{\epsilon}$–SVC and, hence, $\delta^o \|\widehat{W}\| = \|W'\| \geq \|W^o\|$. By the uniqueness [3] of the SVM solutions, it follows that $\widehat{W} = W^o/\delta^o$ and the other equalities are then immediate. □

Turning our attention to the relationship between $\tilde{\epsilon}$–SVC and CH–SVM, recall that writing $X_+ = (X, y + \tilde{\epsilon})$, $X_- = (X, y - \tilde{\epsilon})$ and $\widetilde{W} = (W, -\delta)$, $\tilde{\epsilon}$–SVC minimizes $\|\widetilde{W}\|^2$ subject to $\widetilde{W} \cdot X_-^i + b \geq 1$, $\widetilde{W} \cdot X_+^j + b \leq -1$. As mentioned before, the optimal solution of CH–SVM is given by the closest points X_+^* and X_-^* in the convex hulls $C(\mathcal{D}_\pm)$ of $\mathcal{D}_\pm$ (see [1] for more details). They verify therefore that $X_-^* = \sum \alpha_i^* X_-^i$ and $X_+^* = \sum \beta_j^* X_+^j$, with $\sum \alpha_i^* = \sum \beta_j^* = 1$ and define an optimal vector $\widetilde{W}^*$ and bias b^* as

$$\widetilde{W}^* = X_-^* - X_+^*,\ b^* = \frac{1}{2}\left(\|X_-^*\|^2 - \|X_+^*\|^2\right).$$

Moreover, the maximum margin m^* is given by $m^* = \|\widetilde{W}^*\|/2$. The following proposition is also known [1] and proved using the KKT conditions of each problem.

Proposition 2. *The optimal solution $\widetilde{W}^*$ of CH–SVM is related to the optimal $\widetilde{W}^o$ of $\tilde{\epsilon}$–SVC as*

$$\widetilde{W}^o = \frac{2}{\|\widetilde{W}^*\|^2}\widetilde{W}^* = \frac{1}{m^*}\tilde{W}^*,$$

or, equivalently, $\widetilde{W}^* = 2\widetilde{W}^o/\|\widetilde{W}^o\|^2$. Moreover, $W^o = 2W^*/\|W^*\|^2$, $b^o = 2b^*/\|W^*\|^2$ and $\delta^o = 2\delta^*/\|W^*\|^2$.

CH–SVM is somewhat different formally from ϵ–SVR and $\tilde{\epsilon}$–SVC and although still solvable using quadratic programming tools, it lends itself to algorithms quite different from those standard in SVMs. A good choice is the Schlesinger–Kozinec (SK) algorithm [6]. The starting observation is that for any potential weight $\widetilde{W} = X_- - X_+$, with $X_\pm \in C(\mathcal{D}_\pm)$, its margin $m(W)$ verifies $m(W) \leq \|W\|/2$. Moreover, if $\widetilde{W}^* = X_-^* - X_+^*$ is the optimal weight, we have $m(\widetilde{W}) \leq m(\widetilde{W}^*) = \|\widetilde{W}^*\|/2 \leq \|\widetilde{W}\|/2$. Thus setting $g(\widetilde{W}) = \|\widetilde{W}\|/2 - m(\widetilde{W})$, we have $0 = g(\widetilde{W}^*) \leq g(\widetilde{W})$. The SK algorithm iteratively constructs approximations $\widetilde{W}^t$ to W^* by convex updates $\widetilde{W}^t = (1-\lambda^t)\widetilde{W}^{t-1} + \lambda^t \widetilde{X}_\pm^t$, where λ^t and $\widetilde{X}_\pm^t \in C(\mathcal{D}_\pm)$ are chosen to ensure that $\widetilde{W}^t < \widetilde{W}^{t-1}$ and that approximately (although not true for all iterations) $m(\widetilde{W}^t) < m(\widetilde{W}^{t-1})$ (see [6] for more details). We shall use the SK algorithm in our square penalty experiments.

3 Square Penalty SV Regression and Classification

Recall that square penalty ϵ–SVR seeks to minimize $\|W\|^2 + C\sum(\xi_i^2 + \mu_j^2)$ subject to the restrictions $W \cdot X^i + b - (y^i - \epsilon) + \xi_i \geq 0$, $W \cdot X^i + b - (y^i + \epsilon) - \mu_j \leq 0$. It can be reduced to a hard ϵ–SVR problem by extending the W and X vectors adding $2N$ extra coordinates to them, with N the sample size, and defining

$$\overline{W} = (W, \sqrt{C}\xi_1, \ldots, \cdot\sqrt{C}\xi_N, \sqrt{C}\mu_1, \ldots, \sqrt{C}\mu_N),$$

$$\overline{X}^i_- = (X^i, 0, \ldots, \tfrac{1}{\sqrt{C}}, \ldots, 0, 0, \ldots, 0),\ \ \overline{X}^j_+ = (X^j, 0, \ldots, 0, 0, \ldots, \tfrac{-1}{\sqrt{C}}, \ldots, 0),$$

where the final non–zero coordinate of $\overline{X}^i_-$ is the extra i–th one and the final non–zero coordinate of $\overline{X}^j_+$ is the extra $(N+j)$–th one. We then have $\|W\|^2 + C\sum(\xi_i^2 + \mu_j^2) = \|\overline{W}\|^2$ and the previous restrictions become $\overline{W}\cdot\overline{X}^i_- + b - (y^i - \epsilon) \geq 0$, $\overline{W}\cdot\overline{X}^j_+ + b - (y^i + \epsilon) \leq 0$.

We can similarly reduce square penalty $\tilde{\epsilon}$–SVC to a hard $\tilde{\epsilon}$–SVC problem. Keeping the previously used $X_\pm = (X, y \pm \tilde{\epsilon})$ and $\widetilde{W} = (W, -\delta)$ notation, we consider now the extended weight and vectors

$$\overline{W} = (\widetilde{W}, \sqrt{C}\xi_1, \ldots, \sqrt{C}\xi_N, \sqrt{C}\mu_1, \ldots, \sqrt{C}\mu_N),$$
$$\overline{X}^i_- = (X^i_-, 0, \ldots, \frac{1}{\sqrt{C}}, \ldots, 0, 0, \ldots, 0),\ \overline{X}^j_+ = (X^j_+, 0, \ldots, 0, 0, \ldots, \frac{-1}{\sqrt{C}}, \ldots, 0),$$

for which we have again $\|W\|^2 + \delta^2 + C\sum(\xi_i^2 + \mu_j^2) = \|\overline{W}\|^2$ and the restrictions $\overline{W}\cdot\overline{X}^i_- + b \geq 1$, $\overline{W}\cdot\overline{X}^j_+ + b \leq -1$. Solving the CH–SVM version of $\tilde{\epsilon}$–SVC will give the optimal extended weight $\overline{W}^*$ as $\overline{W}^* = \overline{X}^*_- - \overline{X}^*_+$, with $\overline{X}^*_- = \sum \alpha_i^* \overline{X}^i_-$ and $\overline{X}^*_+ = \sum \beta_j^* \overline{X}^j_+$. In particular we will have

$$\begin{aligned}\overline{W}^* &= (\widetilde{W}^*, \sqrt{C}\xi_1^*, \ldots, \sqrt{C}\xi_N^*, \sqrt{C}\mu_1^*, \ldots, \sqrt{C}\mu_N^*)\\ &= (\sum \alpha_i^* X^i_- - \sum \beta_j^* X^j_+, \frac{\alpha_1^*}{\sqrt{C}}, \ldots, \frac{\alpha_N^*}{\sqrt{C}}, \frac{\beta_1^*}{\sqrt{C}}, \ldots, \frac{\beta_N^*}{\sqrt{C}}),\end{aligned}$$

and, therefore, margin slacks and SV coefficient are directly related as $C\xi_i^* = \alpha_i^*$, $C\mu_j^* = \beta_j^*$. Moreover, as we will shall see next, the size of the optimal α_i^*, β_j^* coefficients determine the tube in which patterns X^j_-, X^j_+ will fall.

Proposition 3. *Set* $\Lambda^* = C\|\overline{W}^*\|^2 = \sum\left\{(\alpha_i^*)^2 + (\beta_i^*)^2\right\} + C\left(\|W^*\|^2 + (\delta^*)^2\right)$. *Then a pattern* (X^i, y^i) *will fall inside the* $\tilde{\epsilon}$ *tube if and only if* $\alpha_i^* < \Lambda^*/2$ *and* $\beta_i^* < \Lambda^*/2$.

Proof. We will bring the extended CH–SVM solution $\overline{W}^*$ back to the ϵ–SVR one retracing the steps already mentioned in the penalty–free case. We go first from $\overline{W}^*$ to the optimal solution $\overline{W}^o$ of $\tilde{\epsilon}$–SVC as $\overline{W}^o = 2\overline{W}^*/\|\overline{W}^*\|^2$. As a consequence, the optimal $\tilde{\epsilon}$–SVC slack variables verify

$$\xi_i^o = \frac{2}{\|\overline{W}^*\|^2}\,\xi_i^* = \frac{2}{\Lambda^*}\,\alpha_i^*,\ \ \mu_j^o = \frac{2}{\|\overline{W}^*\|^2}\,\mu_j^* = \frac{2}{\Lambda^*}\,\beta_j^*. \tag{5}$$

Now, since we have $\delta^o = 2\delta^*/\|\overline{W}^*\|^2 = 2C\delta^*/\Lambda^*$, proposition 1 and (5) imply that the ϵ–SVR slacks are

$$\hat{\xi}_i = \frac{1}{\delta^o}\xi_i^o = \frac{\Lambda^*}{2C\delta^*}\frac{2}{\Lambda^*}\,\alpha_i^* = \frac{1}{C\delta^*}\alpha_i^*,$$
$$\hat{\mu}_j = \frac{1}{\delta^o}\mu_j^o = \frac{\Lambda^*}{2C\delta^*}\frac{2}{\Lambda^*}\,\beta_j^* = \frac{1}{C\delta^*}\beta_j^*.$$

	Linear Penalty			Square Penalty		
Problem	C	σ	ϵ	C	σ	$\tilde{\epsilon}$
flare1	160	250	0.0025	0.125	10	0.04
flare2	30	150	0.001	0.5	12	0.06
flare3	40	175	0.001	3	40	0.05
building1	0.6	125	0.01	0.3	25	0.4
building2	3.2	8	0.01	0.8	6	0.6
building3	6.3	6.5	0.01	0.8	5	0.6
wind power	0.4	32	0.08	5	40	0.2

Table 1. SVM parameters used. For the `flare` and `building` problems only the parameters for the first output are shown.

Furthermore, since $\epsilon = \tilde{\epsilon} - 1/\delta^o$, we have

$$\epsilon + \hat{\xi}_i = \tilde{\epsilon} - \frac{\Lambda^*}{2C\delta^*} + \frac{\alpha_i^*}{C\delta^*} = \tilde{\epsilon} - \frac{1}{C\delta^*}\left(\frac{\Lambda^*}{2} - \alpha_i^*\right),$$
$$\epsilon + \hat{\mu}_j = \tilde{\epsilon} - \frac{\Lambda^*}{2C\delta^*} + \frac{\beta_j^*}{C\delta^*} = \tilde{\epsilon} - \frac{1}{C\delta^*}\left(\frac{\Lambda^*}{2} - \beta_j^*\right).$$

Since $\alpha_i^o = 0$ if and only if $\xi_i^o = 0$, all the regression patterns (X^i, y^i) for which $\alpha_i^o = \beta_i^o = 0$ will be inside an $\hat{\epsilon}$ tube with $\hat{\epsilon} = \tilde{\epsilon} - \Lambda^*/2C\delta^*$. Next, since α_i^* and β_i^* cannot be simultaneously nonzero, patterns $(X^i, y^i \pm \tilde{\epsilon})$ for which either coefficient is $< \Lambda^*/2$ result in regression patterns (X^i, y^i) inside the $\tilde{\epsilon}$–hard tube. On the other hand if, say, $\alpha_i^* > \Lambda^*/2$, the KKT conditions now imply

$$\widehat{W} \cdot X^i + b - y^i = \tilde{\epsilon} + \frac{1}{C\delta^*}\left(\alpha_i^* - \frac{\Lambda^*}{2}\right) > \tilde{\epsilon};$$

that is, (X^i, y^i) will fall outside the $\tilde{\epsilon}$–hard tube, and the same will happen with those (X^j, y^j) for which $\beta_j^* > \Lambda^*/2$. □

We will illustrate next square penalty SVR over several regression problems, comparing its performance to that of linear penalty SVR and of multilayer perceptrons.

4 Numerical Experiments

We have tested the performance of both linear and square penalty SVR methods in two Proben1 regression problems [7] and also in a wind power prediction one. The SVMSeq [13] algorithm was applied in the linear penalty case and the SK algorithm [6] for square penalties. In both cases a Gaussian kernel $k(x, y) = \exp\left(-\|x - y\|^2/2\sigma^2\right)$ was used. The Proben1 problems used were `building`, where hourly electrical energy and hot and cold water consumption in a building are to be predicted, and `flare`, where we want to do the same for the daily

Problem	Proben1 best result	MLP	Linear penalty SVM	Square penalty SVM
flare1	0.5283	0.5472	0.5444	0.5431
flare2	0.3214	0.2732	0.2680	0.2662
flare3	0.3568	0.3423	0.3457	0.3552
building1	0.6450	0.4267	0.4369	0.4556
building2	0.2509	0.2696	0.2418	0.2616
building3	0.2475	0.2704	0.2318	0.2525
Wind Power	–	8.33	8.96	8.68

Table 2. Mean square test errors for `building` and `flare` problems and for wind energy prediction obtained by an MLP and linear and square penalty SVMs. The corresponding best result recorded in the Proben1 database is also given.

Problem	Training set size	Linear penalty SVM	Square penalty SVM
flare1	800	792 (99%)	211 (26.37%)
flare2	800	799 (99.87%)	216 (27.00%)
flare3	800	796 (99.5%)	214 (26.75%)
building1	3156	3156 (100%)	2299 (72.84%)
building2	3156	2988 (94.67%)	1430 (45.31%)
building3	3156	3107 (98.44%)	1243 (39.38%)
Wind Power	1560	583 (37.37%)	315 (20.19%)

Table 3. Initial sample size, number of Support Vectors of linear and square penalty SVMs and corresponding percentages with respect to sample size for the first dependent variable in the `building` and `flare` problems and for wind energy prediction.

number of small, medium and large solar surface flares. Both datasets are given in [7] in three variations, numbered 1 to 3, each one with a different arrangement of the training and test sets. On the other hand, we will also work with a real wind power prediction task, where numerical weather predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF, [5]) at time T are used to provide energy production estimates for the Sotavento wind farm [10] located in Galicia (Spain) on 36 hour periods going from the $T + 12$ to the $T + 48$ hour. The test set was the farm's hourly production in August 2006.

Model performance was measured first by mean square error and the SVM model results were compared with the best ones in the Proben1 database and also with those provided by a single hidden layer multilayer perceptron (MLP). In the wind power problem errors are given as percentages of the farm's installed capacity. As it can be seen from the test results of table 2, all models give similar errors but the square penalty SVM ones are usually slightly larger than those of the linear penalty SVM but comparable to the MLP ones. Notice that, in any case, stopping criteria for linear and square penalty SVMs are different, as SVMSeq performs a gradient ascent over the dual problem while the SK algorithm tries to minimize the function g defined in section 2. On the other hand, the number of support vectors (SVs) obtained using square penalties is in general much smaller than in the case with linear penalties. This is seen in table

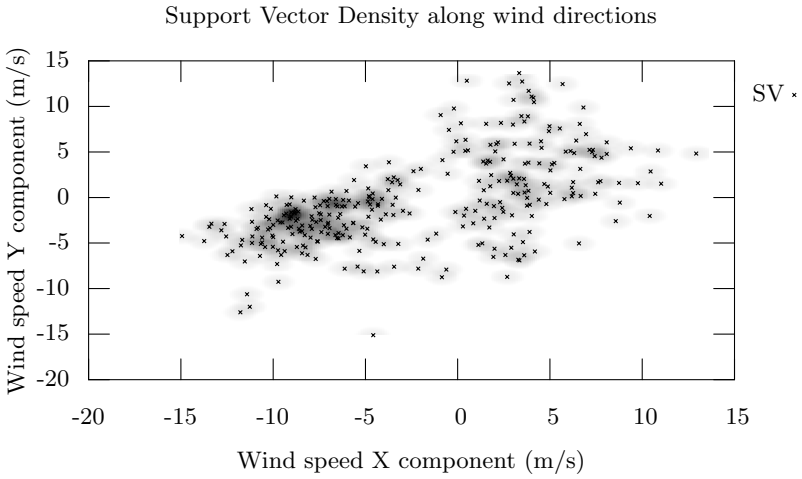

Fig. 1. Placement and density of support vectors plotted along wind speed X and Y components. Darker areas represent a higher support vector density.

3 for the wind energy prediction problem and the first variable to be predicted of each data split for the `building` and `flare` problems (similar results are obtained for the other two dependent variables).

An advantage of SVMs over MLPs for the wind energy prediction problem is the easier interpretation of the resulting model, as the SVs obtained somehow reflect the dominant data and the model's behaviour. For instance, figure 1 depicts the density of support vectors (computed using a Parzen window estimator) along the X and Y wind speed components. On the one hand, the density plot reflects dominant southwest–northeast wind directions in the training database; on the other, it is also clear that model performance over wind speed predictions outside the grey areas is likely to be poor. For instance, no support vectors appear for moderate–to–large wind speeds with northwest and southeast directions; this reflects that wind on these areas has been rare on the training database, but the model will also ignore it in the future.

The effect of the $\tilde{\epsilon}$ tube is depicted for the wind farm training data in figure 2, that shows the distribution of positively (circles) and negatively (crosses) shifted support vectors on a plot with absolute wind speed in the x–axis and energy production (as a percentage of installed power capacity) in the y–axis. As it can be expected, for a given wind speed patterns with negative shifts tend to lie below the positively shifted ones; when this is not the case, it is likely to be due to the presence of outliers. The figure is also somewhat reminiscent of the power curve of wind turbines, which typically have a sigmoidal–like shape with a cut–off for wind speeds above 25 m/s.

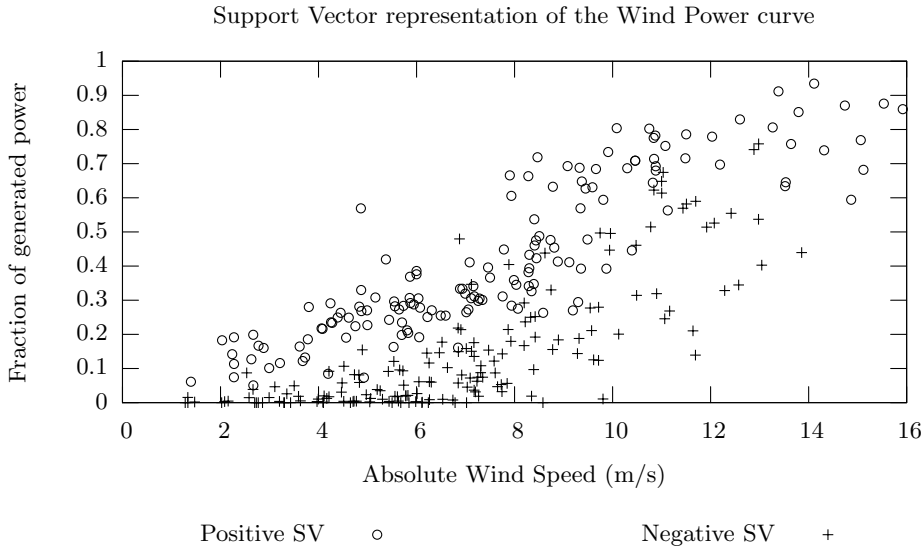

Fig. 2. Positively (circles) and negatively (crosses) shifted SVs over absolute wind (x axis) and percentage of maximum energy output (y axis).

5 Conclusions and Discussion

In this work we have studied support vector regression where tube squared slacks are added as a penalty term to the standard SVM squared weight norm. This has two clear advantages. The first one is that, as it happens with SVM classification, hard and soft SVR can be formally treated in the same way once soft weights and patterns are appropriately extended. The second advantage is that the coefficients of the support vectors obtained have a direct relationship with the width of the tube where these support vectors will fall in. Moreover, and as demonstrated with our numerical experiments, both linear and square penalties seem to give quite similar test errors, while the second models tend to give a smaller number of support vectors (resulting in a faster application on new data).

As pointers to further study, more experimental work is needed for a more precise comparison on linear and square penalty SVR performance. Moreover, the standard SVR formulation has the drawback of having to decide on the extra ϵ parameter on top of the other two usual SVM parameters, namely the penalty factor C and the kernel width σ. For linear penalties, the so–called ν–SVR [8] allows to simultaneously get rid of the C and ϵ parameters by introducing a new parameter ν that, moreover, can be used to control some aspects of the SVM obtained. It may be possible that appropriate square penalty extensions of ν–SVR provide the same benefits. These and similar topics are presently under consideration.

References

1. Bennett, K., Bredensteiner, E.: Geometry in learning. In: Gorini, C., Hart, E., Meyer, W., Phillips T. (eds.). Geometry at Work, Mathematical Association of America, 1997.
2. Bi, J., Bennett, K.: A geometric approach to support vector regression. Neurocomputing 55, 187–220 (2003).
3. Burges, C., Crisp, D.: Uniqueness theorems for kernel methods. Neurocomputing 55, 187–220 (2003).
4. Chang, C., Lin, C.: LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/˜cjlin/LIBSVM.
5. European Centre for Medium-Range Weather Forecasts, http://www.ecmwf.int
6. Franc, V., Hlaváč, V.: An iterative algorithm learning the maximal margin classifier. Pattern Recognition 36, 1985–1996 (2003).
7. Prechelt, L.: Proben1 - A Set of Neural Network Benchmark Problems and Benchmarking Rules, http://digbib.ubka.uni-karlsruhe.de/eva/ira/1994/21.
8. Schölkopf, B., Smola, A., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Computation 12, 1083–1121 (2000).
9. Smola, A., Schölkopf, B.: A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030 (1998).
10. Parque Eólico Experimental Sotavento, http://www.sotaventogalicia.com
11. University of California Irvine: UCI-benchmark repository of machine learning data sets, http://www.ics.uci.edu.
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995).
13. Vijayakumar, S., Wu, S.: Sequential Support Vector Classifiers and Regression. In: Proc. International Conference on Soft Computing (SOCO'99), pp. 610–619 (1999).