Journal of Machine Learning Research 8 (2007) 2467-2495 Submitted 7/06; Revised 4/07; Published 10/07

On the Representer Theorem and Equivalent Degrees of Freedom of SVR

Francesco Dinuzzo FRANCESCO.DINUZZO01@ATENEOPV.IT
Marta Neve MARTA.NEVE@UNIPV.IT
Giuseppe De Nicolao GIUSEPPE.DENICOLAO@UNIPV.IT
Dipartimento di Informatica e Sistemistica
Università di Pavia
Pavia, Italy

Ugo Pietro Gianazza GIANAZZA@IMATI.CNR.IT
Dipartimento di Matematica
Università di Pavia
Pavia, Italy

Editor: Ralf Herbrich

Abstract

Support Vector Regression (SVR) for discrete data is considered. An alternative formulation of the representer theorem is derived. This result is based on the newly introduced notion of pseudoresidual and the use of subdifferential calculus. The representer theorem is exploited to analyze the sensitivity properties of ε-insensitive SVR and introduce the notion of approximate degrees of freedom. The degrees of freedom are shown to play a key role in the evaluation of the optimism, that is the difference between the expected in-sample error and the expected empirical risk. In this way, it is possible to define a C_p-like statistic that can be used for tuning the parameters of SVR. The proposed tuning procedure is tested on a simulated benchmark problem and on a real world problem (Boston Housing data set).

Keywords: statistical learning, reproducing kernel Hilbert spaces, support vector machines, representer theorem, regularization theory

1. Introduction

Although Support Vector Machines are mainly used as classification algorithms, recent years have witnessed a growing interest for their application to regression problems as well. Among the advantages of SVR (Support Vector Regression), there are the sparseness property and the robustness against outliers.

The SVR estimator can be seen as the minimizer of a cost functional given by the sum of an ε-insensitive loss function and a regularization penalty. As such, it is a particular case of a larger class of kernel-based estimators that are obtained by applying regularization theory in Reproducing Kernel Hilbert Spaces (RKHS). Under mild assumptions, the solution of these problems can be written as a linear combination of kernel functions. This kind of result goes under the name of representer theorem. The first result of this type was due to Kimeldorf and Wahba (1979) for squared loss functions, see also Tikhonov and Arsenin (1977) for the application in the context of inverse problems.

The representer theorem was further generalized to differentiable loss functions (Cox and O' Sullivan, 1990; Poggio and Girosi, 1992) and even arbitrary monotonic ones (Schölkopf et al., 2001). Another important issue is the quantitative characterization of the coefficients a_i of the linear combination. For squared losses it is well known that the coefficients are obtained as the solution of a system of linear equations, see for example Wahba (1990) and Cucker and Smale (2001). An explicit characterization of the coefficients as the solution of a system of algebraic equations is still possible if the loss function is differentiable (Wahba, 1998). This result cannot be applied to ε-insensitive SVR because the loss function is not differentiable. The usual computational approach is to reformulate the original variational problem as a constrained minimization one whose dual Lagrangian formulation boils down to a finite dimensional quadratic programming problem (Vapnik, 1995).

Some recent contributions have approached the nondifferentiability issue by resorting to subdifferential calculus. More precisely, Steinwart (2003) has proven a quantitative representer theorem that, without using the dual problem, characterizes the coefficients by means of inclusions, when convex loss functions are considered. Various extensions can be found in De Vito et al. (2004). In particular, besides providing an alternative simpler proof of the quantitative representer theorem, De Vito and coworkers allow for the offset space and cover both regression and classification.

The contribution of the present paper is twofold. First of all, quantitative representation results are worked out for convex loss functions. Then, these results are specialized to SVR in order to study its sensitivity to data and develop a tuning method for its parameters.

Concerning the quantitative representation of the coefficients a_i, the paper provides a simple derivation of the quantitative representer theorem based on Fourier arguments (see Appendix A). Another result is a new formulation of the quantitative representer theorem that replaces inclusions with equations by using the newly introduced notion of pseudoresidual (Theorem 1). This result, not only gives insight into the relation between data and coefficients, but also puts the basis for the subsequent analysis of SVR properties. In particular, we give a complete characterization of the sensitivity of SVR coefficients and predictions with respect to the output data. Past work has focused on sensitivity with respect to the regularization parameter C, see for example Pontil and Verri (1998) and Hastie et al. (2004). As a byproduct of the sensitivity analysis, the degrees of freedom of SVR, defined as the trace of the sensitivity matrix, are found to be equal to the number of marginal support vectors. This analysis is instrumental to the last issue dealt with in the paper, that is the tuning of both the ε and C parameters of the SVR.

In the literature, the tuning of SVR has been addressed using various approaches. The interpretation of SVR as a Bayesian estimator provides a conceptually elegant framework for reformulating parameter tuning as a statistical estimation problem (Gao et al., 2002). The major drawback is the necessity of assuming the validity of the statistical prior underlying the Bayesian interpretation of SVR, an assumption that may not be appropriate in all cases.

As a matter of fact, the great majority of tuning approaches aims at the minimization of the prediction error. A powerful, though computationally expensive solution is to resort to k-fold cross validation. Alternatively, Chang and Lin (2005) strive for the minimization of an upper bound of the leave-one-out absolute error. Other authors have discussed the choice of ε observing that, asymptotically, the optimal ε depends linearly on the measurement error standard deviation (Smola et al., 1998; Kwok and Tsang, 2003). Finally, Schölkopf et al. (2000) have proposed modified SVR schemes that ease the tuning of the parameters.

A tuning method based on the extension of the *GCV* criterion to SVR has been proposed by Gunter and Zhu (2007).

In the present paper, a different approach is pursued which is based on the estimation of the so-called in-sample prediction error (Hastie et al., 2001). See also Cherkassky and Ma (2003) and Hastie et al. (2003) for a discussion on the merits and difficulties of this and other approaches to model selection. Herein, it is shown how the optimism, that is the difference between the in-sample prediction error and the expected empirical risk, depends on the sensitivity of the estimator (Theorem 3). This result opens the way to the estimation of the in-sample prediction error as a function of the measurement error variance and the degrees of freedom. This estimator can be seen as an extension to SVR of the so-called C_p statistic, a well known criterion for linear model order selection. A major advantage of the C_p statistic is that, differently from many other criteria, no assumption is made on the correctness of the model.

The paper is organized as follows. After some preliminaries (Section 2), the major results regarding the representation of the coefficients a_i and the sensitivity analysis of SVR are derived in Section 3, which ends with the definition of the degrees of freedom. The issue of parameter tuning is treated in Section 4, where the estimation of the in-sample prediction error by means of a suitable C_p statistic is addressed. Finally, the proposed parameter tuning procedure is illustrated in Section 5 by means of both a simulated problem and a real-world one. Some concluding remarks (Section 6) end the paper.

2. Preliminaries

Consider the problem of estimating the functional relationship existing between an input vector $\mathbf{x} \in \mathbb{R}^N$ and the output $y \in \mathbb{R}$ given the training set $\mathcal{D} = \{\mathbf{x}_i, y_i\}$ $(i = 1, 2, \ldots, \ell)$, where the input vectors $\mathbf{x}_i$ are all distinct. According to the Support Vector Regression approach, the function $\hat{f}(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ solving the aforementioned problem belongs to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ and minimizes the regularized risk:

$$\hat{f} = \arg\min_{f \in \mathcal{H}} H[f] = \arg\min_{f \in \mathcal{H}} \left(C \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right). \tag{1}$$

The parameter C controls the relative importance given to the empirical risk and the regularization term $\|f\|_{\mathcal{H}}^2$, and must be properly tuned in order to obtain good performance. Among the possible convex loss functions V (quadratic, Laplace, etc) particular attention will be given to the ε-insensitive one:

$$V(y_i, f(\mathbf{x}_i)) = V_\varepsilon(y_i - f(\mathbf{x}_i)) = \begin{cases} 0, & |f(\mathbf{x}_i) - y_i| \leq \varepsilon \\ |f(\mathbf{x}_i) - y_i| - \varepsilon, & |f(\mathbf{x}_i) - y_i| > \varepsilon. \end{cases} \tag{2}$$

Such a function is known to produce sparse solutions, meaning that they depend only on a small number of training examples scattered in the input space. The positive scalar ε measures the extent of the "dead zone" (that is the interval over which the loss function is zero) and should be either fixed according to the desired resolution or tuned using an objective criterion.

The usual approach to the numerical computation of $\hat{f}$ calls for the solution of the dual quadratic programming problem, see for example Vapnik (1995). If the kernel is positive definite, the representer theorem states that the solution $\hat{f}$ can be written as

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{\ell} a_i K(\mathbf{x}_i, \mathbf{x}), \tag{3}$$

where a_i are suitable coefficients. If V is everywhere differentiable with respect to its second argument, it can be shown that

$$a_i = -C\partial_2 V(y_i, \hat{f}(\mathbf{x}_i)),$$

where ∂_2 denotes the partial derivative with respect to the second argument. Conversely, if V is a general measurable function convex with respect to its second argument, it is necessary to resort to subdifferential calculus (for a quick reference to the basic concepts of subdifferential calculus the interested reader may usefully refer to Steinwart (2003) and De Vito et al. (2004)). See also Borwein and Lewis (2000). In particular, Steinwart (2003) and De Vito et al. (2004) (Theorem 2) have shown that

$$a_i \in -C\partial_2 V(y_i, \hat{f}(\mathbf{x}_i)), \tag{4}$$

where, now, ∂_2 is the subdifferential with respect to the second argument. This result goes under the name of quantitative representer theorem. Note that (4) is no longer an equation but just an inclusion.

De Vito et al. (2004) studied also the so called continuous setting, that is measurements are taken on a continuous set rather than being taken as discrete samples. In Appendix A, we provide an alternative concise proof of the quantitative representer theorem based on Fourier arguments. The analysis developed in the next section differs from the representation results by Steinwart (2003) and De Vito et al. (2004) in that we show that the system of inclusions (4) can be replaced by a set of equations.

3. Quantitative Representation and Sensitivity Analysis

Hereafter, it is assumed that the loss function is of the type

$$V(y_i, f(\mathbf{x}_i)) = V(f(\mathbf{x}_i) - y_i),$$

where $V(\cdot)$ is a convex function and is twice differentiable everywhere except in a finite number of points γ_j, $j = 1, \ldots, N$. In the following, $D^-(\gamma)$ and $D^+(\gamma)$ will denote the left and right derivative of $V(\cdot)$ at γ:

$$D^-(\gamma) = \lim_{h \to 0^+} \frac{V(\gamma - h) - V(\gamma)}{-h},$$

$$D^+(\gamma) = \lim_{h \to 0^+} \frac{V(\gamma + h) - V(\gamma)}{h},$$

Letting $I = \{1, 2, \ldots, \ell\}$, define the *pseudoresiduals* as

$$\eta_i := y_i - \sum_{\substack{j \in I \\ j \neq i}} a_j K(\mathbf{x}_i, \mathbf{x}_j).$$

The following result holds

Theorem 1 *The coefficients a_i, $i = 1,\ldots,\ell$, that characterize the solution of problem (1), satisfy a system of algebraic equations*

$$a_i = S_i(\eta_i),$$

where $S_i(\eta_i)$ are monotone nondecreasing Lipschitz continuous functions. Moreover, when

$$\eta_i \in \left[-\left(\gamma_j + CK(\mathbf{x}_i,\mathbf{x}_i)D^+(\gamma_j)\right), -\left(\gamma_j + CK(\mathbf{x}_i,\mathbf{x}_i)D^-(\gamma_j)\right)\right], \quad j = 1,\ldots,N, \tag{5}$$

the functions $S_i(\eta_i)$ are affine and given by

$$S_i(\eta_i) = \frac{\eta_i + \gamma_j}{K(\mathbf{x}_i,\mathbf{x}_i)}.$$

Proof. By the definition of pseudoresidual,

$$\hat{f}(\mathbf{x}_i) - y_i = a_i K(\mathbf{x}_i,\mathbf{x}_i) - \eta_i. \tag{6}$$

Then,

$$a_i = \frac{\eta_i + \hat{f}(\mathbf{x}_i) - y_i}{K(\mathbf{x}_i,\mathbf{x}_i)}.$$

Now, there are two cases depending on whether $V(\cdot)$ is twice differentiable at $\gamma := \hat{f}(\mathbf{x}_i) - y_i$ or not. When $\gamma \neq \gamma_j$, $j = 1,\ldots,N$, $V(\gamma)$ is twice differentiable and its subdifferential is single-valued so that (4) yields

$$a_i = -CV'(\hat{f}(\mathbf{x}_i) - y_i) = -CV'(a_i K(\mathbf{x}_i,\mathbf{x}_i) - \eta_i). \tag{7}$$

Now, the Implicit Function Theorem can be used to prove that, locally, a_i is a monotone nondecreasing Lipschitz continuous function of η_i. In fact, by deriving with respect to η_i,

$$\frac{\partial a_i}{\partial \eta_i} = \frac{CV''(a_i K(\mathbf{x}_i,\mathbf{x}_i) - \eta_i)}{1 + CK(\mathbf{x}_i,\mathbf{x}_i)V''(a_i K(\mathbf{x}_i,\mathbf{x}_i) - \eta_i)}.$$

The denominator is always different from zero because, by convexity, $V'' \geq 0$ whenever it exists. Therefore, locally, a_i is a differentiable function of η_i:

$$a_i = \overline{S}(\eta_i).$$

The function $\overline{S}(\eta_i)$ is monotone nondecreasing and has bounded derivative because

$$0 \leq \frac{\partial a_i}{\partial \eta_i} < \frac{1}{K(\mathbf{x}_i,\mathbf{x}_i)}. \tag{8}$$

Now, let us consider the second case. When γ is fixed as $\gamma = \gamma_j$ for some j, $V(\cdot)$ is not twice differentiable at γ. Then, from (6),

$$a_i = S_i(\eta_i) = \frac{\gamma_j + \eta_i}{K(\mathbf{x}_i,\mathbf{x}_i)}, \tag{9}$$

so that a_i is an affine function of η_i in the interval

$$I_j := [\eta_j^L, \eta_j^R],$$

where

$$\begin{aligned} \eta_j^L &:= -\left(\gamma_j + CK(\mathbf{x}_i, \mathbf{x}_i) D^+(\gamma_j)\right), \\ \eta_j^R &:= -\left(\gamma_j + CK(\mathbf{x}_i, \mathbf{x}_i) D^-(\gamma_j)\right). \end{aligned}$$

On the other hand, recalling the properties of the subdifferential of a convex function,

$$a_i \in \left[-CD^+(\gamma_j), -CD^-(\gamma_j)\right]. \tag{10}$$

Hence, (5) follows from (9) and (10). Finally, since

$$\frac{\partial a_i}{\partial \eta_i} = \frac{1}{K(\mathbf{x}_i, \mathbf{x}_i)} > 0,$$

the functions $S_i(\eta_i)$ are locally monotone nondecreasing also in the second case. Combining this last inequality with the bound (8) that holds in differentiability points, we conclude that the derivative of $S_i(\eta_i)$ is bounded everywhere, possibly except for discontinuity points. Then, in order to prove Lipschitz continuity it suffices to show that $S_i(\eta_i)$ is continuous.

We now conclude the proof showing that the set of discontinuity points is actually empty. In this respect, the only points that must be analyzed are the boundaries of the intervals I_j. In fact, in the interior of I_j, $S_i(\eta_i)$ is infinitely differentiable because it is affine, while, outside, it has the same regularity of $V'(\cdot)$. Hence, it suffices to prove continuity at the left boundary η_j^L of I_j. Consider (9) and take the limit from the right:

$$\lim_{\eta_i \to (\eta_j^L)^+} S_i(\eta_i)\Big|_{\eta_i \in I_j} = \lim_{\eta_i \to (\eta_j^L)^+} \frac{\gamma_j + \eta_i}{K(\mathbf{x}_i, \mathbf{x}_i)} = -CD^+(\gamma_j).$$

Now, observe that, if a_i tends to $-CD^+(\gamma_j)$ from below, then η_i tends to η_j^L from the left. Indeed, taking the limit in (7) for $a_i \to -CD^+(\gamma_j)$ from below, we obtain that $\hat{f}(\mathbf{x}_i) - y_i \to \gamma_j$ from the right (recall that $V'(\cdot)$ is nondecreasing). In turn,

$$\begin{aligned} \lim_{a_i \to \left(-CD^+(\gamma_j)\right)^-} \eta_i &= \lim_{a_i \to \left(-CD^+(\gamma_j)\right)^-} \left(y_i - \hat{f}(\mathbf{x}_i) + a_i K(\mathbf{x}_i, \mathbf{x}_i)\right) \\ &= -\gamma_j - CD^+(\gamma_j) K(\mathbf{x}_i, \mathbf{x}_i) = \eta_j^L \end{aligned}$$

from the left. This proves the continuity of $S_i(\eta_i)$ at the left boundary η_j^L of I_j. ■

Hereafter, it will be assumed that $V = V_\varepsilon$ is the so-called ε-insensitive function. Hence, V_ε is not differentiable only at $\gamma_1 = -\varepsilon$ and $\gamma_2 = +\varepsilon$. The subdifferential of the loss function has a rather simple structure:

$$C\partial V_\varepsilon(\hat{f}(\mathbf{x}_i)-y_i) = \begin{cases} \{-C\} & \hat{f}(\mathbf{x}_i)-y_i < -\varepsilon, \\ [-C,0] & \hat{f}(\mathbf{x}_i)-y_i = -\varepsilon, \\ \{0\} & -\varepsilon < \hat{f}(\mathbf{x}_i)-y_i < \varepsilon, \\ [0,C] & \hat{f}(\mathbf{x}_i)-y_i = \varepsilon, \\ \{C\} & \hat{f}(\mathbf{x}_i)-y_i > \varepsilon. \end{cases}$$

For the subsequent derivation it is useful to define the following sets:

$$\begin{aligned} I_{in} &= \{i \in I : |\hat{f}(\mathbf{x}_i)-y_i| < \varepsilon\}, \\ I_C^+ &= \{i \in I : \hat{f}(\mathbf{x}_i)-y_i > \varepsilon\}, \\ I_C^- &= \{i \in I : \hat{f}(\mathbf{x}_i)-y_i < -\varepsilon\}, \\ I_M^+ &= \{i \in I : \hat{f}(\mathbf{x}_i)-y_i = \varepsilon\}, \\ I_M^- &= \{i \in I : \hat{f}(\mathbf{x}_i)-y_i = -\varepsilon\}, \\ I_{out} &= I_C^+ \cup I_C^- \qquad I_M = I_M^+ \cup I_M^-. \end{aligned}$$

Note that the set I_{in} identifies the data pairs $\{\mathbf{x}_i, y_i\}$ that belong to the so-called ε-tube, whereas I_{out} identifies the data outside the tube. The indices belonging to I_M correspond to data pairs lying on the boundary of the ε-tube, also called marginal support vectors. The union of I_{out} and I_M identifies the so-called support vectors.
In view of Theorem 1, the next corollary follows.

Corollary 1 *For the ε-insensitive loss function,*

$$a_i = S_i(\eta_i) = \begin{cases} -C & \eta_i \le -(\varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)), \\ \frac{\eta_i+\varepsilon}{K(\mathbf{x}_i,\mathbf{x}_i)} & -(\varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)) < \eta_i < -\varepsilon, \\ 0 & -\varepsilon \le \eta_i \le \varepsilon, \\ \frac{\eta_i-\varepsilon}{K(\mathbf{x}_i,\mathbf{x}_i)} & \varepsilon < \eta_i < (\varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)), \\ C & \eta_i \ge (\varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)). \end{cases}$$

Moreover,

- *If $i \in I_{in}$, $|\eta_i| \le \varepsilon$.*
- *If $i \in I_{out}$, $|\eta_i| \ge \varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)$.*
- *If $i \in I_M$, $\varepsilon \le |\eta_i| \le \varepsilon + CK(\mathbf{x}_i,\mathbf{x}_i)$.*

Corollary 1, which is illustrated in Fig. 1, is now used to evaluate the sensitivity of SVR with respect to the data.

Let $m = \#I_M$ denote the number of marginal vectors. We can assume without loss of generality that the indices I are ordered such that $I_M = \{1,\ldots,m\}$, $I_{\bar{M}} := I_{out} \cup I_{in} = \{m+1,\ldots,\ell\}$. Let the matrix $\mathbf{K} := [K(\mathbf{x}_i,\mathbf{x}_j)]$ be partitioned as

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{MM} & \mathbf{K}_{M\bar{M}} \\ \mathbf{K}_{\bar{M}M} & \mathbf{K}_{\bar{M}\bar{M}} \end{pmatrix},$$

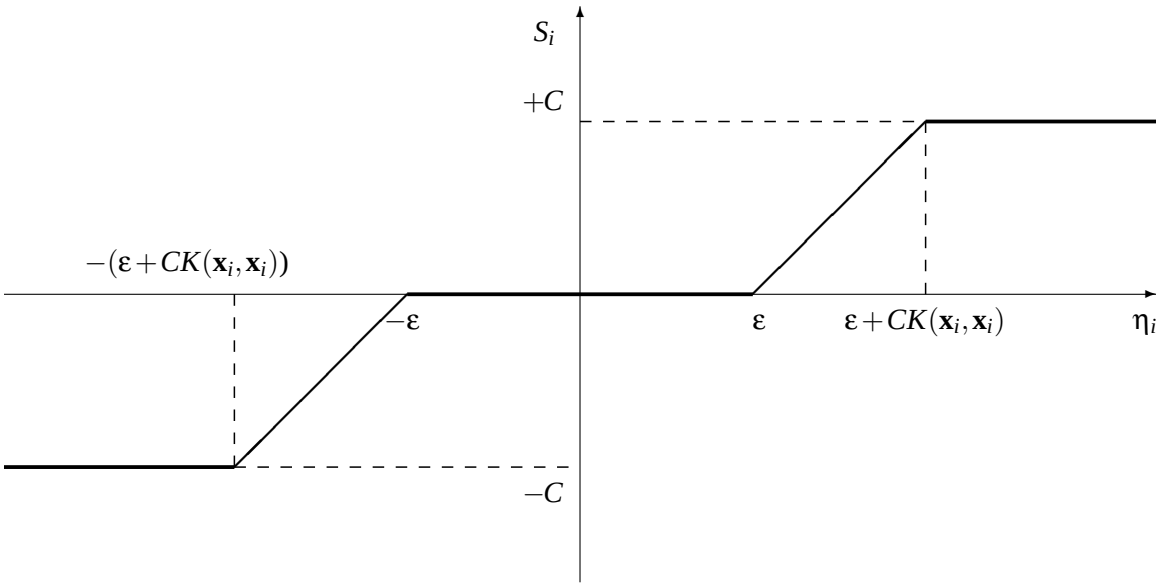

Figure 1: This function S_i gives the dependency of the coefficient a_i on the pseudoresidual η_i when the ε-insensitive loss function is used.

where $\mathbf{K}_{MM} \in \mathbb{R}^{m\times m}$. It is also useful to partition the coefficient vector $\mathbf{a} = (a_1,\ldots,a_\ell)^T$ as $(\mathbf{a}_M^T, \mathbf{a}_{\overline{M}}^T)^T$ and the data vector $\mathbf{y} = (y_1,\ldots,y_\ell)^T$ as $(\mathbf{y}_M^T, \mathbf{y}_{\overline{M}}^T)^T$, where $\mathbf{a}_M, \mathbf{y}_M \in \mathbb{R}^m$. Moreover, define

$$\mathbf{k}(\mathbf{x}) = \begin{pmatrix} K(\mathbf{x},\mathbf{x}_1) \\ \ldots \\ K(\mathbf{x},\mathbf{x}_\ell) \end{pmatrix}.$$

Proposition 1 *Assume that* $\eta_i \neq \pm\varepsilon$ *and* $\eta_i \neq \pm(\varepsilon+CK(\mathbf{x}_i,\mathbf{x}_i))$. *Then,*

$$\frac{\partial \hat{f}}{\partial y}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T \begin{pmatrix} \mathbf{K}_{MM}^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. First of all, by (3)

$$\frac{\partial \hat{f}}{\partial y_k}(\mathbf{x}) = \sum_{j=1}^{\ell} \frac{\partial a_j}{\partial y_k} K(\mathbf{x},\mathbf{x}_j).$$

In view of Corollary 1 and the definition of pseudoresidual η_i,

$$\frac{\partial a_i}{\partial y_k} = S_i'(\eta_i) \left(\delta_{ik} - \sum_{\substack{j\in I \\ j\neq i}} \frac{\partial a_j}{\partial y_k} K(\mathbf{x}_i,\mathbf{x}_j) \right), \forall i \tag{11}$$

where

$$S_i'(\eta_i) = \begin{cases} 0 & i \notin I_M, \\ \frac{1}{K(\mathbf{x}_i,\mathbf{x}_i)} & i \in I_M \end{cases}$$

and δ_{ik} is Kronecker's delta. Hence, $\forall i \notin I_M$, $\forall k$,

$$\frac{\partial a_i}{\partial y_k} = 0. \tag{12}$$

On the other hand, $\forall i \in I_M$, $\forall k$, (11) reads

$$\frac{\partial a_i}{\partial y_k} = \frac{1}{K(\mathbf{x}_i,\mathbf{x}_i)} \left(\delta_{ik} - \sum_{\substack{j \in I_M \\ j \neq i}} \frac{\partial a_j}{\partial y_k} K(\mathbf{x}_i,\mathbf{x}_j) \right),$$

whence

$$\sum_{j \in I_M} \frac{\partial a_j}{\partial y_k} K(x_i,x_j) = \delta_{ik}, \qquad \forall i \in I_M, \forall k \in I. \tag{13}$$

Equations (12) and (13) can be written as

$$\frac{\partial \mathbf{a}_{\bar{M}}}{\partial \mathbf{y}} = 0,$$

$$\mathbf{K}_{MM} \frac{\partial \mathbf{a}_M}{\partial \mathbf{y}} = \begin{pmatrix} I & 0 \end{pmatrix}.$$

Since the vectors $\mathbf{x}_i$ are all distinct, $\mathbf{K}_{MM}$ is a positive definite matrix and therefore

$$\frac{\partial \mathbf{a}}{\partial \mathbf{y}} = \begin{pmatrix} \mathbf{K}_{MM}^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

from which the thesis follows. ■

For linear-in-parameter regression it is usual to define the degrees of freedom of the estimator as the trace of the so-called "hat matrix," that maps the vector of output data into the corresponding predictions. Such degrees of freedom have a number of applications ranging from the computation of confidence intervals to model validation and model order selection, see for example Hastie and Tibshirani (1990) and Hastie et al. (2001).

Let $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_\ell)^T$, where $\hat{y}_i = \hat{f}(\mathbf{x}_i)$ denotes the SVR prediction at $\mathbf{x}_i$. The following Proposition provides the degrees of freedom of SVR. For an alternative proof, based on the dual problem formulation, see Gunter and Zhu (2007).

Proposition 2 *Let the degrees of freedom of the SVR be defined as*

$$q(\mathcal{D}) := tr\left(\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} \right).$$

Then, under the assumption of Proposition 1, $q(\mathcal{D})$ is equal to the number m of marginal support vectors.

Proof. If $\mathbf{x} = \mathbf{x}_i$, the application of Proposition 1 yields

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \begin{pmatrix} I & 0 \\ \mathbf{K}_{\bar{M}M}\mathbf{K}_{MM}^{-1} & 0 \end{pmatrix}, \tag{14}$$

so that $q(\mathcal{D})$ is just equal to m.

Remark 1 *Note that the number m of marginal vectors can be evaluated by looking at the pseudoresiduals* η_i*. More precisely, the i-th observation* y_i *is a marginal vector if*

$$\varepsilon \leq |\eta_i| \leq \varepsilon + CK(\mathbf{x}_i, \mathbf{x}_i).$$

Remark 2 *The assumption on the value* η_i *made in Proposition 1 rules out the marginal support vectors whose coefficient* a_i *is either* 0 *or* $\pm C$*. This corresponds to experimental data that under a suitable infinitesimal perturbation leave the boundary of the* ε*-tube moving either inward (*$a_i = 0$*) or outward (*$a_i = \pm C$*). For such "transition data"* y_i*, the right and left derivatives* $\partial \hat{y}_i / \partial y_i$ *are different, so that the degrees of freedom would not be uniquely defined. Then, the degrees of freedom would range from the minimum to the maximum value of* $tr(\partial \hat{\mathbf{y}} / \partial \mathbf{y})$*. Alternatively, one could assign* $1/2$ *degree of freedom to each transition datum. Given that such pathological situation occurs on a zero-measure set, they will be removed from the analysis without appreciable consequences.*

4. Prediction Error Assessment via C_p Statistic

The goal of any regression method is to achieve good generalization performance. In this section, an index will be derived that assesses the generalization capabilities of SVR. In turn, this index can be used to tune the design parameters of the estimator.

In order to proceed, it is assumed that the training data are given by

$$y_i = f^0(x_i) + v_i, \tag{15}$$

where $f^0(x)$ is the "true function" to be estimated and the measurement error vector $v = [v_1 \dots v_\ell]^T$ is such that $E[v] = 0$, $Var[v] = diag(\sigma_1^2 \dots \sigma_\ell^2)$. Note that $f^0(x_i)$ can be seen as the conditional expectation of y_i given x_i ($f^0(x_i) = E[y_i|x_i]$) and $f^0(x)$ is also known as regression function.

In the following, the generalization performance will be measured in terms of the sum of squared errors. A first type of error is the empirical risk

$$\overline{err} := \frac{1}{\ell}\|y - \hat{y}\|^2.$$

This is not a valid measure of generalization because $\hat{y}$ depends on y. Usually, the generalization capabilities of the estimator are measured by the expected risk. Unfortunately, it is not easy to assess the value of the expected risk without introducing assumptions on the nature of $f^0(x_i)$. As an alternative, one can look for probabilistic upper bounds on the expected risk which may be, in some cases, too loose for an optimal tuning of the SVR parameters. Hereafter, attention will be focused on the so-called in-sample prediction error, that is the expected error associated with a new set of data

$$y_i^{new} = f^0(x_i) + v_i^{new},$$

where v^{new} has the same statistics as v but is independent of it. The in-sample prediction error, see for example Hastie et al. (2001), is defined as

$$Err_{in} := E\left[\frac{1}{\ell}\|y^{new} - \hat{y}\|^2\right],$$

where the x_i are fixed and the expected value is taken with respect to both the distribution of y_i and y_i^{new}. A major motivation for using the in-sample prediction error is that, as shown below, it can be assessed with good accuracy, without introducing undue assumptions on $f^0(x)$.

Remark 3 *Recalling the expression of the cost function (1) it would be tempting to use*

$$Err_{in}^V := E\left[\frac{1}{\ell}\sum_{i=1}^{\ell} V(y_i^{new}, \hat{y}_i)\right]$$

as a measure of generalization performance. In particular, the parameters (C, ε) would be tuned so as to minimize Err_{in}^V. However, it is immediate to see that $Err_{in}^V = 0$ for sufficiently large ε so that a joint tuning of the two parameters is not possible. Conversely, as observed by Hastie et al. (2001), the in-sample prediction error proves useful for model comparison and selection because, although it underestimates the expected risk, in this context the relative size of the error is what matters, see also Efron (1986). Note also that the use of SVR is not necessarily in contrast with square loss minimization, insofar sparsity of the solution is an important feature. On these premises, in the present paper the minimization of the quadratic in-sample prediction error Err_{in} is pursued. A similar choice has been made by Gunter and Zhu (2007) who derive a quadratic-type GCV criterion for SVR.

The empirical and the in-sample prediction error are linked as stated in the following proposition, see for example Hastie et al. (2001). Note that the expectations are taken over the training set.

Proposition 3 *Define the 'optimism' as*

$$op := \frac{2}{\ell} E[\hat{y}^T v].$$

Then,

$$Err_{in} = E[\overline{err}] + op.$$

The index Err_{in} can be approximated by

$$\widehat{Err}_{in} = \overline{err} + \widehat{op},$$

where $\widehat{op}$ is an estimate of op. If $\hat{y}$ is a linear function of y, and $\sigma_i^2 = \sigma^2$, $\forall i$, then the optimism can be expressed as

$$op = \frac{2q\sigma^2}{\ell} \tag{16}$$

(note that in the linear case the degrees of freedom q do not depend on the training data y). In this linear case, $\widehat{Err}_{in}$ is better known as C_p statistic

$$C_p = \overline{err} + \frac{2q\sigma^2}{\ell}. \tag{17}$$

The purpose of the present section is to extend the C_p statistic to Support Vector Regression. The following theorem highlights the relationship between the optimism and the sensitivities $\partial h_i/\partial y_i$ of a generic estimator $\hat{y} = h(y)$. Let us define $y^0 = [f^0(x_1) \dots f^0(x_\ell)]^T$.

Theorem 2 *Assume that*

(i) eq. (15) holds,

(ii) the errors v_i are independent of each other,

(iii) the variances $\sigma_i^2 = Var[v_i]$ are finite,

(iv) the estimator $h(y)$ is such that for $i = 1, \dots, \ell$

$$\lim_{|y_i| \to \infty} \frac{|h_i(y)|}{|y_i|} = 0.$$

Then,

$$op = \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i^2 \int_{\mathbb{R}^\ell} \frac{\partial h_i}{\partial y_i}(y) \prod_{j \neq i}^{\ell} p_j(v_j) \phi_i(v_i) dv,$$

where

$$\phi_i(v_i) = \frac{1}{\sigma_i^2} \int_{v_i}^{+\infty} s p_i(s) ds,$$

and $p_i(v_i)$ denotes the probability density function of v_i.
Moreover, if the errors v_i are Gaussian,

$$op = \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i^2 E\left[\frac{\partial h_i}{\partial y_i}(y)\right]. \tag{18}$$

Proof. By definition,

$$\begin{aligned} op &= \frac{2}{\ell} E[\hat{y}^T v] = \frac{2}{\ell} \sum_{i=1}^{\ell} E[h_i(y) v_i] \\ &= \frac{2}{\ell} \sum_{i=1}^{\ell} \int_{\mathbb{R}^{\ell-1}} \prod_{j \neq i}^{\ell} p_j(v_j) \left(\int_{\mathbb{R}} h_i(y^0 + v) v_i p_i(v_i) dv_i \right) dv^{[-i]}, \end{aligned}$$

where

$$dv^{[-i]} = \prod_{j \neq i}^{\ell} dv_j.$$

Integration by parts of the inner integral yields

$$\int_{\mathbb{R}} h_i(y^0+v)v_i p_i(v_i)dv_i = \sigma_i^2 \left(\int_{\mathbb{R}} \frac{\partial h_i}{\partial v_i}(y^0+v)\phi_i(v_i)dv_i - [h_i(y^0+v)\phi_i(v_i)]|_{-\infty}^{+\infty} \right).$$

Now, we can show that the last term on the right hand side is zero. In fact, for positive v_i we have

$$\phi_i(v_i) = \frac{1}{\sigma_i^2}\int_{v_i}^{+\infty} s p_i(s)ds = \frac{1}{\sigma_i^2}\int_{v_i}^{+\infty} \frac{s^2 p_i(s)}{s}ds \leq \frac{1}{\sigma_i^2 v_i}\int_{v_i}^{+\infty} s^2 p_i(s)ds = \frac{1}{v_i}.$$

For negative v_i, observing that $\int_{v_i}^{+\infty} s p_i(s)ds = -\int_{-\infty}^{v_i} s p_i(s)ds$ (recall that $E[v_i]=0$), a similar argument yields $|\phi_i(v_i)| \leq \frac{1}{|v_i|}$. In conclusions, we have that $|\phi_i(v_i)| \leq \frac{1}{|v_i|}$. Now, the sublinear growth of the estimator (iv) gives

$$\lim_{|v_i|\to\infty} |h_i(y^0+v)\phi_i(v_i)| \leq \lim_{|v_i|\to\infty} \frac{|h_i(y^0+v)|}{|v_i|} = 0.$$

Now, we have

$$\frac{\partial h_i}{\partial v_i}(y^0+v) = \frac{\partial h_i}{\partial y_i}(y^0+v),$$

so that

$$\int_{\mathbb{R}} h_i(y^0+v)v_i p_i(v_i)dv_i = \sigma_i^2 \int_{\mathbb{R}} \frac{\partial h_i}{\partial y_i}(y^0+v)\phi_i(v_i)dv_i.$$

Then,

$$\begin{aligned} op &= \frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i^2 \int_{\mathbb{R}^{\ell-1}} \prod_{j\neq i}^{\ell} p_j(v_j) \left(\int_{\mathbb{R}} \frac{\partial h_i}{\partial y_i}(y^0+v)\phi_i(v_i)dv_i \right) dv^{[-i]} \\ &= \frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i^2 \int_{\mathbb{R}^{\ell}} \frac{\partial h_i}{\partial y_i}(y) \prod_{j\neq i}^{\ell} p_j(v_j)\phi_i(v_i)dv. \end{aligned}$$

Finally, if the errors v_i are Gaussian,

$$\phi_i(v_i) = \frac{1}{\sigma_i^2}\int_{v_i}^{+\infty} \frac{s}{\sqrt{2\pi}\sigma_i} e^{-\frac{s^2}{2\sigma_i^2}} ds = -\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{s^2}{2\sigma_i^2}} \Big|_{v_i}^{+\infty} = p_i(v_i).$$

Therefore,

$$op = \frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i^2 \int_{\mathbb{R}^{\ell}} \frac{\partial h_i}{\partial y_i}(y) \prod_{j\neq i}^{\ell} p_j(v_j)\phi_i(v_i)dv = \frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i^2 E\left[\frac{\partial h_i}{\partial y_i}(y)\right]$$

thus proving the thesis.

Remark 4 *Although linear estimators do not fulfill assumption (iv), the thesis still holds. In fact, recalling that for a linear estimator the degrees of freedom do not depend on y, expression (16) is eventually recovered.*

Note that (18) was already known in the context of Stein's unbiased risk estimators (Stein, 1981). The next theorem derives a simple expression for the optimism of the SVR estimator in a somehow ideal case (see assumption *(v)* below).

Theorem 3 *Assume that*

(i) *eq. (15) holds,*

(ii) *the errors v_i are independent of each other,*

(iii) *the variances $\sigma_i^2 = Var[v_i]$ are finite,*

(iv) *$C < +\infty$,*

(v) *the set I_M of the marginal vectors does not depend on v.*

Then, the optimism of the SVR is

$$op^{SVR} = \frac{2}{\ell} \sum_{i \in I_M} \sigma_i^2.$$

Proof. The proof is based on Theorem 2, whose assumptions *(i)-(iii)* are obviously satisfied. Concerning assumption *(iv)*, consider a vector y and fix all its entries but the i-th one y_i. Then, there exists $\kappa_i > 0$ such that $i \in I_{out}$ whenever $|y_i| > \kappa_i$. Hence, for $|y_i|$ large enough, $|h_i(y)|$ is a finite constant so that assumption *(iv)* of Theorem 2 is satisfied.

Now, observe that, for SVR, the derivatives $\frac{\partial h_i}{\partial y_i}$ are all equal to either 0 or 1, see (14). In particular, $\frac{\partial h_i}{\partial y_i}$ is different from zero if and only if $i \in I_M$. Then, in view of assumption *(v)*,

$$op^{SVR} = \frac{2}{\ell} \sum_{i \in I_M} \sigma_i^2 \int_{\mathbb{R}^\ell} \prod_{\substack{j \in I \\ j \neq i}} p_j(v_j) \phi_i(v_i) dv = \frac{2}{\ell} \sum_{i \in I_M} \sigma_i^2 \int_{\mathbb{R}} \phi_i(v_i) dv_i$$

The thesis is proven by showing that the last integral equals one:

$$\begin{aligned}
\int_{\mathbb{R}} \phi_i(v_i) dv_i &= \frac{1}{\sigma_i^2} \int_{-\infty}^{+\infty} \int_{v_i}^{+\infty} s p_i(s) ds dv_i = \frac{1}{\sigma_i^2} \int_{-\infty}^{+\infty} \int_{1}^{+\infty} v_i^2 z p_i(z v_i) dz dv_i \\
&= \frac{1}{\sigma_i^2} \int_{1}^{+\infty} z \int_{-\infty}^{+\infty} v_i^2 p_i(z v_i) dv_i dz = \frac{1}{\sigma_i^2} \int_{1}^{+\infty} \frac{1}{z^2} \int_{-\infty}^{+\infty} w^2 p_i(w) dw dz \\
&= \int_{1}^{+\infty} \frac{dz}{z^2} = 1.
\end{aligned}$$

■

In practice, it is difficult to guarantee that assumption (v) is satisfied and, in general, it will not. Nevertheless, if the noise variances σ_i^2 are not too large, the result of Theorem 3 could still be used to approximate the true optimism, as shown in the simulated experiment of Section 5.3. For the sake of simplicity, let us consider the homoskedastic case $\sigma_i^2 = \sigma^2$, $\forall i$ and define:

$$\widehat{op}^{SVR} = \frac{2m\sigma^2}{\ell}.$$

This approximated optimism can be used to assess the in-sample error:

$$C_p^{SVR} = \overline{err} + \widehat{op}^{SVR}. \tag{19}$$

This last expression is in very close analogy with the linear case (17), provided that the model order q is replaced by the number m of marginal vectors. Formula (19) provides a further justification for the definition of approximate degrees of freedom given in Proposition 2.
Note that, in the Gaussian case, from Theorem 2 it follows that

$$op^{SVR} = \frac{2\sigma^2}{\ell} \sum_{i=1}^{\ell} E\left[\frac{\partial h_i}{\partial y_i}(y)\right] = \frac{2\sigma^2}{\ell} E\left[\#I_M\right].$$

Therefore, $\widehat{op}^{SVR}$ is an unbiased estimate of the true optimism op^{SVR}.

5. Numerical Examples

In this section the use of the C_p statistic for tuning the SVR parameters (ε, C) is illustrated by means of two numerical examples. Finally, a simulated experiment is used to assess the precision of the optimism estimate $\widehat{op}^{SVR}$ as a function of the noise variance. The SVR solution was obtained by a Finite Newton algorithm implemented in MatLab.

5.1 Simulated Data

The true function to be reconstructed is

$$f^0(x) = e^{\sin(8x)}, \qquad 0 \le x \le 1.$$

The training data (x_i, y_i), $i = 1, \ldots, \ell$, are generated as

$$y_i = y_i^0 + v_i,$$

$$y_i^0 = f^0(x_i),$$

where the errors $v_i \sim N(0, \sigma^2)$, $\sigma^2 = 0.09$, are independently distributed and

$$x_i = \frac{i-1}{\ell - 1},$$

with $\ell = 64$. In order to obtain a statistical assessment of the tuning procedure, $n = 100$ independent data sets were generated according to the above model. A cubic B-spline kernel was adopted:

$$K(x, x') = B_3(x - x').$$

The tuning of the parameters (ε, C) was carried out on a 30×30 equally spaced rectangular grid in the region

$$0.05 \le \varepsilon \le 0.5,$$

$$1 \le \log_{10} C \le 3.$$

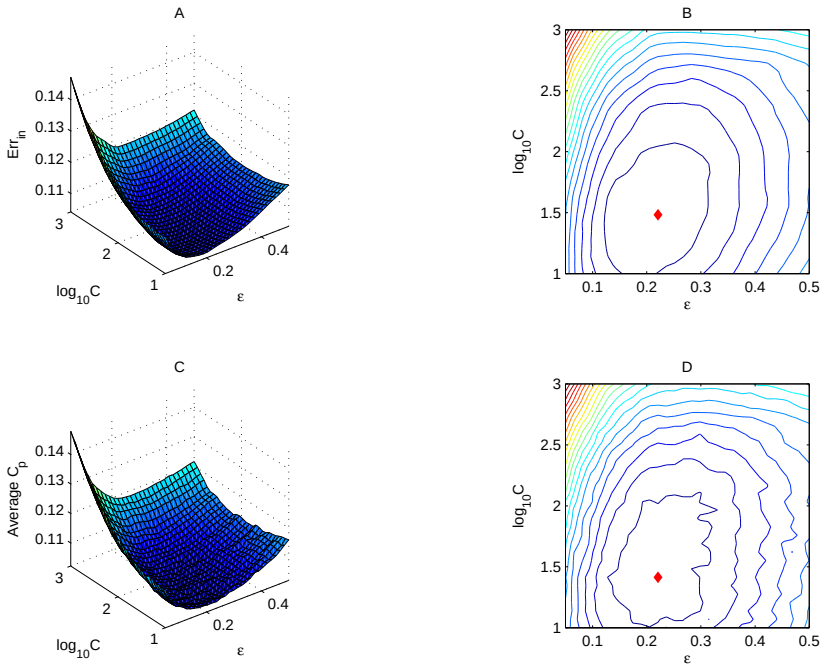

Figure 2: Estimated Err_{in} for the numerical example (Panels A and B) and average C_p over the 100 data sets (Panels C and D).

The choice of a logarithmically spaced C is in agreement with a common practice in Gaussian Processes and Tikhonov regularization methods, see for example De Nicolao et al. (1997) and De Nicolao et al. (2000).

First of all, the in-sample error $Err_{in}(\varepsilon, C)$ was computed as

$$Err_{in}(\varepsilon, C) \simeq \sigma^2 + \frac{1}{n\ell} \sum_{i=1}^{n} \|\hat{y}^{(i)}(\varepsilon, C) - y^0\|^2,$$

where $\hat{y}^{(i)}(\varepsilon, C)$ is the estimate of the vector y^0 obtained from the i-th data set. The function $Err_{in}(\varepsilon, C)$ is shown in Fig. 2. The optimal pair (ε^*, C^*) minimizing $Err_{in}(\varepsilon, C)$ is given by $\varepsilon^* = 0.22069$, $C^* = 30.392$, yielding $Err_{in}(\varepsilon^*, C^*) = 0.10413$.

In order to asses the average performance of the C_p statistic as an estimate of Err_{in}, the SVR estimate was calculated for each pair (ε, C) on the grid and for all the 100 data sets. In Fig. 2C the average C_p over the data sets is plotted against C and ε. The corresponding contour plot is shown in Fig. 2D. The minimal $C_p' = 0.10220$ is obtained in correspondence with $\varepsilon' = 0.22069$, $C' = 25.929$. From Fig. 2, it appears that, on the average, C_p provides a good estimate of Err_{in}. Moreover, $Err_{in}(\varepsilon', C') = 0.10436$ is reasonably close to the optimal $Err_{in}(\varepsilon^*, C^*) = 0.10413$.

In Fig. 3, C_p, Err_{in} and $E[\overline{err}]$ are plotted against C for $\varepsilon = 0.25$. The expected empirical risk $E[\overline{err}]$ was estimated by averaging over the 100 data sets. Then, the optimism op was estimated as the

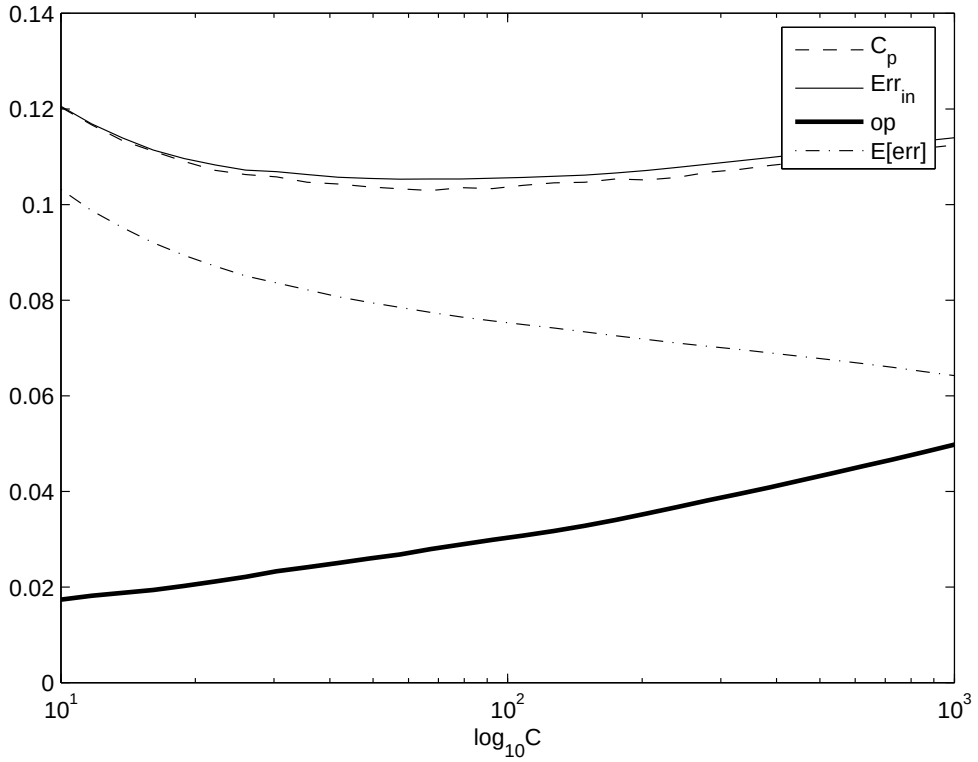

Figure 3: Decomposition of Err_{in} into the sum of $E[\overline{err}]$ and op as a function of C for $\varepsilon = 0.25$. Note that Err_{in} is well approximated by C_p.

difference between the estimates of Err_{in} and $E[\overline{err}]$. Also from this plot it is seen that C_p provides an accurate approximation of Err_{in}.

The real goal of a tuning procedure is obtaining a faithful reconstruction of the true function. A quantitative measure of the predictive performance on a single data set is given by the *RMSE* (Root Mean Square Error) defined as:

$$RMSE^{(i)}(\varepsilon, C) = \frac{1}{\sqrt{n}} \|\hat{y}^{(i)}(\varepsilon, C) - y^0\|.$$

For each of the 100 data sets, the function $f^0(x)$ was estimated using the pair $(\varepsilon^{(i)}, C^{(i)})$ minimizing the C_p statistic for the *i*-th data set. The average of such estimated functions is plotted in Fig. 4A where also the true function $f^0(x)$ is reported for comparison. In order to visualize the variability of the estimates, pointwise ± 2 standard deviations bands are plotted.

For the *i*-th data set the best possible tuning is

$$(\overline{\varepsilon}^{(i)}, \overline{C}^{(i)}) = \arg\min_{\varepsilon, C} RMSE^{(i)}(\varepsilon, C).$$

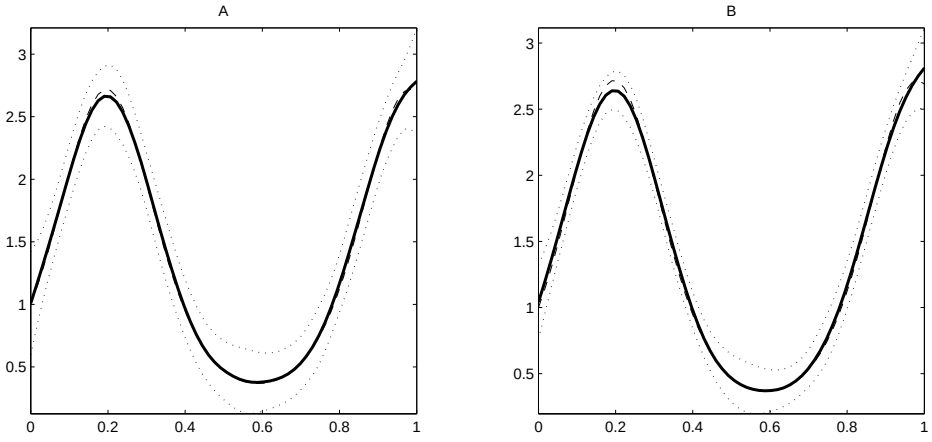

Figure 4: Average of the estimated functions over the 100 data sets: average (thick continuous) and true function (dashed). The results have been obtained by SVR with C_p tuning (Panel A) and SVR with best possible tuning (Panel B). In both cases, the $\pm$ 2 standard deviation bands are reported (dotted black).

Obviously, this cannot be used in practice because y^0 is unknown. Nevertheless, this ideal tuning is interesting because it gives a lower bound on the best achievable performance.

For each of the 100 data sets, the function $f^0(x)$ was estimated using the ideal tuning $(\varepsilon^{(i)}, C^{(i)})$. The average of such estimated functions with pointwise ± 2 SD bands is plotted in Fig. 4B. The comparison with Panel A of the same figure demonstrates that the predictive performance of the C_p tuning scheme is very close to the best achievable performance.

In Fig. 5 the histogram of $RMSE^{(i)}(\varepsilon^{(i)}, C^{(i)})$ (Panel A) is compared with the histogram of the best achievable errors $RMSE^{(i)}(\overline{\varepsilon}^{(i)}, \overline{C}^{(i)})$ (Panel B). Finally, the application of the C_p tuning scheme is illustrated on the first data set. The value of C_p as a function of ε and C is reported in Fig. 6 A-B. On the considered grid, the C_p statistic is minimized by $\varepsilon^{(1)} \simeq 0.28$, $C^{(1)} \simeq 30.4$, yielding $C_p(\varepsilon^{(1)}, C^{(1)}) \simeq 0.103418$. For the sake of comparison, in Fig. 6 C-D the plot of $RMSE^{(1)}(\varepsilon, C)$ is given. The best possible tuning for data set #1 is $\overline{\varepsilon}^{(1)} \simeq 0.28$, $\overline{C}^{(1)} \simeq 25.93$, yielding $RMSE^{(1)}(\overline{\varepsilon}^{(1)}, \overline{C}^{(1)}) \simeq 0.095357$. Using the C_p tuning scheme, a very similar value is obtained: $RMSE^{(1)}(\varepsilon^{(1)}, C^{(1)}) \simeq 0.095727$.

The SVR estimate corresponding to $(\varepsilon^{(1)}, C^{(1)})$ is plotted in Fig. 7 together with the true function $f^0(x)$. The SVR estimate corresponding to the best possible tuning $(\overline{\varepsilon}^{(1)}, \overline{C}^{(1)})$ is plotted for

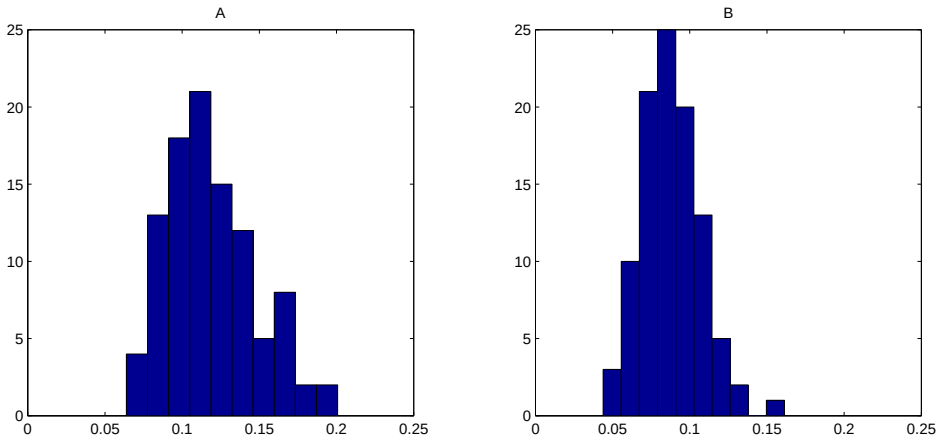

Figure 5: Distribution of the *RMSE* over the 100 data sets using C_p tuning (Panel A) and the best possible tuning (Panel B).

comparison. Taking into account the signal-to-noise ratio of the data it can be concluded that the C_p tuning scheme performs more than satisfactorily.

5.2 Boston Housing Data

To show the effectiveness on real-world data of the tuning procedure based on the C_p statistic, we applied it to the Boston Housing data set from the UCI Repository. The data set consists of 516 instances with 12 input variables (including a binary one) and an output variable representing the median housing values in suburbs of Boston.

The input variables were shifted and scaled to the unit hypercube, while the output variable was first shifted to have zero mean and then scaled to fit into the interval $[-1,1]$. More precisely, letting $m_j = \min_i x_{i,j}$ and $M_j = \max_i x_{i,j}$, the inputs $x_{i,j}$ were transformed into $(x_{i,j} - m_j)/(M_j - m_j)$, while the outputs y_i were transformed into $(y_i - \bar{y}) / \max_i |y_i - \bar{y}|$, where $\bar{y}$ denotes the sample mean.

The data set was randomly split into two parts: 450 instances to be used for training and 56 for testing. Pairs (ε, C) over a 20×20 uniform grid were considered with

$$0 \leq \varepsilon \leq 0.3, \qquad 0 \leq \log_{10} C \leq 4.$$

For each pair (ε, C), the SVR fit solving (1)-(2) was evaluated using a Gaussian RBF kernel with fixed bandwidth ($2\sigma^2_{kernel} = 3.9$). An estimate of the noise variance σ^2 was obtained from

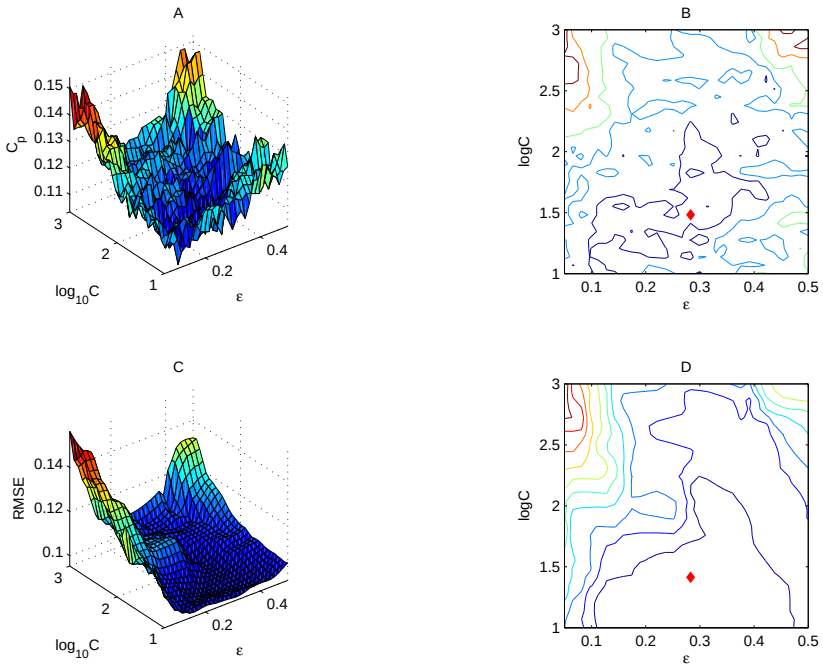

Figure 6: Data set #1: the statistic C_p as a function of ε and C (Panels A and B) and the *RMSE* between the estimate and the true function (Panels C and D).

the residuals generated from a low-bias linear regression. For details on this procedure, see for example Hastie and Tibshirani (1990), page 48, and Loader (1999), page 160. In particular, we used regularized least squares with polynomial kernel of degree 2. The noise variance of the data set was estimated as $\hat{\sigma}^2 = 0.01$ using the estimator

$$\hat{\sigma}^2 = \frac{SSR^L}{\ell - 2\nu_1 + \nu_2},$$

where SSR^L is the sum of squared residuals using the linear estimator, $\nu_1 = tr(H)$, $\nu_2 = tr(H^T H)$, and H is the "hat matrix" of the linear estimator (that is the matrix such that $\hat{\mathbf{y}} = H\mathbf{y}$). Then, the following quantities were evaluated:

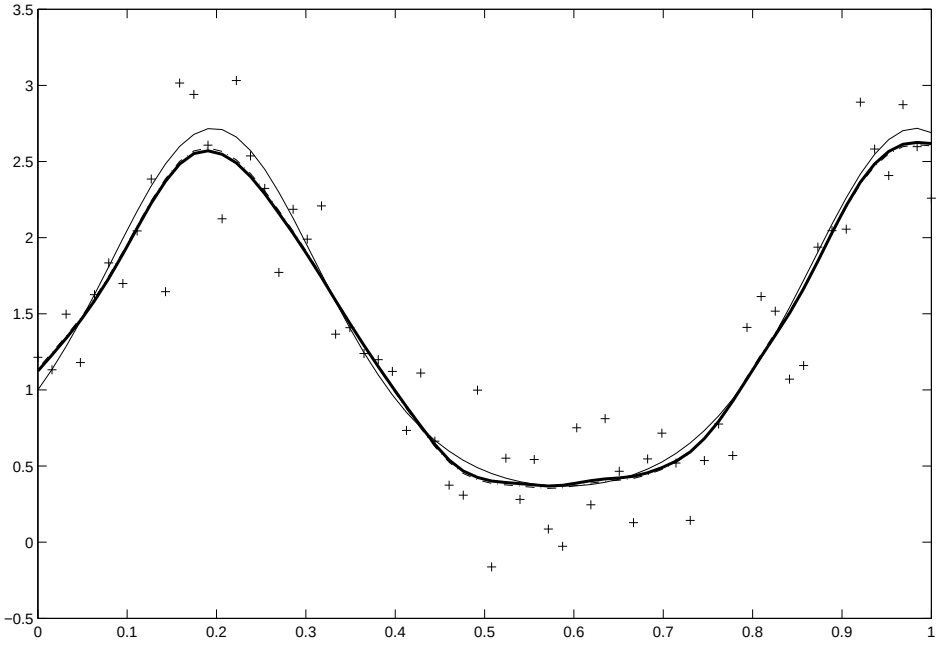

Figure 7: Data set #1: True function (continuous), data (crosses), SVR estimate with C_p tuning (thick continuous) and SVR with best possible tuning (dash-dot).

$$
\begin{aligned}
\overline{err} &= \frac{1}{\ell}\sum_{i=1}^{\ell}\left(y_i - \hat{f}(x_i)\right)^2, \\
\widehat{op}^{SVR} &= \frac{2\hat{\sigma}^2 m}{\ell}, \\
C_p^{SVR} &= \overline{err} + \widehat{op}^{SVR}, \\
GCV^{SVR} &= \frac{\ell^2 \overline{err}}{(\ell - m)^2}.
\end{aligned}
$$

The score GCV^{SVR} was recently proposed as a tuning criterion by Gunter and Zhu (2007). These quantities are plotted in Fig. 8 and Fig. 9 together with the 5-fold cross-validation score whose computation is much heavier and the (quadratic) test error. In the contour plots of Fig. 9, the position of the minimizers are also showed. It can be seen that C_p and *GCV* pick the same value of ε and C. On the considered grid, the minimum value of the test error is 0.01628. The model selected by C_p and *GCV* achieves a test error equal to 0.01670, while the model selected by 5-fold cross-validation achieves 0.01742.

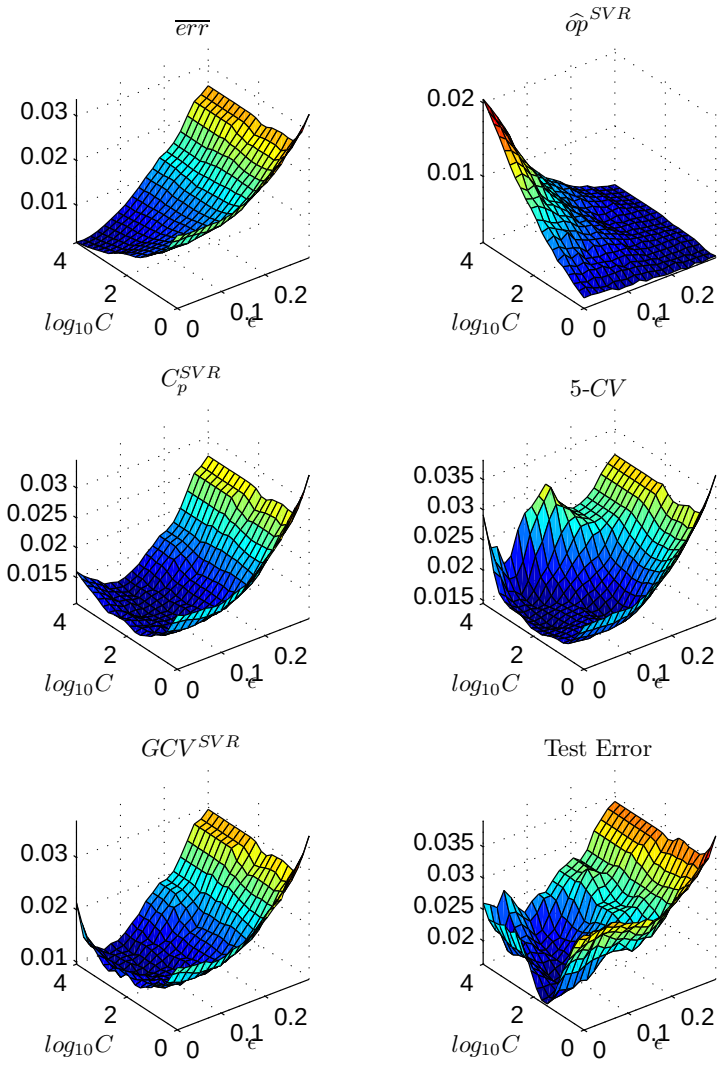

Figure 8: Boston Housing data: empirical risk ($\overline{err}$), optimism estimate ($\widehat{op}^{SVR}$), C_p statistic (C_p^{SVR}), 5-fold cross-validation score (5-CV), Generalized Cross Validation score (GCV^{SVR}), and mean square error on test data (Test Error).

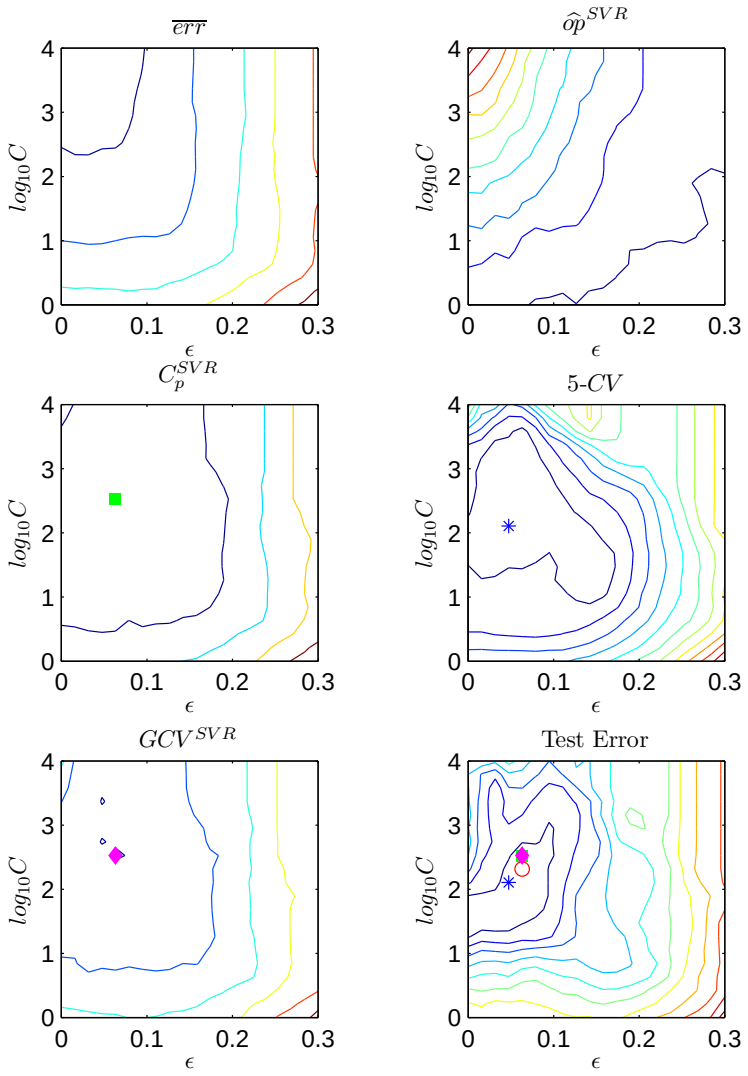

Figure 9: Boston Housing data: contour plots of empirical risk ($\overline{err}$), optimism estimate ($\widehat{op}^{SVR}$), C_p statistic (C_p^{SVR}), 5-fold cross-validation score (5-CV), Generalized Cross Validation score (GCV^{SVR}), and mean square error on test data (Test Error). In the plots of C_p^{SVR}, 5-CV and GCV^{SVR} the minimizer position is marked. In the test error plot, the marks of all minimizers are reported.

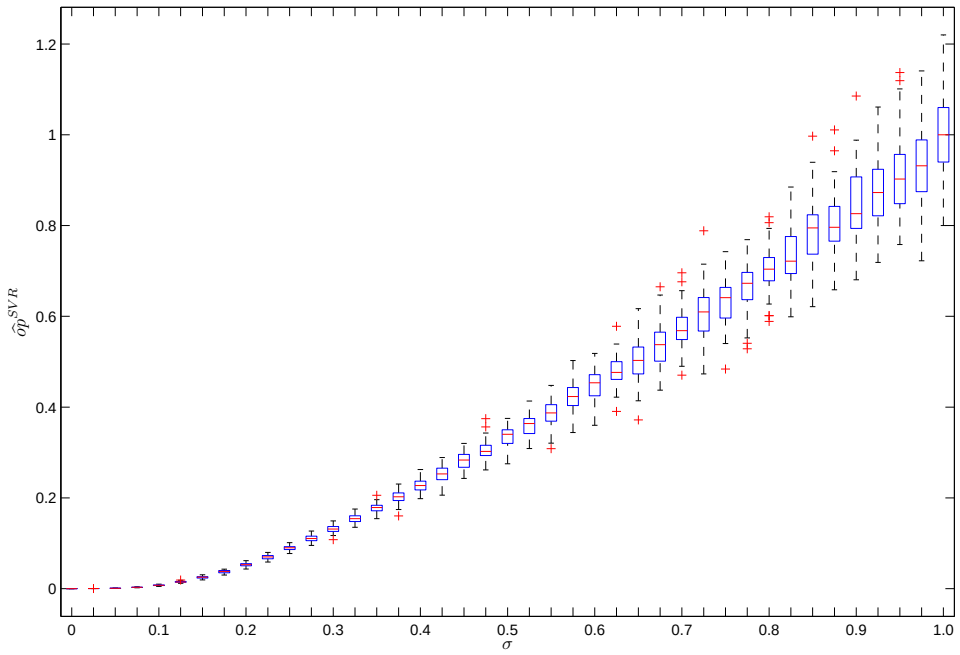

Figure 10: Simulated experiment: boxplots of the estimated optimism $\widehat{op}^{SVR}$ against different values of the noise standard deviation.

5.3 Dependence of $\widehat{op}^{SVR}$ on the Noise Variance

In order to investigate the dependence of the variability of $\widehat{op}^{SVR}$ on the noise variance, we ran a simulated experiment. Specifically, we considered 41 standard deviations in the interval $[0,1]$:

$$\sigma_j = \frac{j-1}{40}, \qquad j = 1,\ldots,41.$$

Next, for each σ_j, 100 independent data sets were generated according to the model

$$\begin{aligned} y_i &= \operatorname{sinc}(3x_i) + v_i, \qquad v_i \sim N(0,\sigma_j^2), \\ x_i &= \frac{2i-101}{99}, \qquad i = 1,\ldots,100. \end{aligned}$$

For each data set, the SVR was computed using the kernel

$$K(x,x') = e^{-\frac{|x-x'|}{4}}$$

with the values of C and ε fixed to $C = 100$, $\varepsilon = 0.1$. For each data set, we evaluated $\widehat{op}^{SVR} = 2\hat{\sigma}^2 m/\ell$. In Fig. 10 the boxplots of $\widehat{op}^{SVR}$ are reported against the considered noise standard devi-

ations (the "+" marks denote the outliers). As expected, both the mean and the variance of $\widehat{op}^{SVR}$ increase with the noise variance.
Since, under Gaussian noise, $\widehat{op}^{SVR}$ is an unbiased estimator, the true optimism op^{SVR}, which is not reported in the plot, coincides with the expected value of $\widehat{op}^{SVR}$. From Fig. 10 it appears that there is a whole range of signal-to-noise ratios such that $\widehat{op}^{SVR}$ estimates op^{SVR} with good precision.

6. Concluding Remarks

In this paper, a novel formulation of the quantitative representer theorem is derived for convex loss functions. More precisely, using the newly introduced notion of pseudoresidual the inclusions appearing in the previous formulations are replaced by equations. This result is exploited in order to study the sensitivity of both the SVR coefficients and predictions with respect to the data. In view of the sensitivity analysis, the degrees of freedom of SVR are defined as the number of marginal support vectors. Such a definition is further justified by the role that the degrees of freedom play in the assessment of the optimism, that is the difference between the in-sample prediction error and the expected empirical risk. A C_p statistic for SVR is defined and proposed as a criterion for tuning both the parameters ε and C. The performance observed on both a simulated benchmark and a real world problem appears more than satisfactory. Among the future developments one may mention the extension of the results of the present paper to kernel based classifiers.

Acknowledgments

This research has been partially supported by the Italian Ministry of University and Research through the FIRB Project "Learning theory and application" and the PRIN Project "New methods and algorithms for identification and adaptive control of technological systems".

Appendix A.

In this appendix, a Fourier series demonstration of the representer theorem is provided. The rationale is inspired by Evgeniou et al. (2000) who prove the representer theorem for differentiable loss functions. Let us assume that the function $K(\mathbf{x},\mathbf{t})$ is such that the bilinear formula holds:

$$K(\mathbf{x},\mathbf{t}) = \sum_{n=1}^{+\infty} \lambda_n \phi_n(\mathbf{x})\phi_n(\mathbf{t}),$$

where $\forall n$, $\lambda_n > 0$, and ϕ_n denote the n-th eigenvalue and eigenfunction of the operator

$$Tf = (f, K(\mathbf{x},\mathbf{t}))_{\mathcal{L}^2}.$$

Then, K is positive definite and a generic function $f \in \mathcal{L}^2$ admits the Fourier expansion

$$f(\mathbf{x}) = \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}). \tag{20}$$

Now, we can build the RKHS $\mathcal{H}$ taking all the functions f such that $\sum_{n=1}^{+\infty} \frac{c_n^2}{\lambda_n}$ is finite and defining the inner product between the two functions $u, v \in \mathcal{H}$, $u = \sum_{n=1}^{+\infty} a_n \phi_n$, $v = \sum_{n=1}^{+\infty} b_n \phi_n$ as

$$(u,v)_{\mathcal{H}} = \sum_{n=1}^{+\infty} \frac{a_n b_n}{\lambda_n},$$

so that the norm is

$$\|f\|_{\mathcal{H}}^2 = \sum_{n=1}^{+\infty} \frac{c_n^2}{\lambda_n}.$$

It is easy to check that the reproducing property holds

$$f(\mathbf{x}) = (f(\mathbf{t}), K(\mathbf{x},\mathbf{t}))_{\mathcal{H}},$$

so that $K(\mathbf{x},\mathbf{y})$ is indeed the reproducing kernel of $\mathcal{H}$. In particular, the reproducing property implies that the series (20) is, in fact, pointwise convergent.
In view of this, solving (1) is equivalent to minimizing the following functional with respect to the coefficient sequence:

$$F[\{c_n\}] = C\sum_{i=1}^{\ell} V\left(y_i, \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i)\right) + \frac{1}{2}\sum_{n=1}^{+\infty} \frac{c_n^2}{\lambda_n}.$$

Noting that the sequence $\{c_n\}$ belongs to $\ell^2(\mathbb{R})$, we can see that the functional F to be minimized maps a subset of ℓ^2 into $\mathbb{R}$. From the necessary condition for optimality, we have $0 \in \partial F$, where ∂F denotes the subdifferential. Exploiting the linearity of the subdifferential with respect to sums of convex functions and the fact that the second term is Gâteaux-differentiable, we obtain:

$$\partial F = \left\{ C\sum_{i=1}^{\ell} \partial V\left(y_i, \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i)\right) + \frac{c_n}{\lambda_n} \right\}.$$

Now, let us recall the following result (see Prop. 5.7 of Ekeland and Temam (1974), where it is given for the more general case of topological vector spaces):

Proposition 4 *Let $\mathcal{H}$, $\mathcal{H}'$ two Banach spaces, V a convex function from $\mathcal{H}$ into $\mathbb{R}\cup\{+\infty\}$, and J a continuous linear operator from $\mathcal{H}'$ into $\mathcal{H}$. Assume that there is $v_0' \in \mathcal{H}'$ such that V is continuous and finite at Jv_0'. Then, for all $v' \in \mathcal{H}'$*

$$(\partial V \circ J)(v') = J^*(\partial V)(Jv'),$$

where $J^* : \mathcal{H} \to \mathcal{H}'$ is the adjoint defined by

$$\left\langle v', J^* v \right\rangle_{\mathcal{H}'} = \left\langle Jv', v \right\rangle_{\mathcal{H}}$$

for all $v \in \mathcal{H}$ and $v' \in \mathcal{H}'$, where $<\cdot,\cdot>_{\mathcal{H}}$ stands for the duality pairing in the Banach space $\mathcal{H}$. ■

Introducing the linear operators $J_i : \ell^2 \to \mathbb{R}$

$$J_i(\{c_n\}) = \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i),$$

we can write

$$\partial V\left(y_i, \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i)\right) = \partial V\left(y_i, J_i(\{c_n\})\right).$$

Notice that the adjoint $J_i^* : \mathbb{R} \to \ell^2(\mathbb{R})$ is given by $J_i^*(t) = \{t\phi_n(\mathbf{x}_i)\}$. In fact,

$$\langle \{c_n\}, J_i^*(t)\rangle_{\ell^2} = \sum_{n=1}^{+\infty} c_n (J_i^*(t))_n = t \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i) = \langle J_i(\{c_n\}), t\rangle_{\mathbb{R}}.$$

In view of Proposition 4,

$$\partial F = \left\{ C \sum_{i=1}^{\ell} \phi_n(\mathbf{x}_i) \partial_2 V\left(y_i, \sum_{n=1}^{+\infty} c_n \phi_n(\mathbf{x}_i)\right) + \frac{c_n}{\lambda_n} \right\}$$

so that the condition $0 \in \partial F$ implies that the optimal sequence $\{\hat{c}_n\}$ must satisfy

$$\hat{c}_n \in -\sum_{i=1}^{\ell} C \partial_2 V\left(y_i, \hat{f}(\mathbf{x}_i)\right) \lambda_n \phi_n(\mathbf{x}_i).$$

It is then possible to write

$$\hat{c}_n = \sum_{i=1}^{\ell} a_i \lambda_n \phi_n(\mathbf{x}_i),$$

where

$$a_i \in -C \partial_2 V(y_i, \hat{f}(\mathbf{x}_i)).$$

Finally, exploiting the bilinear formula for the reproducing kernel, we obtain

$$\hat{f}(\mathbf{x}) = \sum_{n=1}^{+\infty} \hat{c}_n \phi_n(\mathbf{x}) = \sum_{i=1}^{\ell} a_i \sum_{n=1}^{+\infty} \lambda_n \phi_n(\mathbf{x}_i) \phi_n(\mathbf{x}) = \sum_{i=1}^{\ell} a_i K(\mathbf{x}_i, \mathbf{x}).$$

References

J. M. Borwein and A. J. Lewis. *Convex Analisys and Nonlinear Optimization*. Springer, 2000.

M. W. Chang and C. J. Lin. Leave-one-out bounds for support vector regression model selection. *Neural Computation*, 17:1188–1222, 2005.

V. Cherkassky and Y. Ma. Comparison of model selection for regression. *Neural Computation*, 15: 1691–1714, 2003.

D. Cox and F. O' Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann.Stat.*, 18:1676–1695, 1990.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49, 2001.

G. De Nicolao, G. Sparacino, and C. Cobelli. Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica*, 33:851–870, 1997.

G. De Nicolao, G. Ferrari Trecate, and G. Sparacino. Fast spline smoothing via spectral factorization concepts. *Automatica*, 36:1733–1739, 2000.

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

I. Ekeland and R. Temam. *Analyse Convexe et Problémes Variationnels*. Gauthier-Villards, Paris, 1974.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–150, 2000.

J. B. Gao, S. R. Gunn, C. J. Harris, and M. Brown. A probabilistic framework for SVM regression and error bar estimation. *Machine Learning*, 46:71–89, 2002.

L. Gunter and J. Zhu. Efficient computation and model selection for the support vector regression. *Neural Computation*, 19:1633–1655, 2007.

T. J. Hastie and R. J. Tibshirani. Generalized additive models. In *Monographs on Statistics and Applied Probability*, volume 43. Chapman and Hall, London, UK, 1990.

T. J. Hastie, R. J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, Canada, 2001.

T. J. Hastie, R. J. Tibshirani, and J. Friedman. Note on "Comparison of Model Selection for Regression" by Vladimir Cherkassky and Yunqian Ma. *Neural Computation*, 15:1477–1480, 2003.

T. J. Hastie, S. Rosset, R. J. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41:495–502, 1979.

J. T. Kwok and I. W. Tsang. Linear dependency between ε and the input noise in ε-support vector regression. *IEEE Transactions On Neural Networks*, XX, 2003.

C. Loader. *Local Regression and Likehood*. Statistics and Computing. Springer, 1999.

T. Poggio and F. Girosi. A theory of networks for approximation and learning. *Foundation of Neural Networks*, page 91–106, 1992.

M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10:955–974, 1998.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Neural Networks and Computational Learning Theory*, 81:416–426, 2001.

A. J. Smola, N. Murata, B. Schölkopf, and K. Muller. Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Boden, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, Perspectives in Neural Computing*, pages 105–110, Berlin, 1998. Springer.

C. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9: 1135–1151, 1981.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems*. W. H. Winston, Washington, D. C., 1977.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, USA, 1995.

G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, USA, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.