SVRs and Uncertainty Estimates in Wind Energy Prediction

Jesús Prada and José R. Dorronsoro

Universidad Autónoma de Madrid, Cantoblanco, Madrid, Spain
{jesus.prada@estudiante,jose.dorronsoro}@uam.es

Abstract. While Support Vector Regression, SVR, is one of the algorithms of choice in modeling problems, construction of its error intervals seems to have received less attention. On the other hand, general noise cost functions for SVR have been recently proposed. Taking this into account, this paper describes a direct approach to build error intervals for different choices of residual distributions. We also discuss how to fit these noise models and estimate their parameters, proceeding then to give a comparison between intervals obtained using this method. under different ways to estimate SVR parameters as well as the intervals obtained by employing a full SVR Bayesian framework. The proposed approach is shown on a synthetic problem to provide better accuracy when models fitted coincide with the noise injected into the problem. Finally, we apply it to wind energy forecasting, exploiting predicted energy magnitudes to define intervals with different widths.

1 Introduction

Support vector regression, SVR, [1] has been widely used in regression problems such as stock market [2], wind energy [3] or solar radiation [4] forecasting. Classical SVR, however, does not give probability intervals to address the uncertainty in the predictions and, in fact, error interval estimation for SVR has received a somewhat limited attention in the literature. Notice that here approaches such as the well known ones for linear regression under Gaussian models completely break down, not only by the difficulty of ensuring normal random variables but, above all, by the fact that the familiar analytic estimates of the linear coefficients are simply impossible in SVR and, of course, less so, any asymptotic analysis.

In [5], a Bayesian interpretation of SVR is described and then used to propose methods to determine, first, SVR parameters by maximizing an evidence function and, second, to derive probability intervals for predictions. A drawback of these methods is that they modify the classical SVR formulation of the problem to solve, and hence existing SVR software, such as the popular LIBSVM [6] cannot be used, at least without modifying it first.

A more direct approach is proposed in [7], which assumes prediction errors to follow a specific probability distribution that, in turn, is used to define the probability intervals. Zero-mean Gaussian and Laplace families are proposed in [7] as

noise models and fitted by maximum likelihood estimation, MLE, using out-of-sample residuals of SVR models; optimal SVR parameters are obtained simply by cross validation. In this paper, we follow this methodology to give probability intervals under the assumption of both zero-mean Laplace and Gaussian distributions, as well as for their non–zero mean counterparts plus the Beta and Weibull distributions. A difficulty with this approach is that it assumes that the residual distribution is independent of the predicted value and, therefore, probability intervals have exactly the same width for all input instances. General error models for SVR other than the well known ϵ–insensitive loss have been proposed in [8]. This suggest that noise distribution might be different across particular problems and it should be reflected in the particular SVR model to be used. If the assumption is true and the underlying noise distribution is accurately estimated, one should expect a reduction in interval prediction errors. We study if the proposed method can estimate this noise distribution.

Our main contributions can be summarized as follows:

- We enlarge, as mentioned, the noise models considered in [7].
- We discuss Newton–Rapshon maximum likelihood estimates for the Beta and, particularly, Weibull distributions, as well as the definition of uncertainty intervals for them.
- We show on a synthetic problem how the proposed approach is able to pair the models fitted to the residuals with the specific noise injected on the problem targets. We also compare these results to the ones obtained by a statistic test for distribution hypotheses.
- We apply the methods proposed to the estimation of uncertainty intervals for the wind energy prediction of peninsular Spain, where we also consider the use of different intervals according to predicted energy magnitudes, showing how a two group data split results in more accurate intervals.

The rest of this paper is organized as follows. Section 2 briefly reviews both classical and Bayesian SVR formulations. In Section 3 there is an in-depth description of the proposed approach for error interval estimations and experiments are carried in Section 4, where we also consider four publicly available regression datasets besides the already mentioned synthetic and wind energy problems. The paper ends with a short section on conclusions and pointers for further work.

2 Support Vector Regression

2.1 Classical SVR Review

Given a sample $D = \{(x_i, y_i) : 1 \leq i \leq N\}$ of inputs $x_i \in R^n$ and targets $y_i \in R$, the SVR problem is that of minimizing the loss function $L(w, b, \xi)$ defined as

$$L(w, b, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*), \tag{1}$$

over w, b and ξ subject to the constraints $-\xi_i - \epsilon \leq w \cdot x_i + b - y_i \leq \xi_i^* + \epsilon,\ \ \xi_i, \xi_i^* \geq 0$ .This is known as the SVR primal problem. It can also be seen as a variant of the standard L_2 regularized regression where instead of the familiar $z_i^2 = (y_i - w \cdot x_i - b)^2$ square error, we use the ϵ–insensitive loss function [9] $l_\epsilon(\delta) = [\delta]_\epsilon = \max(0, |\delta| - \epsilon)$. i.e., we allow an ϵ-wide, penalty-free "error tube" around the model function $w \cdot x + b$. To solve the primal problem, it is transformed [1] using standard Lagrangian analysis into the so–called dual problem; moreover, instead of working with a simple linear model $w \cdot x + b$, the well known kernel trick is used to arrive to a non–linear model $w \cdot \Phi(x) + b$ where $\Phi(x)$ is a high (and possibly infinite) dimensional projection of the original x. We thus obtain the final model as

$$f(x) = b^* + w^* \cdot \Phi(x) = b^* + \sum \gamma_i^* \Phi(x_i) \cdot \Phi(x) = b^* + \sum_i \gamma_i^* k(x_i, x) \ . \quad (2)$$

We take the usual choice of a Gaussian kernel $k(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$. It is clear that model performance will be highly dependent on the choices of the C, ϵ, γ parameters which, in turn, will affect the residuals $\delta_i = y_i - f(x_i)$ whose distribution we want to estimate. One way to derive optimal C, ϵ, γ values is by standard cross–validation, CV, or validation over a fixed set. A perhaps more principled alternative could be to follow a Bayesian framework to derive their optimal values. We briefly review this approach next.

2.2 A Bayesian Framework for SVR

We assume the x_i, y_i in the sample D to be related through a model $y_i = f(x_i) + \delta_i$ where the δ_i follow random i.i.d. values. In the Bayesian approach of [5], $f = [f(x_1), f(x_2), ..., f(x_N)]$ is taken as the realization of a random field with a known prior probability. The posterior probability of f given D can then be derived by Bayes' theorem as

$$P(f|D) = \frac{P(D|f)P(f)}{P(D)} \propto P(D|f)P(f), \quad (3)$$

where the conditional probability of D given f is $P(D|f) = \prod_{i=1}^N P(y_i - f(x_i)) = \prod_{i=1}^N P(\delta_i)$ and $P(\delta_i)$ is often assumed to be of the exponential form, i.e., $P(\delta_i) \propto \exp\left(-C \cdot l(\delta_i)\right)$ with $C > 0$ a normalizing constant and l a certain loss function. Putting this together, we arrive at

$$P(D|f) = \prod_{i=1}^N P(\delta_i) \propto \exp\left(-C \cdot \sum_{i=1}^N l(\delta_i)\right) = \exp\left(-C \cdot \sum_{i=1}^N l(y_i - f(x_i))\right) . \quad (4)$$

going back to the prior $P(f)$, in [5] it is assumed to be of the form

$$P(f) = (2\pi)^{\frac{N}{2}} |K|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} f^T K^{-1} f\right) , \quad (5)$$

and also the following structure for $K_{i,j}$

$$K_{i,j} = K_{x_i,x_j} = cov[f(x_i), f(x_j)] = \kappa_0 \exp\left(-\frac{\kappa}{2}||x_i - x_j||^2\right) + \kappa_b, \qquad (6)$$

with κ_0, κ and κ_b appropriate positive constants. The parameters in the prior and the likelihood are $\theta = (C, \epsilon, \kappa, \kappa_0, \kappa_b)$. κ_0 is fixed in [5] as the target variance. The optimal values of the other hyperparameters can be inferred by maximizing

$$P(\theta|D) \propto P(D|\theta)P(\theta) \ . \qquad (7)$$

A common assumption now is that $P(\theta)$ is rather insensitive to the values of θ and can thus be ignored.

An optimal θ^* can then derived by maximizing the evidence function $P(D|\theta)$ for which, by Jensen's inequality, we have

$$-\log P(D|\theta) = -\log \int P(D|f,\theta)P(f|\theta)df \leq \int \left[-\log P(D|f,\theta) - \log P(f|\theta)\right] df,$$

and approximating the integral by the sample average we arrive to the problem

$$\min_{\theta} \left\{ C \ \sum_{i=1}^{N} l_{\epsilon,\beta}(y_i - f(x_i)) + \frac{1}{2} f^T K^{-1} f \right\} \ . \qquad (8)$$

[5] derives in this way analytic approximations to $-\log P(D|\theta)$. Moreover, it proposes to replace the standard non–differentiable ϵ–insensitive SVR loss by a smooth version, the so–called soft insensitive loss function, SILF.

Turning our attention to probability intervals, as stated in [7] we have

$$1 - s = \int_{-\infty}^{p_s} p_\zeta(z)dz = \int_{-\infty}^{+\infty} \int_{-\infty}^{p_s - f} p(\delta)d\delta p_{f|D}(f)df, \qquad (9)$$

where

$$p(\delta) = \frac{\exp\left(-C \cdot l_{\beta,\epsilon}(\delta)\right)}{\int \exp\left(-C \cdot l_{\beta,\epsilon}(\delta)\right)d\delta}, \ p(f(x)|D) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{(f(x) - \hat{f})^2}{2\sigma_t^2}\right), \quad (10)$$

where $\sigma_t^2 = K(x,x) - K_{F,x}^T K_{F,F}^{-1} K_{F,x}$ and $K_{F,x}$ the vector containing all $K(x_i, x)$, with $i \in F = \{i | 0 < \alpha_i < C \ or \ 0 < \alpha_i^* < C\}$ and $K_{F,F}$ the corresponding submatrix.

3 Uncertainty Estimates

A natural way to address uncertainty estimates is by assuming that the residual distribution follows some parametric model $p(\delta; \Theta)$ and to use the sample $\Delta = \{\delta_i\}$ to derive a maximum–likelihood (ML) estimate for Θ from the likelihood function $L(\Theta; \Delta) = \prod_1^N p(\delta_i; \Theta)$. This approach, in which the uncertainty

estimates will be independent of the patterns x, is followed in [7] using zero–mean Laplace and Gaussian models. In this work we extend this method to other possible residual models, namely the non–zero mean Gaussian and Laplace distributions (to address possible error biases), and the Beta and Weibull distributions. A drawback of the approach in [7] is the independence between error intervals and patterns; to alleviate this we also propose to split the uncertainty analysis according to the $\hat{f}(x)$ values.While other choices would be possible, we select these distributions because previous works, such as [10], have shown their usefulness to model wind speed and wind power production. We review next their main properties.

3.1 Density Functions

Recall that the general Laplace density with mean μ and standard deviation σ is $p(z) = \frac{1}{2\sigma} e^{-\frac{|z-\mu|}{\sigma}}$ with $z \in R$ and the μ, σ Gaussian is $p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$. The Beta$(\alpha, \beta)$ distribution is given for $x \in [0, 1]$ by

$$p(z) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \alpha, \beta > 0, \tag{11}$$

where $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and $\Gamma(x)$ are the beta and gamma functions respectively. Finally, the Weibull(λ, k) distribution[1] is given by

$$p(z) = \frac{\kappa}{\lambda} \left(\frac{z}{\lambda}\right)^{\kappa-1} e^{-\left(\frac{z}{\lambda}\right)^{\kappa}}, \ z \in [0, \infty), \lambda > 0, \kappa > 0 \ . \tag{12}$$

3.2 Log–likelihood Parameter Estimation

Given an independent sample δ_i to which we want to fit one of the above distributions, we can estimate its parameters Θ by maximizing the log-likelihood, $l(\Theta; \delta_1 ... \delta_n) = \sum_{i=1}^{n} \log p(\delta_i|\Theta)$, where p represents the density function, which we do by solving $\nabla_\Theta l = 0$. For the general Gaussian and Laplace cases we obtain the well known values

$$\hat{\mu}_G = \sum_{i=1}^{n} \frac{\delta_i}{n}, \ \hat{\sigma}_G = \frac{\sum_{i=1}^{n} (\delta_i - \hat{\mu}_G)^2}{n}, \ \ \hat{\mu}_L = m_{\delta_i}, \ \hat{\sigma}_L = \frac{\sum_{i=1}^{n} |\delta_i - \hat{\mu}_L|}{n} \ ; \tag{13}$$

here m_{δ_i} denotes the median of the δ_i residuals [17]. For the Beta distribution, the gradient equations are

$$0 = \frac{\partial l}{\partial \hat{\alpha}} = n \left(\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right) + \sum_{i=1}^{n} \log \delta_i, \tag{14}$$

$$0 = \frac{\partial l}{\partial \hat{\beta}} = n \left(\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\beta)}{\Gamma(\beta)} \right) + \sum_{i=1}^{n} \log (1-\delta_i), \tag{15}$$

[1] We only consider the case $z \geq 0$ in the Weibull distribution as $p(z) = 0$ for $z < 0$

where the δ values are scaled into $[0,1]$. Denoting $\phi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$, we have

$$\begin{aligned} \frac{\sum_{i=1}^{n} \log \delta_i}{n} &= \phi(\alpha) - \phi(\alpha+\beta) = F_1(\alpha,\beta), \\ \frac{\sum_{i=1}^{n} \log (1-\delta_i)}{n} &= \phi(\beta) - \phi(\alpha+\beta) = F_2(\alpha,\beta)\ . \end{aligned} \tag{16}$$

We use the well known Newton–Raphson method to solve (16) which leads to the following iterative scheme [11] to derive a sequence α_j, β_j:

$$\begin{aligned} \frac{\sum_{i=1}^{n} \log \delta_i}{n} &= F_1(\alpha_j,\beta_j) + (\alpha_{j+1}-\alpha_j)\left(\frac{\partial F_1}{\partial \alpha}\right)_{(\alpha_j,\beta_j)} + (\beta_{j+1}-\beta_j)\left(\frac{\partial F_1}{\partial \beta}\right)_{(\alpha_j,\beta_j)} \\ \frac{\sum_{i=1}^{n} \log (1-\delta_i)}{n} &= F_2(\alpha_j,\beta_j) + (\alpha_{j+1}-\alpha_j)\left(\frac{\partial F_2}{\partial \alpha}\right)_{(\alpha_j,\beta_j)} + (\beta_{j+1}-\beta_j)\left(\frac{\partial F_2}{\partial \beta}\right)_{(\alpha_j,\beta_j)} \end{aligned}\ . \tag{17}$$

As initial values (α_0, β_0), pivotal for the efficient convergence of Newton-Raphson's method, we use the ones proposed in [11], namely

$$\alpha_0 = \frac{m_1(m_1-m_2)}{m_2-m_1^2},\ \beta_0 = \frac{\alpha_0(1-m_1)}{m_1}, \tag{18}$$

where m_1, m_2 are the first and second order momenta of the residuals.

Finally, the gradient equations for the Weibull distribution are

$$\begin{aligned} 0 = \frac{\partial l}{\partial \hat{\lambda}} &= -\frac{n}{\lambda} - \frac{\kappa n - n}{\lambda} + \frac{\kappa}{\lambda^{\kappa+1}} \sum_{i=1}^{n} \delta_i^{\kappa} = \frac{\kappa}{\lambda}\left(n - \frac{1}{\lambda^{\kappa}} \sum_{i=1}^{n} \delta_i^{\kappa}\right), \\ 0 = \frac{\partial l}{\partial \hat{\kappa}} &= \frac{n}{\kappa} + \sum_{i=1}^{n} \log \delta_i - n \log \lambda - \sum_{i=1}^{n} \left(\frac{\delta_i}{\lambda}\right)^{\kappa} \log \frac{\delta_i}{\lambda}\ . \end{aligned} \tag{19}$$

The solution of the first equation of (19) is

$$\lambda = \left(\frac{1}{n} \sum_{i=1}^{n} \delta_i^{\kappa}\right)^{\frac{1}{\kappa}}\ . \tag{20}$$

Plugging this λ value into the second equation of (19) we get

$$\frac{\sum_{i=1}^{n} \log \delta_i}{n} = \frac{\sum_{i=1}^{n} \delta_i^{\kappa} \log \delta_i}{\sum_{i=1}^{n} \delta_i^{\kappa}} - \frac{1}{\kappa} = G(\kappa), \tag{21}$$

which we can solve again through Newton–Raphson's, obtaining the iterates

$$\kappa_{j+1} = \kappa_j + \frac{1}{G'(\kappa_j)}\left(\frac{\sum_{i=1}^{n} \log \delta_i}{n} - G(\kappa_j)\right)\ . \tag{22}$$

This time the initial value κ_0 is chosen empirically through experimentation; in our case $\kappa_0 = 1$ seems to ensure a fast convergence.

3.3 Uncertainty Intervals

Given a pre-specified probability $1-2s$, we want to find in the Gaussian and Laplace cases an error interval for which s is the percentage of residuals above and below the upper and lower interval limits. As stated before, we assume the conditional distribution of y given x to depend on x only through the predicted value $\hat{f}(x)$; as a consequence, intervals have the same width for each (x_i, y_i). More precisely, in the zero–mean Gaussian and Laplace cases, if p_s is the upper s–th percentile of the density p, the interval would be $(-p_s, p_s)$ where we can derive p_s solving $1-s = \int_{-\infty}^{p_s} p_\zeta(z)dz = \Phi(p_s)$, i.e., $p_s = \Phi^{-1}(1-s)$, with Φ the distribution associated to p. For a Laplace $p(z) = \frac{1}{2\sigma}e^{\frac{|z|}{\sigma}}$ we simply have $p_s = -\sigma\ln(2s)$. For the non zero μ mean Gaussian and Laplace the interval will be $(\mu-(p_s-\mu), \mu+(p_s-\mu))$. On the other hand, since we restrict the Beta and Weibull densities to positive values, we take for them intervals of the form $(0, p_{2s})$ where $1-2s = \int_0^{p_{2s}} p(z)dz$.

3.4 Hypothesis Testing

Among the advantages of the log–likelihood approach above is the availability of a test for rejecting a model p_0 against another p_1 [12]. In fact, if $\{Z_i\}_{i=1}^t$ are independent random samples following a density $\frac{1}{\sigma}p\left(\frac{z}{\sigma}\right)$), we can define the statistic

$$T(p_0, p_1, \{z_i\}_{i=1}^t) = \frac{\int_0^\infty \tau^{t-1} p_1(\tau z_1)p_1(\tau z_2)...p_1(\tau z_t)d\tau}{\int_0^\infty \tau^{t-1} p_0(\tau z_1)p_0(\tau z_2)...p_0(\tau z_t)d\tau} \ . \tag{23}$$

Then we reject the hypothesis $H_0 : p = p_0$ against $H_1 : p = p_1$ when it is bigger than a threshold c_α associated to a given significance level α. This defines [12] a most powerful test which is invariant under scale transformation. At a given significance level, α, when H_0 is true the probability of rejecting this hypothesis is α, i.e., it holds that $P_0(T(p_0, p_1, \{z_i\}_{i=1}^t) > c_\alpha) = \alpha$. We solve this by numerical integration. Specifically, we use the SAGE function sage.gsl.integration.numerical_integral with its default values, i.e., adaptive integration and absolute and relative error tolerances equal to 10^{-6}.

4 Experiments

We consider three different data scenarios. The first one is a synthetic problem to which we add different types of noise. The second scenario corresponds to four public datasets: **abalone** from Statlog [13], **space_ga** from StatLib [14], and **add10** as well as **cpusmall** from Delve [15]. In the final scenario we deal with a real wind energy prediction problem. In all of them we fit residual models according to Gaussian and Laplace densities with both zero (GAU, LAP) and non–zero (GAUm, LAPm) mean as well as Weibull (WEI) and Beta (BET) distributions. We also consider a version of the Laplace model fitting proposed in [7], LAP*, where we discard those δ_i residuals that exceed $\pm M$ times the residual standard deviation σ_δ; M ranges from 3 to 5 depending on the problem.

Table 1. Interval errors for the artificial data set with s=0.1.

Noise	Test	LAP	LAP*	LAPm	GAU	GAUm	BET	WEI
Noise free	GAUm/WEI	3.1	3.1	3	1.5	1.4	2.5	**1.1**
Laplace	LAP/LAPm	**0.4**	**0.4**	**0.4**	1.4	1.4	1.0	1.5
Gaussian	GAUm/WEI	2.8	2.8	2.9	0.2	**0.1**	1.2	0.5
Beta	BET	3.0	3.0	3.0	1.1	1.2	**0.2**	0.8
Weibull	WEI	3.5	3.5	3.7	1.6	1.4	1.3	**0.2**

4.1 SVR Parameter Selection and Interval Accuracy Metrics

As mentioned, we work with Gaussian SVR models, for which parameter selection is done using two different methods. The first one is either 5–fold cross validation or validation over a fixed dataset and the second one the Bayesian approach of [5], for which there is code available[2]. In the first case we have simply performed a zoomed grid search range to arrive to a (C, ϵ, γ) parameter set giving the smallest mean absolute error, MAE, over the validation set or through CV. As initial point, $(C_0, \epsilon_0, \gamma_0)$, we use the parameters obtained after applying the approach in [16]. We recall that the MAE over $\{(x'_i, y'_i)\}_{i=1}^N$ is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{f}(x'_i) - y'_i| \ . \tag{24}$$

We have used MAE instead of the more common mean square error, MSE. Strictly speaking, the ϵ–insensitive loss function of SVR is the best choice but notice that it depends on the ϵ parameter whose optimal value we want to determine. Our MAE choice avoids this circularity while being close to the ϵ–loss, agreeing with it for large deviations. In all SVR experiments we use the well known LIBSVM software [6].

To test the error err_s^M of the error intervals estimated according to a certain noise model M that corresponds to a pre–specified probability $1-2s$, we compare as in [7] the percentage of the residuals δ_i^{test} lying in the estimated error interval I_s^M derived using M with the expected number, $(1-2s)\times N'$, with N' the test sample size, i.e.,

$$err_s^M = \frac{100}{N'} |\#\ of\ \delta_i^{test} \in I_s^M - (1-2s)\times N'| \ . \tag{25}$$

4.2 Artificial Data and Results

Here we create an artificial dataset $\{(x_i, y_i)\}$ with $y_i = 2\cos x_i + 3\sin(2x_i) + n_i$ where x_i are uniformly distributed over $[0, 2\pi)$ and noise n_i is generated

[2] `http://www.gatsby.ucl.ac.uk/~chuwei/code/bisvm.tar`

Table 2. Interval errors for SVR models with CV parameters for public datasets.

Dataset	s	LAP	LAP*	LAPm	GAU	GAUm	BET	WEI
abalone	s=0.1	**2.8**	3.8	4.0	9.2	12	3.2	2.9
	s=0.05	0.6	**0.0**	0.8	1.9	0.8	1.6	0.5
add10	s=0.1	0.4	0.4	**0.2**	0.6	0.8	0.4	3.0
	s=0.05	**0.9**	**0.9**	**0.9**	1.9	1.9	1.7	1.8
space_ga	s=0.1	9.4	9.0	9.0	3.4	**2.2**	8.6	6.8
	s=0.05	3.0	3.9	2.0	0.7	**0.5**	3.9	4.3
cpusmall	s=0.1	7.0	6.4	**0.2**	14.2	14.8	0.8	4.3
	s=0.05	4.1	3.9	**0.2**	6.9	6.5	**0.2**	2.7

according to four distributions: *1)* Laplace noise with $\mu = 0$ and scale $\sigma = 1$; *2)* Gaussian noise with $\mu = 0$ and $\sigma = 1$; *3)* Beta noise with $\alpha = 1$ and $\beta = 2$; *4)* Weibull noise with scale $\lambda = 1$, and shape $\kappa = 5$. We randomly generate 5,000 x_i values of which 4000 are used for parameter selection and training, and 1000 as test set. The goal here is to test whether intervals built under a concrete noise assumption for the residuals perform better than the others when precisely that kind of noise is present in the data. For this scenario, SVR parameter selection is done by 5–fold CV.

The five rows in Table 1 correspond to data generated according to the four noise models considered as well as noise free targets. Columns from the third on contain the err_s^M for $s = 0.1$ using as noise models M the $LAP, LAP^*, LAPm$, $GAU, GAUm, BET$ and WEI distributions respectively; boldface values highlight the models with smallest err_s^M. As it can be seen, in all noisy cases the underlying noise model yields the smallest error; on the other hand, the Weibull model yields the smallest error for the noise free case, closely followed by $GAUm$. The second column summarizes the results of applying the statistic $T(p_0, p_1)$ for pairwise testing of a model against the others. Again, for each target noise type modeled by p_0 the alternate hypothesis of a different p_1 can be rejected in all cases except that of Gaussian noise, where a Weibull model cannot be rejected.

4.3 Public Datasets Results

Here we consider the three approaches mentioned above to build error intervals, namely *1)* SVR with parameters obtained by CV and uncertainty intervals as in Section 3.3; *2)* SVR with Bayesian parameters as described in Section 2.2 and uncertainty intervals as in Section 3.3. *3)* direct Bayesian intervals obtained through the procedure in [5]; we refer to it as BSVR in what follows. For each one of the public datasets we employ 1000 instances as test and the rest as training.

Table 3. Summary of interval errors and best noise models for public datasets.

Dataset		CV	BAYESIAN	BSVR
abalone	Best	LAP/WEI	LAP*	-
	Mean	**1.70**	3.55	4.95
add10	Best	LAPm	WEI	-
	Mean	**0.55**	1.40	4.00
space_ga	Best	GAUm	GAUm	-
	Mean	**1.35**	2.00	1.80
cpusmall	Best	LAPm	LAPm	-
	Mean	**0.20**	1.75	2.5

Columns in Table 2 from the third on contain for each dataset the errors err_s^M for $s = 0.1$ and $s = 0.05$ associated to noise models $LAP, LAP^*, LAPm$, $GAU, GAUm, BET$ and WEI, respectively, when parameters of the SVR are set through CV. Errors (not shown) obtained when Bayesian SVR parameters are used or using BSVR approach are higher; this can be observed in Table 3 that summarizes the best residual model and the mean of the $err_{0.1}^M$ and $err_{0.05}^M$ errors. As it can be seen, CV error means are about one percentage point below those of the other methods; moreover, either the GAU or LAP models, with or without means, yield the most accurate intervals.

4.4 Wind Energy Results

Finally, in this section we consider SVR uncertainty estimates in the prediction of wind energy. We use as input features the numerical weather predictions, NWP, of surface wind, temperature, pressure and 100 meter wind provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) with a 0.25° resolution for a rectangular area that contains the Iberian peninsula. Since the ECMWF forecasts are given at three hour intervals we interpolate them linearly to hourly values. The prediction target here is the hourly total wind energy production of peninsular Spain. We work with data from January 1st 2011 to December 31th 2013. Notice that the problem has a natural time structure; because of this we use data from 2011 as training data, that from 2012 as a validation set and that from 2013 as test. In our experience this data structuring yields better model parameters than those obtained applying random cross-validation. We consider five different approaches, namely:

1. M_1^C**:** compute residuals on the validation set of a SVR model with parameters found by a grid search using the fixed validation set previously described, and define a unique uncertainty interval for the entire test as in Section 3.3.

Table 4. Interval errors intervals for wind energy prediction.

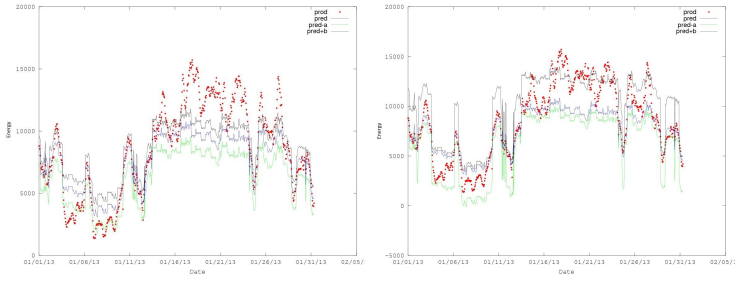
model	s	LAP	LAP*	LAPm	GAU	GAUm	BET	WEI	BSVR
M_1^C	0.1	6.0	5.9	5.0	5.6	5.5	5.0	**4.5**	-
	0.05	5.4	5.3	3.9	3.3	3.4	3.8	**3.2**	-
M_2^C	0.1	5.1	4.8	2.4	3.2	**2.2**	4.7	2.6	-
	0.05	5.4	5.3	3.8	2.9	**1.0**	2.3	2.7	-
M_4^C	0.1	8.7	8.3	**7.4**	**7.4**	**7.4**	9.0	7.7	-
	0.05	7.6	7.5	**5.3**	6.7	**5.3**	6.2	6.3	-
M_1^{B1}	0.1	8.2	8.1	7.8	6.1	5.8	6.9	**5.5**	-
	0.05	7.8	7.7	5.5	4.5	**4.3**	5.2	4.4	-
M_1^{B2}	0.1								**6.9**
	0.05								**5.7**

2. M_2^C**:** compute the residuals as in M_1 but split the validation set V into two groups V_1 and V_2 of equal size as follows: find the median y_m^V of the estimates $\hat{f}(x_i)$ in V and let $V_1 = \{(x_i, y_i) : \hat{f}(x_i) \leq y_m^V\}$ and $V_2 = V - V_1$. We divide then the residuals' set R into R_1 and R_2, where $R_1 = \{\delta_i : (x_i, y_i) \in V_1\}$ and $R_2 = R - R_1$ and finally fit two different error models to the residuals in R_1 and R_2 and build two different uncertainty intervals I_1, I_2 as in Section 3.3.
3. M_4^C**:** Same method as in M_2^C but this time with 4 subsets and intervals.
4. M_1^{B1}**:** Same method as in M_1^C but computing the residuals of the validation set using as SVR parameters the ones obtained by the Bayesian approach described in Section 2.2.
5. M_1^{B2}**:** Compute uncertainty intervals using the Bayesian procedure in [5] in its entirety.

Notice that procedure M_1^C yield a constant uncertainty interval for all input patterns x, while methods M_2^C and M_4^C yield either two or four intervals depending on the relationship between $\hat{f}(x)$ and y_m^V. In principle, these intervals adjusted to the magnitude of energy forecasts are a rather sensible choice in wind energy predictions, as the prediction errors are very dependent on forecast values. A comparison between M_1^C and M_2^C intervals with $s = 0.1$ for January 2013 is shown in Figure 1.

Columns in Table 4 contain errors err_s for $s = 0.1$ and $s = 0.05$ associated to each of the models above. For the M_i^C and M_1^{B1} models we give err_s^M values for all uncertainty models considered; these are omitted for the M_1^{B2} model, as it follows a different approach; boldface values highlight the smallest errors. Table 5 summarizes that information showing for each approach the best residual model and the corresponding $err_{0.1}$ and $err_{0.05}$ errors. As it can be seen, best results are

Fig. 1. M_1^C and M_2^C uncertainty intervals for January 2013 data of wind energy problem with $s = 0.1$. Figure shows real production (prod), prediction given by the model (pred), lower bound of prediction interval (pred-a) and upper bound (pred+b)



clearly achieved with the M_2^C approach using $GAUm$ models after dividing into two subsets the residuals of an SVR whose parameters are obtained by validation over 2012 data. This best performance is followed by that of the M_1^C and M_1^{B1} approaches; on the other hand, dividing the residuals into 4 groups results in a lower accuracy. This is probably due to working with too many groups which makes it difficult to find a core of common patterns inside them. Finally, the purely Bayesian approach M_1^{B2} yields again the lowest accuracy.

5 Conclusions and Further Work

Support Vector Regression, SVR, is one of the most used tools in non–linear regression and modeling problems and, as such, uncertainty estimates of SVR predicted values are of great importance. A particularly clear example is SVR–based wind energy prediction, whose intermittency and wide fluctuations make necessary to define appropriate levels of rolling reserve; good uncertainty estimates are an obvious tool for this. In this work we have broadly followed the approach in [7], considering noise model distributions that are fitted to the residuals of SVR models whose C, ϵ and γ parameters have been found directly by CV or validation over a fixed set, or under a Bayesian perspective; as in [7], we have also considered the entirely Bayesian approach to define uncertainty intervals proposed in [5]. We have enlarged this set up by adding to the noise models considered in [7] non–zero mean versions of Gaussian and Laplace noise, as well as Beta and Weibull variants.

A first general conclusion is that, in agreement with [7], purely Bayesian interval estimates are poorer than those obtained by fitting noise distributions to residual values; moreover, interval error estimates are more accurate when SVR parameters are chosen by CV or validation over a fixed set. Since this is SVR specific only to the extent that SVRs are the underlying model, it suggests that direct residual–based fitting of error models should also be a useful approach when non–linear regressors are built under other alternative paradigms.

Table 5. Summary of interval errors for wind energy prediction.

	M_1^C	M_2^C	M_4^C	M_1^{B1}	M_1^{B2}
Best	WEI	GAUm	GAUm	WEI	-
s=0.1	4.5	**2.2**	7.4	5.5	6.9
s=0.05	3.2	**1.0**	5.3	4.4	5.7

Moreover, we have shown over a synthetic example how this approach is able to resolve the true underlying noise model; this is the case for the four noise distributions considered, even when taking into account that the ϵ–insensitive SVR loss does not entirely corresponds to any of them. Of course, the true noise model depends entirely on the sample data and not, in principle, on the loss function used. On the other hand, the loss function somehow addresses a particular noise structure, which suggests that perhaps shifted versions of the Gaussian and Laplace densities would yield more accurate uncertainty intervals. For instance, the ϵ–insensitive loss $[\delta]_\epsilon$, that corresponds to a uniform noise density $p_\epsilon(\delta) = 1/2(1+\epsilon)$ when $|\delta| \leq \epsilon$, and $p_\epsilon(\delta) = e^{|\delta|-\epsilon}/2(1+\epsilon)$ when $|\delta| > \epsilon$, could possibly yield tighter uncertainty intervals. Similarly, this also suggests that loss functions other than the ϵ–insensitive one should be considered to build an SVR model when noise from a particular distribution is suspected. Of course, not every conceivably noise density could be so addressed, but SVR models for Beta and Weibull–inspired losses are considered in [8]. In this vein, other candidates could be the Cauchy, Logistic or Voigt distributions.

Furthermore, the proposed technique for building error intervals is not exclusive for SVR models, with the approach being independent of the model chosen to solve the regression problem and the noise assumptions presumed. Thus, testing accuracy of intervals result of applying other regression models, such as Random Forests or Neural Networks, is a clear line of further work.

Finally, a drawback of the residual fitting approach is that error intervals are built independently of the $\hat{f}(x)$ regressor values. This is particularly so in problems such as wind energy prediction, where model errors are usually much higher for large energy values. As mentioned in the Introduction, the well known analytic estimates for linear regression under Gaussian models that yield error estimates dependent on x are simply impossible for SVR.

Despite this fact, the splitting techniques we use that yield two energy level–based uncertainty regimes give much better results than a single interval procedure. However, this is still rather coarse and a finer grain split of the sample residuals into four regimes but using a single SVR model gives worse results. Because of this, it might be worth trying to consider distinct SVR models for each residual regime and fit noise distributions to the residuals of each model, instead of working with an unique SVR model for the entire dataset. We are currently working on these and other related issues.

Acknowledgments. This paper was developed with partial support from Spain's grants TIN2013-42351-P and S2013/ICE-2845 CASI-CAM-CM and also of the Cátedra UAM–ADIC in Data Science and Machine Learning. Authors also gratefully acknowledge the use of the facilities of Centro de Computación Científica (CCC) at UAM. They also thank Red Eléctrica de España for kindly supplying wind energy data.

References

1. Shawe-Taylor J., Cristianini N.: An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press (2000)
2. Yang, H., Chan, L., King, I.: Support vector machine regression for volatile stock market prediction. In: Intelligent Data Engineering and Automated Learning—IDEAL 2002, pp. 391-396, Springer (2002)
3. Kramer, O., Gieseke, F.: Short-term wind energy forecasting using support vector regression. In: Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011, pp. 271-280, Springer (2011)
4. Gala Y., Fernandez, A., Díaz J., Dorronsoro J.R.: Support vector forecasting of solar radiation values. In: Hybrid Artificial Intelligent Systems, pp. 51–60, Springer (2013)
5. Chu, W., Keerthi, S.S., Ong, C.J.: Bayesian support vector regression using a unified loss function. In: IEEE Transactions on Neural Networks, vol. 15, pp. 29–44, IEEE (2004)
6. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. In: ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, pp. 27, ACM (2011)
7. Lin C., Weng R.: Simple Probabilistic Predictions for Support Vector Regression. National Taiwan University, Taipei (2004)
8. Hu, Q., Zhang, S., Xie, Z., Mi, J., Wan, J.: Noise model based ν-support vector regression with its application to short-term wind speed forecasting. In: Neural Networks, vol. 57, pp. 1–11, Elsevier (2014)
9. Pontil, M., Mukherjee, S.,Girosi, F.: On the noise model of support vector machines regression. In: Algorithmic Learning Theory, pp. 316–324, Springer (2000).
10. Celik, A. N.: A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey. In: Renewable energy, vol. 29, pp. 593–604, Elsevier (2004)
11. Gnanadesikan, R., Pinkham, R.S., Laura P. Hughes.: Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics. In: Technometrics, vol. 9, pp. 607–620, Taylor & Francis Group (1967)
12. Lehmann E.L., Romano J.P.: Testing statistical hypotheses. Springer Science & Business Media (2006)
13. UCI Machine Learning Repository, `https://archive.ics.uci.edu/ml/datasets.html`
14. StatLib Datasets Archive, `http://lib.stat.cmu.edu/datasets/`
15. Delve Datasets, `http://www.cs.utoronto.ca/~delve/data/datasets.html`
16. Cherkassky, V., Ma, Y.: Practical selection of SVM parameters and noise estimation for SVM regression. In: Neural networks, vol. 17, pp. 113–126, Elsevier (2004)
17. Johnson, N.L., Kotz, S. and Balakrishnan, N.: Continuous Univariate Distributions, Vol. 2. John Wiley & Sons (1995).